Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment

Author(s): Kevin Arceneaux, Alan S. Gerber and Donald P. Green

Source: *Political Analysis*, Winter 2006, Vol. 14, No. 1 (Winter 2006), pp. 37-62

Published by: Cambridge University Press on behalf of the Society for Political Methodology

Stable URL: http://www.jstor.com/stable/25791834

Advance Access publication September 16, 2005

Political Analysis (2006) 14:37–62
doi:10.1093/pan/mpj001

Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment

Kevin Arceneaux
Department of Political Science, Temple University,
453 Gladfelter Hall, 1115 West Berks Street, Philadelphia, PA 19122
e-mail: kevin.arceneaux@temple.edu (corresponding author)

Alan S. Gerber and Donald P. Green
Yale University, Institution for Social and Policy Studies,
P.O. Box 208209, 77 Prospect Street, New Haven, CT 06520
e-mail: alan.gerber@yale.edu
e-mail: donald.green@yale.edu

In the social sciences, randomized experimentation is the optimal research design for establishing causation. However, for a number of practical reasons, researchers are sometimes unable to conduct experiments and must rely on observational data. In an effort to develop estimators that can approximate experimental results using observational data, scholars have given increasing attention to matching. In this article, we test the performance of matching by gauging the success with which matching approximates experimental results. The voter mobilization experiment presented here comprises a large number of observations (60,000 randomly assigned to the treatment group and nearly two million assigned to the control group) and a rich set of covariates. This study is analyzed in two ways. The first method, instrumental variables estimation, takes advantage of random assignment in order to produce consistent estimates. The second method, matching estimation, ignores random assignment and analyzes the data as though they were nonexperimental. Matching is found to produce biased results in this application because even a rich set of covariates is insufficient to control for preexisting differences between the treatment and control group. Matching, in fact, produces estimates that are no more accurate than those generated by ordinary least squares regression. The experimental findings show that brief paid get-out-the-vote phone calls do not increase turnout, while matching and regression show a large and significant effect.

1 Introduction

Randomized experimentation assists social scientists in two ways. First, the estimates themselves are of substantive interest. Social scientists have used random assignment to obtain insights into the effects of interventions ranging from police raids of crack houses (Sherman and Rogan 1995) to school vouchers (Howell and Peterson 2002) to relocation of public housing residents (Katz et al. 2001). Political scientists in particular have made

Authors' note: We thank the anonymous reviewers, Chris Achen, Jake Bowers, Greg Huber, and Jasjeet Sekhon for their helpful comments. Data and files for replication are available at the *Political Analysis* Web site.

extensive use of random assignment in order to gauge the effects of campaign activity on voting behavior (e.g., Gerber and Green 2000; for a summary see Green and Gerber 2004).

Second, experimental results provide a useful benchmark for evaluating the success with which nonexperimental methods recover causal parameters. The comparison of experimental and observational results was pioneered in economics. In his seminal essay on this approach, LaLonde (1986, pp. 617–618) explains the logic behind this comparison:

> The data from an experiment yield simple estimates of the impact of economic treatments that are independent of any model specification. Successful econometric methods are intended to reproduce these estimates. The only way we will know whether these econometric methods are successful is by making the comparison.

This type of methodological investigation has special importance for political scientists, who, for practical or ethical reasons, are frequently unable to conduct randomized experiments in real-world settings. If observational methods could be shown to reproduce experimental results, political scientists need not bear the costs of conducting randomized interventions but could rely instead on less expensive and more widely available observational data.

A method that has attracted special attention as a potential substitute for experimentation is matching. Matching has been proposed as a nonparametric solution to problems of bias that arise in observational studies (Rubin 1973; Rosenbaum and Rubin 1983, 1985). Matching compares individuals in a nonrandomly generated "treatment group" to similar individuals in a nonrandomly produced "comparison group." The matching process identifies treated individuals who share the same background characteristics as untreated individuals. It is hoped that after matching on covariates, any remaining difference between groups can be attributed to the effect of the treatment.

As those who employ matching concede, matching on observed characteristics leaves open the possibility of unobserved differences between groups. Since unobserved differences can generate biased parameter estimates, the question is whether matching actually works in practice. To date, studies evaluating the performance of matching estimators using experimental benchmarks have obtained mixed results (Dehejia and Wahba 1999; Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura et al. 1998; Smith and Todd 2003). In the field of labor economics, from which most of the empirical applications have been drawn, matching sometimes produces estimates that coincide with experimental benchmarks, but sometimes it does not. Moreover, the performance of matching has not been clearly superior to other statistical approaches, such as linear regression (Bloom et al. 2002; Glazerman et al. 2003).

Matching has nonetheless attracted increasing interest in political science, and recent published works express great enthusiasm for this method. For example, Barabas (2004, p. 692) uses matching to assess the effects of participation in group deliberations. This approach is advanced on the grounds that "matching techniques reduce bias by adjusting estimates of the treatment effect as if the whole study were a randomized experiment." Imai (2005, p. 295) employs matching to estimate the effects of voter mobilization campaigns on the grounds that matching is "known to effectively reduce bias." At a time when political scientists are actively exploring the use of this technique, it is important to extend the LaLonde-style evaluation of matching to political science applications.

This essay reports the results of a randomized field experiment designed to facilitate a comparison between experimental and observational approaches. This experiment examined whether nonpartisan phone calls encouraging people to vote succeeded in raising voter turnout. In our experiment, not everyone assigned to receive a phone call

could be reached by canvassers. Because only some of the people assigned to the experimental treatment group actually received treatment, we are confronted with a selection problem: exposure to the treatment may be correlated with unobserved causes of voting. This selection problem can be overcome by using instrumental variables estimation, as described below. The instrumental variables estimator, which takes advantage of the fact that people were randomly assigned to treatment and control groups, is the standard way to gauge average treatment effects using experimental data (Angrist et al. 1996). (Although instrumental variables estimation is often applied to non-experimental data, those applications rely on strong *substantive* assumptions; by contrast, experimental applications rely on a *procedural* assumption, namely, that the units of observation were randomly assigned to treatment and control groups.) This estimator generates consistent estimates and provides our experimental benchmark.

Treating the experimental data as though they were observational, we use matching to compare those who were actually reached by the calling campaign to those who were not, an approach pioneered by Imai (2005). Matching attempts to remedy the selection problem by comparing people with exactly the same background characteristics. At the end of the exercise, we compare the results of this observational approach to the experimental benchmark in order to assess the performance of the matching estimator in this application.

This study was designed to provide favorable conditions for the matching estimator. The data set contains an extremely large randomized control group (nearly two million cases), allowing us to find exact covariate matches for more than 90% of those who received phone calls and close inexact matches for almost everyone else. The data set also includes a great deal of information about subjects' past voting behavior, demographic characteristics, and geographic location.[1]

The study was also designed to assess whether matching correctly detects variations in the effectiveness of experimental treatments. The voter mobilization script was read by callers from two different phone banks.[2] One of the two phone banks made fewer attempts to reach respondents, so the selection problem is especially severe for this treatment condition. When instrumental variables regression is used to estimate the effect of each phone bank's calls, the estimates are small and not significantly different from one another. The estimates generated by matching, by contrast, are significant for both phone banks and mistakenly indicate that the phone bank with the lower contact rate was substantially more effective than the other phone bank. In other words, the performance of matching deteriorates as the selection problem becomes more severe, a pattern that suggests that matching fails to account for unmeasured differences between treatment and control subjects.

Finally, the experimental sample can be defined in alternative ways in order to assess the sensitivity of matching to omitted variables. By including subjects with unlisted phone numbers, we simulate what happens when researchers overlook a variable that determines whether a person is contacted by a campaign. Instrumental variables regression is robust to this change in sample definition, and the inclusion of unlisted numbers scarcely affects the estimates. Matching, on the other hand, is highly sensitive to this change, and bias becomes more severe when the sample expands to include those with unknown phone numbers.[3]

[1]Because the data set contains information about subjects' voting behavior in two prior elections, it is even richer than the data analyzed by Imai (2005), which contained information about just one prior election.
[2]The authors listened in on calls from both phone banks to verify that callers read the script in the same way.
[3]Unlisted numbers refer to numbers that were unknown to Voter Contact Services, the firm that provided the registration and voting data used here.

