The importance of statistical methodology for analyzing data from field experimentation: Evaluating voter mobilization strategies

Abstract

Field experimentation is making its way back into the toolkit of political scientists. Gerber and Green have led this important methodological development that is likely to improve causal inferences in political science research. However, they believe that field experiments only require "rudimentary data analysis." Countering this claim, I use Gerber and Green's voter mobilization data (2000) to show that statistical methods are essential to address complications that invariably arise in field experiments. I demonstrate how incomplete randomization of treatment assignment led to the authors' puzzling finding that get-out-the-vote calls discourage voters from going to the polls, reducing turnout by 5 percent. An application of matching, which is more appropriate given the incomplete randomization, reveals that telephone canvassing increases turnout by about 5 percent. My analysis also finds that mail canvassing is a significant cost-effective alternative, and that appeals related to civic engagement are more effective than the original analysis indicated.

1 Introduction

Under the leadership of Gerber and Green, field experimentation is making its way back into the toolkit of political scientists (Gerber and Green, 2000, 2001; Green and Gerber, forthcoming). In their research, Gerber and Green have taken advantage of many modern experimental techniques that have been developed since Gosnell (1927) and others first used this methodology more than half a century ago. However, Gerber and Green believe that field experiments only require "rudimentary data analysis." Using the study of the effects of campaign contributions on political access as an example, they write

> Rather than launch a complex multivariate analysis of the flow to and from donations and access, the researcher may obtain an unbiased assessment of the average treatment effect merely by cross-tabulating the size of contribution. Rudimentary data analysis replaces scores of regressions, freeing the researcher from the scientific and moral hazards of data mining (Green and Gerber, forthcoming, p.6).

If field experiments work perfectly – with perfect random selection of a large sample and completely randomized assignment of treatment among individuals – and, in addition, the empirical relationships are unambiguously strong, then sophisticated statistical analysis may be unnecessary. However, precisely because field experiments take place in the real world, such perfection is rarely achieved in practice. I demonstrate that assuming otherwise is often as misleading as the optimistic assumptions of observational research that Gerber and Green reject (Gerber *et al.*, 2002).

This paper shows the value of statistical methods by analyzing the Gerber and Green's data from their field experiment of voter mobilization strategies (Gerber and Green, 2000). Their research is an influential study of an important topic and offers a unique opportunity to examine the appropriateness of different statistical techniques for analyzing data from field experiments. My reanalysis confirms their finding that personal canvassing is the most effective mobilization method. However, I find that telephone and mail canvassing can also increase turnout and that appeals to community values and civic duty are an effective message for mobilization.

The original analysis shows that get-out-the-vote calls encouraging people to vote *decrease* turnout by 5 percent on average, which Gerber and Green described as "one of the most surprising results to emerge from our experiment" (p.660). Such a finding of course throws into question why so many millions of dollars are spent on these calls. Applying a different methodology that is more appropriate for these data, I demonstrate that telephone calls *increase* turnout by about 5 percent. The new evidence supports the results of earlier experimental studies on telephone canvassing (Eldersveld, 1956; Adams and Smith, 1980; Miller, Bositis, and Baer, 1981). Moreover, it corroborates with the result of a recent new experiment conducted by Green and Gerber (2001) that also confirms the effectiveness of get-out-the-vote calls.

Secondly, Gerber and Green dismissed the utility of mailings as a voter mobilization strategy. However, this conclusion was based on their assumption that everyone who was sent postcards received and read them. This assumption led to a comparison of different strategies that underestimated the relative effectiveness of postcards. Since the data do not contain any information about who actually read the postcards, the effectiveness of mail canvassing should be evaluated in terms of intention-to-treat effects rather than treatment effects. This equal basis of comparison shows that mail canvassing can be a cost-effective alternative to personal visits.

Finally, with regard to the messages used to encourage voting, the original analysis suggests that a close election message was on average from 50 to 150 percent more effective than appeals related to civic duty or neighborhood solidarity for mobilization. If true, this result would call into question a large body of prominent research on civic engagement (e.g. Skocpol and Fiorina, 1999; Putnam, 2000) as well as existing empirical literature on political participation (e.g. Brady, Verba, and Schlozman, 1995; Blais, 2000, ch.5). After correcting the problems of the original analysis, I find that civic duty and neighborhood solidarity messages are often more effective at mobilizing voters than a close election message. This suggests that a sense of civic duty is part of a voter's decision to cast their ballots.

The point of my paper is not to discourage field experiments as infeasible. In fact, I believe that randomized field experiments provide a great opportunity for political scientists to make valid causal inferences. However, the problems of the Gerber and Green study

highlight how difficult it is to implement perfect experimental design in the real world. Thus, careful statistical analysis is often necessary to adjust for unintended, but not entirely unforeseeable, complications that arise in field experiments.

Overview of methodology In Section 3, I replicate the Gerber and Green study (2000) using the same instrumental variables method employed in the original analysis. In the process of replication, I find that the authors incorrectly defined the treatment assignment and control groups. Since the estimation of causal effects necessarily involves the comparison of these two groups, many of their estimates are inappropriate. However, correcting the definitions produces even more implausible results about the effect of telephone calls. This calls for further investigation about the data and the assumption that underlies the instrumental variables method Gerber and Green use in their statistical analysis.

I demonstrate that the incomplete randomization of treatment assignment produced this unexpected result. Statistical tests suggest that unlike personal visits and postcard mailings, the treatment assignment for telephone calls was not completely randomized. In particular, Gerber and Green were more likely to assign a phone call to those who live in single-voter households and those who did not vote in the previous election. Since these individuals were less likely to vote, the original analysis underestimated the treatment effect of telephone calls. This provides direct evidence of how instrumental variable estimation produces inaccurate results when randomization is incomplete.

In order to overcome the problems of instrumental variable estimation, Section 4 applies an alternative statistical method, matching, which does not require the assumption of completely randomized treatment assignment. The method of matching literally matches each observation in the treatment group with an observation in the control group whose observed characteristics are the same. Thus, it allows analysis of control and treatment groups that are different only with respect to whether they received treatment. When important covariates are available, matching effectively reduces the bias resulting from non-random treatment without the functional form assumptions of usual regression analysis.

As the number of available covariates increases, however, it becomes difficult to match along all covariates. In such situations, the propensity score of Rosenbaum and Rubin (1983)

provides a more general approach for matching. The propensity score, defined as the probability of receiving the treatment, is a scalar variable that measures the similarity among a group of observations in terms of observed covariates. Hence, the propensity score facilitates the use of matching in multivariate settings. Matching with the propensity score has become standard in other fields when estimating the causal effects of non-random treatment.[1] Here, the statistical technique helps overcome complications of the field experiment and produces more plausible conclusions than the original analysis.

The method of instrumental variables is useful for estimating treatment effects as long as the treatment assignment is completely randomized. In fact, I show that when the treatment is randomly assigned, as it was for personal canvassing, the estimates based on the method of matching largely agree with those based on instrumental variable estimation. However, instrumental variable estimation is very sensitive to the lack of complete randomization and can produce misleading results under certain circumstances.

2 Importance of randomized field experiments

A central goal of scientific inquiry is to make causal inferences. In the physical sciences, experiments are essential for this purpose. In contrast, for many social sciences including political science, analysis of observational data and comparative case studies has been the more common approach, and relatively few researchers design and conduct experiments. Recently, Green and Gerber (forthcoming) have characterized the state of the discipline as resembling "monocrop agriculture, efficiently generating prodigious quantities of nonexperimental research but deeply vulnerable to an experimental intrusion that could vitiate the entire enterprise" (p.24). They advocate field experiments, which are attempts of randomized interventions into real world settings, as the best way to answer many important questions in political science.

Indeed, the experimental approach can often provide more insight into causal processes with fewer arbitrary assumptions than would be necessary in observational studies. In political science as well as in economics, a growing number of researchers conduct experiments in a laboratory.[2] Although laboratory experiments offer greater control, conclusions based

on such studies are difficult to generalize and hence often lack external validity. In contrast, field experimentation combines real world settings with a significant level of control over experimental design and produces more generalizable results.

Establishing causality involves the comparison between what actually occurred and what might have happened under different circumstances. The fundamental problem of causal inference is that we never observe the counterfactual scenario (Neyman, 1923; Rubin, 1974; Holland, 1986; King and Zeng, 2001).[3] For example, in order to measure the causal effect of British colonial rule on the economic development of India in the post-colonial era, one needs to know the economic growth of India if it had not been ruled by the British empire. Answering such counterfactual questions is often difficult, but doing so is necessary to address important research topics in political science.

