

**Comparing Experimental and Matching Methods using a Large-Scale Field Experiment on
Voter Mobilization**

Kevin Arceneaux
Alan S. Gerber
Donald P. Green

Yale University
Institution for Social and Policy Studies
P.O. Box 208209
77 Prospect St.
New Haven, CT 06520

kevin.arceneaux@yale.edu
alan.gerber@yale.edu
donald.green@yale.edu

Preliminary Draft

June 22, 2004

Please do not cite without authors' permission. Comments and suggestions welcome.

Abstract

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 studies. In an effort to develop estimators that can approximate experimental results using observational data, scholars have given increasing attention to matching and closely related methods such as propensity score matching. In this paper, 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. The experimental findings show that brief paid get-out-the-vote phone calls do not increase turnout, while matching estimators show a significant positive effect.

1. Introduction

By furnishing unbiased estimates of causal parameters, 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, Kling, and Liebman 2001). Political scientists in particular have made 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 enjoys a rich intellectual history in economics. In his seminal essay on this approach, LaLonde (1986, pp.617-618) explains:

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 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.

An observational method that has attracted special attention in recent years is

matching. Matching has been proposed as a nonparametric solution to problems of bias that emerge in observational studies (Rosenbaum and Rubin 1983, 1985). This method compares individuals in a non-randomly generated “treatment group” to similar individuals in a non-randomly 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 readily concede, matching on observed characteristics leaves open the possibility of unobserved differences between groups. The question is whether matching actually works in practice. To date, studies evaluating the performance of matching estimators using experimental benchmarks (Dehejia and Wahba 1999; Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura, Smith, and Todd 1998; Smith and Todd 2003) have obtained mixed results. In the field of labor economics, where most of the empirical applications have been drawn, matching often fails to reproduce experimental results.

Matching has nonetheless attracted increasing interest in political science, and as more political scientists consider using this method, it is important to extend the empirical evaluation of matching to political science applications. This paper 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 succeed in raising voter turnout. Rather than comparing results from different samples, as is often done in studies that evaluate matching, we compare experimental and matching estimators applied to the same sample.

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, exposure to the treatment may be correlated with unobserved causes of voting. This selection problem can be overcome by using the instrumental variables estimator proposed by Angrist, Imbens, and Rubin (1996), which is 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. 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 campaign to those who were not. 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.

It should be stressed that our 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 the vast majority of those who received phone calls. In some instances, the rate of exact matches is 100 percent. The data set also includes a great deal of information about subjects' past voting behavior, demographic characteristics, and geographic location. Indeed, in one of the two states where this experiment took place, the voter file supplies information about whether people voted in each of the last ten elections.

Finally, the sample can be defined in alternative ways in order to shed light on the conditions under which matching performs well. By including subjects with unlisted phone numbers¹, we more closely approximate the way matching might be applied to public opinion surveys, which draw their samples by means of random digit dialing. The use of unlisted numbers also highlights an important difference between experimental and nonexperimental estimation approaches. Instrumental variables estimation is robust to changes in sample definition, and the inclusion of unlisted numbers scarcely affects the IV results. Matching, on the other hand, is sensitive to this change, and bias becomes more severe when the sample expands to include those with unknown phone numbers.

The paper is structured as follows. In the next section, matching estimation is discussed and placed in context of the existing literature. In the third section the data are described. In the fourth section we assess the performance of matching by comparing matching estimates to the experimental benchmark. The empirical results show that even

under what appear to be highly favorable conditions, matching generates biased estimates of the treatment effect, especially when people with unlisted numbers are included in the sample.

2. Matching Estimation

2.1. Theory

In order to estimate causal parameters using observational data, Rosenbaum and Rubin (1983) propose matching participants in the treatment group with similar nonparticipants in a comparison group. Matching is done with the help of a balancing score. Rosenbaum and Rubin (1983: 42) define “[a] balancing score, $b(x)$, [as] a function of the observed covariates x such that the conditional distribution of x given $b(x)$ is the same for the treated ($z = 1$) and control ($z = 0$) units....”

$$x \perp z \mid b(x). \tag{1}$$

Rosenbaum and Rubin note that the best-case scenario occurs when the analyst can match individuals with the exact same set of characteristics in x in treated and control groups. With sufficient data, as is the case in this study, exact matching on covariates is feasible.

