Voter Persuasion in Compulsory Electorates: Evidence from a Field Experiment in Australia

Patrick Lam and Kyle Peyton*

This version: December 13, 2013

Abstract

Most of the literature on grassroots campaigning focuses on mobilizing potential supporters to turn out to vote. The actual ability of partisan campaigns to boost support by changing voter preferences is unclear. We present the results of a field experiment the Australian Council of Trade Unions (ACTU) ran during the 2013 Australian Federal Election. The experiments were designed to minimize the conservative (the Coalition) vote as part of one of the largest and most extensively documented voter persuasion campaigns in Australian history. Union members who were identified as undecided voters in over 30 electorates were targeted with appeals by direct mail and phone banks. Because of the presence of compulsory voting in Australia, we are able to identify the effects of voter persuasion independently of voter turnout. We find that direct mail, the most extensively used campaign strategy in Australia, has little effect of voter persuasion. Direct human contact, on the other hand, seems to be an effective tool for voter persuasion. Among undecided voters who actually receive direct contact via phone call, we find a ten percentage point decrease in the Coalition vote. From a methodological standpoint, we use various methods to account for multiple treatment arms, measured treatment noncompliance in one of the treatments, and missing outcome and covariate data. The field experiment also provides a good lesson in conducting and saving broken experiments in the presence of planning uncertainty and implementation failures.

*Patrick Lam (patrickkennethlam@gmail.com) is a visiting fellow at the Institute for Quantitative Social Science at Harvard University, where he received his Ph.D. in November 2013. Kyle Peyton (kpeyto1@gmail.com) is a Research Fellow at the Melbourne Institute of Applied Economic and Social Research, University of Melbourne. This project was funded by the Australian Council of Trade Unions (ACTU). Many people at the ACTU worked passionately to ensure the design and execution of this project was successful. These include Paul Erickson, James Booth, Daniel Mookhey and George Simon. We thank Gary King for overall advice and guidance on the project. We thank David Gow, Yusaku Horiuchi, Kosuke Imai, Jack Vowles and participants at the 2nd Annual Conference of the Australian Society for Quantitative Political Science for helpful comments. The authors conducted this research under the umbrella of Beecher Analysis Group.

1 Introduction

Political parties and interest groups spend millions of dollars every year on campaigns designed to influence political behavior. Evaluation of campaign effectiveness is one of the most active research areas in the social sciences and one of the few that has benefited from evidence based on numerous randomized field experiments. The earliest documented experiments were designed to increase voter turnout in the United States (Gosnell, 1927; Eldersveld, 1956) and this tradition has been part of many subsequent elections. Since the landmark Get out the Vote (GOTV) experiment by Gerber and Green (2000) this has become the most active area of experimental research in political science and has resulted in considerable spillover effects that have changed the way political organizations run campaigns and evaluate their effectiveness.

Implicit in many campaigns is the assumption that targeted voter turnout helps one side or another at the polls, yet most of the evidence on the effectiveness of campaign strategies derives from research on Get out the Vote (GOTV) experiments designed to increase voter turnout (Gerber and Green, 2008; Issenberg, 2012). The focus on voter turnout as opposed to persuasion at least partly reflects the measurement challenges involved when the outcome of interest is private information. Voter turnout is usually publicly available; in the United States, for example, turnout data are available at the individual level. The secret ballot, however, ensures that measuring who an individual actually votes for requires a post-election survey, which can be expensive to run and presents additional methodological challenges.

Although the literature on voter persuasion is less developed than the one on voter turnout, a few studies have attempted to quantify the relationship between partisan campaigns and candidate or party support. Many (though not all) of these studies have found that grassroots campaigns both persuade and mobilize voters (Hillygus, 2005; Rosenstone and Hansen, 1993; Vavreck, Spiliotes and Fowler, 2002; Wielhouwer and Lockerbie, 1994; Kramer, 1970). However, the use of observational data in these studies makes establishing causality difficult. In terms of field experiments, Miller and Robyn (1975) conclude in their experiment on campaign mailing and turnout and candidate preference in a barely contested Democratic congressional primary that mailing had no effect on change in behavior. Adams and Smith (1980) come to a similar conclusion looking at a field experiment of persuasion calls in a special election for a Washington D.C. city council seat. Gerber (2004) finds that negative direct mail messages increase the vote share of challengers. Nickerson (2005) finds no evidence that phone calls influenced voter preferences in the 2002 Michigan gubernatorial race using survey data. Arceneaux (2005) studied the effects of door-to-door canvasing on precinct-level election outcomes on ballot measures Kansas City, but the experiment lacked sufficient power to discern significant effects. Finally, Arceneaux (2007) finds that both door-to-door canvassing and commercial phone bank calls can have strong effects on voter preferences in a Democratic primary for a county commissioner seat in New Mexico.

All of these field experiments suffer from drawbacks. First, the experiments are almost always localized within relatively minor elections or ballots. While smaller elections have certain advantages such as better control over implementation and less noise from outside influences such as the media or rival political parties and groups, they also suffer from an external validity problem. It is unclear how any of the measured persuasion effects - changing a voter's vote from one party/candidate to another - translate to larger-scale national elections, which are more salient to voters and ultimately more influential in driving the political agenda and policy-making.

Second, the field experiments are all located within the United States, an electorate without compulsory voting and one of the lowest turnout rates in the world. This makes it difficult to untangle the interdependent outcomes of voter turnout and voter preferences. Consider a model of voting where a potential voter's propensity to turnout and propensity to vote for a candidate are a function of many factors. Assume that the potential voter may or may not turn out and vote if he supports a candidate and stays home if he is undecided. Now suppose we observe that the potential voter receives a campaign treatment and subsequently turns out to vote for the candidate. Also, suppose that the unobserved counterfactual in the absence of receiving treatment is that the potential voter stays home. In this scenario, it is unclear whether the treatment had an effect only on turnout (convincing an existing supporter to go vote), an effect only on voter preference (convincing an undecided voter to support the candidate), or both (convincing an undecided voter to support the candidate and then convincing him to go vote). The ability of a voter to make both a choice about turnout and about voting makes it difficult to disentangle a turnout effect from a persuasion effect.

In a compulsory electorate where not voting comes with an economic cost, voter turnout becomes a second order issue as almost everybody votes. Thus, there are basically no turnout effects and we can identify the persuasion effect independently. In this paper, we present results from the first field experiment on voter persuasion conducted in a compulsory electorate that we are aware of. The field experiment was conducted on union members in over 30 federal House of Representatives districts (herein referred to as electorates) in Australia during the 2013 federal election. In Section 2, we discuss the background of the election, the field experiment, and the organization that conducted it. In Section 3, we outline the planned experimental design and discuss the implementation failures that arose in practice. Section 4 outlines the estimation strategy and statistical methods we employed in light of the methodological challenges that arose and also presents our results. Overall, we find that using the same strategies proposed by the GOTV literature produces similar results for voter persuasion. That is, direct human contact is an effective method for influencing voting behavior but we should be less sanguine about direct mail and robocalls.

2 Background

Australia is a multi-party parliamentary system and the political landscape is dominated by the Australian Labor Party (ALP) and the smaller Australian Greens on the left and the "Coalition" consisting of the Liberal Party, the National Party, the Country Liberal Party, and the Liberal National Party on the right. The years preceding the 2013 federal election marked a tumultuous period for the ALP, which was the party in power. Kevin Rudd, elected by landslide in 2007, resigned as party leader in 2010 after losing the support of the ALP and was replaced by Julia Gillard, Australia's first female prime minister.[1] Gillard called an election less than a month after becoming prime minister[2] and in August 2010 won 72 of 150 seats in the House of Representatives. This was four seats short of the 76 required for a majority government and resulted in the first hung parliament since 1940. The period between 2011 and 2013 was marked

[1] Under the Labor Party's rules, the Parliamentary leader of the Labor Party was elected by and from the Federal Parliamentary Labor Party (known as "Caucus"), with each Labor member of the House of Representatives and the Senate holding a vote.

[2] In Australia the maximum time that can pass between federal elections is three years but there is no minimum and an election can be called at any time. Australia is a federal constitutional monarchy in which legislative power is vested in the bicameral Parliament of Australia, under the Queen of England as Head of State. Section 51 of the Australian Constitution sets out the powers and responsibilities of the Commonwealth; remaining powers are retained by six States. An election is "called" when the Prime Minister, who is generally the leader of the majority party or coalition in the House of Representatives, petitions the Governor General as the Queen's representative to issue an electoral writ. Only once the writ is issued by the Governor General is an election confirmed.

by continued speculation within the Labor Party about Prime Minister Gillard's leadership, with supporters of Kevin Rudd agitating to return the former Prime Minister to the leadership. In February 2012 and March 2013, Gillard "spilled" the Labor leadership, a process in which the position of ALP leader was declared vacant and opened for re-election. At the first spill, Gillard was re-elected by a resounding majority of the caucus; at the second spill, Gillard was re-elected unopposed after Rudd declined to contest the election. Despite this, Gillard's leadership remained in question and in June 2013, following a drop in support for the ALP in published opinion polling, Gillard called a third spill which saw Kevin Rudd re-elected as leader by a narrow majority of the caucus. Rudd then led the ALP to the 2013 election, held on September 7, where the Government was defeated by a conservative Coalition led by Tony Abbott, leader of the center-right Liberal Party of Australia.

The experiment described here was conducted by the Australian Council of Trade Unions (ACTU), the peak body of the union movement in Australia, during the 2013 election. The body was established in 1927 and is composed of 46 affiliated unions. One of the primary political goals of the ACTU is to lobby for workers rights in Australia. Although not an explicitly partisan organization, the ACTU tends to support progressive political parties (primarily the ALP and the Greens). During the 2013 federal election, the ACTU was concerned a Coalition government would legislate policy aimed at diminishing workers rights and launched a campaign with the objective of minimizing the conservative vote in target electorates. The campaign in the run-up to the September 7, 2013 election involved three parts:

1. An initial phone survey of union members conducted by a commercial call center from the beginning of May to the beginning of August 2013 to establish voting intentions and identify volunteers.
2. A randomized persuasion effort conducted by volunteers on union members in 32 target electorates. This effort ran from August 20 all the way up to the election.
3. A followup phone survey to establish voting behavior on a small subset of union members from September 13 to October 2.