More formally, let $Y_i(T_i)$ be the potential outcome under the treatment status, T_i, for unit i. Here, T_i is an indicator variable which is equal to 1 if this unit received the treatment and 0 otherwise. The treatment effect for unit i can be defined as

$$Y_i(T_i = 1) \; - \; Y_i(T_i = 0). \tag{1}$$

If a unit belongs to the treatment group, we only observe $Y_i(T_i = 1)$, and the potential outcome if the same unit had not received the treatment, $Y_i(T_i = 0)$, is unknown. Therefore, the validity of causal inference rests entirely on how reliably we can estimate the potential outcome under a counterfactual scenario. Given that we cannot repeat the identical experiment on the same unit in the same environment, the only way to achieve this goal is to form an appropriate control group which is similar to the treatment group in all characteristics except for the treatment status.

In field experimentation, randomized interventions play a critical role in obtaining such a control group. By giving a treatment to randomly selected units in a sample, all characteristics of the treatment and control groups, except for whether they received the treatment, become similar in distribution as the sample size increases. As a whole, the two groups are essentially identical if there is a large sample, even though each unit is different in its characteristics. The greatest advantage of randomization is that it adjusts unobserved as well as observed characteristics of the two groups. Thus, if treatment is indeed completely random,

we can simply use the mean difference of the observed outcome between the treatment and control groups as an unbiased estimate of the average treatment effect.[4] A serious limitation encountered in observational studies is that researchers do not possess the powerful tool of randomized interventions (Achen, 1986).

In many field experiments, the distinction between assignment of treatment and receipt of treatment is critical because researchers often can randomize the former, but not the latter. In the field, not everyone assigned the treatment by researchers actually receives it. In addition, some of those who are not assigned the treatment may receive one. In the absence of random treatment, the estimation of treatment effects in equation (1) requires additional assumptions and statistical adjustments.

The difficulty of estimating treatment effects in such situations leads many to estimate another causal quantity, known as the Intention-To-Treat (ITT) effect. Unlike the treatment effect, the ITT effect does not take into account whether those assigned the treatment actually received it. That is, the ITT effect represents the effect of treatment assignment rather than treatment itself. Precisely for this reason, the ITT effect is relatively easy to estimate so long as the treatment assignment is randomized. Formally, let Z_i be the treatment assignment indicator which is equal to 1 if unit i is assigned the treatment and 0 otherwise. Then, the ITT effect for unit i can be defined as

$$Y_i(T_i, Z_i = 1) \; - \; Y_i(T_i, Z_i = 0), \tag{2}$$

where T_i can take either 0 or 1.

The ITT effect can be a key quantity of interest for evaluating policy effectiveness and may differ substantially from treatment effects (Sommer and Zeger, 1991). Consider a hypothetical example where an international organization plans an AIDS prevention campaign in an African country. The first proposal is to distribute educational pamphlets to local high schools. The second proposal is to build health counseling centers around the country. And, a third proposal is to put up an educational message on roadside billboards. The first policy will have the greatest treatment effect if those pamphlets are actually read by students at school. However, it is questionable whether school teachers will read them to students. Moreover, it is possible that the youth in schools are those who are less likely to be infected

with AIDS in the first place. The second policy also can be effective if people who are at risk of being infected visit the office for counseling. It is likely, however, that most of those who go to the counseling center have already been infected by the disease so that this policy would have little effect to prevent AIDS. In this example, one would expect the ITT effects of these two proposals to be low although their treatment effects are reasonably large. In contrast, the billboard advertisements may have a higher ITT effect because they are more likely to be read by the target population. Thus, policy-makers may prefer the billboard advertisement even if it has the smallest treatment effect. One can readily think of many other situations where there would be a substantial difference between ITT effects and treatment effects.

While ITT effects can be useful for policy makers, academic researchers are often more interested in estimating treatment effects in order to learn about underlying causal processes. Unfortunately, in field experiments it is significantly more difficult to estimate treatment effects than ITT effects. Indeed, for some cases when the information about who actually received the treatment among those assigned is unavailable, one can only estimate ITT effects (e.g. Wantchekon, 2002). In other cases when it is possible to estimate treatment effects, statistical techniques based on additional assumptions are necessary.[5] A goal of this paper is to show the importance of choosing appropriate statistical methodology when estimating treatment effects with data from field experiments.

3 Replicating the New Haven voter mobilization study

In this section, I replicate and extend Gerber and Green's analysis of the New Haven voter mobilization study. Gerber and Green designed and conducted an experiment where registered voters who live in randomly selected households of New Haven were encouraged to vote in the 1998 general election by means of personal canvassing, telephone calls, and postcard mailings. The authors then examined voting records and analyzed which strategies had increased voter turnout. In addition to the voting record of the 1998 election, the data set includes covariates that describe the following characteristics of each registered voter: the number of registered voters in the household (one or two), age, party affiliation, voting record in the last general election (1996), and ward of residence in New Haven (Wards 2 to 30, which

omits the ward with a heavy student population).

3.1 Problems with experimental design

Inefficient experimental design The experimental design of the original study is unusually complicated because over 40 percent of voters in the sample were assigned more than one treatment. Table 1 replicates Table 2 of Gerber and Green (2000, p.655) and shows the substantial overlap of different treatment assignments. For example, 383 people were assigned 3 mailings, a phone call, and a personal visit. Further variation in the nature of the treatment was possible because each treatment has three different appeal messages (civic duty, neighborhood solidarity, and close election).[6] Altogether, this produced a total of 37 different treatment combinations and their corresponding potential outcomes.

[Table 1 about here.]

This inefficient experimental design makes the estimation of treatment combinations difficult unless one makes strong assumptions such as no interaction effect. For example, get-out-the-vote calls may not effectively increase turnout for those voters who already have received a personal visit. Moreover, the timing of contact differs from one canvassing method to another and this variation was not randomized. Such systematic differences in when each treatment was administered also raise the risk of post-treatment bias. Therefore, I focus on the marginal effects of three treatments rather than their interaction effects in order to avoid these additional complications.[7]

Incorrect definitions of treatment assignment and control groups Gerber and Green used odd definitions of the treatment assignment and control groups, and as such did not directly estimate their causal quantities of interest. For example, when estimating the marginal effect of personal canvassing, the authors used the treatment assignment group that includes those who were also assigned other treatments such as telephone calls and postcard mailings (the upper two rows in Table 1). Their control group also includes those voters who were assigned other treatments (all categories in the bottom two rows in Table 1). In order to correctly estimate the treatment and ITT effects, the appropriate control group

should consist solely of the 10,800 voters who were assigned no treatment and hence received no intervention.

This implies that both the ITT and treatment effects reported in the Gerber and Green study are confounded by the effects of other treatments.[8] In designing experiments, an appropriate control group is critical to ensure internal validity (Campbell and Stanley, 1963). In principle, it is advisable to avoid the assignment of multiple treatments in field experiments. Although factorial designs like the one used in the Gerber and Green study are feasible in laboratory experiments, additional complications such as non-compliance and limited randomization make it difficult to estimate multiple overlapping treatment effects in field experiments.

Mismatch between randomization and data analysis Another problem is that the authors performed the randomization of treatment assignment based on households while their unit of analysis at the estimation stage was an individual voter. Furthermore, in two-voter households, at most one voter was contacted by a telephone call, but the authors did not record which voter was actually reached. If one of the two voters in a two-person household was contacted, both voters in the household were coded as if they had been reached by a canvasser. The mismatch between randomization and data analysis is likely to cause the underestimation of the treatment effect because many two-person household voters in the treatment group actually did not receive the treatment. This may also have contributed to the surprising finding of the original study that telephone calls reduce voter turnout.[9]

3.2 Analysis with complete randomization assumption

Estimation of ITT effects With the correct definitions of treatment assignment and control groups, I first reanalyze the original data set by applying the same statistical method as employed in Gerber and Green (2000). In order to assess the relative effectiveness of different mobilization strategies, I first reestimate average Intention-To-Treat (ITT) effects. Under the assumption of complete randomization, the treatment assignment is independent of any observed and unobserved individual characteristics. That is, those assigned treatment are in all other respects identical to those not assigned treatment. Therefore, the average

ITT effect can be estimated by simply calculating the difference of the sample means of the treatment assignment and control groups. Namely,

$$\widehat{\text{ITT}} = \frac{\sum Y_i Z_i}{N_1} - \frac{\sum Y_i (1 - Z_i)}{N_0}, \tag{3}$$

where $N_1 = \sum Z_i$ is the size of the treatment assignment group and $N_0 = \sum(1 - Z_i)$ is the size of the control group. In the case of telephone calls, for example, $N_1 = 958$ and $N_0 = 10,800$. If the treatment assignment is completely random, this formula gives an unbiased estimate of the ITT effect for each treatment.

[Table 2 about here.]