In practice, it is rare to find a control group with a large enough reservoir of observations to allow exact matching on all covariates. Some analysts exactly match on a handful of covariates deemed important and use inexact matches for remaining covariates (Imbens 2003). A variant of this approach is to use the covariates to generate predicted probabilities of being treated – so-called propensity scores. Rather than matching directly on the covariates, matches are made on the basis of propensity scores. This approach requires

the researcher to construct a selection model such that the propensity scores are balanced as defined in Equation (1). One advantage of exact matching over propensity score matching is that it automatically produces balance (Rosenbaum and Rubin 1983), thereby eliminating the search for an adequate propensity score model.

In the next section, we will survey studies that have evaluated nonexperimental estimators, such as matching. All of these studies use some variant of propensity score matching due to data constraints. Because the data we use allow us to employ 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. Exact matching greatly simplifies the process of obtaining estimates.

2.2. Literature Evaluating the Performance of Matching

The question of whether nonexperimental estimators can eliminate biases associated with observational data has stimulated a large and growing literature. LaLonde (1986) was the first to evaluate such estimators by comparing results from a randomized field experiment with results from observational data, and his work has served as a model for subsequent studies. He used data from a randomized field experiment evaluating the National Supported Work (NSW) program, comparing individuals in the treatment group from this experiment to individuals in a comparison group drawn from survey data (see also Fraker and Maynard

1987). While LaLonde did not consider matching estimators in his seminal study, others have used his method when evaluating matching.

The literature provides a mixed picture regarding the effectiveness of matching. 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), but meta-analyses of this literature caution that the performance of matching has been mixed. In their meta-analysis Glazerman, Levy, and Myers (2002) find that while matching may reduce bias somewhat, it does not provide much improvement beyond OLS regression. Heckman and colleagues suggest that matching may even exacerbate bias:

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, et al. 1998).

Heckman, Ichimura, and Todd (1997, 1998) and Heckman, Ichimura, Smith, and Todd (1998) suggest that matching performs better under certain circumstances. Two conditions that seem to favor, but by no means assure, better performance: 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 2003: 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 *using the same data*. By matching treated individuals to untreated people from the same population using an extensive set of covariates, we are able to test matching methods under the best possible conditions.

3. Data

In this paper 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 non-partisan 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 non-partisan 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. 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.

After the election, public voting records on each individual were obtained, allowing us to assess whether the GOTV phone appeals stimulated voter turnout. This measurement approach has been used by a number of political scientists studying turnout (Adams and Smith 1980; Gerber and Green 2000).

Table 1 summarizes the data. 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 have 2002 voter records available we removed 1,565 observations, bringing the treatment group total to 59,972 and the control group total to 1,845,348. An additional 569,607 households that had unlisted phone numbers in the voter files were randomly assigned to treatment and control groups. Since individuals in these households cannot be reached by phone, including them in the sample exacerbates the selection problem while at the same time expanding the number of observations available for matching.

[Table 1 about here]

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, age, the number of registered voters in a household, and the state house district. Because the randomization occurred within competitive and uncompetitive strata, the randomization check was carried out at this level. As expected, the chi-squares for each stratum are nonsignificant. For respondents with listed phone numbers, the tests break down as follows: Iowa noncompetitive, $p = 0.38$; Iowa competitive, $p = 0.69$; Michigan noncompetitive, $p = 0.60$; Michigan competitive, $p = 0.23$. For respondents with unknown phone numbers, the tests are: Iowa noncompetitive, $p = 0.25$; Iowa competitive, $p = 0.75$; Michigan noncompetitive, $p = 0.65$; Michigan competitive, $p = 0.72$.