Before each part of the campaign, every union within the ACTU had a choice of whether to participate in the next part of the campaign. In the initial phone survey, 22 unions decided to participate and calls were placed to over 190,000 union members in 150 electorates. There are over 470,000 union members in the 22 unions in these 150 electorates, so contact was attempted on just under half of these union members. The members selected to be called were not chosen randomly because unions opted into the campaign at different points in time. Of the 190,000 members called, just over 115,000 completed the initial phone survey. One of the questions asked was "If the federal election were held today, for which Party would you be most likely to vote in the House of Representatives?" We use this as the measure of voting intentions since the ballot for House of Representatives is by far the more salient ballot. Union members who responded with "Undecided" were classified as undecided voters. There were just over 38,000 undecided voters from the phone survey.

At the second stage of the campaign, 11 of the 22 unions remained willing participants and 32 electorates deemed marginal[3] by the ACTU were selected as targets. The randomized field experiment was conducted on members of the participating unions in these 32 electorates with six different treatment arms (described in Section 3). This experimental population consisted of nine unions and approximately 51,000 union members who were "not identified as partisan in the initial phone survey."[4] In the third stage of the campaign, a small subset of 8,000 union members were chosen for a post-election phone survey of voting behavior. As described in Section 3.3 the eventual number of members in the final analyses is significantly smaller and only eight unions were represented.

3 Description of Experiment

To our knowledge, this was the most extensive direct voter persuasion campaign in Australian history and the first experimental campaign of its type. In addition to the actual treatments being experimental, the implementation and planning of the campaign was also experimental. As such, each step of the campaign experienced great uncertainty as to the scope and budget of the remaining parts of the campaign. On the implementation side, the field teams also experienced much uncertainty as this was the first time they would implement an organized campaign with randomized strategies. Unsurprisingly, there were some implementation failures. The field experiment can be considered a "broken experiment" due to the planning uncertainty

[3]The target electorates were identified by the ACTU data analytics team as being within approximately 2,000 votes in favor of the conservatives. According to ACTU estimates, the average target electorate contained approximately 10,000 to 12,000 union members. Between 6,000 and 7,000 were members of unions participating in the ACTU campaign.

[4]By "not identified as partisan in the initial phone survey", we mean that they were either identified as undecided voters in the initial phone survey or they were not reached by the initial phone survey. Thus, the initial experiment population included partisans who were simply not called in the initial survey, although we limit the results to identified undecided voters later.

and implementation failures.

3.1 Treatments

The ACTU was interested in the following treatments[5] for the persuasion campaign:

1. **Direct contact:** A phone call from a campaign volunteer and a home visit, both with a script urging the voter to not vote for the Coalition.
2. **Glossy mail:** A series of colorful postcards designed by the ACTU Campaign Team. The messages were designed to target salient campaign issues derived from predictive models based on a survey of union membership.
3. **Phone Call:** A phone call from a campaign volunteer with a script urging the voter to not vote for the Coalition.
4. **Formal mail and robocall:** A signed letter from the president of the ACTU personally addressed to the union member urging the member not to vote for the conservatives and a robocall containing a message from the President of the ACTU urging the voter not to vote for the Coalition.
5. **Glossy mail and robocall:** The same series of colorful postcards from the **glossy mail** treatment and a robocall containing a message from the President of the ACTU urging the voter not to vote for the Coalition.

In addition, there was a sixth treatment arm of control observations consisting of union members who did not receive any contact from the ACTU during the campaign. The **direct contact** treatment is simply the **phone call** treatment with an added home visit, while **glossy mail and robocall** is simply **glossy mail** with an added robocall. The treatments were all framed to minimize the Coalition vote rather than maximizing any progressive votes since the ACTU does not explicitly endorse any political party.

To implement the direct contact and phone call treatments, we had to take into account the number of volunteer staff available for each. Direct contact home visits required volunteers living in the same electorate whereas phone call treatments did not have the same exact restriction, although there was a strong preference for volunteers to call union members within their

[5]The direct contact and glossy mail methods are two strategies that reflect the most commonly used methods of voter persuasion used. The direct contact strategy involves either a personal phone call from a campaign volunteer, a home visit or both. The glossy mail strategy involves a series of direct mail messages delivered to the voter by the post. The mail messages are typically crafted around salient campaign issues and urge the voter to vote for a particular party on the basis of the party's position on one of these issues. Gerber and Green (2008) present a meta-analysis of numerous field experiments conducted in the United States that show the most effective tactics for Get Out the Vote Campaigns are phone calls conducted by campaign volunteers and home visits. The results for direct mail and robocalls are less certain and in all likelihood close to zero. A crucial distinction, however, is between the type of mail used. Anecdotal evidence suggests that a formal letter addressed to the resident is a more effective tool than the typical glossy mail used by political campaigns (Issenberg, 2012).

electorate. We initially received an estimated number of volunteers available to door knock in 16 of the 32 electorates over the two weekends prior to the election. The other 16 electorates had no door-knocking volunteers to implement the full direct contact treatment. We subsequently divided the 32 target electorates[6] into two groups: Group A and Group B.

Group A contained the experimental population of 25,871 union members who were "not identified as partisan in the initial phone survey" and lived in the 16 electorates with door-knocking volunteers.[7] Group A was then designated to have three treatment arms: direct contact, glossy mail, and control. The randomization of treatment in Group A was performed in the following way:

1. Block on electorate.
2. Within each electorate, determine the number of direct contact recipients by multiplying the number of door-knocking volunteers by 30.[8] Cap this number at one-half of the experimental population in the electorate.
3. Within each electorate, subtract the number of direct contact recipients from the total number of voters in the experimental population in the electorate and divide this number by two to get the number of glossy mail recipients and the size of the control group respectively.
4. Within each electorate, randomize the three treatment arms according to the numbers in 2, 3, and 4.

Group B contained the experimental population of 24,532 union members that lived in the other 16 electorates without door-knocking volunteers. Group B was designated to have four treatment arms: phone calls, formal mail and robocall, glossy mail and robocall, and control. The randomization of treatment in Group B followed a similar process to group A:

1. Block on electorate.
2. Within each electorate, divide the experimental population by four to get the number of phone call recipients. Cap this number at 1,000, which was a rough estimate of the maximum number of calls that can be made per electorate.

[6]Tables 11 and 12 in Appendix A show the electorates in Group A and Group B as well as the electoral vote-share swings from the 2008 election to the 2013 election.

[7]Note that in the final analyses we only look at undecided voters in the initial phone survey and drop all voters (including potential partisans) that were not initially surveyed. However, at this point in the design stage, we were unsure of the budget and scope of the post-election survey, so we designed a much larger experimental population.

[8]30 was an estimate of how many houses a door-knocking volunteer would be able to reach over two weekends. This was an overestimate, but we felt it better to give an overestimate and have the number come up short rather than underestimate and have door-knockers start knocking on doors of union members that were not supposed to receive the direct contact treatment.

3. Within each electorate, divide the number of formal mail and robocall, glossy mail and robocall, and control observations evenly among the remaining number of voters in the experimental population in the electorate.
4. Within each electorate, randomize the four treatment arms according to the numbers in 2 and 3.

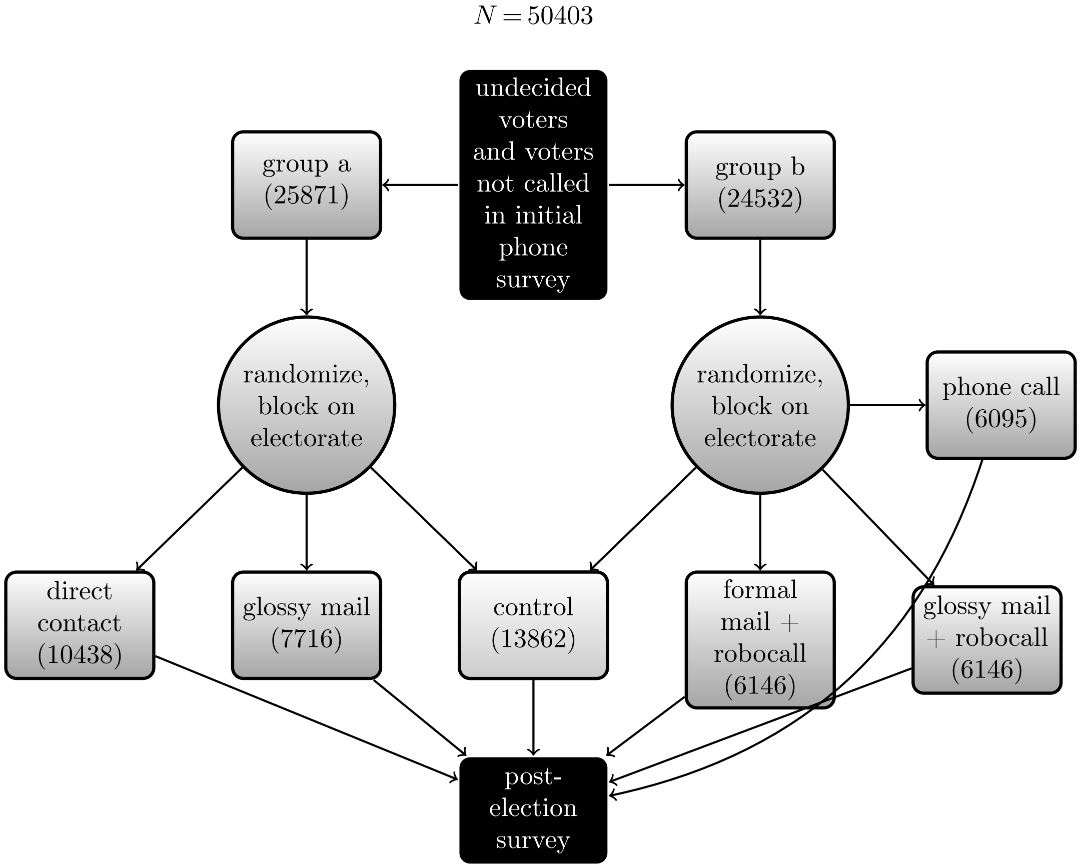
This design was made in conjunction with the ACTU field campaign. The field campaign was a decentralized operation that spanned across target electorates in Australia and relied on local support networks for both phone call and direct contact treatments. Volunteers implementing the phone call treatments were instructed to first call the union members identified in the initial phone survey as undecided. They would then proceed down the list of union members scheduled to receive a phone call in order of "predicted undecidedness" using a simple predictive model with information that we had on all the members and the initial survey. All treatments involving direct mail were delivered by Australia Post and did not require the use of campaign volunteers. Robocalls were likewise automated.