The essay is structured as follows. In the next section, matching estimation is discussed and placed in the context of the existing literature. In the third section the data are described. In the fourth section we define the treatment parameter to be estimated, derive the IV estimator, estimate the experimental benchmark, and show it to be highly robust to variations in sample definition and model specification. We assess the performance of matching by comparing matching estimates to the experimental benchmark and find that matching significantly overestimates the effectiveness of voter mobilization calls. Matching's performance is shown to depend on seemingly innocuous factors, such as sample definition and the rate at which calls are completed. Matching's performance is especially poor when the sample is expanded to include unlisted phone numbers and when the treatment is administered in ways that lower the proportion of the treatment group that is successfully contacted. This essay does not demonstrate that matching fails in general or that matching is a "bad idea." Rather, in this particular application, matching produces biased results. The performance of matching in other applications remains an open question that requires further investigation and careful case-by-case assessment of the plausibility of the key assumptions underlying the method.

2 Matching Estimation

Matching estimation compares the voting rates of the treated to the untreated. Each treated subject is matched to an untreated subject exhibiting the same observable characteristics. (See the appendix for an illustration of the mechanics of matching and a discussion of the complications created when multiple people share the same observable characteristics.) The difference between the outcome for the treated subjects and matched counterparts is used to estimate the treatment effect.

Matching raises two issues, one practical and the other theoretical. The practical issue is whether the data set at hand contains untreated subjects whose covariate values match those who were treated. Few social science data sets have a large enough reservoir of observations to allow exact matching on all covariates. Even our data set falls short of exactly matching all of the treated observations; depending on the way the sample is defined, we match either 90.7% or 93.7% exactly. In order to match the remaining observations, we used the approach recommended by Imbens (2003), who matches exactly on the covariates deemed to be most important (such as past voting) and allows for inexact matches on the remaining covariates (such as age). Our procedure for generating inexact matches is described in the appendix. Two features of this procedure warrant emphasis. First, the estimates we present below show that the exact matching estimates are almost identical to the inexact matching estimates. Second, the empirical criterion by which matching is assessed, balance,[4] is excellent in both cases. Mean levels of age, for example, are almost identical in the treated group and their matched counterparts. With exact matching, balance is, by definition, perfect; the treated group has the same distribution of covariates as the control group. When inexact matches are permitted, the balancing tests reported in the appendix show that the covariate balance is nearly perfect.

The second issue is whether matching on covariates isolates the effect of the treatment. This question has attracted a great deal of scholarly interest. Following LaLonde (1986), several scholars have investigated whether nonexperimental estimators can eliminate biases associated with observational data. The literature provides a mixed picture

[4]Balance is a term of art used to describe the similarity in the distribution of the covariate values in the treated group and their matched counterparts.

regarding the effectiveness of matching, specifically, propensity score matching.[5] Some studies have reported that propensity score matching is able to recover estimates similar to the experimental benchmark (e.g., Dehejia and Wahba 1999; but see Smith and Todd 2001 and Dehejia 2005), but meta-analyses of this literature caution that the performance of matching has been mixed. In their meta-analysis Glazerman et al. (2003) find that while matching may reduce bias somewhat, it does not provide much improvement beyond OLS regression. Heckman and colleagues offer the following assessment:

> In general, matching is not guaranteed to reduce bias and may increase it (see Heckman and Seigelman [1993] and Heckman, LaLonde, and Smith [1999]). Moreover, matching is open to many of the same criticisms that have been directed against traditional econometric estimators because the method relies on arbitrary assumptions (Heckman, Ichimura et al. 1998, p. 1019).

On the other hand, Heckman, Ichimura, and Todd (1997, 1998) and Heckman, Ichimura et al. (1998) suggest that matching performs better under certain circumstances. Two conditions that seem to favor, but by no means assure, better performance obtain when the treatment and comparison groups come from identical data sources, and the data contain a "rich set of variables" that affect both participation in the treatment group and outcomes of interest (Smith and Todd 2005, p. 6). The present study was designed with these two propositions in mind. Whereas the previous literature compares the experimental benchmark with matching estimates derived from different data sets, we estimate the treatment effects using the experimental data and see if matching can recover that estimate within a single data set. By matching treated individuals to untreated people from the same population using an extensive set of covariates, we test matching methods under conditions that are thought to improve its performance.

3 Data

In this essay we draw on data from a large-scale field experiment in which individuals were randomly assigned to treatment and control groups. The field experiment was conducted in Iowa and Michigan before the 2002 midterm elections. The congressional districts of each state were divided into "competitive" and "uncompetitive" strata. Within each stratum, households containing one or two registered voters were randomly assigned to treatment and control groups. For two-person households, just one representative from each household was assigned to treatment or control; if there was another voter in the household, he or she was ignored for purposes of calling and statistical analysis. Only one type of treatment was used: get-out-the-vote (GOTV) phone calls.

Two national phone banks were hired to read a nonpartisan GOTV message to individuals in the treatment group. The script read as follows:

> Hello, may I speak with (name of person) please? Hi. This is (caller's name) calling from Vote 2002, a nonpartisan effort working to encourage citizens to vote. We just wanted to remind you that elections are being held this Tuesday. The success of our democracy depends on whether we exercise our right to vote or not, so we hope you'll come out and vote this Tuesday. Can I count on you to vote next Tuesday?

Members in the control group were not called.

[5] All of the studies discussed in this section use some variant of propensity score matching due to data constraints. Because the data we use allow us to employ exact and near-exact matching, we are able to sidestep a number of difficult issues that arise in the application of propensity score matching. We do not have to formulate a propensity score model, choose among alternative statistical criteria for assessing balance, or stipulate a procedure for choosing among near matches. This greatly simplifies the process of obtaining estimates.

Table 1 Summary of treatment, control, and contacted groups

	Not contacted	*Contacted*	*Subtotal*	*Unlisted Phone Number*	*Overall Total*
Assigned to control group	1,845,348	0	1,845,348	545,076	2,390,424
Assigned to treatment group	34,929	25,043	59,972	24,531	84,503
Total	1,880,277	25,043	1,905,320	569,607	2,474,927

Note. Random assignment was performed within strata in each state, which accounts for the fact that the treatment and control groups for the sample as a whole have slightly different rates of listed and unlisted numbers.

Respondents were coded as contacted if they listened to the script and replied to the question, "Can I count on you to vote next Tuesday?" regardless of whether they answered yes or no (or volunteered some other answer). After the election, public voting records on each individual were obtained, allowing us to assess whether the GOTV phone appeals stimulated voter turnout.

Table 1 summarizes the data and provides an overview of how the variables were coded. Using the list of registered voters in Iowa and Michigan, a total of 60,000 households with listed phone numbers were randomly assigned to be called; the corresponding control group contains 1,846,885 randomly assigned households with listed phone numbers. Because a handful of small counties in the Michigan subsample did not provide 2002 voter records, we removed 1565 observations, bringing the treatment group total to 59,972 and the control group total to 1,845,348. Exclusion of these counties does not bias the results, as county is uncorrelated with treatment assignment.

The voter file also contained a large number of names without phone numbers. Obviously these people could not be called, and their presence in the voter file does nothing to improve the accuracy of our experiment. From a methodological standpoint, however, including voters with unlisted numbers serves two purposes. The first is to assess how matching performs in the presence of unobserved heterogeneity. Those who analyze survey data typically have no information about whether campaigns have access to respondents' phone numbers. If they were to use matching to gauge the effects of phone canvassing on voter turnout, they would not be able to distinguish between those who were called but not reached and those who could not be called because their numbers were unknown. By adding unlisted numbers into the sample, we simulate the behavior of the matching estimator under these conditions.[6] The second reason to include unlisted numbers is that it provides a more direct comparison between our results and those reported by Imai (2005), who applied matching to a sample that contained both listed and unlisted numbers.

A comparison of the covariate distributions in the treatment and control group shows that random assignment created experimental groups with very similar observable characteristics. Table 2 shows that there are only minor differences in age, household size, or past voting rates. The same is true when we randomly allocate unlisted numbers to treatment and control groups (Table 2B). As a randomization check, we used logistic regression to predict treatment based on vote in 2000, vote in 1998, age, number of registered voters in a household, gender, newly registered voter, and state house district. Because the

[6]Across all four of the strata, an additional 569,607 households that had unlisted phone numbers in the voter files were randomly assigned to treatment and control groups.