[Table 2 about here]

In the analyses that follow, we vary two aspects of the matching process: 1) the number of covariates used in matching and 2) the inclusion or exclusion of households with unlisted phone numbers. The voter registration lists from which are subjects were drawn include a great deal of information about past voting history, demographic characteristics, and geographic location. By varying the number of matching covariates we may track the effects of model specification on the bias associated with matching. For each data set we match on

three different sets of covariates: small, medium, and large. Table 3 lists the variables. The small set of covariates includes only a handful of demographic variables, the state of residence, and whether the individual is a newly registered voter. The medium set of covariates adds previous voting behavior in 2002 and the competitiveness of the district, and the large set adds previous voting behavior in 1998, specific geographic location, and additional demographic information. Additional demographics, competitiveness of congressional district, geographic location, and previous voting behavior increase the precision with which voting behavior is explained.

[Table 3 about here]

In order to illustrate the value of these covariates, Table 4 displays crosstabulations of vote in 2002 by the two previous elections. Among those who voted in the 1998 and 2000 elections, 83.2 percent voted in the 2002 elections, as compared to 19.5 percent 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.

[Table 4 about here]

Covariates such as previous voting behavior, household size, date of registration, party registration, and geography arguably provide an adequate selection model for participation in GOTV phone experiments (Imai 2003). It is possible, however, that attributes besides the background characteristics available in the voter files predict both phone contact and voting.

This is the empirical question that a randomized experiment allows us to answer.

4. Analysis

4.1. Establishing the Experimental Benchmark

Because some individuals either refused to listen to the GOTV message or did not answer the phone, only 41.8 percent 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. Angrist, Imbens, and Rubin (1996) derive a statistical solution to this selection problem, which Gerber and Green (2000) apply to randomized voter mobilization experiments. Following Gerber and Green (2000), we note that the population can be divided into two groups, those who are reachable by phone and those who are not. Let α be the proportion of the population that is reachable. Let p_{nr} be the probability that a nonreachable person votes, and let p_r be the probability that a reachable person votes in the absence of the experimental treatment. Let t be the effect of the treatment on those who receive it. Thus, $p_r + t$ is the probability that a reachable person votes after exposure to the treatment.

In the control group, we do not observe whether someone is reachable; we only observe the voting rate for the whole group. Given the terms defined above, the expected voting rate for the control group $V_{T=0}$ is:

$$V_{T=0} = \alpha p_r + (1 - \alpha) p_{nr}. \quad (3)$$

When treatment and control groups are formed randomly, both groups have the same expected proportions of reachable and nonreachable people. The expected voting rate for the treatment

group $V_{T=1}$ is therefore:

$$V_{T=1} = \alpha(p_r + t) + (1 - \alpha)p_{nr}. \quad (4)$$

The parameter α can be estimated by dividing the number of people in the treatment group who are contacted by the number of people assigned to the treatment group. Using Equations (4) and (5) to solve for t suggests an estimator:

$$\hat{t} = \frac{\hat{V}_{T=1} - \hat{V}_{T=0}}{\hat{\alpha}}. \quad (5)$$

This estimator is equivalent to an instrumental variables (IV) estimator for the treatment effect, where an indicator for random assignment to the treatment group is used as an instrument for actual contact by the phone bank.

The IV estimates (which in this case are equivalent to two-stage least squares estimates) are displayed in Table 5. 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 percent confidence region extends from -0.6 to 1.3. Including unlisted phone numbers in the analyses increases the sample size but diminishes the proportion of people contacted in the treatment group ($\hat{\alpha}$). 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 percent 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 implies that the matching estimates below should be biased slightly *downward*. As we will see, this small bias due to sampling is offset by strong upward biases associated with the matching estimator.

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.