3.2 Designed versus Achieved

Figure 1 illustrates the planned experimental design with sample sizes for each treatment. There were two general implementation failures, one that was major and another that was smaller and somewhat expected. The major implementation failure was that the door-knocking campaign did not reach nearly the scale originally envisaged and very little data was collected. As a result, interactions with less than 100 households out of a planned campaign of over 10,000 door-knocks were recorded. This meant that one part of the direct contact treatment in Group A was basically not implemented.[9] However, the other part of the direct contact treatment, the volunteer phone calls, actually did proceed. Many union members in Group A and Group B were called, so we decided to discard the door-knocking part of the direct contact treatment and relabel the direct contact treatment in Group A as simply a phone call treatment. Note that the new phone call treatment in Group A is identical in implementation to the phone call treatment in Group B. So despite the massive failure of the door-knocking campaign, we were able to save the experiment by simply reducing the number of treatment arms and combining the direct contact treatment into the phone call treatment.

[9] Although there may have been a few door-knocks that were not recorded, this number was likely to be small and so we proceed with the experiment under the assumption that households without a recorded door-knock were not door-knocked.

Figure 1: Field experiment design



The minor implementation failure was that not every union member scheduled to receive a phone call actually did. This was expected since our initial estimate in the design phase of the number of phone calls possible was an overestimate by design. Those that were assigned to a phone call but were not called automatically became a part of the control group, although we never survey any of them in the post-election survey, so they were basically dropped from the experiment. Since phone calling was in order of "undecidedness", those that failed to receive the phone calls were likely union members who were more partisan, making this minor implementation failure a non-issue given how we do the post-election survey described in Section 3.3.

Despite the two implementation failures, we consider the experiment generally to be a success. We were able to save the experiment by combining the now identical phone call and direct contact treatments, which we will label from here on out as simply the phone call treatment. The mail and robocall treatments in both Group A and Group B were implemented successfully as planned and nobody assigned to the control group received a treatment.

Table 1 shows the difference between the designed and achieved and implemented treatment assignments across both groups. Since most of the phone call volunteers lived in Group A

Table 1: Designed versus Achieved Treatments

Treatment	Designed	Achieved	Difference
Phone call	10438	6397	4041
Glossy mail	7716	7716	0
Control	7717	7717	0

(a) Group A

Treatment	Designed	Achieved	Difference
Phone call	6095	1919	4167
Formal mail + robocall	6146	6146	0
Glossy mail + robocall	6146	6146	0
Control	6145	6145	0

(b) Group B

electorates, Group A had a higher proportion of phone calls implemented. Volunteers were instructed to call union members from their own electorates first before calling members from other electorates. Note that although 6397 and 1919 union members were called in Group A and Group B respectively, not all union members were reached via a phone call. Only 2107 and 574 members in the two groups were confirmed to have picked up the phone and engaged with the volunteer caller. The rest of the union members either had a message left on their voicemails, picked up the phone but refused to engage the volunteer, had bad phone number information on file, or their status was unknown due to data entry error. Thus, there is a treatment compliance problem in the sense that many union members who were assigned to treatment and had the treatment assignment successfully implemented nevertheless did not actually receive the treatment. We will revisit this compliance problem in detail in the analyses.

3.3 Post-Election Survey

To measure voting behavior on election day as our outcome of interest, the same commercial call center used for the initial pre-election phone survey was used to conduct a post-election survey within a month after the election. Given budget and timing[10] limitations, we were unable to survey the entire experimental population. The decision was made to sample only

[10]The decision was made that a post-election survey of voting behavior would only be valid and relevant for a limited window of time following the election.

8,000 union members within the experimental population. Given this limitation, we decided to first include in the survey all individuals who were identified as undecided voters in the initial phone survey. For practical reasons, we excluded those who were assigned the phone call treatment and subsequently refused the phone call, indicated no interest in receiving phone calls, or were recorded as having bad phone number information. To get to 8,000 post-election calls, we also chose to survey those who were successfully door-knocked in Group A and about 700 more who received either mail or phone call treatments. Ultimately, we decided to drop these individuals from the analysis so that the final set of union members includes only those identified as undecided voters in the initial phone survey and who were not door-knocked. During the implementation of the post-election survey, one union and a geographic subset of another union asked to be removed from the calling list, so members from those groups who were not already called were dropped from the analyses as well. Finally, we dropped one electorate (Swan) which only had three union members after the other exclusions were made. The final set of union members used for analyses includes 5,781 union members from 8 unions in 31 different target electorates. Table 2 shows the number of union members in our final sample by treatment assignment.

Table 2: Final experimental sample from post-election surveys by original treatment assignment

	Phone Call	Glossy mail	Glossy mail + robocall	Formal mail + robocall	Control	Total
Group A	1184	1164			1115	3463
Group B	274		685	678	681	2318
Total	1458	1164	685	678	1796	5781

The post-election survey consisted of eight questions. In addition to a few covariates (described in Section 3.4), the main variable measured as the outcome of interest was a question asking which party the union member supported in the 2013 election.[11] Not surprisingly, many contacted members did not wish to divulge their vote choice. The survey non-response consisted of individuals who could not be reached for the post-election survey, individuals who were contacted but refused to participate in the survey, and individuals who were contacted and participated in the survey but chose to keep their voting behavior a secret. In total, 2,079 of the

[11]The exact wording of this question was "Thinking about the Federal Election held on Saturday September 7, which Party did you vote for?" Although somewhat vague in whether it refers to the House election or the Senate election on the same ballot, we believe this question actually captures the idea of party support in a way that a more specific question may not.

5,781 individuals in our final sample chose to reveal their election voting behavior.

3.4 Background Covariates

For our sample of union members, we obtained a very limited set of background covariates (denoted by X) either provided by the unions beforehand or asked of the union members in the post-election survey. From records provided by individual unions beforehand, we have recorded:

- **Electorate**: the member's home electorate. We initially received member home information in the form of a set of latitude and longitude coordinates, which we then mapped onto the Australian electorates.
- **Union**: the member's union
- **Gender**: male or female. The gender variable originally contained a large number of missing values (around 15 percent missing in our final sample), which we imputed using what we believe to be a very accurate procedure.[12]
- **Age**: the member's age, which we derived from dates of birth provided by the unions. This variable is highly missing (around 70 percent missing in our final sample). Some union records were not complete for dates of births of their members and other unions failed to provide information on age or dates of birth for any of their members. Since we believe age to be an important predictor of voting behavior, we impute age statistically using an imputation method described in our methodology section.

We supplemented the union-recorded covariates with a few covariates from the post-election survey. These variables are fully observed for all individuals who completed the post-election survey, including all who gave their vote information. However, they are missing for all individuals in our sample who did not complete the post-election survey. The variables we extracted from the post-election survey are:[13]

1. **Past party**: "Thinking back to past federal elections, which Party have you supported most often?"
2. **Labor contact**: "Do you recall receiving any of the following forms of contact in the lead up to the election from the Labor Party?" We code this variable to be four categories: 1)

[12] For all our members, we received first name information and for many, we also received their titles (Mr., Mrs., Dr., etc.). In our final sample, we imputed gender first using the gender-specific titles. For the rest of the missing genders, we then used the first names and matched them to a dataset of births and names in the US from 1930-2012 provided freely by the US social security administration (SSA) office (http://www.ssa.gov/OACT/babynames/limits.html). For each first name in the SSA database, we took the gender that had the higher count of people and then matched these names/genders to the first names of everybody in our sample. We were left with only seven members whose names were not matched. For these seven, we took a best guess based on name endings and a quick search on Google.

[13] The two variables of Labor contact and Liberal contact are not necessarily pre-treatment since the parties may have contacted the members either before or after our treatments were implemented. However, because our treatments were block randomized, these covariates are unlikely to be strongly correlated with our treatments. We use them in some of our initial findings as a way to improve efficiency without inducing post-treatment bias.

direct contact either with door knock or phone call, 2) indirect contact with either email, mail, or robocall, 3) other, and 4) not sure.

3. **Liberal contact**: "Do you recall receiving any of the following forms of contact in the lead up to the election from the Coalition?" We code this variable in the same way as Labor contact.

4. **Age range**: We asked respondents to put themselves into one of four age ranges: 1) 18-25, 2) 26-40, 3) 41-55, 4) 56 and above. We use this variable to supplement in the age imputations in our analyses.

4 Methodology and Results

Our final sample here consists of 5,781 union members who were previously identified as undecided voters. Due to the nature of the experimental population and the data, we do not claim the sample is representative of Australia as a whole or even representative of union members in Australia. The sample is precisely a convenience sample of undecided union members in eight unions across 31 target electorates. Our quantities of interest will be relegated simply to sample quantities rather than population quantities. While there are clear external validity concerns, we believe that the benefits and insights from running a field experiment in this unique situation of a national election in a compulsory voting electorate outweigh the concerns.

The primary objective of the ACTU voter persuasion campaign was to minimize the right-of-center Coalition vote, defined here as a vote for the Liberal Party, the National Party, the Country Liberal Party, or the Liberal National Party. Let $Y_i = 1$ if individual i voted for a Coalition party and $Y_i = 0$ if i voted for a non-Coalition party or independent candidate. Let $Z_i = 1$ if individual i is assigned the treatment of interest and $Z_i = 0$ otherwise.[14] Since we have four types of treatment, we create four binary treatment assignment variables: Z^p for the phone call treatment, Z^g for the glossy mail treatment, Z^{gr} for the glossy mail and robocall, and Z^{fr} for the formal mail and robocall treatment.[15] Note that for each assignment variable, the zeroes contain both individuals who are in the control group and individuals who are in one of the other treatment groups.

[14] We use the letter Z to denote treatment assignment here rather than the more conventional T in order to stay consistent with notation in later sections where we discuss treatment noncompliance.

[15] For simplicity, depending on the context, we use the notation Z without a superscript for the rest of the paper to denote either any one of the four treatment assignment variables or a matrix of all four treatment assignment variables (which can be equivalently expressed as one multi-category treatment assignment vector). For a specific treatment assignment vector, we use Z with the appropriate superscript.