Table 2 shows the result of the ITT analysis with the correct definitions of the treatment and control groups for three mobilization methods.[10] First, the adjusted analysis confirms the result of the Gerber and Green study that personal canvassing is the most effective method for increasing voter turnout. The adjusted ITT effect is even larger than the original estimate. Second, get-out-the-vote calls have a significant negative effect on voter turnout. Using the appropriate treatment assignment and control groups does not change the odd finding of the original article that telephone canvassing reduces turnout.

Finally, mail canvassing also mobilizes voters. Gerber and Green assumed that all voters who were sent postcards received and read them (Gerber and Green, 2000, fn.10, p.659). This assumption seems unrealistic given the possibility that many cards did not reach a person due to changes of address or were discarded unread as junk mail. Since the dataset contains no information about who actually read postcards, we can only estimate the ITT effect of mail canvassing. Hence, the relative effectiveness of mail canvassing should be evaluated using estimated ITT effects for other canvassing methods.

Gerber and Green incorrectly used their estimated ITT effects for mail canvassing as estimated treatment effects and reached the conclusion that "even if the effective marginal costs of canvassing were doubled, face-to-face mobilization would still be cost effective" (Gerber and Green, 2000, p.661). In contrast, the adjusted ITT analysis in Table 2 makes the appropriate comparison of the ITT effects across the three mobilization strategies. The evidence indicates that sending three postcards is quite influential. Given the relatively low

cost of sending postcards compared to visiting each voter's residence, policy-makers might reasonably prefer to use postcard mailings as a feasible and cost-effective method to raise turnout levels.

Using instrumental variables to estimate treatment effects Moving from the estimation of ITT effects to treatment effects necessitates attention to compliance with treatment assignment. In field experiments, non-compliance often occurs because researchers cannot force everyone assigned a treatment to receive it. The Gerber and Green study is no exception. Table 3 shows that for telephone calls and personal visits, only slightly more than 25 percent of those assigned a treatment actually received the treatment. This non-compliance occurred mostly because voters were not home when they were visited or telephoned. Among 217 assigned both treatments, there were only 27 people who actually received both treatments. This made it difficult to estimate the interaction effect with precision.

[Table 3 about here.]

Instrumental variable (IV) estimation is a well-known statistical method that identifies average treatment effects by focusing on those who would receive a treatment only if assigned (Bloom, 1984; Permutt and Hebel, 1989; Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996). Although some argue that the average treatment effect for the entire population is a more meaningful quantity of interest, it is difficult to reliably estimate it given the high non-compliance rate in the Gerber and Green study.[11] An 'instrument' is a variable that satisfies an assumption referred to as the *exclusion restriction*; the instrument influences the outcome variable only through its effect on the treatment variable.[12] That is, the instrument cannot have any direct effect or indirect effect through variables other than the treatment variable. When analyzing data from randomized field experiments, the random assignment of treatment can be used as an instrument. In the Gerber and Green study, the fact that voters were assigned telephone canvassing via random numbers generated by a computer should not affect anything other than the probability of their receiving get-out-the-vote calls. Formally, the exclusion restriction can be written as

$$Y_i(T_i = t, Z_i = 1) \;=\; Y_i(T_i = t, Z_i = 0) \quad \text{for} \quad t = 0, 1. \tag{4}$$

The IV estimator is not unbiased, but it consistently estimates average treatment effects for compliers when treatment assignment is completely randomized. Gerber and Green employ this approach to estimate the marginal treatment effects of telephone calls and personal visits for the subgroup of those who received an assigned treatment. The ITT effect divided by the compliance rate gives the IV estimate of complier average treatment effect. Namely,

$$\widehat{\text{IV}} = \frac{\widehat{\text{ITT}}}{\widehat{\text{CPR}}} \quad \text{where} \quad \widehat{\text{CPR}} = \frac{\sum T_i Z_i}{N_1}, \tag{5}$$

where $\widehat{\text{CPR}}$ is the estimated compliance rate as appear in Table 3.

[Table 4 about here.]

Table 4 presents the IV estimates of the average treatment effects of telephone calls and personal canvassing for compliers. Gerber and Green found that telephone calls have a significant negative effect of 5 percent on turnout.[13] Even more worrisome is the observation that their inappropriate use of overlapping treatments obscured greater problems. Correcting the definitions of treatment assignment and control groups makes the effect even larger, reaching negative 12 percent.[14] These IV estimates suggest that get-out-the-vote calls encouraging people to vote actually discourage them from casting their ballots. This implausible result should raise concern about the data and statistical analysis.

The large negative effect on individuals in single-voter households drives this odd finding about telephone canvassing. The overall treatment effect combines estimates for single-voter and two-voter households. Hence, Gerber and Green's overall estimate of negative 5 percent is largely due to the significant negative effect of 14 percent found for the subgroup of single-voter households. This divergence between the subsamples would have been apparent if Gerber and Green had conducted a separate analysis of telephone calls to correspond with their analysis of visits.[15] Similarly, the adjusted IV estimate for single-voter households is a negative 27 percent, which leads to the overall effect for the total sample of negative 12 percent. These large negative effects for single-voter households contrast with slightly positive effects for two-person households. Such a wide gap between the two subgroups calls for further investigation.

In contrast to telephone calls, the IV estimates for personal visits seem more reasonable. The adjusted IV analysis supports the conclusion of the Gerber and Green study that personal

canvassing is a very effective way to mobilize voters. This conclusion holds regardless of which definitions one employs, or which subgroup one examines.

3.3 Sources of negative finding for telephone calls

Why are the IV estimates for telephone calls counter-intuitively negative? While the IV estimation is very useful in many situations, the valid use of this technique critically relies on a key assumption that treatment assignment is completely randomized. In the Gerber and Green study, this assumption was violated, and the violation led to their odd finding that telephone canvassing reduces turnout.

Incomplete randomization of treatment assignment Gerber and Green tried to randomize the treatment assignment, but unforeseen, and until now unrecognized, complications arose. While the treatment assignment for personal visits and postcard mailings appears to be well randomized, the randomization is incomplete for the sample that was assigned to receive telephone calls. In principle, statistical tests based on the observed data can never guarantee that the treatment assignment is completely randomized since it is always possible that unobserved variables are unbalanced. Nevertheless, diagnostics to test the completeness of randomization are essential given that the validity of IV estimation relies on the assumption of completely randomized treatment assignment.

I begin with an analysis of single-voter households. For this subgroup, the incomplete randomization of treatment assignment for telephone calls is apparent. Only 42 percent of the treatment assignment group voted in the last election whereas 47 percent of the control group voted (p-value for this mean difference is 0.05). The randomization for single-voter households appears to be incomplete even with the incorrect definitions of treatment assignment and control groups used in the original analysis. When compared with the control group, the treatment assignment group includes significantly more individuals who abstained in the last election (the mean difference is statistically significant at the 0.05 level). Since those who voted in the last election are 40 percent more likely to vote in the current election, this difference will contribute to the under-estimation of the treatment effect of phone calls for single-voter households.

Furthermore, voters in a single-person household were more likely to be assigned a phone call than those in a two-person voter household (p-value is 0.11). While this difference may be subtle, its consequence can be serious given the strong predictive power of this variable. In fact, voters in single-voter households are 10 percent less likely to go to the polls than those who live with their family, which is enough to cause large selection bias. Unlike telephone canvassing, similar analyses of observed covariates show little indication of incomplete randomization for personal canvassing and postcard mailings.

Finally, the particular experimental design of the Gerber and Green study allows us to examine the balance of *unobserved* variables using the observed data. In this study, both personal visits and telephone calls are assigned with randomly selected appeal messages: civic duty, neighborhood solidarity, and close election. For telephone calls, the message of neighborhood solidarity was not used. If the assignment of treatments as well as the selection of appeal messages are random, one should see no systematic difference in compliance rates among different messages.[16] This is because the complete randomization prevents one message from being assigned to a group of people who are more likely to receive the treatment. Since the probability of each voter being at home when called or visited depends on their unobserved characteristics as well as their observed ones, this test allows us to check the balance of unobserved characteristics of voters.

The results of this analysis strongly indicate that the treatment assignment for telephone calls is not completely randomized, while that for personal canvassing is well randomized. For telephone calls, those who were assigned the close election message are on average about 10 percent more likely to answer a phone call than those who were assigned the civic duty appeal. This mean difference is statistically significant using the two sample z test (p-value is less than 0.01). In contrast, for personal canvassing, one finds no systematic variation in compliance rates among different appeal messages; the Pearson's χ^2 test shows that one cannot reject the null hypothesis of equal compliance rate for all three appeal messages (p-value is 0.71). In sum, the data exhibits poor randomization for phone calls and good randomization for visits. This result holds even when looking at the incorrect definitions of treatment assignment and control groups used in the original analysis.[17] This may explain the odd result of the IV estimates for different messages of telephone canvassing.