[Table 5 about here]

4.2. 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 contacted ($V_{T=1}^*$) with those who were not ($V_{T=0}^*$). Regressing the vote in 2002 on phone contact is tantamount to the following estimator of the treatment effect:

$$\hat{t}^* = \hat{V}_{T=1}^* - \hat{V}_{T=0}^* \quad (6)$$

Ignoring sampling variability, $V_{T=1}^*$ and $V_{T=0}^*$ may be expressed as

$$\hat{V}_{T=1}^* = p_r + t, \quad (7)$$

$$\hat{V}_{T=0}^* = \gamma p_{nr} + (1 - \gamma)V_{T=0}, \quad (8)$$

where γ = proportion of untreated group who are in the randomized control group.

Therefore, the OLS estimator can be expressed as

$$\hat{V}_{T=1}^* - \hat{V}_{T=0}^* = t + (1 - \alpha\gamma)(p_r - p_{nr}). \quad (9)$$

The term in equation (9) to the right of the plus sign represents the bias in the OLS estimator.

When $p_r = p_{nr}$, this term cancels, and the bias is zero. In other words, when reachable and unreachable people do not differ in their propensities to vote, OLS generates unbiased estimates of the average treatment effect. To illustrate the bias in the OLS estimator, we reanalyzed the data using OLS, as shown in Table 6. 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.

When additional covariates (X) are included in this model, the relevant question is whether $p_{r|X} = p_{nr|X}$. In other words, regression will produce unbiased estimates if, conditional on linear functions of the covariates, the reachable and nonreachable have similar voting propensities. In our data, the inclusion of covariates reduces the size of the estimated treatment effects, but they still remain significantly greater than zero. The estimate of 2.8 for the sample without unlisted numbers has a t-ratio of more than 8, and the estimate of 5.0 for the entire sample has a t-ratio of more than 16. Including higher order polynomials and hundreds of interactions of the covariates leaves the estimates virtually unchanged, 2.8 and 4.8, respectively, with t-ratios of 10 and 17.²

Thus, the OLS results contrast with the instrumental variables results in two ways. First, the OLS results are significantly larger. The 95 percent confidence intervals of the OLS and IV estimates do not overlap. Second, 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 no effect on the instrumental variables estimator.

[Table 6 about here]

The logic underlying matching is similar to OLS. Matching compares those contacted by phone with uncontacted people who share identical background characteristics. The assumption is that for people with identical background characteristics $p_r = p_{nr}$. This assumption is somewhat less demanding than the corresponding assumption imposed by OLS, which requires that $p_r = p_{nr}$ conditional on linear functions of X (where X may include interactions and polynomials). Matching will produce unbiased results if the requirement that $p_r = p_{nr}$ 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 conditioning on the observables eliminates the unobserved differences that may cause reachable and nonreachable people to vote at different rates.

4.3. Comparing Matching Estimates to the Experimental Benchmark

By making exact matches on covariates – rather than imperfect matches on a vector of propensity scores – we compare treated and untreated voters who share the same observed attributes. If controlling for observables is sufficient to eliminate bias, matching should produce estimates that coincide with the experimental benchmark described above.

Table 7 summarizes the results of our matching analyses.³ 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.⁴ Table 7 also reports OLS estimates for purposes of comparison given the analogous assumptions that OLS and matching make about p_r and p_{nr} . The top panel of the table shows estimates obtained using only those subjects with listed phone numbers. Matching overestimates the impact of brief non-partisan GOTV phone calls in every instance, providing estimates ranging from 2.8 to 3.7. Augmenting the list of control variables reduces but does not eliminate bias. Even in the large covariate set, the estimates generated by matching are seven times larger than the experimental estimates.

[Table 7 about here]

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 7, however, reveals that matching becomes more biased when we include people with unlisted phone numbers. Estimates range from 4.4

to 6.1. Even when matching is performed using a large set of covariates, the estimated treatment effects remain eleven times the size of the experimental benchmark.⁵ Evidently, those with unlisted phone numbers have lower propensities to vote for reasons that are not captured fully by the covariates.