To simplify the setup first, let us consider the experiment to be four different experiments where each of four datasets contain observations for only one treatment of interest and all the control observations. In this setup, $Z_i = 0$ denotes an individual in the control group who did not receive any of the four treatments. Also assume no treatment noncompliance, so an individual who is assigned a treatment actually receives a treatment. For any of the four treatments of interest, let $Y_i(z)$ denote the potential outcome for individual i if he receives treatment arm z, where $Y_i(1)$ is the vote outcome for individual i if i receives the treatment and $Y_i(0)$ is the vote outcome for individual i if i receives control. The treatment effect for individual i is simply

$$TE_i = Y_i(1) - Y_i(0)$$

Since an individual only receives either treatment or control, we can only observe at most one of the potential outcomes, a problem commonly referred to as the *fundamental problem of causal inference* (Holland, 1986). In this paper, our main quantity of interest is the (sample) average treatment effect: $\frac{1}{n}\sum_{i=1}^{n} TE_i$.

Up to this point, we have defined treatment simply as either a phone call or one of the mail treatments. However, it is important to give a precise definition of the treatments, especially in the presence of treatment noncompliance. There are several different possible definitions of the treatments of interest. From the ACTU perspective, the relevant treatments of interest may be the the phone call campaign or the mail campaigns themselves. On the other hand, from a more general perspective, the relevant treatments may not be the campaigns themselves, but whether or not an individual *receives and processes* the phone calls or *reads* the mail. To capture the different definitions of the treatments, we adopt the conventional language used in assessing experiments in the presence of treatment noncompliance. We define the treatment effects of the ACTU campaigns as intention-to-treat (ITT) effects. So in our notation, $Y_i(Z_i = 1) - Y_i(Z_i = 0)$ is an individual ITT effect. The effects of actually cooperating in receiving phone calls or reading pieces of mail are the complier average causal effects (CACEs). Both types of treatment effects are of interest. For example, the ACTU or any other political organization may be interested in the overall effect of running this type of campaign, since individual receptiveness of the treatments is largely out of their control, so the ITT effects may be of interest. On the other hand, the CACEs may be of interest since if the ITT effects and the CACEs differ significantly, then organizations may devote more resources to reduce the noncompliance rate.

4.1 Methodological Challenges

We face several methodological challenges given the state of our experiment.

- Multiple treatments
- Missingness in covariates (age)
- Treatment noncompliance
- Non-response/missingness in our outcome variable

We deal with the first two rather simply. With multiple treatments, we first separate the data into four separate experiments as described above to examine the treatment effects for each treatment separately. However, separating the data results in loss of efficiency. In our more complicated parametric models, we combine the datasets back into a single dataset with four separate treatment variables and adjust our models to allow for more than two treatment arms. Due to missingness in the age covariate, we include an age imputation step in our models. The covariates from the post-election survey are too highly missing given survey non-response in our sample to impute.

4.1.1 Treatment Noncompliance

If the treatment effects of interest are the ITT effects, then treatment noncompliance is not a problem. However, if we are interested in the CACEs, then treatment noncompliance becomes an issue. We define compliance here as whether or not a union member receives the treatment in full that the ACTU intended. We assume only one-sided noncompliance, so a union member who is assigned control cannot receive any of the treatments by definition. For the phone call treatments, we define compliance to mean that the union member picked up the phone call and allowed the volunteer to deliver the message in full. Examples of noncompliance would be if the union member has incorrect phone number information on file, never picks up the phone, or picks up the phone but refuses to engage the volunteer or allow the volunteer to finish delivering the message. Table 3 shows the distribution of call results for union members assigned to the phone call treatment in our sample. Union members are classified as compliers if they were recorded as having answered the phone or had meaningful interactions with the volunteers.[16] The result "Unknown" here is simply data entry error where we did not receive a call result coding even though calls were placed. We assume these to be calls placed that were not answered.

TABLE 3: Call Results for Union Members Assigned to Phone Call Treatment

	Call Result	Compliance Status
Answered	409	compliers
Meaningful	383	compliers
Asked for More Info	17	non-compliers
Left Message	208	non-compliers
No Answer	320	non-compliers
Unknown	121	non-compliers
Total	1458	54% compliers

Compliance status is a pre-treatment covariate for each individual. Let D_i denote whether or not individual i received the phone call treatment, so $D_i = 1$ if i was assigned the phone call treatment and the call result was either "answered" or "meaningful" and $D_i = 0$ otherwise. Using the potential outcomes notation, let $D_i(z^p)$ be the call result status for individual i when $Z_i^p = z^p$. A complier is defined as an individual i where $D_i(1) = 1$ and $D_i(0) = 0$. A non-complier is defined as an individual i where $D_i(1) = D_i(0) = 0$. We use the principal stratification framework (Frangakis and Rubin, 2002), and use $S(z_i^p, D_i(z_i^p))$ to denote the strata for groups of individuals with the observed treatment assignment z^p and treatment received status $D(z^p)$ and let C_i denote binary compliance status. In our setup, there are three possible principal strata: $S(1,1)$, $S(1,0)$ and $S(0,0)$. For union members that were assigned the phone call treatment and received the treatment $S(1,1)$, they are classified as compliers ($C_i = 1$). For union members that were assigned the phone call treatment but did not receive the treatment $S(1,0)$, $C_i = 0$. For all union members who either received other treatments or received control $S(0,0)$, their compliance status is unknown since we do not observe what they would have done if they had received the phone call treatment. Therefore, C is a latent variable and our methods attempt to impute C for these individuals in $S(0,0)$. The problem that treatment noncompliance poses is that while Z^p is randomly assigned, D is not randomly assigned. There may be observed or unobserved characteristics that affect the propensity for compliance.

For the mail or mail and robocall treatments, define compliance as receiving and reading the

[16] Meaningful meant that the volunteer felt that the voter's mind was changed by the conversation. Since this was ultimately a subject coding, we decided to code both answered calls and meaningful calls as the member having complied.

pieces of mail and (if applicable) picking up the robocall and listening to the entire message. So we can imagine a spectrum of noncompliance or partial compliance for these treatments. One can receive and read the mail but fail to listen to the robocall or vice versa. One can also receive the mail but throw it into the trash before reading it. Unfortunately, for the mail and mail and robocall treatments, we do not observe and are unable to measure compliance.

In the rest of the paper, the notation for D, $S(Z^p, D)$, and C refer only to treatment received, principal strata, and compliance status relating to the phone call treatment. Compliance status on the phone call treatment may be correlated with compliance status on the mail treatments. For example, union members who answer the phone and receive the phone treatment may be more likely to answer a robocall or read their mail more carefully. Those not reached for the phone call treatment may be less likely to be reached for mail and robocalls as well. Therefore, compliance status C on the phone call treatment may highly inform the fully unobserved compliance status on the other treatments. Our models incorporate this by allowing for information about C on the phone call treatment to interact with the other mail treatments.

4.1.2 Non-Response: Missingness in the Outcome Variable

Missingness in the voting outcome variable also poses a problem since the missingness is not guaranteed to be balanced across treatments. The propensity of a union member to give their vote information in the post-election survey may be correlated with certain variables or even the outcome itself. At best, only using observations with observed outcomes will lead to less efficient estimates and change the target sample or population of interest. At worst, it can lead to biased estimates if the missingness is influenced by the treatment assignment itself. We only observe the outcome for approximately 36% of surveyed members in the phone call treatment. The non-responses are due to inability to reach the union member in the post-election survey, refusal to participate in the survey, incomplete surveys, or refusal to divulge vote information in the survey. We consider all types of non-response to be the same for simplicity. Table 4 shows the extent of non-response among the different treatment assignments.

The distribution of non-response suggests that missingness does not vary significantly across treatment groups. However, missingness may still vary across treatments conditional on observed or unobserved covariates or on the outcome itself.[17] We conduct our analyses by making three

[17] There is some imbalance across treatment groups conditional on the covariates we have in the dataset.

Table 4: Vote Outcome Response by Treatment Assignment

	Phone Call	Glossy mail	Glossy mail + robocall	Formal mail + robocall	Control
Gave Vote (observed)	540	404	246	263	626
Unreached	490	413	235	235	648
Refused to Participate	268	229	132	118	335
Incomplete Survey	21	16	8	3	22
Refused to Divulge	139	102	64	59	165
Proportion Observed	0.37	0.35	0.36	0.39	0.35

different assumptions about the behavior of the missingness and present three different sets of results. The first missing outcome assumption we make is that the outcomes are missing completely at random (MCAR). This assumption implies that the missingness is unrelated to the outcome or any other variable and is akin to flipping a coin. A second, less restrictive assumption about the missingness is the missing at random (MAR) assumption (Rubin, 1976), which implies that the missingness is independent of the outcomes conditional on some observed covariates and treatment assignment. This assumption would be valid if the variables correlated with the missingness patterns were included in our covariates. The third assumption that we make about missingness is the latent ignorability assumption (Frangakis and Rubin, 1999; Barnard et al., 2003), which states that the missingness is independent of the outcomes conditional on the observed covariates, treatment assignment, and latent compliance status. This third assumption suggests that an individual's compliance status informs the pattern of missingness. For example, we may think that compliers (who are more likely to answer phones and are more receptive to ACTU phone calls) may also be more receptive to the post-election survey and more willing to divulge vote information.

In our view, this third assumption is quite realistic for our experiment. It may be the case that none of the assumptions (which are untestable) actually hold and missingness is correlated with the outcome even after conditioning on everything observable. In that situation, all of our models would still be biased. However, suppose we make the assumption that all of our treatments can only affect the Coalition vote in the negative direction and that non-response is correlated with voting for the Coalition. The combination of these two reasonable one-sided effects assumptions suggests that any bias is simply attenuation bias and our estimates will

However, this may be due to either systematic relationships between missingness and treatment assignment or simply due to a relatively small sample size and a relatively large set of covariate strata to check balance on since many of our covariates are factors.

be underestimates of the treatment effects. Consider a simple numerical example in Table 5. Suppose we have 100 individuals and we observe all potential outcomes for each individual as in Table 5a. The treatment effect is -0.1 under fully observed outcomes. Now suppose that Coalition voters are less likely to respond than Progressive voters.[18] In Table 5b, the response rate for Coalition voters is 10% and the response rate for Progressive voters is 50%. Using only the observed outcomes, we can see that our estimated treatment effect is now -0.06. Using this very simple example, we can see that even if missingness is dependent upon the actual outcomes, if we make reasonable assumptions about the direction of the treatment effect and missingness, the bias induced by missing outcomes will only underestimate our treatment effects.