Randomization of treatment assignment in field experiments is not as easy to accomplish as one may expect. In practice, it is almost impossible to completely randomize every aspect of each treatment. In the Gerber and Green study, personal canvassing was conducted over four weeks while telephone canvassing took place over 3 days including election day. Although a visit right before the election would have a greater effect than a visit a month before the election day, the timing of contact was not randomized. If people they attempt to contact right before the election are particularly difficult to reach, one may underestimate the effect of canvassing. Likewise, the effect of different canvassers, if not randomized, can confound the treatment effect of different canvassing methods. These examples illustrate the difficulty of randomization and potential confounding effects that threaten internal validity of field experiments.

Why are the IV estimates for telephone calls biased? The sensitivity of the IV estimation to the violation of the exclusion restriction is well documented in the statistical literature (e.g. Angrist *et al.*, 1996, p.450). In particular, the bias due to incomplete randomization is worsened when unbalanced variables are good predictors of the outcome variable and when a large number of non-compliers exist. Equation (6) illustrates this fact that the bias of the IV estimate is larger when the bias of the ITT estimate due to incomplete randomization is larger and/or when the compliance rate is lower, *ceteris paribus*.

$$\frac{\widehat{\text{ITT}}}{\widehat{\text{CPR}}} = \frac{\text{ITT} + \text{bias}}{\widehat{\text{CPR}}}. \tag{6}$$

The Gerber and Green study fits both conditions. First, the unbalanced covariates, the voting record in the previous election and the number of registered voters in a household, predict turnout well, which suggests that the bias in the estimated ITT effect is large. Furthermore, the compliance rate of this field experiment is low (around 25 percent). Indeed, this low compliance rate implies that if the ITT effect is biased by 3 percent, then the bias of the IV estimate can be as large as 12 percent. Therefore, the combination of a large bias in the ITT estimate and low compliance rate led to the puzzling finding of IV estimation that get-out-the-vote calls have a significant negative impact on voter turnout.[18]

If one successfully randomizes the treatment assignment, the method of instrumental

variables can give estimated treatment effects that are consistent in large sample. However, as the analysis of this section suggests, making this assumption in practice requires careful experimental design and its successful implementation. I have shown that the lack of complete randomization for the assignment of telephone calls led to biased causal inference about the effects of telephone calls in the Gerber and Green study.

4 Analysis without complete randomization assumption

The previous section showed that Gerber and Green's IV estimation was inappropriate for telephone canvassing given the incomplete randomization of treatment assignment. This calls for more general statistical methods to estimate the effects of non-random treatments. I apply the method of matching to reduce the bias caused by non-random treatment. Matching is particularly useful for field experiments when randomization of treatment assignment is incomplete and important covariates are available.

The basic idea of matching follows the logic of causal inference described in Section 2. The goal is to construct a control group as similar to the treatment group as possible. The method of matching does this by finding a pair of subjects who have exactly the same observed characteristics except that one receives the treatment and the other does not. Matching, thus, reduces bias due to incomplete randomization in an intuitive way and also has the advantage of not requiring the linearity and other specification assumptions of regression.

The intuition behind matching resembles the traditional comparative case study method which dates back to John Stuart Mill (1930/1843). Both approaches call for comparing cases that are very similar to each other except for the primary causal variable. This facilitates the evaluation of main causal effects in isolation by reducing the possibility of confounding effects from other variables. Although the comparative method has largely been used for qualitative studies, with the method of matching, quantitative and historical case studies can rest on a common ground of causal inference.

4.1 Selection bias due to non-compliance

In field experiments, the actual treatment group, as opposed to the treatment assignment group, is often different from the control group in its characteristics because of selection bias due to non-compliers who do not receive their assigned treatment. Non-compliance leads to selection bias when individuals who comply with treatment assignment have characteristics significantly different from those who do not comply.

[Table 5 about here.]

Table 5 illustrates the imbalance of observed covariates between the actual treatment (or complier) group and the control group for telephone calls and personal visits. For telephone calls, individuals who were home and received the phone call were on average 9 years older than members of the control group. In other words, it was difficult to contact young voters. Since young voters are significantly less likely to vote, this leads to selection bias. Also, the turnout for the last election is almost 20 percent higher in the treatment group than in the control group, while there were 15 percent less newly registered voters in the treatment group than in the control group. Moreover, the ratio of registered Democrats in the treatment group is 5 percent greater than that in the control group. Similar differences between the treatment and control groups are apparent for the data on personal canvassing.

The wide gap between the two groups indicates a significant selection bias that calls for statistical adjustment. The treatment group is older, more Democratic, and has a better past voting record than the control group. Estimates of treatment effects will be biased, unless one properly adjusts for these systematic differences between the treatment and control groups. Next, I show how matching with the propensity score reduces this selection bias.

4.2 Method of matching

Our goal is to estimate the average treatment effect for the treated,[19]

$$E\{Y(T=1) - Y(T=0) \,|\, T^A = 1\} \;=\; E\{Y(T=1) \,|\, T^A = 1\} - E\{Y(T=0) \,|\, T^A = 1\} \quad (7)$$

where T^A is the indicator variable of actual treatment status. For the treated, we can directly estimate the average outcome under the treatment, $E\{Y(T=1) \,|\, T^A = 1\}$, from the data

by computing the mean of their observed outcome. However, without extra assumptions and information, we cannot identify the counterfactual outcome, $E\{Y(T=0)\,|\,T^A=1\}$. Our strategy is to use the control group to estimate this key quantity.

The assumption of matching is that the counterfactual outcome for the treated can be computed from those individuals in the control group who have the same observed characteristics. That is, the counterfactual outcome, $Y(T=0)$, is mean independent of the actual treatment status, T^A, conditioning on X (Heckman *et al.*, 1998). Formally,

$$E\{Y(T=0)\,|\,T^A=1,X\} \;=\; E\{Y(T=0)\,|\,T^A=0,X\}. \tag{8}$$

The implication of this assumption is that the method of matching most effectively reduces bias when important covariates are included. The omitted variable bias is possible if X does not contain the variables which affect T^A as well as $Y(T=0)$. The bias due to omitted variables can be reduced, however, if those variables are highly correlated with X. The advantage of matching is that this conditional independence assumption does not require parametric functional forms common to usual regression analysis.

The method of matching then averages equation (8) over the distribution of covariates, X, to obtain an unbiased estimate of average treatment effect for the treated.

$$\begin{aligned} & E\{Y(T=1)-Y(T=0)\,|\,T^A=1\} \\ = \;& E_X[E\{Y(T=1)\,|\,T^A=1,X\}-E\{Y(T=0)\,|\,T^A=0,X\}]. \end{aligned} \tag{9}$$

Unfortunately, the application of matching becomes practically impossible as the dimensionality of X increases. The use of the propensity score, defined as the conditional probability of receiving a treatment, aids statistical analysis in such situations. Rosenbaum and Rubin (1983) show that conditioning on the propensity score, $e(X)=P(T^A=1|X)$, is equivalent to conditioning on all observed covariates, X. Thus, instead of equation (9), one only needs to average over the distribution of the propensity score to obtain an unbiased average treatment effect,

$$\begin{aligned} & E\{Y(T=1)-Y(T=0)\,|\,T^A=1\} \\ = \;& E_{e(X)}[E\{Y(T=1)\,|\,T^A=1,e(X)\}-E\{Y(T=0)\,|\,T^A=0,e(X)\}]. \end{aligned} \tag{10}$$

Now, one needs to find the closest value of a scalar variable, the propensity score, instead of looking for a match on the entire vector of X.

In most cases, one must estimate the propensity score by modeling the actual receipt of treatment given observed covariates. The logistic regression can serve this purpose although other methods such as neural network and classification models can also be used. Whatever method is used, the estimated propensity score carries little substantive interpretation and it should be regarded as a tool to create a control group similar to the treatment group in terms of observed characteristics. If the propensity score is estimated properly, it should balance observed covariates between the treatment and matched control groups. One has to change the model specification and reestimate the propensity score until this balance is achieved.[20] Thus, one can reliably check the validity of model specification, and it often matters little what method is used to estimate the propensity score.

Matching with the propensity score is known to be very effective in reducing bias caused by non-random treatment. Unlike the randomization of treatment, however, the method of matching can only balance *observed* characteristics of the treatment and control groups. Hence, there is always potential bias due to omitted variables. Estimates based on matching are also biased when one cannot find appropriate matches because the treatment group is too different from the control group. This does not apply here because as shown later I can find exact or close match in the control group for most voters in the treatment group. When the covariates measuring important characteristics of subjects concerned are available, matching with the propensity score is a powerful method for reducing bias. In fact, there is empirical evidence that it produces more reliable estimates of treatment effects than various instrumental variable estimators (Dehejia and Wahba, 1999). When treatment assignment is not completely random and important covariates are available, as in Gerber and Green's study, matching with the propensity score is more appropriate than the method of instrumental variables.