Table 7 also shows that the matching estimates track the OLS estimates quite closely. In some cases, the matching estimates are smaller; in other cases, larger. The absolute difference between any pair of OLS and matching estimates never exceeds 0.4. It appears that the distinction between parametric and nonparametric estimators is of little consequence in this application. Instead, the key distinction is 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).

4.4. Two Further Tests

The inadequacy of OLS and matching is driven home by two further analyses. The first, reported in Table 8, disaggregates the data according to the phone bank that conducted the calls. In this study, 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.

[Table 8 about here]

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 banks 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. The results above suggest that as the selection problem worsens, the upward bias of these estimators increases. Thus, we should expect to see larger estimates for the phone bank that had a lower contact rate.

As expected, the matching estimates vary according to the severity of the selection problem. The selection problem – by definition – is more severe for the low contact phone bank, and the matching estimates range from 4.7 to 8.1. For the high contact phone bank, the bias is smaller, with estimated treatment effects ranging from 1.4 to 4.5. Although the actual effects of both phone banks are close to zero, matching finds the low contact phone bank to have significantly stronger effects than the high contact phone bank. Holding constant the list of covariates used to address bias, the performance of nonexperimental estimators deteriorates as the selection problem becomes more acute.

[Table 9 about here]

The obvious response to this problem is to search for better covariates. Table 9, however, suggests that even marked improvements in the quality of one's covariates may fail to eliminate bias. The data from Iowa contain information about each voter's participation in

the previous *ten* annual elections. Indeed, we selected Iowa for study because it furnishes unusually extensive vote history. If one were unaware of the benchmark experimental results, one might be tempted to think that this long list of covariates would suffice to control for the selection problem in this application. Table 9 shows that matching severely overestimates the treatment effects even when matches are formed based on extensive information about prior behavior. The estimated treatment effects vary from 2.5 to 4.4, depending on whether unlisted numbers are included. (Dropping age from the match criteria dramatically increases the number of matched observations but leaves the estimates unchanged). These estimates are significantly greater than the corresponding IV estimates for Iowa, which are 0.6 (SE=0.6) excluding unlisted numbers and 0.2 (SE=0.8) including them.

5. Discussion

Using voting behavior as an application, we assessed the performance of matching under conditions that previous scholars have identified as especially favorable to the success of matching methods. The availability of a very large control group allowed us to match exactly on covariates, freeing us from the many choices that arise in propensity score matching.⁶ Both Iowa and Michigan maintain voter files with a great deal of information on registered voters' background characteristics and previous voting behavior. Nevertheless, exact matching fails to eliminate bias and seems to offer little improvement over OLS regression.

The failure of matching to reduce bias in this case stems from the strong assumptions

that this method makes 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 failed to remedy the selection problem.

Given the large number of characteristics we were able to take into account, the failure of the selection model should serve as a cautionary note to those applying matching to data for which there is no experimental counterpart. The disappointing performance of matching in this application serves as a reminder that matching should not be used without careful consideration of the validity of the substantive assumptions on which the method relies. Even when these assumptions seem plausible, residual caution must remain. In the absence of an experimental study to confirm the matching results, one cannot know whether one has overlooked an important source of bias. As this study demonstrates, large datasets and long lists of covariates may provide a false sense of security.

More broadly, this paper illustrates the useful methodological role that experiments can play in evaluating observational methods. Political science has rarely used the LaLonde (1986) approach to gauge the usefulness of its large and growing stock of statistical tools. As a result, the discipline lacks an empirically grounded sense of the conditions under which various methodological approaches provide a sound basis for causal inference.

Appendix: Exact Matching Routine

We matched control group observations to treated group observations that shared the exact same values on covariates. As illustrated in Table A1, multiple observations in the treated group are matched to treatment group observations, if possible. As the case in the bottom row demonstrates, near matches are rejected. Because the control group in the field experiment analyzed in this paper contains over one million observations, it was possible to find enough exact matches in the control group to 90 to 100 percent of the treated group.