Table 2 Covariate balance between randomly assigned treatment and control groups

	Strata							
	Iowa A		*Iowa B*		*Michigan A*		*Michigan B*	
Covariate	*Treatment*	*Control*	*Treatment*	*Control*	*Treatment*	*Control*	*Treatment*	*Control*
A. Unlisted Phone Numbers Excluded								
Age	55.8	55.8	53.5	53.5	52.0	52.2	50.9	50.8
Household size	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
Newly registered voter	4.9	4.8	4.6	4.8	11.6	11.7	13.4	13.3
Vote in 2000	73.2	73.4	78.0	78.1	56.7	56.4	59.3	59.5
Vote in 1998	57.4	57.2	59.4	59.9	22.7	23.1	25.9	25.8
Gender (female=1)	55.9	56.3	55.3	55.5	54.6	55.2	53.5	54.1
N	15,000	85,931	15,000	289,163	14,972	1,153,072	15,000	317,182
B. Unlisted Phone Numbers Included								
Age	52.9	52.9	50.8	50.8	51.6	51.7	50.5	50.5
Household Size	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
Newly Registered Voter	6.4	6.3	5.9	6.2	11.0	11.0	12.5	12.3
Vote in 2000	67.1	67.3	71.9	71.8	57.1	56.9	60.1	60.3
Vote in 1998	50.6	50.8	53.0	53.5	21.8	22.3	25.7	25.5
Gender (female=1)	55.6	55.8	55.0	55.5	55.1	55.5	53.8	54.3
N	24,000	137,490	24,000	462,663	18,222	1,403,814	18,281	386,457

Note. Numbers in cells are means for age and household size and percentages for the other variables. Age is in years. Household size varies from one to two voters. Vote in 1998 and 2000 is coded as 1 for voters and 0 otherwise.

Table 3 Voting in 2002 as a function of previous voting behavior

		Voted in 2000		
	Voted in 2002	*Did not vote*	*Voted*	*Total*
Did not vote in 1998	% Voted	19.5	61.3	40.3
	Total *N*	860,415	852,203	1,712,618
Voted in 1998	% Voted	47.6	83.2	78.4
	Total *N*	103,409	658,900	762,309
	Grand Total	963,824	1,511,103	2,474,927

randomization occurred within competitive and uncompetitive congressional districts in each state, the randomization check was carried out within these four strata. As expected, the chisquares for each stratum are nonsignificant. For respondents with unknown phone numbers, the tests are: Iowa noncompetitive, $p = 0.57$; Iowa competitive, $p = 0.75$; Michigan noncompetitive, $p = 0.65$; Michigan competitive, $p = 0.72$. For respondents with unknown phone numbers, the tests are: Iowa noncompetitive, $p = 0.25$; Iowa competitive, $p = 0.75$; Michigan noncompetitive, $p = 0.57$; Michigan competitive, $p = 0.69$.

The voter registration lists from which subjects were drawn include a great deal of information about past voting history, demographic characteristics, and geographic location. Readers familiar with analyses of National Election Study data may be concerned that conventional predictors of voter turnout, such as education and income, are not included among the set of covariates. It should be noted, however, that the influence of these background attributes, which change slowly or not at all, is mediated to a large extent by past voting behavior (Plutzer 2002). The models estimated here correctly predict 74% and 75% of the vote/abstain outcomes for the listed and unlisted samples, in both cases producing a 47% reduction in error. Comparable figures based on an extensive list of regressors drawn from the American National Election Studies are 73% and 45% (Rosenstone and Hansen 1993, Table D-5). Thus the voting models estimated here are arguably as predictive as voting models using a standard battery of social psychological and demographic variables.

In order to illustrate the value of these covariates, Table 3 displays cross tabulations of vote in 2002 by the two previous elections. Among those who voted in the 1998 and 2000 elections, 83.2% voted in the 2002 elections, as compared to 19.5% of those who did not vote in either the 1998 or 2000 election. In other words, these two covariates account for a great deal of variation in voting propensities. Adding in other demographic and geographic information increases the range of predicted probabilities from less than 0.01 to 0.97.

Covariates such as previous voting behavior, household size, being a newly registered voter, gender, and geography arguably provide an adequate selection model for participation in GOTV phone experiments (Imai 2005). It is possible, however, that attributes besides the background characteristics available in the voter files predict both phone contact and voting. Ordinarily the analyst of observational data will not know whether the covariates are adequate to solve the selection problem and could only speculate about the direction and magnitude of bias. In this application, we have the luxury of being able to assess the adequacy of observational approaches by comparing matching estimates to an experimental benchmark.

4 Analysis

4.1 *Specifying the conditions under which instrumental variables estimation and matching estimation produce consistent results*

Because some individuals either refused to listen to the GOTV message or did not answer the phone, only 41.8% of the treatment group subjects with listed phone numbers were contacted (see Table 1). The failure to treat a portion of the assigned treatment group creates a potential selection problem. We now consider two possible approaches to this problem: instrumental variables estimation and matching.

The instrumental variables estimator and matching estimator have the same estimand: the average treatment effect for those who are treated (ATT). This section, which draws on the expositions found in Angrist et al. (1996) and Smith and Todd (2005), provides a formal definition of this estimand and shows the conditions under which standard experimental comparisons of treatment and control groups, adjusted by the proportion of the treatment group actually treated, produce a consistent estimate of the average treatment effect for the treated.[7] We next describe the matching estimator and the critical assumption required for matching to provide an unbiased estimate of the ATT. The remainder of the essay is an empirical investigation of whether matching is unbiased in this application.

4.1.1 Experimental benchmark

The goal of the experiment is to estimate the causal effect of the treatment. For each individual i let Y_0 be the outcome when i is not exposed to the treatment, and Y_1 be the outcome when i is exposed to the treatment. The treatment effect is defined as:

$$Y_1 - Y_0. \tag{1}$$

The basic problem in estimating the causal effect of a treatment is that, because each individual is either treated or not, the data are not available to compute Eq. (1). For each person, only Y_1 or Y_0 is observed. Random assignment solves this "missing data" problem by creating two groups of individuals that are similar prior to application of the treatment. The randomly assigned control group then can serve as a proxy for what the outcome measures would have been for individuals in the treatment group if the treatment had not been applied to them.

Sometimes only a subset of the group assigned to the treatment group receives the treatment. In our application, not all those assigned to get phone calls were in fact reached. To distinguish these cases, let Z_i equal 1 when individual i is assigned to the treated group, 0 otherwise, and let D_i equal 1 when i is actually treated, and 0 otherwise. The average effect of the treatment on the treated is defined as:

$$\text{ATT} = \text{E}((Y_1 - Y_0) \mid \text{D} = 1), \tag{2}$$

where E( ) represents the expected value for the population of subjects.

[7]These assumptions are presented formally and discussed in Angrist et al. (1996, pp. 446–448).

The instrumental variables (IV) estimator has been proposed as a way to provide a consistent estimate of the ATT. The familiar expression for the instrumental variables estimator is:

$$B_{IV} = \frac{COV(Y,Z)}{COV(D,Z)}. \tag{3}$$

This IV estimator is equivalent to a 2SLS regression of Y on D, with Z used as an instrument in the first stage. The IV estimator can be rewritten as:

$$\frac{\widehat{ITT}}{\hat{c}}, \tag{4}$$

where $\widehat{ITT}$ is the estimated "intent to treat" effect, the average outcome for the treatment group minus the average outcome for the control group, and $\hat{c}$ equals the observed share of the treatment group that is actually treated (Angrist et al. 1996), which is sometimes referred to as the "contact rate."

Angrist et al. (1996) show that under a set of sufficient conditions,

$$\frac{E[(Y_i \mid Z = 1) - (Y_i \mid Z = 0)]}{E[D_i \mid Z = 1]} = E[(Y_1 - Y_0) \mid D = 1] = ATT. \tag{5}$$

The left-hand side is the average effect of being placed into the treatment group divided by the probability that you are treated given that you are placed into the treatment group. This ratio is consistently estimated by the ratio of the sample analogues to these quantities, the ITT estimate and the observed contact rate. The equality means that this ratio is equal to the ATT. Therefore, under the conditions outlined in Angrist et al. (1996), the IV estimator generates an asymptotically unbiased estimate of the ATT. For an intuitive explanation for this derivation, see Gerber and Green (2000, pp. 657–658).