4.3 Application of matching to the voter mobilization study

Next, I apply the method of matching with the propensity score to the New Haven voter mobilization study. The first step in the procedure of matching is to find an exact match in

the control group for each individual voter in the treatment group. Each matched pair should be identical in all observed characteristics. The original data set includes many important covariates such as past voting behavior, party affiliation, age, and ward of residence. This is an ideal situation for applying matching adjustment. The past voting record is for the 1996 general election, and the ward of residence variable represents 29 small geographical areas in New Haven. The ability to match on these variables allows further bias reduction by balancing some unobserved variables including race and income that may be correlated with past voting behavior, party affiliation, and the neighborhood of voters' residence.

For telephone calls, more than half of the treatment group can be matched with an individual in the control group who has exactly the same values for all covariates. That is, for 126 out of 242 voters in the treatment group, there is at least one voter in the control group of 10,800 voters who lives in a household with the same number of registered voters, is exactly the same age, has the same party affiliation, lives in the same ward of New Haven, and has the same voting record in the previous election. Similarly, for personal visits, I am able to find an exact match for about 45 percent of individuals in the treatment group. Although the method of matching only uses a subset of the control group, the comparison of the treatment group with a matched sample gives a more reliable estimate of treatment effect.

For about half of the treatment groups for telephone calls and personal visits, I could not find an exact match in the control group.[21] For these voters, I use a procedure referred to as "nearest propensity score matching" and pair each treated voter with another voter in the control group whose propensity score is the closest (Rosenbaum and Rubin, 1985b).[22] Before matching on the propensity score, I match on the most important variable, the number of voters in household. This variable is important because the randomization of treatment assignment was performed on households. Also, poor randomization of this variable is a likely cause of the bias in the IV estimation that was discussed in Section 3. Those individuals who were exactly matched will by definition have the same propensity score. If there is more than one voter in the control group with exactly the same propensity score, then I randomly select one of them.[23]

[Table 6 about here.]

I estimate the propensity score with logistic regression using all available covariates (shown in Table 5) and their first order interactions. In addition, the model includes dummy variables for ward of residence and age squared. A standard evaluation for model specification is a t test of mean difference and F test of variance ratio, both of which measure how well covariates are balanced between the treatment and control groups. Table 6 shows that matching on the estimated propensity score successfully balances the covariates. For all the observed covariates, mean differences between the two groups are negligible and their variances are approximately the same.

[Figure 1 about here.]

Figure 1 compares the distribution of the propensity score of the treatment group with the corresponding distribution of the control group before and after matching adjustment. The propensity score is a scalar summary of similarity between the treatment and control groups, which is measured in terms of observed characteristics. While the difference of the distribution between the two groups is substantial before matching, they are almost identical after matching. The similarity of the propensity score distribution between the treatment and control groups indicates that matching successfully balances the observed covariates between the two groups.

The effectiveness of the matching method illustrates an advantage of field experiments. In many observational studies, it is often difficult to conduct the matching adjustment because the treatment group is too different from the control group. For such cases, even the propensity score may prove inadequate if the distribution of the propensity score for the treatment group is not overlapping with that for the control group (Rosenbaum and Rubin, 1985a; Heckman *et al.*, 1998). In contrast, field experiments avoid this problem by creating a control group that is a representative sample of the relevant population. Hence, there is sufficient similarity between the treatment and control groups to allow for statistical comparison via matching. In the New Haven voter mobilization study, the treatment assignment, though not completely random, produced a control group such that matching with the propensity score can effectively balance the distribution of all covariates.

5 Effectiveness of voter mobilization strategies

After matching with the propensity score, I estimate the average treatment effects of telephone calls and personal canvassing. Since the compliance record for postcard mailings is unknown, I estimate the ITT effect for this treatment. When analyzing data that has been subjected to matching, the average treatment/ITT effects can be estimated by simply calculating the mean difference of the outcome variable between the treatment and control groups.[24] Furthermore, because the control group is much larger than the treatment group for telephone calls and personal visits, I also conduct one-to-three matching in order to improve the efficiency of estimation by using a larger sample. That is, for each voter in the treatment group, I find three individuals in the control group whose propensity score is closest. Although one-to-three matching reduces the standard error, it often results in greater incomplete matching, and thus can be less effective in reducing bias. For postcard mailing, I cannot conduct one-to-three matching because the size of the treatment assignment group is large (7,369 voters as opposed to the control group of 10,800 voters).

[Table 7 about here.]

Table 7 presents the estimated treatment effects of telephone calls and personal visits as well as the estimated ITT effect for postcard mailings. For both one-to-one and one-to-three matching, personal canvassing is more effective than telephone calls. This result confirms the finding of the Gerber and Green study that personal canvassing has the greatest treatment effect. Moreover, the matching estimates are consistent with the IV estimates reported in Table 4. Similarly, the estimated overall ITT effect for mail canvassing based on one-to-one matching is close to the estimate reported in Table 2. The agreement between the two methods implies that in the New Haven mobilization study, matching and IV estimation give substantively similar results only when the treatment assignment is well randomized.

The most important finding of the new analysis is that telephone calls *increase* turnout by around 5 percent on average, reversing Gerber and Green's key conclusion. While not as effective as personal visits, telephone canvassing offers a significant alternative mobilization strategy. Even with the incorrect definitions of treatment assignment and control groups

used in the original article, the method of matching produces a significant positive effect of 6.1 percent (standard error is 1.5).[25] This positive estimate agrees with the results of another recent experimental study by Green and Gerber (2001) which concludes that "Phone canvassing increased turnout by an average of 5 percentage-points. This finding, based on six experiments involving nearly 10,000 people, is statistically significant" (p.2).[26] Since making a phone call costs much less than visiting a home, a get-out-the-vote phone call is often the most cost-effective mobilization strategy.

[Table 8 about here.]

Another substantive finding is that across all canvassing methods, the messages emphasizing civic duty and neighborhood solidarity are more effective than the message that tells voters the upcoming election is a close race. Table 8 presents the estimated effect of 3 appeal messages. This contrasts with the finding in Gerber and Green's original analysis that the close election message was between 50 to 150 percent more effective than using the appeal to voter's sense of civic duty or community. This new finding gives empirical support to the thesis of civic engagement. Citizens vote at least in part because they feel obligation from a sense of civic duty. Connections formed by neighborhood and local activities may also be an important vehicle for higher turnout.

Nevertheless, the finding about the relative importance of messages may not apply to other districts or elections. It is important to keep in mind that New Haven and Connecticut are a Democratic stronghold. For example, in 1998 general election for which the original study was conducted, Christopher J. Dodd, the democratic candidate for Senate, won the election with 65 percent of the total vote. In the 3rd Congressional district to which New Haven belongs, Rosa L. DeLauro of Democratic party won the race by more than 70 percent of the total vote. Therefore, the close election message may not have been a convincing way to mobilize voters in New Haven. Field experiments in other geographic regions would be useful for generalizing the conclusion.

Comparison of matching and IV estimates For telephone canvassing, matching gives more plausible estimates than IV estimation. In Table 4, the instability of IV estimates was

apparent from the discrepancy between a negative 27 percent turnout from phone calls for single-voter households and a positive 4 percent for two-voter households. In contrast, the estimates based on matching are much more reliable. The estimated average treatment effect using one-to-one matching is 6.1 percent for voters in single-person households and 1.3 percent for those in two-person households.[27] The smaller effect for voters in two-person households is expected because for telephone calls, only one of the two voters in those households are actually contacted while Gerber and Green coded both of them as treated.

For personal canvassing, matching estimates of treatment effects on voters in single-person households and those in two-person households are 7.3 and 9.8 percent, respectively.[28] The results are comparable with the estimates for the overall sample shown in Table 7. In sum, the method of matching gives more reliable results than the IV estimation when the treatment assignment is not completely random.

6 Concluding remarks

As Gerber and Green argue, field experimentation has many advantages over observational studies. However, it also faces practical complications that often lead to incomplete randomization of treatment assignment and a large number of non-compliers. Although some of these complications can be avoided by planning a better experimental design, other problems will need to be addressed with statistical methods when analyzing the data.

Using the New Haven voter mobilization study of Gerber and Green (2000), I demonstrated that matching with the propensity score allows us to overcome incomplete randomization of treatment assignment and deal with the existence of non-compliers at the same time.[29] In contrast, the original analysis suffered from its failure to properly address these complications. I showed that the application of instrumental variable estimation led to the odd finding that phone calls, a widely used mobilization method, significantly reduce turnout. On the contrary, my analysis showed that phone calls and postcard mailings are effective methods to increase turnout with relatively low cost, even if not as effective as personal canvassing.