[Table A1 about here]

In order to execute this matching procedure, we wrote a program in Stata, which is available on the Internet for download (website). The program matches control group observations to treated group observations that share the same values on covariates of interest (as shown in Table A1), calculates the treatment on treated effect for the matched observations, and estimates the appropriate standard errors. The program was benchmarked using a simulated dataset in which the answer was known. It was also tested against other matching programs available for Stata. While these programs were designed to do propensity score matching (hence the need to write the exact matching program), it was possible to construct a dataset in which the propensity score vector contained exact matches. Our program successfully replicated results.

References

- Adams, William C. and Dennis J. Smith. 1980. "Effects of Telephone Canvassing on Turnout and Preferences: A Field Experiment." *Public Opinion Quarterly*, 44 (Autumn): 389-95.
- 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): 444-55.
- 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): 1053-62
- Fraker, Thomas and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources*, 22 (2): 194-227.
- 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): 653-64.
- Glazerman, Steven, Dan M. Levy, and David Myers. 2002. "Nonexperimental Replications of Social Experiments: A Systematic Review." *Mathematica Policy Research*, Inc.
- Green, Donald P., and Alan S. Gerber. 2004. *Get Out The Vote! How to Increase Voter Turnout*. Washington, D.C.: Brookings Institution Press.
- Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, 66 (5): 1017-98.
- _____, _____, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies*, 65 (2): 261-94.
- _____, _____, and _____. 1997. "Matching as and Econometric Evaluator Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies*, 64 (4): 605-54.
- Howell, William G., and Paul E. Peterson. 2002. *The Education Gap: Vouchers and Urban Schools*. Brookings Institution Press: Washington, DC.
- Imai, Kosuke. 2003. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." Unpublished manuscript, August 19th.

- Imbens, Guido. 2003. "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review." Technical Report, Department of Economics, U.C. 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: 604-20.
- Rosenbaum, Paul R. and Donald B. Rubin. 1985. "The Bias Due to Incomplete Matching." *Biometrics*, 41 (1): 103-16.
- _____, and _____. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41-55.
- 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. 2003. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*. Forthcoming.
- _____, and _____. 2001. "Reconciling Conflicting Evidence on the Performance of Matching Estimators." *American Economic Review, Papers and Proceedings* 91 (2): 112-18.

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.

Table 2: Covariate Balance between Treatment and Control Groups

A. Unlisted Phone Numbers Excluded

Covariate	Strata							
	Iowa A		Iowa B		Michigan A		Michigan B	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
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
Number of Observations	15,000	85,931	15,000	289,163	14,972	1,153,072	15,000	317,182

Numbers in cells are means.

Table 2 continued

B. Unlisted Phone Numbers Included

Covariate	Strata							
	Iowa A		Iowa B		Michigan A		Michigan B	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
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
Number of Observations	24,000	137,490	24,000	462,663	18,222	1,403,814	18,281	386,457

Numbers in cells are means.

Table 3: Covariates Used in Matching Analyses

Covariates Used in Matching		
Small	Medium	Large
Age, Household size, Newly registered voter, State	Age, Household size, Newly registered voter, Competitiveness, Vote in 2000, State	Age, Household size, Newly registered voter, Competitiveness, Vote in 2000, Vote in 1998, County, State Senate district, State House district, Gender

Table 4: 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

Table 5: Effect of Phone Calls on Voter Turnout in Iowa and Michigan, 2002
Two-Stage Least Squares Estimates