If a set of assumptions listed by Angrist et al. (1996) is met, the IV estimator provides a consistent estimate of the ATT. For the experiment reported here, there do not appear to be any plausible arguments suggesting that the assumptions are not satisfied, let alone indications of sufficient deviations to cause meaningful distortions of the experimental estimate. Of the five assumptions presented in Angrist et al. (1996), three are satisfied by the design of the experiment: the treatment groups are formed by random assignment (assumption 2, random assignment); some attempted contacts were successful (assumption 4, nonzero causal effect of Z on D); if individuals were assigned to the treatment group, they were sometimes treated, but if assigned to the control group, they were excluded from treatment (assumption 5, monotonicity). The two remaining Angrist et al. (1996) assumptions were also apparently satisfied by the experiment. For each household, only a single subject was selected for assignment to either treatment or control group, which eliminated the possibility that the experimental effect from a treatment could spill over to another member of the household (assumption 1, stable unit treatment value).[8]

[8]While a theoretical possibility, spillover effects outside the household are likely to be limited since the contact occurs very close to the election, the treated comprise less than 2% of the population, and the treatment represents only a tiny fraction of overall campaign activity occurring in the subject's vicinity. It should be noted that regression analyses of treatment effects typically ignore the possibility of these effects. Further, this assumption is also made by matching estimators. It is nevertheless possible measure any indirect mobilization effects experimentally (Nickerson 2005).

A subject assigned to the treatment group was provided the treatment but not otherwise disturbed in a manner that might alter the outcome (assumption 3, exclusion restriction). While our experimental design satisfies assumption 3, other designs might inadvertently violate it. For example, the assumption could be violated if prior to the election an experimenter administered a political survey to the treatment group but not the control group.

The results for IV are based on large sample properties of the estimator. A study of the small sample bias in similar applications with much smaller samples than that found here shows that finite sample considerations are inconsequential (Gerber and Green 2005).

Matching Estimates. When does matching produce biased estimates of the ATT? The success of matching hinges on whether untreated subjects who share the observed characteristics of the treated differ from the treated in unobserved ways that are related to the outcome variable. This will depend on why the untreated were not treated and what variables are available for matching.

Stated formally, to provide an unbiased estimate of the ATT, matching requires the conditional independence of Y_0 and D:

$$Y_0 \perp D \mid X, \tag{6}$$

where X is a set of observed covariates. This assumption implies the equality of the average values of Y_0 given X:

$$\mathrm{E}(\mathrm{Y}_0 \mid \mathrm{X}, \mathrm{D} = 1) = \mathrm{E}(\mathrm{Y}_0 \mid \mathrm{X}, \mathrm{D} = 0). \tag{7}$$

Recall that $\mathrm{ATT} = \mathrm{E}(\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{D} = 1) = \mathrm{E}(\mathrm{Y}_1 \mid \mathrm{D} = 1) - \mathrm{E}(\mathrm{Y}_0 \mid \mathrm{D} = 1)$. The quantity $\mathrm{E}(\mathrm{Y}_1 \mid \mathrm{D} = 1)$ can be estimated using the outcomes from the treated group. To estimate $\mathrm{E}(\mathrm{Y}_0 \mid \mathrm{D} = 1)$, the law of iterated expectation and Eq. (7) can be used to express $\mathrm{E}(\mathrm{Y}_0 \mid \mathrm{D} = 1)$ as:

$$\mathrm{E}(\mathrm{Y}_0 \mid \mathrm{D} = 1) = \mathrm{E}_{\mathrm{X}|\mathrm{D}=1}\mathrm{E}(\mathrm{Y}_0 \mid \mathrm{D} = 1, \mathrm{X}) = \mathrm{E}_{\mathrm{X}|\mathrm{D}=1}\mathrm{E}(\mathrm{Y}_0 \mid \mathrm{D} = 0, \mathrm{X}), \tag{8}$$

where the notation $\mathrm{X} \mid \mathrm{D} = 1$ indicates that an expectation is taken over the conditional distribution of X given $\mathrm{D} = 1$. This equality implies that the ATT can be estimated using the outcomes from the treated subjects and a weighted average of the outcomes of the untreated, where the weights are set equal to the proportion of times the particular combination of X values occurs in the treated group. In the simplest example, where a single untreated subject is matched to each of the treated subjects, the ATT can be calculated as the average difference in Y for each pair.

When there are many observed characteristics it may not be possible to find exact matches. In this case matching can be performed using a univariate summary of subject characteristics called the "propensity score." The propensity score is defined as $\Pr(\mathrm{D} = 1 \mid \mathrm{X})$; this quantity is often estimated from the data using probit or logit. Treated subjects are then matched to untreated subjects with similar propensity scores. If Y_0 is independent of X, then Y_0 is also independent of $\Pr(\mathrm{D} = 1 \mid \mathrm{X})$, which is a function of X (Rosenbaum and Rubin 1983; Smith and Todd 2005). If $\Pr(\mathrm{D} = 1 \mid \mathrm{X}) = \mathrm{g}(\mathrm{X})$, the argument provided above can be repeated with g(X) replacing X. Thus propensity score matching can be used to approximate exact matching. Note that when subjects share the same values of X, they will also have the exact same propensity score.

Empirical studies that employ matching justify the procedure by explaining that when the key variables determining whether a subject is treated are available to the analyst, then

matching is an effective method. In the best case, matching will consistently estimate the ATT. However, this best case cannot be assumed. Proponents of matching concede that it is not reliable in general but works only when the necessary variables are available to the researcher (Dehejia 2005). Unfortunately, it is not clear how to tell whether or not a particular set of variables comprises the "key" variables. We find that although covariates reduce bias somewhat, even an expansive set of subject characteristics nevertheless produces matching estimates that are more than five times as large as the experimental benchmark. When examining the empirical application described below, the reader should consider whether, absent an experimental benchmark, the list of covariates appears to be sufficient for unbiased estimation using matching.

4.2 *Estimating the experimental benchmark*

The IV estimates (which in this case are equivalent to two-stage least squares estimates) are displayed in Table 4. Controlling for the two design strata (state and competitiveness), two-stage least squares generates an estimated treatment effect of 0.4 percentage points for the sample with listed phone numbers. Due to the large sample size, the standard error of this estimate is just 0.5, which means that the 95% confidence region extends from -0.6 to 1.3. These estimates closely resemble other experimental estimates using similar treatments (brief phone calls from commercial phone banks) in similar kinds of elections (federal midterm elections); see Gerber and Green (2005). Including unlisted phone numbers in the analyses increases the sample size but diminishes the proportion of people contacted in the treatment group. The instrumental variables estimator, however, remains consistent because the unlisted phone numbers were randomly assigned to the treatment and control group. As expected, the expanded sample generates similar results: an estimate of 0.0% with a standard error of 0.6. The slight decline in the estimated treatment effect that occurs when one includes unlisted numbers is attributable to chance, as these numbers were randomly assigned to treatment and control groups. The fact that the numbers decline at all means that the control group votes at a slightly higher than expected rate, which depresses both the experimental estimates and the matching estimates reported below.

The results change only trivially when controls are introduced for past voting behavior, age, or other covariates. The estimated treatment effects are unaffected, and the standard errors diminish slightly. Moreover, there are no significant interactions across state, competitiveness stratum, or phone bank. In sum, the experimental benchmark in this application is a robust number that is perhaps slightly greater than but not statistically distinguishable from zero. Finally, the results are essentially unchanged when we restrict the sample such that every member of the assigned treatment group has at least one counterpart in the control group with identical background characteristics. This "exactly matching" sample will be useful later because it corresponds to the sample used to generate exact matching estimates. For the moment, the central conclusion is that the experimental benchmark is close to zero and robust to sample definition and model specification.[9]

4.3 *Properties of Alternative Estimators: OLS and Matching*

Suppose one were to ignore this study's experimental design, treating the data instead as observational. This approach involves comparing the voting rates of those who were

[9]Although our experiment, involving more than a million people, generates an estimate that is not statistically distinguishable from zero, the effect is probably positive. An even larger study might well show a substantively trivial, though statistically significant effect.