TABLE 5: A Simple Example of Attenuation Bias Under One-Sided Assumptions

	Under Treatment	Under Control	Treatment Effect
Coalition voters	50	60	
Progressive voters	50	40	
Proportion Coalition voters	0.5	0.6	-0.1

(A) Fully Observed Outcomes

	Under Treatment	Under Control	Treatment Effect
Coalition voters	5	6	
Progressive voters	25	20	
Proportion Coalition voters	0.17	0.23	-0.06

(B) Observed Outcomes with 10% response rate for Coalition voters and 50% response rate for Progressive voters

We present a set of three results in the next sections starting from the simplest assumptions about outcome missingness and treatment noncompliance. Let R_i be a binary response indicator for Y_i, where $R_i = 1$ if Y_i is observed and $R_i = 0$ if Y_i is missing, and let $R_i(z)$ denote the potential response indicator for $z \in \{0, 1\}$ for any of our four treatment assignment indicators.

4.2 Assumptions

We begin by stating the assumptions we make for our results. Table 6 reviews the notation that we use in the paper. The first two assumptions we make are standard for most experiments.

[18]Since union members know that the ACTU is conducting the post-election survey and the ACTU generally sides with the Progressive parties, it may be reasonable to assume that Coalition voters would be less likely to divulge vote information or participate in ACTU surveys.

TABLE 6: Notation Used in the Paper

Variable	Description
Y	outcome variable for Coalition vote
X	matrix of covariates
Z^p	binary treatment assignment for phone call
Z^g	binary treatment assignment for glossy mail
Z^{gr}	binary treatment assignment for glossy mail + robocall
Z^{fr}	binary treatment assignment for formal mail + robocall
D	binary treatment received indicator for phone call treatment only
C	binary indicator for latent compliance status for phone call treatment only
$S(Z^p, D)$	principal strata for phone call treatment
R	binary response indicator for whether outcome is observed

Assumption 1 (Randomization of Treatment):

$$(Y_i(1), Y_i(0), R_i(1), R_i(0)) \perp\!\!\!\perp Z_i | X_i$$

Assumption 2 (Stable Unit Treatment Value Assumption): i) treatment assignment in one unit does not affect the potential outcomes of another unit, and ii) there is only one version of each treatment.

Assumption 1 is satisfied since all treatment assignments were randomized. Assumption 2 implies that union members do not discuss or affect each other's vote choices after receiving a certain treatment. It also implies that each of the types of treatment were identical across all individuals who received them.[19]

For our three sets of results, we make three different sets of assumptions about missing outcomes and treatment noncompliance.

Assumption 3a (Outcomes are Missing Completely at Random and there is no treatment noncompliance)

$$p(R(1), R(0))|Y(1), Y(0), Z, X) = p(R(1), R(0))$$

Assumption 3b (Outcomes are Missing at Random and there is no treatment non-

[19]SUTVA may technically be violated in our experiment if for example, the volunteer phone calls were of different quality across different volunteers. However, we define treatment loosely here to simply mean the mail or phone call campaigns, in which case these concerns are mitigated by definition.

compliance)

$$p(R(1), R(0))|Y(1), Y(0), Z, X) = p(R(1), R(0)|Z, X)$$

Assumption 3c (Latent Ignorability of outcomes with treatment noncompliance for phone calls)

$$p(R(1), R(0))|Y(1), Y(0), Z, X, C) = p(R(1), R(0)|Z, X, C)$$

For the phone call treatment noncompliance, we also make several standard assumptions that are commonly assumed in the literature.

Assumption 4 (One-sided noncompliance; no always-takers)

$$D_i(0) = 0 \text{ for all } i$$

Assumption 5 (Monotonicity; no defiers)

$$D_i(1) \geq D_i(0) \text{ for all } i$$

Assumption 6 (Compound exclusion for noncompliers)

$$p(Y(1), R(1)|Z, X, C = 0) = p(Y(0), R(0)|Z, X, C = 0)$$

Assumptions 4 and 5 are satisfied by the experimental design and implementation. No union member who was not assigned a campaign phone call received one. Assumption 6 generalizes the standard exclusion restriction and allows us to focus on estimating CACE. The assumption states that for noncompliers (never-takers), treatment assignment has no effect either on the outcome or the probability of outcome non-response. This seems to be a plausible assumption for both outcome and outcome non-response since noncompliers are unlikely to cooperate or receive calls regardless of whether they get a phone call treatment or not. However, possible violations of this assumption include scenarios where a phone call angers a union member enough that he refuses to let the phone call volunteer finish the call at the treatment stage (noncompliance) and also subsequently increases the probability of the member voting Coalition (out of anger) or lowers the probability of response in the post-election survey. Such a situation is unlikely, especially since we removed callers who indicated they did not want to be called from the post-election call list.

4.3 Results with Assumptions 1, 2, 3a: MCAR Outcomes with No Treatment Noncompliance

If we assume the outcomes are missing completely at random we can use only the complete cases from the post-election survey without inducing bias in the estimated treatment effects. We subset the data here to only those who completed the post-election survey and revealed their vote information. We consider each of the four treatments separately by comparing those who were assigned each treatment to the relevant control group.[20] Since we first assume no treatment noncompliance in this set of results, the treatment effects are simply average treatment effects (ATEs). However, we still refer to them as ITT effects for the purpose of consistency.

Given randomized treatment, a simple difference in means estimator is unbiased for our ITT effects. Figure 2 shows our difference in means estimates of the ITT effects along with 95% confidence intervals. Individuals assigned to receive a phone call are approximately 7.5 percentage points less likely to vote for the Coalition than those who were not assigned any treatment. This effect is moderately large and statistically significant. On the other hand, although all mail treatments have negative point estimates, none of them are statistically different from zero.

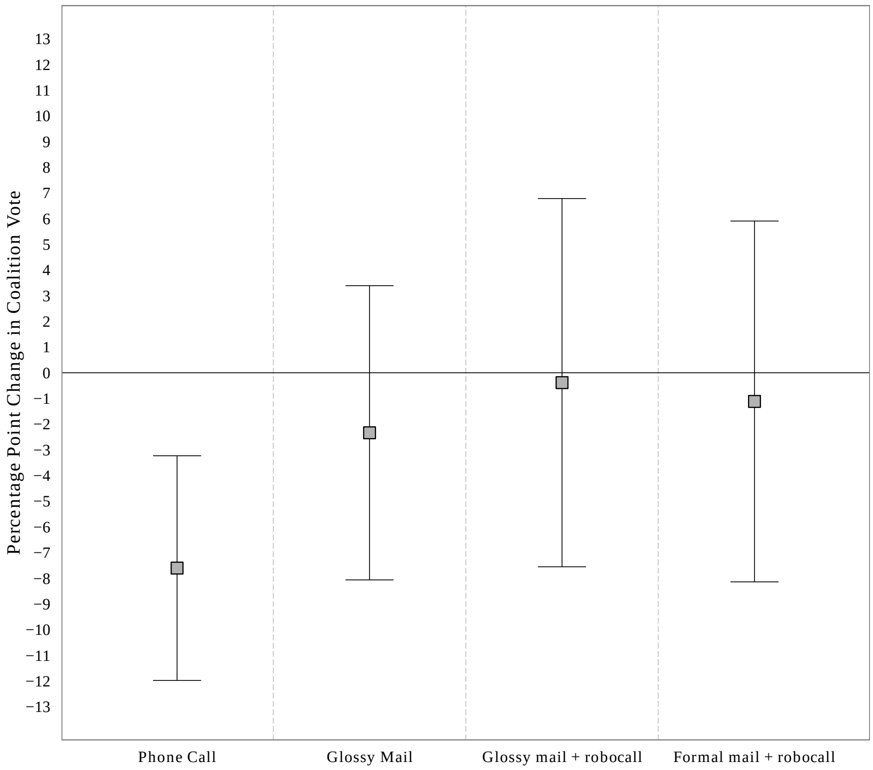
One concern with the simple difference in means estimates is that although randomization assures balance on covariates in expectation, within any single randomization and with a relatively small sample size, the randomization may have simply been "unlucky" and treatment imbalance may exist. As a result, we also use regression adjustment and present OLS estimates with robust standard errors (Samii and Aronow, 2012).[21] We use the following regression specification:

$$Y_i = \beta_0 + \beta_1 Z_i^p + \beta_2 Z_i^g + \beta_3 Z_i^{gr} + \beta_4 Z_i^{fr} + \gamma X_i + \epsilon_i$$

[20] For the phone call treatments that occurred in both Group A and Group B, the relevant control group is all control observations. For the glossy mail treatment, the control group is limited to controls in Group A. For the glossy mail and robocalls and formal mail and robocalls treatments, the control group is controls in Group B.

[21] For a discussion on the merits and pitfalls of regression adjustment for experiments, see Freedman (2008*a*,*b*); Lin (2013).

Figure 2: Simple Difference in Means Estimates of ITT Effects of Treatments on the Coalition Vote



Notes: Point estimates denote the average difference between treatment and control groups in the sample of complete cases. Estimates are displayed within 95% confidence intervals.

where X_i is a matrix of covariates that includes past party, labor contact, liberal contact, electorate, union, gender, and age range.[22] $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$ are the regression adjusted ITT effect estimates. Table 7 compares the different estimates of the ITT effects using simple difference in means and regression adjustment. With regression adjustment, the ITT for the phone call treatment is slightly smaller but still statistically significant. Formal mail becomes slightly more effective, although all of the mail treatments are still statistically different from zero.

We also conduct a sensitivity analysis using exact matching on all covariates since they are all categorical. Using exact matching allows us to condition on the covariates without making functional form assumptions. The trade-off with exact matching is that it discards observations without common support and changes the target sample/population of interest. Exact matching on all covariates significantly decreases the number of observations from over

[22]We use age range here instead of the age variable since age range is fully observed for our sample of completed post-election surveys.