Covariates*	Sample Excludes Unlisted Numbers		Sample Includes Unlisted Numbers	
	Coefficient (robust s.e.)	Coefficient (robust s.e.)	Coefficient (robust s.e.)	Coefficient (robust s.e.)
Phone Contact	0.4 (0.5)	0.4 (0.4)	-0.0 (0.6)	0.0 (0.5)
State Dummy (1=Iowa)	7.4 (0.4)	-0.6 (0.3)	3.3 (0.3)	-1.2 (0.2)
Competitiveness Dummy in Michigan	4.9 (0.1)	4.0 (0.1)	5.0 (0.1)	3.8 (0.1)
Competitiveness Dummy in Iowa	6.1 (0.2)	5.1 (0.2)	5.8 (0.1)	4.7 (0.1)
Household Size		7.8 (0.1)		9.1 (0.1)
Age		0.5 (0.002)		0.5 (0.002)
Female		-1.0 (0.1)		-1.0 (0.1)
Newly Registered		2.5 (0.1)		5.4 (0.1)
Vote in 2000		41.9 (0.1)		43.6 (0.1)
Missing Values in Female Dummy		-31.9 (0.2)		-27.8 (0.2)
Constant	46.1 (0.1)	-11.5 (0.2)	43.9 (0.1)	-16.9 (0.2)
Number of Observations	1,905,320	1,905,320	2,474,927	2,474,927
F	5,649.20	60,394.42	3,855.41	85,139.61
Adjusted R ²	0.01	0.24	0.01	0.26

Vote 2002 = $\alpha + \beta_1 \text{ contact} + \beta_2 \text{ MI competitiveness} + \beta_3 \text{ IA competitiveness} + \beta_4 \text{ state dummy} + \text{covariates}$
Instruments: Random assignment to treatment group

**Table 6: Effect of Actual Contact on Voter Turnout in Iowa and Michigan, 2002
Ordinary Least Squares Estimates**

Covariates*	Sample Excludes Unlisted Numbers		Sample Includes Unlisted Numbers	
	Coefficient (robust s.e.)	Coefficient (robust s.e.)	Coefficient (robust s.e.)	Coefficient (robust s.e.)
Phone Contact	6.2 (0.3)	2.8 (0.3)	10.7 (0.3)	5.0 (0.3)
State Dummy (1=Iowa)	6.7 (0.3)	-1.0 (0.3)	2.5 (0.3)	-1.6 (0.2)
Competitiveness Dummy in Michigan	4.8 (0.1)	4.0 (0.1)	4.9 (0.1)	3.8 (0.1)
Competitiveness Dummy in Iowa	6.4 (0.2)	5.2 (0.2)	6.1 (0.1)	4.9 (0.1)
Household Size		7.8 (0.1)		9.1 (0.1)
Age		0.4 (0.002)		0.5 (0.002)
Female		-1.0 (0.1)		-1.0 (0.1)
Newly Registered		2.5 (0.1)		5.4 (0.1)
Vote in 2000		41.9 (0.1)		43.6 (0.1)
Missing Values in Female Dummy		-31.9 (0.2)		-27.8 (0.2)
Constant	46.1 (0.1)	-11.5 (0.2)	44.0 (0.1)	-16.9 (0.2)
Number of Observations	1,905,320	1,905,320	2,474,927	2,474,927
F	5,745.62	60,407.23	4,141.81	85,184.28
Adjusted R ²	0.01	0.24	0.01	0.26

Table 7: Matching and OLS Analyses for Contacted Individuals

	Small Covariates		Medium Covariates		Large Covariates	
	Exact Matching	OLS	Exact Matching	OLS	Exact Matching	OLS
Sample Excludes Unlisted Numbers						
Treatment Effect ^a (s.e.)	3.7 (0.3)	3.7 (0.3)	3.0 (0.3)	2.8 (0.3)	2.8 (0.4)	2.7 (0.3)
N ^b	25,043	1,905,320	25,043	1,905,320	22,711	1,905,320
Matched ^c	100%		100%		90.7%	
R ²		0.09		0.24		0.28
Sample Includes Unlisted Numbers						
Treatment Effect (s.e.)	6.1 (0.3)	6.5 (0.3)	4.7 (0.3)	4.9 (0.3)	4.4 (0.3)	4.4 (0.3)
N	25,043	2,474,927	25,039	2,474,927	23,467	2,474,927

Matched	100%	99.9%	93.7%
R²	0.09	0.25	0.30

^aFor the OLS results the treatment effect is the slope coefficient on the contact variable included in the regression:

$$\text{Vote 2002} = \alpha + \beta_1 \text{contact} + \sum \gamma_i \text{controls}_i$$

^bFor 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.