Table 4 Experimental benchmark estimates of the effect of phone calls on turnout in Iowa and Michigan, 2002

	All observations				*Sample containing exact matches only*	
	Sample excludes unlisted numbers		*Sample includes unlisted numbers*		*Excluding unlisted numbers*	*Including unlisted numbers*
Covariates	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*
Phone contact	0.4 (0.5)	0.5 (0.4)	−0.0 (0.6)	0.3 (0.5)	0.5 (0.6)	0.2 (0.7)
State dummy(1=Iowa)	7.4 (0.4)	2.6 (1.3)	3.3 (0.3)	3.6 (1.1)		
Competitiveness dummy in Michigan	4.9 (0.1)	−1.8 (0.3)	5.0 (0.1)	−1.4 (0.3)		
Competitiveness dummy in Iowa	6.1 (0.2)	−0.7 (0.7)	5.8 (0.1)	−1.6 (0.6)		
Household size		7.0 (0.1)		8.0 (0.1)		
Age		0.3 (0.002)		0.3 (0.002)		
Female		−1.2 (0.1)		−1.2 (0.1)		
Newly registered		5.5 (0.1)		8.1 (0.1)		
Vote in 2000		37.1 (0.1)		38.2 (0.1)		
Vote in 1998		21.7 (0.1)		22.2 (0.1)		
Missing values in female dummy		−32.1 (0.2)		−29.2 (0.2)		
Constant	46.1 (0.1)	[a]	43.9 (0.1)	[a]		
N	1,905,320	1,905,320	2,474,927	2,474,927	499,836	781,780
F	5,649.20	4,128.26	3,855.41	5,786.29	1,085.00	1,319.74
Adjusted R^2	0.01	0.29	0.01	0.30	0.01	0.01

Note. Two-stage least squares estimates: Vote 2002 = a + β_1 contact + β_2 MI competitiveness + β_3 IA competitiveness + β_4 state dummy + $\Sigma\gamma_i$ covariates$_i$.
Instrument: Random assignment to treatment group.
[a]Dummy variables for state-house district, state-senate district, and county are included but not shown to save space.

Table 5 Biased OLS estimates of the effect of actual contact on turnout in Iowa and Michigan, 2002

	All observations				*Sample containing exact matches only*	
	Sample excludes unlisted numbers		*Sample includes unlisted numbers*		*Excluding unlisted numbers*	*Including unlisted numbers*
*Covariates**	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*
Phone contact	6.2 (0.3)	2.7 (0.3)	10.7 (0.3)	4.4 (0.3)	2.7 (0.3)	4.2 (0.3)
State dummy(1=Iowa)	6.7 (0.3)	2.4 (1.3)	2.5 (0.3)	3.3 (0.1)		
Competitiveness dummy in Michigan	4.8 (0.1)	−1.8 (0.3)	4.9 (0.1)	−1.5 (0.3)		
Competitiveness dummy in Iowa	6.4 (0.2)	−0.6 (0.7)	6.1 (0.1)	−1.5 (0.6)		
Household size		7.0 (0.1)		8.0 (0.1)		
Age		0.3 (0.002)		0.3 (0.002)		
Female		−1.2 (0.1)		−1.2 (0.1)		
Newly registered		5.5 (0.1)		8.1 (0.1)		
Vote in 2000		37.1 (0.1)		38.2 (0.1)		
Vote in 1998		21.7 (0.1)		22.2 (0.1)		
Missing values in female dummy		−32.1 (0.2)		−29.2 (0.2)		
Constant	46.1 (0.1)	[a]	44.0 (0.1)	[a]		
N	1,905,320	1,905,320	2,474,927	2,474,927	243,736	309,535
F	5,745.62	4,129.75	4,141.81	5,791.01	51.45	52.44
Adjusted R^2	0.01	0.29	0.01	0.30	0.01	0.01

Note. Entries are OLS estimates.

[a]Dummy variables for state-house district, state-senate district, and county are included but not shown to save space.

contacted with those who were not. To illustrate the bias in the OLS estimator, we reanalyzed the data using OLS, as shown in Table 5. In both samples that exclude and include unlisted numbers, the OLS estimates are large and positive (6.2 and 10.7) without covariates other than controls for the experimental strata. The inclusion of covariates reduces the size of the estimated treatment effects, but they still remain significantly greater than zero. The estimate of 2.7 for the sample without unlisted numbers has a t-ratio of more than 8, and the estimate of 4.4 for the entire sample has a t-ratio of more than 14. Including higher-order polynomial and hundreds of interactions of the covariates leaves the estimates virtually unchanged, 2.8 and 4.6, respectively, with t-ratios of 10 and 17.[10]

Thus the multivariate OLS results contrast with the instrumental variables results in two ways. First, the OLS results are significantly larger. The 95% confidence intervals of the OLS and IV estimates do not overlap. Second, the OLS results change markedly when covariates are included. Indeed, the success of OLS estimation hinges on whether the covariates eliminate bias. The IV estimates, on the other hand, scarcely change with the inclusion of covariates, which are statistically independent of the randomly assigned instrumental variable. Third, the OLS estimates are sensitive to the way in which the sample is defined. The inclusion of unlisted numbers dramatically increases the OLS estimate of the treatment effect. By contrast, the inclusion of unlisted numbers has negligible effects on the instrumental variables estimates.

The logic underlying matching is similar to OLS. Matching compares those contacted by phone with uncontacted people who share identical background characteristics. In contrast to the assumption stated in Eq. (7), OLS imposes the somewhat more demanding assumption:

$$E(Y_0 \mid f(X), D = 1) = E(Y_0 \mid f(X), D = 0), \tag{9}$$

where $f(X)$ refers to a linear function of the variables in X, which may include interactions and polynomials. Matching will produce unbiased results if the requirement in Eq. (7) is satisfied after conditioning on *any* function of X, whether linear or not. However, this assumption remains a strong one, because it is not clear whether there remain unobserved differences that may cause reachable and nonreachable people to vote at different rates.

4.4 *Comparing Matching Estimates to the Experimental Benchmark*

We conduct two sets of matching analyses: one in which we exactly match on the covariates and discard treated observations that do not find matches, and another in which we find inexact matches for those treated group observations that do not find exact matches. Our procedures for exact and inexact matching are described in the appendix. In light of the alternative conventions for estimating standard errors, we report two sets of standard errors for the matching estimates. The first is based on the algorithm proposed by Abadie and Imbens (2004), and the second relies on the method used by Becker and Ichino

[10]These OLS models included the following covariates: age, age squared, household size, gender, newly registered, previous vote in 1998 and 2000, state, competitiveness, and fixed effects for county, state senate district, and statehouse district. These covariates were interacted up to four times.

Table 6 Sensitivity of OLS and Matching Estimates to Changes in Sample Definition

	OLS	*Exact matching*	*Inexact matching*
Sample excludes unlisted numbers			
Treatment effect[a]	2.7	2.8	2.9
(SE)	(0.3)	(0.3)[b]	(0.1)[b]
(SE)		(0.4)[c]	(0.3)[c]
N[d]	1,905,320	22,711	25,028
Matched[e]		90.7%	99.9%
R^2	0.28		
Sample includes unlisted numbers			
Treatment effect[a]	4.4	4.4	4.4
(SE)	(0.3)	(0.3)[b]	(0.1)[b]
(SE)		(0.3)[c]	(0.3)[c]
N[d]	2,474,927	23,467	25,034
Matched[e]		93.7%	99.9%
R^2	0.30		

[a]For the OLS results, the treatment effect is the slope coefficient on the contact variable included in the regression: Vote 2002 = a + β_1 contact + $\Sigma\gamma_i$ covariates$_i$.
[b]Standard errors for treatment effects estimated with Abadie and Imbens (2004) method (see appendix for details).
[c]Standard errors for treatment effects estimated with Becker and Ichino (2002) method (see appendix for details).
[d]For the matching analysis, this indicates the number of contact group individuals who were matched to the control group; for the OLS analysis it indicates the total number of observations.
[e]Percent of contacted group with at least one identical match in the control group.

(2002). In this application the two sets of standard errors lead to identical substantive conclusions.

Table 6 summarizes the results of our matching analyses.[11] In this table, treated individuals were matched to a comparison group that comprises subjects randomly assigned to the control group and members of the treatment group who were not contacted.[12] Table 6 also reports OLS estimates for purposes of comparison given the analogous assumptions that OLS and matching make.

We begin by summarizing the top panel of the table, which reports the estimates obtained using only those subjects with listed phone numbers. Matching overestimates the impact of brief nonpartisan GOTV phone calls. Exact matching generates an estimated treatment effect of 2.8; inexact matching, 2.9.[13] Both estimates have t-ratios of at least 7. There is no overlap between the 99% confidence intervals formed around the experimental benchmark and the 99% confidence intervals formed around the matching estimates.