After their analysis, Gerber and Green (2000, p.662) reached a rather pessimistic conclusion that "The question is whether the long-term decay of civic and political organizations

has reached such a point that our society no longer has the infrastructure to conduct face-to-face canvassing on a large scale." In contrast, my findings allow greater optimism for how to re-invigorate democracy. A simple phone call or postcard appeal to voters asking them to vote for the sake of their community can make a difference.

The Gerber and Green study I reanalyzed was one of the few field experiments conducted in the discipline in more than half a century. As more experience with field experiments accumulates, political scientists will learn how to use this promising methodology even more effectively. Nonetheless, there will always be unforeseen complications in field experiments. The real world is a messy place, and only with statistical methods continuously adapted to the problem at hand are we able to make valid causal inferences.

Afterward

The problems highlighted in this paper led to the discovery of further complications in the Gerber and Green study. As a consequence of my analysis and through discussions with Donald Green about an earlier version of this paper, it has become apparent that the original data set includes coding mistakes for telephone canvassing. As it turns out, the phone bank the authors hired for telephone canvassing mixed up the list of voters with that for another experimental study by Gerber and Green (2001) about voters in West Haven, a town next to New Haven.

As a consequence, a few hundred voters who were analyzed as if they had received get-out-the-vote calls may not have actually received a call. It is also possible that some received messages different from the designated message. Some voters in the New Haven list received a blood donation message, which Gerber and Green had planned to use for voters in West Haven. Moreover, in their original article, Gerber and Green noted that the neighborhood solidarity message was not used for telephone calls. However, the new data corrections suggest that more than 1,500 voters could have received this message. This correction altered the coding of treatment assignment and led to changes of which individuals were included in the control group. For example, the size of the control group was reduced from 10,800 to 10,582 voters.

These coding changes in the assignment as well as receipt of telephone calls affect almost all analyses and tables reported in the original article because of multiple overlapping treatments. Gerber and Green have now recalculated all of their estimates, which can be found at Donald Green's website.[30] Their correction shows that phone calls still reduce turnout on average, but the effect is less than half of the original result (-2.0 percent with a standard error of 2.2). This slightly negative IV estimate of treatment effect for telephone calls, however, is the weighted average of a significant negative effect of 9.1 percent for single-voter households and a positive effect of 2.9 percent for two-voter households.[31] The recoded data set generates the same wide gap between single-voter and two-voter households that was problematic in the original data. As in the original analysis, incomplete randomization produces the large negative effect for single-voter households, which in turn leads to the implausible conclusion that telephone canvassing reduces turnout.

Note that this coding mistake in the implementation of the experiment is distinct from the problem of incorrect definition of treatment and control groups that was discussed in Section 3.1. In their reanalysis of the recoded data, Gerber and Green continue to mix individuals incorrectly with overlapping treatments in the control group. Analyzing the recoded dataset, with the correct definition of treatment assignment and control groups, I find that the IV estimate of telephone canvassing is negative 10 percent overall.[32] IV estimation with the recoded data set produces the same puzzling finding that telephone calls significantly reduce turnout.

Similar to the main analysis of this paper, use of matching leads to a positive finding for telephone calls that reverses the negative effect predicted by IV estimates. Based on reanalysis of the corrected coding with matching, telephone canvassing is still found to increase turnout by 6.4 percent on average and the effect of personal canvassing is 9.5 percent. The findings for the messages are also similar. Civic duty and neighborhood solidarity messages are more effective for personal visits and postcard mailings than the close election appeal, although the difference becomes much smaller for telephone calls.[33]

Thus, despite the recent changes in the dataset, the analysis and conclusions presented in this paper remain valid. More importantly, the discovery of additional complications confirms the difficulty of implementing perfect experimental design in the field. In recognition that

even one of the best existing experimental studies may encounter such problems, it is still advisable to use the best available statistical methods even when data are generated by a field experiment.

Notes

[1] For recent examples, see Rubin and Thomas (1996) in the statistical literature, Heckman, Ichimura, Smith, and Todd (1998) in the econometric literature, and Ming and Rosenbaum (2000) in the biostatistics literature.

[2]Recent examples of laboratory experiments include Iyengar and Kinder (1987), Ansolabehere and Iyengar (1995), Nelson *et al.* (1997), Morton and Williams (1999), and Bottom *et al.* (2000). An interesting "experiment" combined with survey is Sniderman *et al.* (1991). Comprehensive reviews of literature include Kinder and Palfrey (1993) and Lupia (forthcoming). See also articles in a forthcoming special issue of *Political Analysis*, vol.10, no.4, on "Experimental Methods in Political Science."

[3]For causal inference which does not rely on counterfactuals, see Dawid (2000).

[4]Even in laboratory experiments randomization of treatment is often desirable because without randomization experimenters can at most control observed characteristics.

[5] Recently, a variety of statistical techniques have been proposed, which reflects the lack of a unified approach to this difficult problem (e.g. Rosenbaum and Rubin, 1983; Manski, 1990; Angrist *et al.*, 1996; Balke and Pearl, 1997; Imbens and Rubin, 1997).

[6] Gerber and Green (p.656 2000) note that for telephone calls they did not use the neighborhood solidarity message.

[7]Sometimes, the interaction effect can be of interest. However, when designing experiments with multiple overlapping treatments, one needs to take into account the possibility that the compliance rate of two treatments will be much lower than that of one treatment alone. Table 1 implies that only 217 individuals were assigned both a personal visit and telephone call. Since many of these voters were not home to receive both treatments, the small size of treatment group, only 27 voters, makes it difficult to estimate the interaction effect of the two treatments with reasonable precision.

[8] The direction of bias due to this incorrect definition is unclear. On one hand, the control

group of Gerber and Green includes those who received other treatments. For example, more than half of the voters in the control group used by the authors for personal canvassing were assigned telephone calls and/or postcard mailings. This will lead to the under-estimation of the treatment effect for personal visits. On the other hand, many voters in the treatment assignment group were assigned the other treatments. This will in turn lead to the over-estimation of the treatment effect for personal canvassing if other treatments are effective.

[9]The same procedure was used in another study of Gerber and Green (2001) which also reported a negative effect of telephone canvassing. For personal canvassing, Gerber and Green were able to identify individual voters who were actually contacted in two-voter households. Nevertheless, even in this case there exists a possibility that one voter who talked to a canvasser in person influences the voting behavior of the other person in the same household. This spill-over effect, if present, violates the exclusion restriction assumption. Gerber and Green in their recent experiment estimate that this spill-over effect could cause as much as a 5.7 percent increase in the *untreated* household member (Green *et al.*, 2002).

[10]The estimates are overall ITT effects averaging over different messages.

[11]One may estimate the non-parametric bounds of the average treatment effect developed by Manski (1990) and Balke and Pearl (1997). The bounds for visits and phone calls were $[-27.9\%, 43.9\%]$ and $[-28.1\%, 46.6\%]$, respectively, and they are not very informative due to many non-compliers.

[12]Another important assumption is that of the strict monotonicity condition that excludes the possibility of defiers and requires at least one complier. Defiers, a type of non-compliers, receive the treatment only if they are not assigned the treatment. Such non-compliance behavior is impossible in the Gerber and Green study, and hence the monotonicity condition is satisfied. See Angrist *et al.* (1996).

[13]The two-stage probit regression used in their original article also indicates that the effect of telephone canvassing is negative 5 percent and statistically significant. This uses the turnout of the control group, 44.5 percent, as a base line (Gerber and Green, 2000, p.660).

[14]Standard errors of the adjusted IV estimates are larger because correct definitions use smaller treatment assignment and control groups.

[15] In their original article, Gerber and Green report the results of the separate analysis for

personal canvassing, but not for telephone calls (Gerber and Green, 2000, fn.8, p.658).

[16]This relies on the fact that the message itself is unlikely to affect compliance since all messages have the identical opening script. Moreover, all scripts are relatively short, and the authors note that for telephone calls the scripts lasted only for 30 seconds (Gerber and Green, 2000, p.656).

[17]The mean difference for telephone canvassing is 5 percent, which is significant at the 0.01 level. For personal canvassing, there is no significant difference across messages.

[18]It is also important to note the finite sample bias and inefficiency of IV estimation. In particular, the small size of each treatment group in the New Haven mobilization study suggests the importance of finite sample consideration. There exists a considerable amount of empirical evidence in political science as well as economics and statistics that IV estimates have poor small sample properties (e.g. Bartels, 1991; Bound, Jaeger, and Baker, 1995; Imbens and Rubin, 1997). Improving these aspects of IV estimation is currently one of the most important topics in the literature (Angrist and Krueger, 1995; Angrist, Imbens, and Krueger, 1999).

