A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark

Kevin Arceneaux
Associate Professor
Temple University
Department of Political Science
Institute for Public Affairs, Faculty Affiliate
453 Gladfelter Hall
1115 West Berks St.
Philadelphia, PA 19122

Alan Gerber
Charles C. and Dorathea S. Dilley Professor
Yale University
Department of Political Science
Center for the Study of American Politics, Director
77 Prospect St, Room 292
New Haven, CT 06511

Donald Green
A. Whitney Griswold Professor
Yale University
Department of Political Science
Institution for Social and Policy Studies, Director
77 Prospect St, Room 106
New Haven, CT 06511

June 16, 2010

Forthcoming in Sociological Methods and Research

Abstract

In recent years, social scientists have increasingly turned to matching as a method for drawing causal inferences from observational data. Matching compares those who receive a treatment to those with similar background attributes who do not receive a treatment. Researchers who use matching frequently tout its ability to reduce bias, particularly when applied to data sets that contain extensive background information. Drawing on a randomized voter mobilization experiment, we compare estimates generated by matching to an experimental benchmark. The enormous sample size enables us to exactly match each treated subject to forty untreated subjects. Matching greatly exaggerates the effectiveness of pre-election phone calls encouraging voter participation. Moreover, it can produce nonsensical results: matching suggests that another pre-election phone call that encouraged people to wear their seat belts also generated huge increases in voter turnout. This illustration suggests that caution is warranted when applying matching estimators to observational data, particularly when one is uncertain about the potential for biased inference.

1. INTRODUCTION

Recent years have witnessed a surge in the use of matching, particularly among social scientists. One indication of the growing visibility of this technique can be found via a simple topic search of the ISI database. Between 1995 and 2000, the term "propensity score matching" appears only twice in social science journal articles, whereas it appears 189 times between 2001 and 2008. Moreover, researchers use matching to estimate causal effects across diverse settings and questions, illustrating its broad appeal. For example, Green and Ensminger (2006) use matching to study the long-term behavioral effects of marijuana use among African-Americans; Glick et al. (2006) use matching to assess whether countries' regulation of capital flows affects their susceptibility to currency crises; and Mithas et al. (2006) use matching to estimate whether firms that use certain customer relations strategies enjoy better marketing effectiveness.

Although the idea of comparing observations with identical (or nearly identical) background attributes is not new (Campbell and Stanley 1963; Cook and Campbell 1979), it has grown in prominence in the wake of Rosenbaum and Rubin (1985) and, especially, Dehejia and Wahba (1999), who employ the National Supported Work (NSW) field experiment to show that matching can recover from observational data causal estimates that coincide with the NSW experimental benchmark (see also Shadish, Clark, and Steiner 2009). Dehejia and Wahba's study generated considerable enthusiasm for matching. Researchers frequently suggest that matching is a superior way to estimate causal effects from observational data (Gilligan and Sergenti 2008, p.91; Morgan and Harding 2006, p.3; VanderWeele 2006, p.95; Titus 2007, p.487), with some going even

further to argue that matching "mimic[s] randomization" (Hahs-Vaughn and Onwuegbuzie 2006, p.31; see also Barabas 2004, p.692) and that matching is "sufficient to obtain the causal effect of the treatment" when the distribution of covariates is exactly matched in the treatment and comparison groups (Gilligan and Sergenti 2008, p.96).

Although matching on measured pre-treatment covariates possesses some attractive properties – it is intuitive, relies less on parametric assumptions, and calls the analyst's attention to issues such as common support – the method rests on the assumption that there are no hidden biases. Matching may mimic the logic of randomization, but it does not transform an observational study into an experimental one. This concern has led scholars to investigate the conditions under which matching is most likely to obtain unbiased causal estimates. A series of studies focusing on job training evaluations (Heckman, Ichimura, Smith, and Todd 1996, 1998; Heckman, Ichimura, and Todd 1997; Smith and Todd 2005) conclude that matching only performs well when the data satisfy three criteria: 1) treatment and comparison groups measure outcomes in identical ways, 2) treatment and control observations are sampled from similar locations or contexts, and 3) "the data contain a rich set of variables that affect both" the outcome and the probability of being treated (Smith and Todd 2005, p.309). Smith and Todd (2005) criticize Dehejia and Wahba's (1999) supportive evidence for matching on the grounds that their data did not meet any of these criteria (for a response, see Dehejia 2005).

In this paper, we show that while these criteria may be a useful conceptual guide for observational researchers, the inherent uncertainty surrounding the crucial third criterion makes it difficult to apply in practice. The challenge facing researchers who

analyze observational data is the selection problem. Matching requires that assignment to treatment and comparison groups be ignorable after accounting for relevant covariates. Because one can never know a priori all of the relevant covariates, possessing a "rich" set of covariates that seem sufficient to satisfy the ignorability assumption does not ensure that there is no omitted variable bias. We demonstrate this simple but key point using a large-scale field experiment that seems to satisfy Smith and Todd's criteria. The experiment was conducted in 2004 to probe the effect of get-out-the-vote (GOTV) phone calls on people's propensity to vote. As is often the case in field experimentation, not all subjects in the randomly assigned treatment group were actually treated. Noncompliance with experimental assignment creates a useful opportunity to evaluate the accuracy of matching methods. We have two methods for estimating the average effect of the treatment on the treated (ATT). We estimate the experimental benchmark using random assignment as an instrumental variable (see Angrist, Imbens, and Rubin 1996). Comparing treated individuals to untreated individuals via matching, we obtain an observational estimate and see whether it recovers the experimental benchmark (cf. Arceneaux, Gerber, and Green 2006; LaLonde 1986; Shadish, Clark, and Steiner 2009). Because the experiment has nearly 2.8 million subjects, we are able to match exactly on covariates and ensure that the treated and comparison groups are perfectly balanced in terms of observed attributes.