[11]The authors used *Stata* 8/SE to perform exact matching, corroborating the results using two other programs. See the appendix for details.
[12]This matching was done with replacement, although due to the large number of subjects in the comparison group, less than 1% of the treatment group was matched repeatedly to the same observation in the comparison group.
[13]We have also applied the propensity score matching to these data using a more nuanced set of covariates, which included the year a person registered to vote and their voter participation in each primary, special, and general election since 1998. In order to obtain propensity scores, we used Sekhon's (2004) GenMatch program, which uses a genetic algorithm to choose an optimal propensity score model. Due to the size of our data set and the memory demands of the program, we divided the sample randomly into 100 samples. The mean estimate based on 50 jackknife samples is 2.2 for the listed sample and 3.6 for the sample that includes unlisted numbers. In other words, matching on a fuller set of covariates using propensity scores produces results that are slightly smaller but similar to the matching results reported in the tables.

The matching estimates are not only statistically distinguishable from the experimental benchmark; they have quite different substantive implications. The experimental benchmark (b = 0.4) suggests that this calling campaign generated one additional voter for every 250 completed calls. At 50 cents per completed call,[14] each additional vote cost $125. The matching estimate (b = 2.8), on the other hand, implies that the campaign generated one additional voter for every 36 calls, which means that each vote cost $18. If the matching estimates were to be believed, they would imply that nonpartisan calls by commercial phone banks rank among the most cost-effective voter mobilization tactics. In fact, just the opposite is true.

If matching on observables were sufficient to eliminate bias, the inclusion of observations with unlisted phone numbers should improve the performance of this estimator by increasing the pool of potential matches. Table 6, however, reveals that matching becomes more biased when we include people with unlisted phone numbers. Both exact and inexact matching generate an estimated treatment effect of 4.4 with standard errors ranging from 0.1 to 0.3.[15] Again, the estimates indicate the sensitivity of matching to unobserved heterogeneity. Lacking information about whose phone numbers are listed, analysts who use survey data to assess the effects of voter mobilization calls would obtain biased estimates from matching—even when voters are targeted at random, as in this experiment. Imai (2005, Table 9), for example, reports a treatment effect of 6.5 (SE = 3.2) when applying matching to a sample that included unlisted numbers.

It should be stressed that we obtain these biased estimates in spite of the fact that, by any diagnostic criteria, our covariates are perfectly balanced. Most applications of matching place great emphasis on the procedures used to achieve balance among the covariates. Our results illustrate that these diagnostic criteria are insufficient. Despite perfect balance, one obtains biased estimates in this application.

Table 6 also shows that the matching estimates track the OLS estimates quite closely. The absolute difference between any pair of OLS and matching estimates never exceeds 0.2. It appears that the distinction between parametric and nonparametric estimators is inconsequential in this application. Although there are reasons to believe that age bears a quadratic relationship to voting (see Gerber and Green 2000), the linearity assumptions of the OLS are innocuous here. Despite the fact that the OLS regression omitted the quadratic effect of age, the OLS results are nearly identical to the matching results. The same may be said for interaction effects among the independent variables; these too are ignored by the OLS model, yet the matching estimator generates results that are scarcely different. The key distinction is not between parametric and nonparametric estimators but rather between estimators that address the selection problem by use of randomization (instrumental variables) as opposed to estimators that grapple with selection by use of covariates (OLS and matching).

4.5 *Varying the Severity of the Selection Problem*

The inadequacy of OLS and matching in this application can be further demonstrated by varying the severity of the selection problem. Researchers are often in the position of

[14]This figure is lower than the costs incurred in this experiment. We paid the two phone banks approximately 60 cents per completed call, plus an $800 setup charge. In addition, we paid $5000 for the voter lists. Thus a total of $20,800 was spent to complete 25,000 calls, which generated roughly 100 votes.

[15]The results in Table 6 remain unchanged when noncompliers in the treatment group are excluded from the analysis.

Table 7 Sensitivity of OLS and matching estimates to the severity of selection problem

	OLS		*Exact matching*		*Inexact matching*	
	Low-contact phone bank	*High-contact phone bank*	*Low-contact phone bank*	*High-contact phone bank*	*Low-contact phone bank*	*High-contact phone bank*
A. Excluding unlisted phone numbers						
Treatment effect[a]	4.7	1.2	4.7	1.4	4.8	1.4
(SE)	(0.4)	(0.3)	(0.5)[b]	(0.4)[b]	(0.1)[b]	(0.1)[b]
(SE)			(0.5)[c]	(0.5)[c]	(0.5)[c]	(0.5)[c]
N[d]	1,875,338	1,875,330	9,324	13,334	10,266	14,762
Matched[e]			90.8%	90.3%	99.9%	99.9%
R^2	0.29	0.29				
B. Including unlisted phone numbers						
Treatment effect[a]	6.4	3.0	6.0	3.2	6.2	3.0
(SE)	(0.4)	(0.3)	(0.4)[b]	(0.4)[b]	(0.1)[b]	(0.1)[b]
(SE)			(0.5)[c]	(0.5)[c]	(0.5)[c]	(0.4)[c]
N[d]	2,432,644	2,432,707	9,630	13,796	10,268	14,765
Matched[e]			93.8%	93.4%	99.9%	99.9%
R^2	0.30	0.30				

Note. The high-contact phone bank successfully contacted 49.3% of the 29,982 people it attempted, and the low-contact phone bank successfully contacted 34.2% of the 29,990 people it attempted to reach.

[a]For the OLS results, the treatment effect is the slope coefficient on the contact variable included in the regression: Vote 2002 = a + β_1 contact + $\Sigma\gamma_i$ covariates$_i$.

[b]Standard errors for treatment effects estimated with Abadie and Imbens (2004) method (see appendix for details).

[c]Standard errors for treatment effects estimated with Becker and Ichino (2002) method (see appendix for details).

[d]For the matching analysis this indicates the number of contact group individuals who were matched to the control group; for the OLS analysis it indicates the total number of observations.

[e]Percent of contacted group with at least one identical match in the control group.

working with observational data of varying quality in an effort to estimate causal effects. We now show how two studies that involve the same treatment may nevertheless generate radically different estimated effects using observational methods. In our experiment, individuals in the treatment group were randomly assigned to one of two phone banks. One phone bank, henceforth called the "high-contact" phone bank, made more attempts to reach subjects than the other and completed 14,773 contacts. The other phone bank completed only 10,270 contacts. The selection problem is arguably more severe for the low-contact phone bank. The people who were reached by the low-contact phone bank may have had greater propensities to vote than people reached by the high-contact phone bank.

When the experimental data are analyzed using instrumental variables estimation, neither phone bank is found to have an effect. Excluding unlisted numbers, the low-contact phone bank's effect was 0.6 (SE = 0.9); the high-contact phone bank's effect was 0.2 (SE = 0.6). Including unlisted numbers produces effect estimates of 0.0 in both cases. The question is whether matching renders estimates that coincide with this experimental baseline.

As shown in Table 7, the matching estimates vary according to the severity of the selection problem. For the sample with listed phone numbers, the exact matching estimate and the OLS estimate are both 4.7. The corresponding estimates for the high-contact phone

bank are 1.2 and 1.4. The same pattern holds for the sample that includes unlisted numbers. Exact matching reveals a treatment effect of 6.0 for the low-contact phone bank and 3.2 for the high-contact phone bank. Although the actual effects of both phone banks are small and insignificant and although both phone banks read identical scripts and had very similar callers, matching implies that the low-contact phone bank had significantly stronger effects than the high-contact phone bank.

Had this study not had an experimental benchmark, the observational analysis would have suggested the mistaken recommendation to rely on low-contact phone banking strategies. According to the matching estimates, the low-contact approach generates one vote for every 21 contacts, which is more than three times more efficient than the high-contact strategy and highly cost effective in relation to other voter mobilization tactics. These matching estimates, however, provide a misleading basis for evaluating the two phone banking strategies. In fact, the low-contact strategy is not significantly better than the high-contact strategy.

In sum, least squares, exact matching, and inexact matching generate similar estimates. When unlisted numbers are excluded, the estimates suggest that approximately 2.8 votes are generated per 100 contacts. When unlisted numbers are included, worsening the selection problem, this estimate jumps to 4.4 per 100 contacts. The estimate continues to grow when we focus attention on the low-contact phone bank, 4.7 per 100 contacts excluding unlisted numbers and approximately 6.0 votes per 100 contacts when unlisted numbers are included in the sample. These estimates overshoot the experimental benchmark (4 or 5 votes per 1000 contacts) by a factor ranging from 6 to 12.

5 *Discussion*

This essay illustrates the useful methodological role that experiments can play in evaluating observational methods. Political scientists have rarely used the LaLonde (1986) approach to gauge the usefulness of the large and growing stock of statistical tools for political science applications. The recent surge of interest in experimentation provides the discipline with an opportunity to develop an empirically grounded sense of the conditions under which various methodological approaches provide a sound basis for causal inference.