[19]The method of instrumental variable estimates the average treatment effect for compliers Since in the New Haven mobilization study all non-compliers are never-takers, this IV estimand is essentially equivalent to the average treatment effect of for the treated.

[20] In addition, it is advisable to check the model fit.

[21]When one cannot find an exact match for all individuals of the treatment group, bias due to incomplete matching arises. However, the nearest propensity score matching like the one used in this paper will give the optimal results (Rosenbaum and Rubin, 1985a).

[22]Another standard adjustment technique is the method of blocking with the propensity score (Rosenbaum and Rubin, 1984). In this case, matching is more desirable than blocking because we have a large control group, more than 10,000 voters, from which we can select individuals whose characteristics are very similar to those in the treatment group.

[23]The procedure of nearest propensity score matching is as follows. First, I randomly sort the control group. Then, I start with voter in the treatment group whose propensity score is the highest and match with a voter in the control group whose propensity score is the closest. This is because voters with higher propensity score tend to be more difficult to

match. When there is variation due to this random selection, I obtain estimates by averaging 25 independently matched samples.

[24]If matching does not remove imbalance, one may run a regression for further adjustment.

[25]The estimated effect of personal canvassing is 9.2 (standard error is 1.7).

[26]After I sent Donald Green an earlier version of this paper, he forwarded to me these findings from a recent working paper.

[27]Standard errors are 6.6 and 5.7 percent, respectively.

[28]Standard error is 3.6 in both cases.

[29]Imai and van Dyk (2002) and Imbens (2000) have extended the propensity score to non-binary treatments such as continuous treatments and the interaction effect of multiple treatments. I expect this generalization to widen the potential applications of propensity score in observational studies.

[30]URL for this website is `http://research.yale.edu/vote/CORRECTEDAPSRREPLICATIONTABLES.HTM`, which was initially posted on June 20, 2002 and updated on July 10, 2002.

[31]Standard errors are 3.7 and 2.6, respectively.

[32] The number of those assigned a phone call is 805, and the compliance rate is 31 percent (= 247/805).

[33]For telephone canvassing, the average treatment effect of civic duty, neighborhood solidarity, and close election messages are 6.5, NA, and 6.3, respectively. For personal visits, they are 9.9, 11.9, and 6.4, respectively. For mail canvassing, they are 2.1, 2.1, and -0.1 percent, respectively.

References

Achen, C. H. (1986). *The Statistical Analysis of Quasi-experiments.* University of California Press, Berkeley.

Adams, W. C. and Smith, D. J. (1980). Effects of telephone canvassing on turnout and preferences: A field experiment. *Public Opinion Quarterly* **44**, 389–395.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* **90**, 431–442.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics* **14**, 57–67.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91**, 444–455.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business and Economic Statistics* **13**, 225–235.

Ansolabehere, S. and Iyengar, S. (1995). *Going Negative: How Political Advertisements Shrink and Polarize the Electorate.* The Free Press, New York.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92**, 1171–1176.

Bartels, L. M. (1991). Instrumental and "quasi-instrumental" variables. *American Journal of Political Science* **35**, 777–800.

Blais, A. (2000). *To Vote or Not to Vote: The Merits and Limits of Rational Choice Theory.* University of Pittsburgh Press.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* **8**, 225–246.

Bottom, W. P., Eavey, C. L., Miller, G. J., and Victor, J. N. (2000). The institutional effect on majority rule instability: Bicameralism in spacial policy decisions. *American Journal of Political Science* **44**, 523–540.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90**, 443–450.

Brady, H. E., Verba, S., and Schlozman, K. L. (1995). Beyond ses: A resource model of political participation. *American Political Science Review* **89**, 271–294.

Campbell, D. T. and Stanley, J. C. (1963). *Experimental and Quasi-experimental Designs for Research.* RAND McNally, Chicago.

Dawid, A. P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association* **95**, 407–424.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053–1062.

Eldersveld, S. J. (1956). Experimental propaganda techniques and voting behavior. *American Political Science Review* **50**, 154–165.

Gerber, A. S. and Green, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review* **94**, 653–663.

Gerber, A. S. and Green, D. P. (2001). Do phone calls increase voter turnout?: A field experiment. *Public Opinion Quarterly* **65**, 75–85.

Gerber, A. S., Green, D. P., and Kaplan, E. H. (2002). The illusion of learning from observational research. *Presented at the Economic Science Association conference* .

Gosnell, H. F. (1927). *Getting-Out-the-Vote: An experiment in the stimulation of voting.* University of Chicago Press, Chicago.

Green, D. P. and Gerber, A. S. (2001). Getting out the youth vote: Results from randomized field experiments. *Unpublished manuscript, Yale University* .

Green, D. P. and Gerber, A. S. (forthcoming). *Political Science: State of the Discipline (eds. Katznelson, I. and Milner, H. V.)*, vol. III, chap. Reclaiming the Experimental Tradition in Political Science. W. W. Norton, New York.

Green, D. P., Gerber, A. S., and Nickerson, D. W. (2002). Getting out the youth vote in local elections: Results from six door-to-door canvassing experiments. *Unpublished manuscript, Yale University* .

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica* **66**, 1017–1098.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945–960.

Imai, K. and van Dyk, D. A. (2002). Causal inference with general treatment regimes. *Technical Report, Harvard University* .

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706–710.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–475.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* **25**, 305–327.

Iyengar, S. and Kinder, D. R. (1987). *News that matters: Television and American Opinion.* The University of Chicago Press, Chicago.

Kinder, D. R. and Palfrey, T. R., eds. (1993). *Experimental Foundations of Political Science.* University of Michigan Press.

King, G. and Zeng, L. (2001). How factual is your counterfactual? *Technical Report, Harvard University* .

Lupia, A. (forthcoming). New ideas in experimental political science. *Political Analysis* .

Manski, C. F. (1990). Non-parametric bounds on treatment effects. *American Economic Review, Papers and Proceedings* **80**, 319–323.

Mill, J. S. (1930/1843). *A System of Logic, Ratiocinative and Inductive: Being A Connected View of the Principles of Evidence and the Methods of Scientific Investigation.* Longman, London.

Miller, R. E., Bositis, D. E., and Baer, D. L. (1981). Stimulating voter turnout in a primary: Field experiment with a precinct committeeman. *International Political Science Review* **2**, 445–460.

Ming, K. and Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56**, 118–124.

Morton, R. and Williams, K. (1999). Information assymetries and simultaneous versus sequential voting. *American Political Science Review* **93**, 51–67.

Nelson, T. E., Clawson, R. A., and Oxley, Z. M. (1997). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review* **91**, 567–583.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science* **5**, 465–480.

Permutt, T. and Hebel, J. R. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics* **45**, 619–622.

Putnam, R. D. (2000). *Bowling Alone: The Collapse and Renewal of American Community.* Simon and Schuster, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics* **41**, 103–116.

Rosenbaum, P. R. and Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33–38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics* **52**, 249–264.

Skocpol, T. and Fiorina, M. P., eds. (1999). *Civic Engagement in American Democracy.* Brookings, New York.

Sniderman, P. M., Brody, R. A., and Tetlock, P. E. (1991). *Reasoning and Choice: Explorations in Political Psychology.* Cambridge University Press, Cambridge.

Sommer, A. and Zeger, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10**, 45–52.

Wantchekon, L. (2002). Clientelism and voting behavior: Evidence from a field experiment in Benin. *Working Paper, New York University* .

		Postcard mailing			
		none	**once**	**twice**	**3 times**
Personal visit	**Telephone call**	217	385	352	383
		(0.7%)	(1.3%)	(1.2%)	(1.3%)
	No call	2,686	519	625	627
		(9.1%)	(1.8%)	(2.1%)	(2.1%)
No visit	**Phone call**	958	1,451	1,486	1,522
		(3.3%)	(4.9%)	(5.1%)	(5.2%)
	No call	10,800	2,406	2,588	2,375
		(36.8%)	(8.2%)	(8.8%)	(8.1%)

Table 1: *Distribution of treatment assignment for sample total of 29,380 registered voters in New Haven. The table shows the substantial overlap of different treatment assignments. The figures represent the number of registered voters in each treatment assignment and control group with their ratio as a percentage of the total in parentheses. A box highlights the control group, and the two treatment assignment groups of interest are underlined.*

	adjusted ITT analysis		*Gerber & Green*	
Treatment	ITT	s.e.	ITT	s.e.
Personal visit	3.9 %	1.1	2.4 %	0.7
Telephone call	−2.9	1.7	−1.5[a]	0.7
Postcard mailing[b]				
once	0.4	1.1	0.6	0.3
twice	0.8	1.1	1.2	0.5
three times	2.6	1.1	1.7	0.8

Table 2: *Estimated average Intention-To-Treat (ITT) effects on voter turnout with the assumption of complete randomization. The table shows the differences of estimates due to different definitions. The adjusted ITT estimates use the correct definitions of treatment assignment and control groups. The right column displays the original results as reported in Gerber and Green (2000).*

[a]The effect of telephone calls was not reported by Gerber and Green and is the author's calculation based on their method.