^cPercent of contacted group with at least one identical match in the control group.

Table 8: Matching Analysis, by Phone Bank*

	Small Covariates		Medium Covariates		Large Covariates	
	Low Contact	High Contact	Low Contact	High Contact	Low Contact	High Contact
A. Excluding Unlisted Phone Numbers						
Treatment Effect (s.e.)	5.6 (0.5)	2.1 (0.4)	4.9 (0.5)	1.5 (0.4)	4.7 (0.5)	1.4 (0.5)
N^a	10,270	14,773	10,266	14,768	9,324	13,334
Matched^b	100%	100%	99.9%	99.9%	90.8%	90.3%
B. Including Unlisted Phone Numbers						
Treatment Effect (s.e.)	8.1 (0.5)	4.5 (0.4)	6.6 (0.5)	3.3 (0.4)	6.0 (0.5)	3.2 (0.5)
N	10,270	14,773	10,269	14,770	9,630	13,796
Matched	100%	100%	99.9%	99.9%	93.8%	93.4%

*The “high contact” phone bank successfully contacted 49.3 percent of the 29,982 people it attempted, and the “low contact” phone bank successfully contacted 34.2 percent of the 29,990 people it attempted to reach.

^aFor 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.

^bPercent of contacted group with at least one identical match in the control group.

Table 9: Matching and OLS Analyses, Including Vote History from 1992 to 2001

	Excluding Unlisted Numbers		Including Unlisted Numbers	
	Exact Matching	OLS	Exact Matching	OLS
Including Age				
Treatment Effect^a	2.5	2.8	4.4	4.5
(s.e.)	(0.9)	(0.3)	(0.8)	(0.3)
N^b	4,471	405,094	5,434	648,153
Matched^c	32.6%		39.7%	
R²		0.34		0.39
Excluding Age				
Treatment Effect	2.5	2.9	4.4	4.5
(s.e.)	(0.5)	(0.3)	(0.4)	(0.3)
N	12,170	405,094	12,501	648,153
Matched	88.9%		91.3%	
R²		0.34		0.38

^aFor the OLS results the treatment effect is the slope coefficient on the contact variable included in the regression:

$$\text{Vote 2002} = \alpha + \beta_1 \text{contact} + \Sigma \gamma_i \text{controls}_i$$

^bFor 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.

^cPercent of contacted group with at least one identical match in the control group.

Table A1: Exact Matching Procedure Example

Treated Subjects				Untreated Subjects			
Age	Gender	Precinct	Previous Vote	Age	Gender	Precinct	Previous Vote
30	1	10	1	55	1	16	0
45	0	15	1	45	0	15	1
19	0	12	0	19	0	12	1
32	1	16	1	56	1	14	0
55	1	16	0	28	1	12	0
42	0	15	1	18	1	12	0
70	1	10	0	19	0	12	0
24	1	12	0	21	0	14	1
21	0	14	1	21	0	14	1
34	1	14	0	25	0	10	1
62	0	10	0	62	0	10	1

End Notes

¹ Unlisted numbers refer to numbers that were unknown to Voter Contact Services, the firm that provided the registration and voting data used here.

² These OLS models included the following covariates: age, age squared, household size, gender, newly registered, previous vote in 2000, state, and competitiveness. These covariates were interacted up to four times.

³ The authors used STATA 8 to perform exact matching, corroborating the results using two other programs.

⁴ 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.

⁵ The results in Table 6 remain unchanged when noncompliers in the treatment group are excluded from the analysis.

⁶ Studies evaluating the performance of matching estimators typically have treatment and control groups on the order of three percent the size of the treatment group in our data (e.g., Dehejia and Wahba 1999; Heckman, Ichimura, and Todd 1997, 1998; Heckman, et al. 1998). Had our data 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 over the construction of the propensity score model, how to assess balance, and the type of matching method to use.