Table 7: Difference in Means and Regression Adjusted Estimates of ITT Effects

	Difference in Means	Regression Adjusted
Phone Call	-7.60	-5.18
	(-11.98, -3.23)	(-9.37, -0.99)
Glossy Mail	-2.34	-1.77
	(-8.06, 3.39)	(-6.65, 3.10)
Glossy Mail + Robocall	-0.38	-0.67
	(-7.55, 6.78)	(-7.05, 5.70)
Formal Mail + Robocall	-1.12	-2.83
	(-8.14, 5.91)	(-8.75, 3.09)

Notes: Dependent variable is the binary indicator for Coalition vote. Regression estimates use the White robust "HC2" standard errors. Estimates are displayed with approximate 95% confidence intervals.

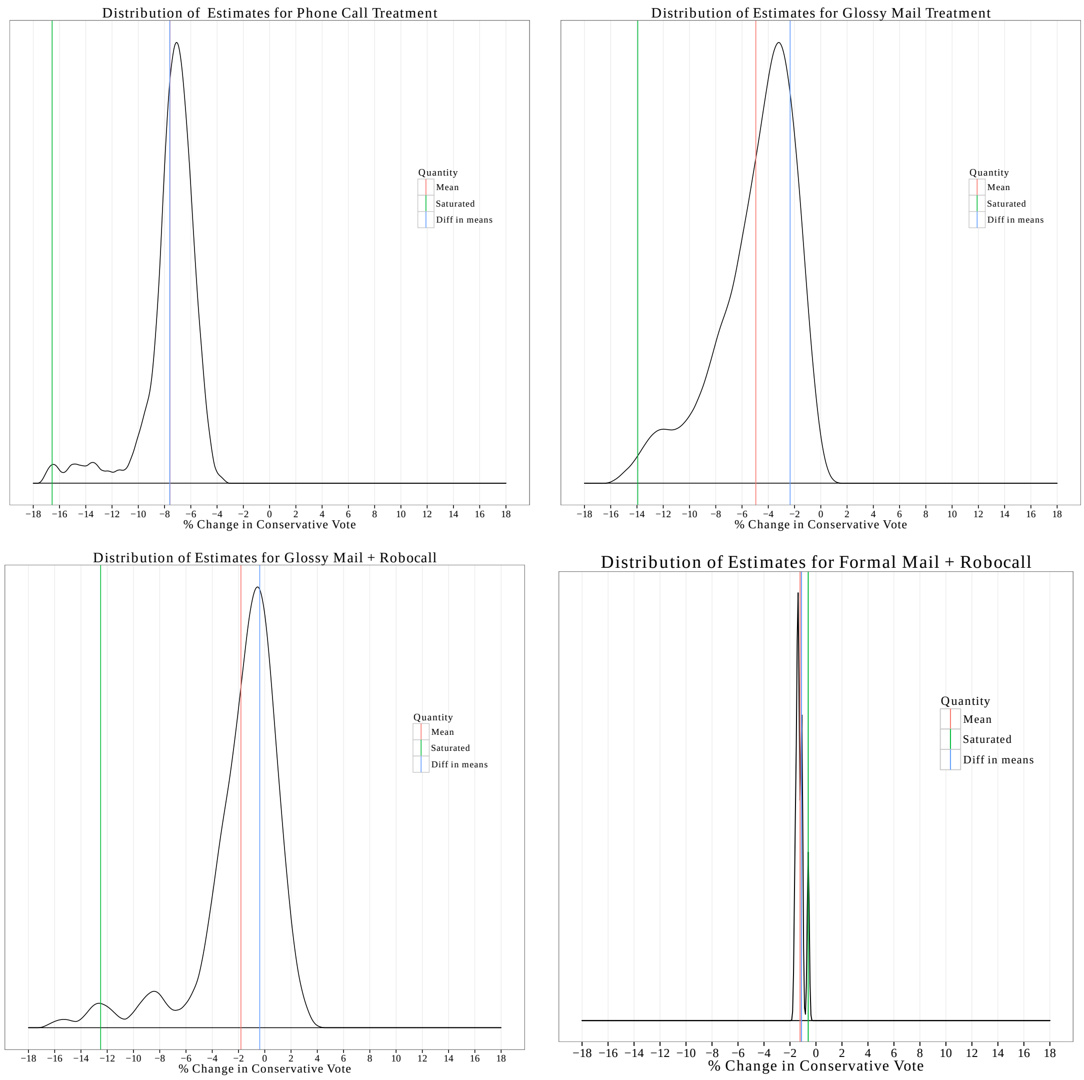
2,000 complete cases to less than 300 complete cases with common support. We vary the trade-off between sample size and exact matching by exact matching on all 127 possible combinations of covariates. Figure 3 shows the distributions of estimated ITT effects from exact matching on all the different covariate combinations. The results suggest the phone call treatment ITT effects are consistently negative. The mail treatments are distributed closer to zero, confirming our previous results about the ineffectiveness of mail treatments.

4.4 Results with Assumptions 1, 2, 3b: MAR Outcomes with No Treatment Noncompliance

We now relax the missing completely at random assumption to allow the outcome variable to be missing at random. This implies that the observed covariates can predict the missingness pattern and the missingness pattern is independent of the outcome itself after conditioning on the covariates and treatment assignment. We no longer restrict our analysis to the complete cases; instead, the entire sample of 5,781 undecided voters is used. The matrix of covariates X then includes electorate, union, gender, and age. We also add a column of 1s to the X matrix for the intercept. We drop the covariates from the post-election survey since missingness is too high to be informative. We also include an imputation model for the partially missing age variable, which we also assume to be missing at random conditional on the other covariates and Y.

To get correct inferences in the presence of data missing at random, we need to first impute the missing data using information from the observed covariates and then draw inferences using this imputed data. We use a Bayesian model to impute the missing values of Y and X_{age},

Figure 3: Distribution of ITT Effects from Exact Matching on 127 Different Covariate Combinations



Notes: The red lines are the means of the distributions. The green lines are the ITT estimates exact matching on all seven covariates (saturated model). The blue lines are the original difference in means estimates.

simulating from the posterior via a Gibbs sampler.

First, we model the age variable with a Bayesian linear regression of age on covariates and the outcome to get parameters ξ, γ_{age} and σ^2_{age}.

$$p(X_i^{age}|Y_i, X_i^{eg}, \xi, \gamma_{age}, \sigma^2_{age}) \quad \sim \quad \mathcal{N}(Y_i\xi + X_i^{eg}\gamma_{age}, \sigma^2_{age})$$

where X_i^{eg} is the covariate vector without the age and union variables (leaving only the electorate and gender variables and 1 for the intercept). We omit union as a predictor of age because age is completely missing for some unions.

Second, we model vote choice given the covariates using a probit model.

$$\begin{aligned} Pr(Y_i = 1|Z_i, X_i, \beta, \gamma) &= \Phi(\beta_p Z_i^p + \beta_g Z_i^g + \beta_{gr} Z_i^{gr} + \beta_{fr} Z_i^{fr} + X_i \gamma) \\ &= \Phi(Z_i \beta + X_i \gamma) \end{aligned}$$

where Z_i without a superscript is a vector of binary treatment assignments and β and γ are vectors of coefficients.

Improper multivariate Gaussian uniform priors centered at zero an are used for all the coefficients and an Inverse Gamma prior with shape and scale equal to 0.0005 is used for σ^2_{age}. The algorithm works by the method of data augmentation where missing data are "imputed" at each iteration. We then use the imputations to calculate ITT effects (ITT$_p$, ITT$_g$, ITT$_{gr}$, and ITT$_{fr}$). Computational details are presented in Appendix B.

Column 3 of Table 8 shows the estimated ITT effects using the Bayesian model assuming MAR. Point estimates are posterior means with 95% central credible intervals. The results from the Bayesian model with the MAR assumption shows that the phone call treatment once again has an effect of decreasing the Coalition vote by about seven percentage points. The effects for glossy mail and robocall and formal mail and robocall are not stronger. This suggests the robocalls have an added effect and that perhaps formal mail works slightly better than glossy mail. However, as before, only the phone call treatment is significant and the mail treatments have effects indistinguishable from zero according to standard interpretations. Note that the credible intervals from the Bayesian model are generally tighter than the confidence intervals from before, reflecting the fact that we are no longer limiting the sample to complete cases.

4.5 Results with Assumptions 1, 2, 3c, 4-6: Latent Ignorability with Treatment Noncompliance

The final set of results brings in treatment noncompliance on the phone call treatment and a less restrictive missing data assumption of latent ignorability. The difference between the latent ignorability assumption and the MAR assumption is that latent ignorability conditions on latent compliance status on the phone call treatment, which is a more reasonable assumption as previously discussed. Our model is similar to those in Barnard et al. (2003) and Horiuchi, Imai and Taniguchi (2007).

Table 8: Difference in Means, Regression Adjusted, and Bayesian (MAR) Estimates of ITT Effects

	Difference in Means (1)	Regression Adjusted (2)	Bayesian (3)
Missingness	MCAR	MCAR	MAR
ITT_p	-7.60	-5.18	-7.14
	(-11.98, -3.23)	(-9.37, -0.99)	(-10.56, -3.51)
ITT_g	-2.34	-1.77	-1.83
	(-8.06, 3.39)	(-6.65, 3.10)	(-6.34, 2.53)
ITT_{gr}	-0.38	-0.67	-2.23
	(-7.55, 6.78)	(-7.05, 5.70)	(-8.04, 3.52)
ITT_{fr}	-1.12	-2.83	-3.12
	(-8.14, 5.91)	(-8.75, 3.09)	(-8.58, 2.01)

Notes: Dependent variable is the binary indicator for Coalition vote. Regression estimates use the White robust "HC2" standard errors. For difference in means and regression adjusted, estimates are displayed with approximate 95% confidence intervals. For Bayesian specifications, point estimates are posterior means with 95% central credible intervals.

The model follows the same setup as the previous model, except that we also impute the latent compliance status for those in $S(0,0)$, whose compliance status is unknown. Compliance here refers only to compliance on the phone call treatment because we do not observe compliance on the mail treatments. The observations assigned to the phone call treatment either belong to the strata of compliers $S(1,1)$ or noncompliers $S(1,0)$. We first model the baseline compliance rate using a probit model.

$$Pr(C_i = 1 | X_i, \gamma_C) = \Phi(X_i \gamma_C)$$

where X_i includes a column of 1s, union, electorate, gender, and age.