[b]Gerber and Green used ITT estimates of mail canvassing as treatment effects.

	Personal visit	Telephone call	Visit and call
Compliance rate	28.1 %	25.3 %	12.4 %
Size of treatment group	756	242	27

Table 3: *Compliance rates and size of treatment groups. The table shows that the compliance rate in the Gerber and Green study is very low. The compliance rate represents the ratio of those who received treatment among those assigned treatment. The size of treatment group is the number of those who actually received the treatment.*

	Telephone call		Personal visit	
	adjusted IV	*Gerber & Green*[a]	*adjusted IV*	*Gerber & Green*
Overall treatment effect	−11.6 %	−4.7 %	13.9 %	8.7 %
	(6.8)	(2.3)	(3.8)	(2.6)
Single-voter households	−26.8	−13.7	13.3	9.9
	(10.6)	(4.0)	(5.4)	(3.7)
Two-voter households	3.7	1.6	15.3	8.4 %
	(8.6)	(2.7)	(5.3)	(3.6)

Table 4: *The instrumental variable (IV) estimates of average treatment effects on voter turnout. The table shows that the surprising finding about telephone calls is driven by the large negative effects for single-voter households. The adjusted IV estimates use the correct definitions of treatment assignment and control groups. The right columns use the incorrect definitions used in Gerber and Green (2000). Standard errors are in parentheses.*

[a]To obtain their estimates for telephone calls, Gerber and Green used the two-stage least squares regression, which follows the logic similar to the one presented in Section 3.2 (see Angrist and Imbens, 1995). Gerber and Green did not report the separate analysis of telephone calls for different household types. The estimates in the table are the author's calculation based on their method.

	Telephone call			Personal visit		
Variables	mean diff	t-stat	var ratio	mean diff	t-stat	var ratio
Age	9.01	7.00	1.12	3.22	4.66	0.96
Voted in 96 election	0.19	6.41	0.82	0.04	2.10	0.99
Newly registered voter	−0.16	−7.56	0.51	−0.07	−4.81	0.80
Registered Democrat	0.05	1.95	0.89	0.03	1.76	0.95
Registered Republican	0.01	0.40	1.11	−0.01	−1.55	0.80
Number of voters in household	0.03	0.79	1.00	−0.00	−0.17	1.00

Table 5: *Imbalance of observed covariates between treatment and control groups prior to adjustments. The table shows the imbalance of covariates caused by non-compliance. The mean of each covariate for the control group is subtracted from that for the treatment group. The mean differences and their t statistics are reported. The variance ratios are calculated by dividing the variance of the treatment group by that for the control group.*

	Telephone call			Personal visit		
Variables	mean diff	t-stat	var ratio	mean diff	t-stat	var ratio
Age	1.63	0.90	0.99	0.28	0.30	1.01
Voted in 96 election	−0.02	−0.51	1.05	−0.02	−0.63	1.01
New registered voter	−0.01	−0.28	0.94	0.01	0.52	1.04
Registered Democrat	−0.01	−0.21	1.02	−0.02	−1.04	1.06
Registered Republican	−0.02	−0.45	0.94	0.02	0.67	1.05
Number of voters in household	exactly matched			exactly matched		
Ward of residence	54 % matched			54 % matched		
Exact match	52 % matched			47 % matched		

Table 6: *Balance of observed covariates between treatment and control groups after one-to-one matching. The table shows that matching effectively balances the observed covariates. The mean of each covariate for the control group is subtracted from that for the treatment group. The mean differences and their t statistics are reported. The variance ratios are calculated by dividing the variance of the treatment assignment group by that for the control group.*

	one-to-one	*one-to-three*	*Gerber & Green*
Telephone call	3.9%	7.2%	−4.7%
	(4.4)	(3.4)	(2.3)
Personal visit	8.9	10.5	8.7
	(2.6)	(2.1)	(2.6)
Postcard mailing	1.2		1.2
(ITT effect)	(0.8)		(0.5)

Table 7: *Estimated average treatment/ITT effects of 3 mobilization methods on voter turnout using the method of matching. The table shows that the estimated positive effect of telephone calls reverses the original negative finding. Results based on one-to-one and one-to-three matching are reported. The right column displays the estimates Gerber and Green reported in their original article. The estimates for postcard mailings are overall ITT effects. Standard errors are in parentheses.*

	Matching			*Gerber & Green*
	Calls	Visits	Postcards	Visits
Civic duty	4.7%	8.8%	1.8%	9.1%
	(6.1)	(4.2)	(1.4)	(4.3)
Neighborhood solidarity		13.2	2.2	5.1
		(4.5)	(1.4)	(4.1)
Close election	2.9	4.5	−0.5	12.1
	(6.4)	(4.7)	(1.4)	(4.2)

Table 8: *Estimated average treatment/ITT effects of 3 appeal messages on voter turnout using one-to-one matching. Estimates based on matching indicate that the civic duty and neighborhood solidarity appeals are more effective than the close election message. The right column reports the IV estimates of Gerber and Green. Standard errors are in parentheses.*

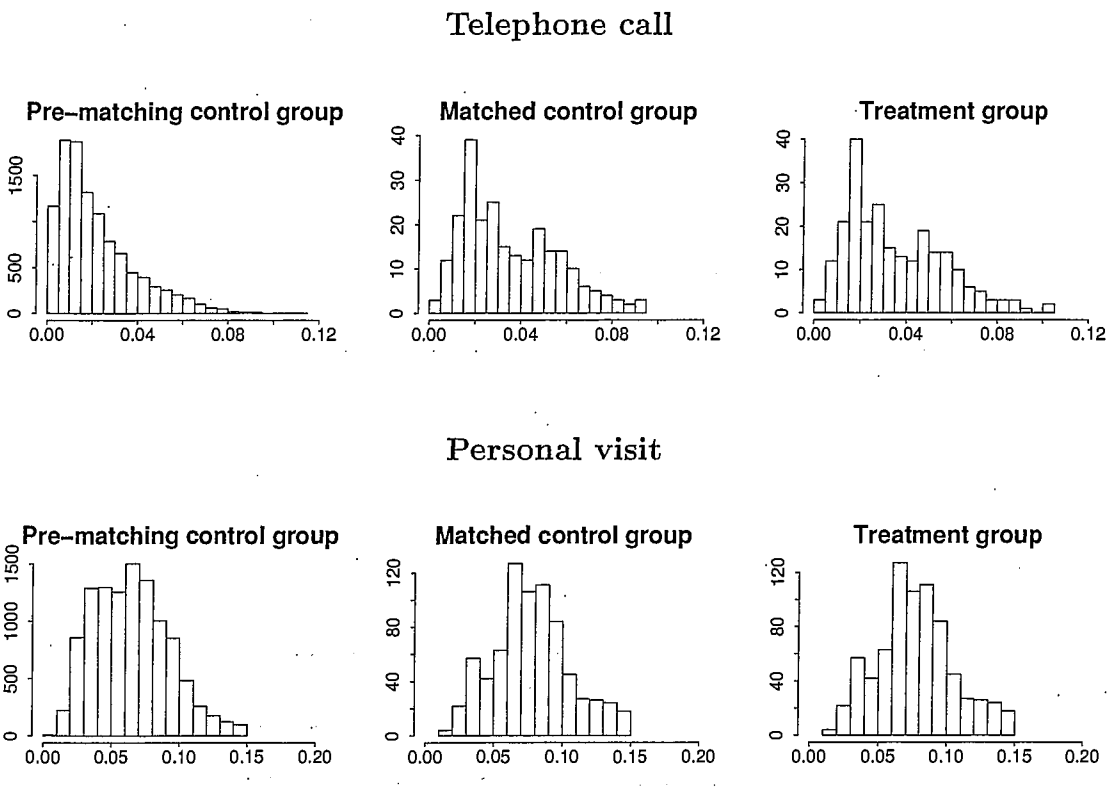

Figure 1: *Distributions of the propensity score for treatment, control, and matched control groups for telephone calls (first row) and personal visits (second row). The figure shows the similarity between the treatment and matched control groups. The vertical axis represents the number of registered voters. The graphs are based on one-to-one nearest propensity score matching.*