Using a voter mobilization experiment, we assessed the performance of matching under conditions that previous scholars have identified as favorable to the success of matching methods. The availability of a very large control group allowed us to find exact matches for more than 90% of the treated observations and near-exact matches for the remainder.[16] The voter files used here contained more information about voters' past behavior than previous research using matching to assess the effects of voter mobilization calls. Nevertheless, exact matching failed to eliminate bias and seemed to offer little improvement over OLS regression. Not only did matching fail to recover the experimental benchmark, it also turned out to be sensitive to subtle variations in sample definition and contact rates. In every instance, matching exaggerated the cost-effectiveness of nonpartisan voter mobilization calls from commercial phone banks.

[16]Studies evaluating the performance of matching estimators typically have treatment and control groups on the order of 3% the size of the treatment group in our data (e.g., Dehejia and Wahba 1999; Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura et al. 1998). Had our N been this size, we would have been precluded from using exact matching. Propensity score matching would have added additional sources of discretion and uncertainty. We would have been confronted with decisions regarding the construction of the propensity score model, how to assess balance, and the type of matching method to use.

The failure of matching in this application stems from the strong assumptions that this method imposes regarding the interchangeability of reachable and nonreachable individuals. Answering the phone and listening to a brief GOTV appeal evidently reveals information about an individual's propensity to vote beyond what can be predicted based on their background characteristics or previous voting behavior. In hindsight it is clear that matching fails to remedy the selection problem.

Unfortunately, researchers seldom have the luxury of comparing their observational results to experimental benchmarks. Instead, they must determine *ex ante* whether their matching models are likely to generate unbiased results. This exercise may be more difficult than is typically assumed. Most applications of matching focus on the empirical problem of achieving balance among the covariates. Statistical assessments of balance are the main diagnostic tools by which the performance of matching is assessed. Balance is clearly important, but the question remains: How does one know whether matched observations are balanced in terms of the *unobserved* causes of the dependent variable? Because the issue concerns unmeasured variables, it can be addressed only by means of a theoretical investigation of the selection process and the possible biases associated with it. This theoretical investigation will inevitably involve a certain amount of guesswork about the sources of bias and the extent to which they distort the matching estimates. In that regard, matching is no different from other observational estimation approaches.

To say that uncertainty surrounds the application of matching to observational data is not to say that matching is incapable of generating unbiased estimates. Given the right set of covariates, the causal effects of phone calls can be properly estimated using this observational approach. The problem is that researchers rarely know whether the covariates at their disposal are adequate.

One practical consequence of this uncertainty is that the standard errors associated with observational estimators are biased downward. The nominal standard errors associated with matching estimates, for example, are reported based on the implicit assumption that the biases associated with matching are known with certainty. Clearly this assumption is false. The actual standard errors associated with estimators that are potentially susceptible to bias are often much larger (Gerber et al. 2004). For instance, in Table 5, the nominal mean-squared error of the exact matching estimates in the upper panel is $(.3)^2 = .09$, but when we take bias into account, this figure rises to $(.3)^2 + (2.8 - 0.5)^2 = 5.38$. In other words, in this application error due to bias overshadows error due to sampling variability.

For these reasons, political scientists should be cautious about the claims that are made with regard to matching. Matching cannot be credited with "adjusting estimates of the treatment effect as if the whole study were a randomized experiment" (Barabas 2004, p. 692). The advantages of matching instead have do with its nonparametric properties, which allow the researcher to estimate treatment effects without making restrictive assumptions about the functional form through which the covariates affect the dependent variable. Whether this approach leads to more accurate estimates will obviously depend on the application. In some applications, the parametric assumptions about linearity and additivity will be adequate. In other applications, the inadequacy of these assumptions will introduce the sort of bias that matching can correct. The point to bear in mind, however, is that modeling the relationship between the covariates and the outcome variable is just one part of a larger inference problem. The findings reported here illustrate the fact that nonparametric methods may relax these assumptions about the covariates without appreciable gains in accuracy.

Appendix

Matching Procedure

When performing exact matching, we matched treated observations to control group observations that shared the exact same values on covariates. As illustrated in Fig. A1, multiple observations in the treated group are matched to treatment group observations, if possible. For example, Observation 9 matches two observations, 108 and 109, in the comparison group. When n multiple matches are found for a given treated observation, the voting rates among the comparison observations are averaged together; this procedure is equivalent to weighting each of these comparison-group matches by a factor of $1/n$. (A less efficient alternative to this procedure is to match each treated observation to one or more randomly chosen exact matches.) When all of the matches have been found, unmatched cases are discarded, and the average voting rate in the treated group is compared to the average voting rate in the comparison group using the weighting procedure described below.

Code to implement exact matching was implemented in *Stata* 8/SE. Our program matches control group observations to treated group observations that share the same values on covariates of interest (as shown in Fig. A1), calculates the treatment-on-treated effect for the matched observations, and estimates the appropriate standard errors. (The bias corrections proposed by Abadie and Imbens 2004 do not apply to exact matching.) The program was checked using a simulated data set in which the proper estimates were calculated by hand. The program was also tested against other matching programs available for *Stata*. While these programs were designed to perform propensity score matching (hence our need to write an exact matching program), it was possible to construct a hypothetical data set in which propensity score matching was equivalent to exact matching. Our program successfully replicated these results.

Although we draw exact matches from an extremely large comparison group, we are not able to match 100% of the treated group. We discard between 7 and 9% of the treated group in the exact matching analyses. In order to allay concerns that excluding these cases introduces bias in our matching estimates, we simulated constricting the size of the comparison group to see how the percentage of the treatment group matched affects the matching estimates. In this simulation, we randomly sampled s number of observations from the comparison group, calculated exact matching estimates, and recorded the percent

	Treated Subjects				
	Age	Gender	Precinct	Previous Vote	Vote 2002
1	30	1	10	1	0
2	45	0	15	1	1
3	19	0	12	0	1
4	32	1	16	1	1
5	55	1	16	0	0
6	42	0	15	1	1
7	70	1	10	0	0
8	24	1	12	0	1
9	21	0	14	1	0
10	34	1	14	0	1
11	62	0	10	0	0

	Untreated Subjects				
	Age	Gender	Precinct	Previous Vote	Vote 2002
101	55	1	16	0	0
102	45	0	15	1	1
103	19	0	12	1	1
104	56	1	14	0	0
105	28	1	12	0	0
106	18	1	12	0	1
107	19	0	12	0	1
108	21	0	14	1	0
109	21	0	14	1	0
110	25	0	10	1	0
111	62	0	10	1	0

Fig. A1 Exact matching procedure example.

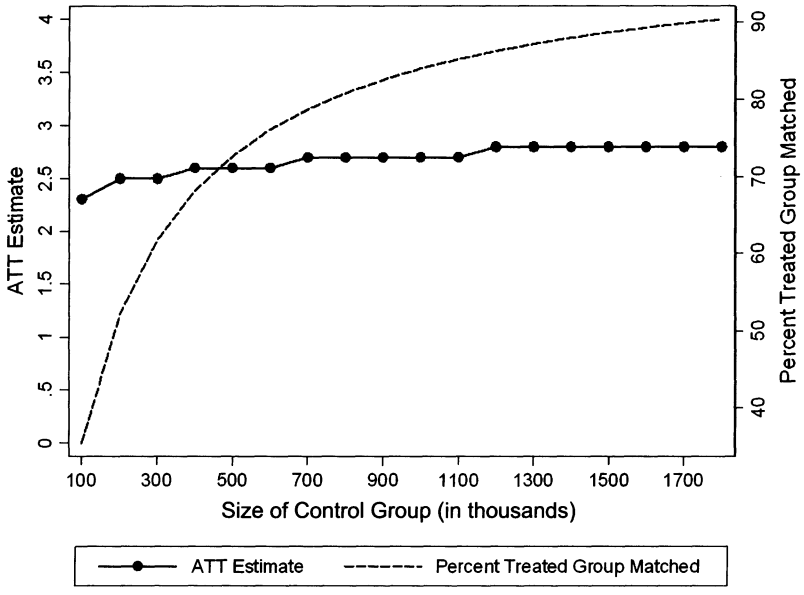

Fig. A2 Matching estimates and comparison group size.

of the treated group that found exact matches. We began s at 100,000 and incremented it to 1.8 million in 100,000 intervals. For each s, we averaged over 100 simulations. The results of this exercise are graphically depicted in Fig. A2. These average estimates are quite stable across the range of percent treated group matched. Although exact matching discards a portion of the observations in the treatment group, the exclusion of these cases appears to be inconsequential.