Second, our age model remains a Bayesian linear regression of age on the outcome and covariates excluding union, but we add in compliance status as a covariate.

$$p(X_i^{age} | Y_i, C_i, X_i^{eg}, \xi, \zeta, \gamma_{age}, \sigma_{age}^2) \sim \mathcal{N}(Y_i \xi + C_i \zeta + X_i^{eg} \gamma_{age}, \sigma_{age}^2)$$

Third, our model on vote choice given the covariates changes to the following. Define Z_i^m to be a binary indicator for whether i is assigned any of the three mail treatments.

$$\begin{aligned}
Pr(Y_i = 1|Z_i, X_i, C_i, \alpha, \beta, \gamma) &= \Phi(\eta_i) \\
\eta_i &= \alpha_1 C_i Z_i^p + \alpha_2 C_i (1 - Z_i^p) + \beta_g Z_i^g C_i + \beta_{gr} Z_i^{gr} C_i \\
&\quad + \beta_{fr} Z_i^{fr} C_i + \beta_m Z_i^m + X_i \gamma
\end{aligned}$$

To see how the covariates and compliance status affect the outcome in this model, consider Table 9 where we show the values of η_i given treatment assigned and latent compliance status on the phone call treatment.

TABLE 9: Values of η_i Given Treatment Assigned and Compliance Status on Phone Call Treatment

	$C_i = 1$	$C_i = 0$
$Z_i^p = 1$	$\alpha_1 + X_i\gamma$	$X_i\gamma$
$Z_i^g = 1$	$\alpha_2 + \beta_g + \beta_m + X_i\gamma$	$\beta_m + X_i\gamma$
$Z_i^{gr} = 1$	$\alpha_2 + \beta_{gr} + \beta_m + X_i\gamma$	$\beta_m + X_i\gamma$
$Z_i^{fr} = 1$	$\alpha_2 + \beta_{fr} + \beta_m + X_i\gamma$	$\beta_m + X_i\gamma$
Controls	$\alpha_2 + X_i\gamma$	$X_i\gamma$

First, note that α_1 is the difference between complying and not complying to the phone call treatment on the outcome. Note that for noncompliers, the outcome is the same regardless of whether they are assigned the phone call treatment or control, reflecting the compound exclusion restriction. α_2 represents the baseline difference in outcomes between compliers and noncompliers. Recall that noncompliance is only on the phone call treatment, so β_m reflects the fact that compliance status on the phone call treatment does not perfectly correlate with unobserved compliance status on mail. We consider the correlation between the two compliance statuses to be likely positive, but not perfect. Finally, each mail treatment specific β reflects an additional effect of possibly reading the mail. By setting up the outcome model in this way, we hope to capture the fact that compliance status on phone calls may somewhat inform the unobserved mail compliance statuses. Priors for the model remain the same as the previous model. Computational details are presented in Appendix C.

Column 4 in Table 10 shows the posterior mean point estimates and 95% central credible intervals for ITT effects and the CACE for the phone call treatment. We see again that the ITT effect for phone calls drops slightly to around a 6 percentage point difference. Using latent compliance status as a covariate reduces the width of the credible interval as we get more information about the individuals. The CACE for the phone call treatment is about a 10 percentage point decrease in Coalition voting. This is a significant effect that warrants great attention. For individuals who actually pick up the phone and receive the message from the volunteers, their likelihood of voting for the Coalition goes down by 10 percentage points. The fact that the CACE is larger than the ITT effect makes sense, as the ITT reflects both individuals who receive the message and individuals who do not. The results for the treatment effects for phone call compliers who receive a mail treatment (not shown) are slightly larger than for those noncompliers for the phone call treatment. This is suggestive evidence that compliance status on phone calls informs unobserved compliance status on mail treatments.

TABLE 10: Difference in Means, Regression Adjusted, and Bayesian Estimates of ITT and CACE Effects

	Difference in Means (1)	Regression Adjusted (2)	Bayesian (3)	Bayesian (4)
Missingness	MCAR	MCAR	MAR	LI
Noncompliance	No	No	No	Yes
ITT_p	-7.60	-5.18	-7.14	-5.94
	(-11.98, -3.23)	(-9.37, -0.99)	(-10.56, -3.51)	(-8.80, -3.06)
ITT_g	-2.34	-1.77	-1.83	-1.73
	(-8.06, 3.39)	(-6.65, 3.10)	(-6.34, 2.53)	(-5.71, 2.24)
ITT_{gr}	-0.38	-0.67	-2.23	-1.65
	(-7.55, 6.78)	(-7.05, 5.70)	(-8.04, 3.52)	(-6.56, 3.11)
ITT_{fr}	-1.12	-2.83	-3.12	-2.33
	(-8.14, 5.91)	(-8.75, 3.09)	(-8.58, 2.01)	(-7.11, 2.30)
CACE_p				-9.86
				(-14.31, -5.36)

Notes: Dependent variable is the binary indicator for Coalition vote. Regression estimates use the White robust "HC2" standard errors. For difference in means and regression adjusted, estimates are displayed with approximate 95% confidence intervals. For Bayesian specifications, point estimates are posterior means with 95% central credible intervals.

5 Conclusion

In this paper we described a series of field experiments designed to evaluate the effectiveness of various campaign strategies used by the ACTU during the 2013 election. To our knowledge this is the first large scale voter persuasion experiment conducted by an interest group in a compulsory voting system. The results, however, are similar to those found in GOTV experiments and persuasion campaigns by interest groups and political parties in non-compulsory voting systems.

We find that volunteer phone calls are the most effective means of influencing voter behavior. The phone call campaign itself reduced the Coalition vote by about five to seven percentage points. However, looking at the actual people who receive the intended phone call treatment, we find that the effect increases to about ten percentage points. This is a strong effect and suggests that campaigns should put energy and effort into actually reaching people personally. We fail to find sufficient evidence to conclude that direct mail has a consistent effect on voter behavior. There is some suggestive evidence that formal mail is better than glossy mail, but none of the mail treatment effects are significant.

This paper has made a few contributions. First, this is the first field experiment we know of that was conducted in a compulsory voting system. This allows us to separate persuasion effects from turnout effects. Second, this is also the first partisan persuasion experiment that we know of conducted in a national federal election. Though smaller elections offer the ability to control different aspects of the experiment, a large election on the national stage has implications that far exceed smaller elections. On the methodological front, we adapt existing methods to help deal with problems of combining multiple treatment arms, missing outcome and covariate data, and treatment noncompliance.

References

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

Arceneaux, Kevin. 2005. "Using Cluster Randomized Field Experiments to Study Voting Behavior." *Annals of the American Academy of Political and Social Science* 601:169–179.

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* 2:43–65.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill and Donald B. Rubin. 2003. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York (with Discussion)." *Journal of the American Statistical Association* 98(462):299–311.

Eldersveld, Samuel. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Political Science Review* 50(1):154–165.

Frangakis, Constantine E. and Donald B. Rubin. 1999. "Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes." *Biometrika* 86(2):365–379.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(1):21–29.

Freedman, David A. 2008*a*. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* 2(1):176–196.

Freedman, David A. 2008*b*. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40(2):180–193.

Gerber, Alan S. 2004. "Does Campaign Spending Work? Field Experiments Provide Evidence and Suggest New Theory." *American Behavioral Scientist* 47(5):541–574.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Convassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94(3):656–663.

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

Gosnell, Harold. 1927. *Getting-Out-the-Vote: An Experiment in the Stimulation of Voting.* Chicago: University of Chicago Press.

Hillygus, Sunshine D. 2005. "Campaign Effects and the Dynamics of Turnout Intention in Election 2000." *Journal of Politics* 67(1):50–68.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Horiuchi, Yusaku, Kosuke Imai and Naoko Taniguchi. 2007. "Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment." *American Journal of Political Science* 51(3):669–687.

Issenberg, Sasha. 2012. *The Victory Lab: The Secret Science of Winning Campaigns.* Crown Publishers.

Kramer, Gerald H. 1970. "The Effects of Precinct-Level Canvassing on Voting Behavior." *Public Opinion Quarterly* 34:560–572.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *Annals of Applied Statistics* 7(1):295–318.

Miller, Roy E. and Dorothy L. Robyn. 1975. "A Field Experimental Study of Direct Mailing Congressional Primary Campaign: What Lasts Until Election Day?" *Experimental Study of Politics* 5:1–37.

Nickerson, David W. 2005. "Partisan Mobilization Using Volunteer Phone Banks and Door Hangers." *Annals of the American Academy of Political and Social Science* 601:41–65.

Rosenstone, Steven J. and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America.* New York: Macmillian.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63(3):581–592.

Samii, Cyrus and Peter M. Aronow. 2012. "On Equivalencies Between Design-based and Regression-based Variance Estimators for Randomized Experiments." *Statistics and Probability Letters* 82:365–370.

Vavreck, Lynn, Constantine J. Spiliotes and Linda L. Fowler. 2002. "The Effects of Retail Politics in the New Hampshire Primary." *American Journal of Political Science* 46(3):595–610.

Wielhouwer, Peter W. and Brad Lockerbie. 1994. "Party Contacting and Political Participation 1952-90." *American Journal of Political Science* 38(1):211–229.

Appendix A

TABLE 11: Election results for Group A target electorates with two party preferred (TPP) percentages

Electorate	State/Territory	ALP Percentage	Coalition Percentage	Swing from 2010
Adelaide	SA	53.95	46.05	-3.57
Barton	NSW	49.69	50.31	-7.17
Bonner	QLD	46.31	53.69	-0.87
Braddon	TAS	47.44	52.56	-10.04
Brisbane	QLD	45.72	54.28	-3.15
Corangamite	VIC	46.06	53.94	-4.22
Deakin	VIC	46.82	53.18	-3.78
Hasluck	WA	45.13	54.87	-4.30
Herbert	QLD	43.83	56.17	-4.00
Hindmarsh	SA	48.11	51.89	-7.97
La Trobe	VIC	45.99	54.01	-5.67
McMahon	NSW	55.32	44.68	-2.49
Moreton	QLD	51.55	48.45	0.42
Page	NSW	47.48	52.52	-6.71
Parramatta	NSW	50.57	49.43	-3.80
Solomon	NT	48.60	51.40	0.35

Notes: Data are taken from the Australian Electoral Commission's Virtual Tally Room. Coalition refers to either the Liberal or National Party. The two-party preferred results show the percentage received for the Australian Labor Party (ALP) or Coalition candidate in each electorate. The swing is the difference between the percentage of first party preference votes received by the incumbent (ALP) at the 2013 election and the percentage of the first preference votes received by their party at the 2010 election.