Nevertheless, we sought to further allay concerns by reducing the number of the treated group observations excluded from the analyses by including inexact matches. The inexact matching estimates were obtained by first finding exact matches and then finding close matches for the unmatched treated group observations. We developed an incremental strategy to find these close matches. In the first stage, the age variable was recoded into three-year categories and the unmatched treated group observations were matched on the new age category variable along with the other covariates. In successive stages, we continued to match on the three-year age categories and other covariates, but in each stage we excluded geographic variables. Geographic variables were excluded in order of their size, beginning with state-house district, then state-senate district, and ending with county. Note that at each stage treated group observations continued to be exactly matched on important covariates, including state, competitiveness strata, gender, household size, newly registered, and past voting behavior.

For the analyses that combined observations from phone banks (see Table 6), there were only 15 treated group observations in the listed phone number sample that did not have any matches in the comparison group at the end of this process; there were only 9 treated group observations in the unlisted phone number sample that did not (see Table A1 for details). To ensure that the inclusion of inexact matches did not introduce imbalance, we regressed phone contact on all of the covariates using the matched sample. Joint tests of significance found no evidence of imbalance (chi-square[261] = 105.08, p = 1.00 and chi-square[261] = 45.83, p = 1.00 for the listed and unlisted phone number sample, respectively). The inexact matching process worked in a similar fashion for the analyses of

Table A1 Inexact matching process

	Phone banks combined		*Low-contact phone bank*		*High-contact phone bank*	
	Listed	*Unlisted*	*Listed*	*Unlisted*	*Listed*	*Unlisted*
Before inexact matching						
Number of unmatched treated group observations following exact matching	2,332	1,576	946	640	1439	977
First stage: Age in three-year categories						
Number of unmatched treated group observations that found matches	1,523	1,083	627	462	931	647
Number of treated group observations that remain unmatched	809	493	319	178	508	330
Second stage: Exclude state-house district						
Number of remaining unmatched treated group observations that found matches	206	139	93	60	116	84
Number of treated group observations that remain unmatched	603	354	226	118	392	246
Third stage: Exclude state-house and state-senate districts						
Number of remaining unmatched treated group observations that found matches	151	96	54	28	102	67
Number of treated group observations that remain unmatched	458	258	172	90	290	179
Final stage: Exclude state-house district, state-senate district, and county						
Number of remaining unmatched treated group observations that found matches	437	249	168	88	279	171
Number of treated group observations that remain unmatched	15	9	4	2	11	8

the sample segmented by phone bank (see Table A1). Joint tests of significance also found no evidence of imbalance in these matched samples ($p = 1.00$).

Estimating Quantities of Interest

To calculate the average treatment effect for the treated in the matching analyses, we weight the comparison group observations using the method presented in Abadie and Imbens (2004). $\Gamma_M(i)$ refers to the set of indices for all the exact matches in the comparison group associated with treated unit i. T denotes whether an individual received treatment $\{0,1\}$. X_i refers to the covariate pattern that describes unit i in the treated group, and X_l refers to a list of comparison group observations that share a particular covariate pattern.

$$\Gamma_M(i) = \{l = 1, \ldots, N \mid T_l = 0, X_i = X_l\}. \tag{A1}$$

The number of elements in $\Gamma_M(i)$ is denoted by $\#\,\Gamma_M(i)$. Once each treated observation is matched with comparison group observations that share the same covariate pattern, weights are calculated in the following fashion:

$$K_M(i) = \sum_{l=1}^{N} 1\{i \in \Gamma_M(l)\} \frac{1}{\#\Gamma_M(l)}. \tag{A2}$$

Let Y_i denote the voting behavior of unit i. The matching estimator for the average treatment on treated effect that Abadie and Imbens propose is

$$\tau = \frac{1}{N_1} \sum_{i=1}^{N} (T_i - (1 - T_i)K_M(i))Y_i, \tag{A3}$$

where N_1 = the number of observations in the treated group. For exact matching, our implementation of their estimator differs in that we include only treated group observations that find matches in the comparison group.

Because there is currently no agreed-upon method to estimate standard errors for matching estimates, we employ two different estimators. The first uses the following variance formula derived by Abadie and Imbens (2004):

$$\hat{V} = \frac{1}{N_1^2} \sum_{i=1}^{N} (T_i - (1 - T_i)K_M(i))^2 \sigma^2_{T_i}(X_i), \tag{A4}$$

where

$$\sigma^2_{T_i}(X_i) = \frac{1}{2N_1} \sum_{i:T_i=0} \left(\frac{1}{\Gamma_M(i)} \sum_{l \in \Gamma_M(i)} (Y_i - Y_l - \hat{\tau})^2 \right). \tag{A5}$$

The second method uses the variance formula offered by Becker and Ichino (2002), which for our purposes can be written as

$$\tilde{V} = \frac{1}{N_1} Var(Y_i(1)) + \frac{1}{N_1^2} \sum (K_M(i))^2 Var(Y_i(0)). \tag{A6}$$

We do not bootstrap the standard errors because Abadie and Imbens (2004, p. 3) note that the properties of resampling methods are unclear in regard to matching estimates, which "are highly non-smooth functions of the data."

References

Abadie, Alberto, and Guido Imbens. 2004. "Large Sample Properties of Matching Estimators for Average Treatment Effects." Unpublished manuscript, Harvard University.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434):444–455.

Barabas, Jason. 2004. "How Deliberation Affects Policy Opinions." *American Political Science Review* 98(4):687–702.

Becker, Sascha O., and Andrea Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Scores." *Stata Journal* 4:358–377.

Bloom, Howard S., Charles Michalopoulos, Carolyn J. Hill, and Ying Lei. 2002. "Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" Working paper, Manpower Demonstration Research Corporation.

Dehejia, R., and S. Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448):1053–1062.

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics* 125:355–364.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Personal Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94(3):653–664.

Gerber, Alan S., and Donald P. Green. 2005. "Do Phone Calls Increase Voter Turnout? An Update." *Annals of the American Academy of Political and Social Science* 601(September):142–154.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. "The Illusion of Learning from Observational Research." In *Problems and Methods in the Study of Politics*, eds. Ian Shapiro, Rogers Smith, and Tarek Massoud. New York: Cambridge University Press, pp. 251–273.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Nonexperimental versus Experimental Estimates of Earnings Impacts." *Annals of the American Academy of Political and Social Science* 589(1):63–93.

Green, Donald P., and Alan S. Gerber. 2004. *Get Out the Vote! How to Increase Voter Turnout.* Washington, DC: Brookings Institution Press.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5):1017–1098.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(2):261–294.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluator Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64(4):605–654.

Howell, William G., and Paul E. Peterson. 2002. *The Education Gap: Vouchers and Urban Schools.* Washington, DC: Brookings Institution Press.

Imai, Kosuke. 2005. "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99(2):283–300.

Imbens, Guido. 2003. "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review." Technical Report, Department of Economics, University of California—Berkeley.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment." *Quarterly Journal of Economics* 116:607–654.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76:604–620.

Nickerson, David W. 2005. "Measuring Interpersonal Influence." PhD dissertation, Department of Political Science, Yale University.

Plutzer, Eric. 2002. "Becoming a Habitual Voter: Inertia, Resources, and Growth in Young Adulthood." *American Political Science Review* 96(1):41–56.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Rosenbaum, Paul R., and Donald B. Rubin. 1985. "The Bias due to Incomplete Matching." *Biometrics* 41(1):103–116.

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

Rubin, Donald B. 1973. "Matching to Remove Bias in Observational Studies." *Biometrica* 29:153–183. Correction: 1974. 30:728.

Sekhon, Jasjeet S. 2005. "Multivariate and Propensity Score Matching Software." (Available from http://jsekhon.fas.harvard.edu/matching/.)

Sherman, Lawrence W., and Dennis P. Rogan. 1995. "Deterrent Effects of Police Raids on Crack Houses: A Randomized, Controlled Experiment." *Justice Quarterly* 12(4):755–781.

Smith, Jeffrey, and Petra Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Matching Methods?" *American Economic Review, Papers and Proceedings* 91(2):112–118.

Smith, Jeffrey, and Petra Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods?" *Journal of Econometrics* 125(1-2):305–353.