Table 12: Election results for Group B target electorates with two party preferred (TPP) percentages

Electorate	State/Territory	ALP Percentage	Coalition Percentage	Swing from 2010
Banks	NSW	48.17	51.83	-3.28
Bass	TAS	45.96	54.04	-10.78
Capricornia	QLD	49.23	50.77	-4.45
Dawson	QLD	42.42	57.58	-5.15
Eden-Monaro	NSW	49.39	50.61	-4.85
Flynn	QLD	43.47	56.53	-2.95
Forde	QLD	45.62	54.38	-2.75
Franklin	TAS	55.09	44.91	-5.73
Gilmore	NSW	47.35	52.65	2.67
Greenway	NSW	52.98	47.02	2.10
Lindsay	NSW	47.01	52.99	-4.11
Longman	QLD	43.08	56.92	-5.00
Lyons	TAS	48.78	51.22	-13.51
Macquarie	NSW	45.52	54.48	-3.22
Robertson	NSW	47.00	53.00	-4.00
Swan	WA	43.47	56.53	-4.00

Notes: Data are taken from the Australian Electoral Commission's Virtual Tally Room. Coalition refers to either the Liberal or National Party. The two-party preferred results show the percentage received for the Australian Labor Party (ALP) or Coalition candidate in each electorate. The swing is the difference between the percentage of first party preference votes received by the incumbent (ALP) at the 2013 election and the percentage of the first preference votes received by their party at the 2010 election.

Appendix B: Algorithm for Bayesian Model with MAR Outcome Data and No Treatment Noncompliance

The Gibbs sampling algorithm starts with initial values for the parameters ($\xi^{(0)}, \gamma_{age}^{(0)}$, $\sigma^{2(0)}$, $\gamma^{(0)}$, $\beta^{(0)}$) and missing data ($Y^{(0)}$, $X^{age(0)}$). For observations with Y or X^{age} observed, $Y^{(0)}$ and $X^{age(0)}$ are set to their observed values. We then proceed via the following steps at iteration t.

1. For observations j with age observed, set $X_j^{age(t)} = X_j^{age(t-1)}$. For observations i with age missing, sample $X_i^{age(t)}$ from the Normal distribution with mean $Y_i^{(t-1)}\xi + X_i^{eg}\gamma_{age}^{(t-1)}$ and variance $\sigma_{age}^{2(t-1)}$. However, we also observe in the post-election survey that some respondents give an age range who otherwise have age missing. We take these age ranges to be constraints.[23] For any $X_i^{age(t)}$ that does not satisfy an observed age range constraint, redraw $X_i^{age(t)}$ from the observed empirical distribution of ages within that age range.

2. For observations j with vote outcome observed, set $Y_j^{(t)} = Y_j^{(t-1)}$. For observations i with vote outcome missing, sample $Y_i^{(t)}$ from the Bernoulli distribution with probability $\Phi(Z_i\beta^{(t-1)} + X_i\gamma^{(t-1)})$.

3. Draw the ITT effects as follows.

$$\begin{aligned}
\text{ITT}_p^{(t)} &= \frac{\sum_i Z_i^p Y_i^{(t)}}{\sum_i Z_i^p} - \frac{\sum_i (1-Z_i^p) Y_i^{(t)}}{\sum_i (1-Z_i^p)} \\
\text{ITT}_g^{(t)} &= \frac{\sum_i Z_i^g Y_i^{(t)}}{\sum_i Z_i^g} - \frac{\sum_i (1-Z_i^g) Y_i^{(t)}}{\sum_i (1-Z_i^g)} \text{ for all } i \text{ in Group A} \\
\text{ITT}_{gr}^{(t)} &= \frac{\sum_i Z_i^{gr} Y_i^{(t)}}{\sum_i Z_i^{gr}} - \frac{\sum_i (1-Z_i^{gr}) Y_i^{(t)}}{\sum_i (1-Z_i^{gr})} \text{ for all } i \text{ in Group B} \\
\text{ITT}_{fr}^{(t)} &= \frac{\sum_i Z_i^{fr} Y_i^{(t)}}{\sum_i Z_i^{fr}} - \frac{\sum_i (1-Z_i^{fr}) Y_i^{(t)}}{\sum_i (1-Z_i^{fr})} \text{ for all } i \text{ in Group B}
\end{aligned}$$

4. Given the updated outcome variable $Y_i^{(t)}$ and updated age covariate $X_i^{age(t)}$, perform the Bayesian probit regression for the outcome model and Bayesian Normal linear regression for the age model again to get new draws of the model parameters ($\xi^{(t)}$, $\gamma_{age}^{(t)}$, $\sigma^{2(t)}$, $\gamma^{(t)}$, $\beta^{(t)}$)

We run this algorithm to 5,000 iterations with 100 iterations of burn-in. The typical convergence checks were performed as well.

[23] We constrain all ages to be between 18 and 83, the youngest and oldest age in our sample.

Appendix C: Algorithm for Bayesian Model with Latent Ignorability and Treatment Noncompliance

The algorithm first starts with initial values for the parameters ($\gamma_C^{(0)}, \xi^{(0)}, \zeta^{(0)}, \gamma_{age}^{(0)}$, $\sigma_{age}^{2(0)}$, $\gamma^{(0)}$, $\alpha^{(0)}$, $\beta^{(0)}$) and missing data ($Y^{(0)}$, $X^{age(0)}$). For observations with Y or X^{age} observed, $Y^{(0)}$ and $X^{age(0)}$ are set to their observed values. For observations in $S(1,1)$, $C^{(0)}$ is set to 1. For observations in $S(1,0)$, $C^{(0)}$ is set to 0. We then proceed via the following steps at iteration t.

1. For observations j with age observed, set $X_j^{age(t)} = X_j^{age(t-1)}$. For observations i with age missing, sample $X_i^{age(t)}$ from the Normal distribution with mean $Y_i^{(t-1)}\xi + C_i\zeta + X_i^{eg}\gamma_{age}^{(t-1)}$ and variance $\sigma_{age}^{2(t-1)}$. For any $X_i^{age(t)}$ that does not satisfy an observed age range constraint, redraw $X_i^{age(t)}$ from the observed empirical distribution of ages within that age range.

2. For observations j with vote outcome observed, set $Y_j^{(t)} = Y_j^{(t-1)}$. For observations i with vote outcome missing, sample $Y_i^{(t)}$ from the Bernoulli distribution with probability $\Phi(\nu_i)$.

3. For observations j in $S(1,1)$ or $S(1,0)$, set $C_j^{(t)} = C_j^{(t-1)}$. For observations i in $S(0,0)$, sample $C_i^{(t)}$ from the Bernoulli distribution with probability

$$Pr(C_i = 1|Z_i^p = 0) = \frac{\lambda_i[Y_i\omega_i + (1-Y_i)(1-\omega_i)]}{\lambda_i[Y_i\omega_i + (1-Y_i)(1-\omega_i)] + (1-\lambda_i)[Y_i\pi_i + (1-Y_i)(1-\pi_i)]}$$

where

$$\begin{aligned}
\lambda_i &= Pr(C_i = 1) \\
&= \Phi(X_i\gamma_C) \\
\\
\omega_i &= Pr(Y_i = 1|C_i = 1, Z_i^p = 0) \\
&= \Phi(\alpha_2 + \beta_g Z_i^g + \beta_{gr} Z_i^{gr} + \beta_{fr} Z_i^{fr} + \beta_m Z_i^m + X_i\gamma) \\
\\
\pi_i &= Pr(Y_i = 0|C_i = 1, Z_i^p = 0) \\
&= \Phi(\beta_m Z_i^m + X_i\gamma)
\end{aligned}$$

4. Draw the ITT effects as follows.

$$\begin{aligned}
\text{ITT}_p^{(t)} &= \frac{\sum_i Z_i^p Y_i^{(t)}}{\sum_i Z_i^p} - \frac{\sum_i (1-Z_i^p) Y_i^{(t)}}{\sum_i (1-Z_i^p)} \\
\text{ITT}_g^{(t)} &= \frac{\sum_i Z_i^g Y_i^{(t)}}{\sum_i Z_i^g} - \frac{\sum_i (1-Z_i^g) Y_i^{(t)}}{\sum_i (1-Z_i^g)} \text{ for all } i \text{ in Group A} \\
\text{ITT}_{gr}^{(t)} &= \frac{\sum_i Z_i^{gr} Y_i^{(t)}}{\sum_i Z_i^{gr}} - \frac{\sum_i (1-Z_i^{gr}) Y_i^{(t)}}{\sum_i (1-Z_i^{gr})} \text{ for all } i \text{ in Group B} \\
\text{ITT}_{fr}^{(t)} &= \frac{\sum_i Z_i^{fr} Y_i^{(t)}}{\sum_i Z_i^{fr}} - \frac{\sum_i (1-Z_i^{fr}) Y_i^{(t)}}{\sum_i (1-Z_i^{fr})} \text{ for all } i \text{ in Group B}
\end{aligned}$$

5. Draw the CACE effect for the phone call treatment as follows.

$$\text{CACE}_p^{(t)} = \frac{\sum_i Z_i^p Y_i^{(t)} C_i^{(t)}}{\sum_i Z_i^p C_i^{(t)}} - \frac{\sum_i (1 - Z_i^p) Y_i^{(t)} C_i^{(t)}}{\sum_i (1 - Z_i^p) C_i^{(t)}}$$

6. Given the updated outcome variable $Y_i^{(t)}$, updated age covariate $X_i^{age(t)}$, and updated compliance status $C_i^{(t)}$, perform the Bayesian probit regression for the outcome and compliance models and Bayesian Normal linear regression for the age model again to get new draws of the model parameters $(\gamma_C^{(t)}, \xi^{(t)}, \zeta^{(t)}, \gamma_{age}^{(t)}, \sigma_{age}^{2(t)}, \gamma^{(t)}, \alpha^{(t)}, \beta^{(t)})$

We run this algorithm to 5,000 iterations with 100 iterations of burn-in. The typical convergence checks were performed as well.