Our results suggest that even when researchers possess what appear to be optimal data, they should take seriously the threat of hidden bias. We conclude by showcasing the diagnostic value of sensitivity tests, which few researchers use in practice. Although one can never be sure if estimates derived from observational data are free from bias,

sensitivity tests help researchers assess how well their inferences hold up under increasing levels of hidden bias.

2. GOTV AND VOTER TURNOUT

In recent years, social scientists have displayed an intense interest in studying voter turnout (e.g., DellaVigna and Kaplan 2007; Gentzkow 2006; Gerber and Green 2000; Hastings, et al. 2007), and scholars have also begun to apply matching methods in this area (Hastings, et al. 2007; Imai 2005). A growing literature that studies the effects of GOTV campaigns – much of it drawing on field experiments – consistently finds that encouraging people to vote via a personal conversation increases the likelihood that they do so (Gerber and Green 2000; Green, Gerber, and Nickerson 2003). In this paper, we pay particular attention to the effects of GOTV phone calls. Initial studies found that, unlike door-to-door canvassing, phone calls from commercial phone banks do little to boost turnout (Green and Gerber 2004). Subsequent research, however, suggests that phone calls can boost turnout when phone bank callers are paid extra to read the script in an unhurried and conversational manner (Nickerson 2006, 2007; Nickerson, Friedrichs, and King 2006). We conduct an experiment that seeks to replicate and extend these empirical findings. Before discussing the experimental design, though, we explicate the assumptions underlying experimental and observational approaches to estimating causal effects.

3. MODEL

Using the Neyman (1923) and Rubin (1974) potential outcomes modeling framework, we characterize the dependent variable, y_i, as a pair of potential outcomes for subject i: y_{i1} denotes the subject's voting behavior if exposed to our GOTV phone call, and y_{i0} denotes the subject's response if not exposed to this phone call. Thus, there are four possible types of subjects: those who will abstain from voting regardless of whether they are treated or not (y_{i1}=0, y_{i0}=0), those who vote if treated and abstain otherwise (y_{i1}=1, y_{i0}=0), those who abstain if treated and vote otherwise (y_{i1}=0, y_{i0}=1), and those who vote regardless of whether they are treated (y_{i1}=1, y_{i0}=1). We further distinguish between those who are reachable by phone and those who are not. Let x_i =1 for those who answer the phone if called, and x_i =0 for those who do not answer if called. In the terminology of Angrist, Imbens, and Rubin (1996), the former group is called Compliers, and the latter group is called Never Takers. Because one cannot receive a call from our phone bank without being placed in the treatment group, there are no Always Takers and no Defiers.[1] For ease of exposition, we will refer to Compliers as "reachable" by phone and Never Takers as "unreachable."

[Table 1 about here]

Therefore, there are eight possible combinations of y_i and x_i, which is to say eight possible kinds of subjects. Table 1 describes each of the eight possible voter types. Each type comprises a share π_j of the total subject population, with $\sum_j^8 \pi_j = 1$. When we speak of the treatment-on-treated effect, we refer to the causal effect of a phone call

[1] Subjects could receive calls and political messages from other campaigns. So long as these campaigns direct their messages with equal probability to our treatment and control groups, our experimental results remain unbiased.

among those who are reachable. From Table 1, we see that the average effect of the treatment on the treated[2] is

$$\mathrm{E}\left[y_{i1} - y_{i0} \mid x_i = 1\right] = \frac{\pi_6 - \pi_7}{\pi_5 + \pi_6 + \pi_7 + \pi_8}. \quad (1)$$

Empirically, we are limited by the fact that we do not observe y_{i1} and y_{i0} for the same individuals. Instead, one outcome is observed, and the other remains counterfactual. In order to estimate the treatment-on-treated effect, a researcher may conduct a randomized experiment. Suppose that the researcher randomly assigns subjects to the treatment group (z_i =1) or the control group (z_i =0). No one from the control group is called. Among those assigned to the treatment group, some are reachable (z_i =1, x_i =1) and others are not (z_i =1, x_i =0). A randomized experiment provides estimates of several useful quantities, even when some of the subjects assigned to receive treatment prove to be unreachable. For example, when the assignment of subjects to treatment and control groups is random, the reachable fraction of the population (α) is estimated in a consistent manner by the proportion of the assigned treatment group that was in fact reached by phone.

$$\operatorname*{plim}_{N_t \to \infty} \hat{\alpha} = \pi_5 + \pi_6 + \pi_7 + \pi_8 \quad (2)$$

The researcher also observes the voting rate in the assigned treatment group (z_i =1) and in the assigned control group (z_i =0). As the number of control group observations $N_c \to \infty$, the observed voting rate in the assigned control group ($\hat{V}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} y_{i0}$) may be expressed as

[2] Because our experimental design permits only one-way crossover (i.e., the treatment group may be untreated, but the control group cannot receive the treatment), the ATT is the same in this case as the average treatment effect among compliers.

$$\operatorname*{plim}_{N_c \to \infty} \hat{V}_c = \pi_3 + \pi_4 + \pi_7 + \pi_8 \quad (3)$$

because unreachable people, by definition, vote as though they were not called. Similarly, as the number of observations increases, the voting rate in the assigned treatment group ($\hat{V}_t$) converges in probability to

$$\operatorname*{plim}_{N_t \to \infty} \hat{V}_t = \pi_3 + \pi_4 + \pi_6 + \pi_8 . \quad (4)$$

Combining equations 2, 3, and 4, the estimator

$$\operatorname*{plim}_{N \to \infty} \frac{\hat{V}_t - \hat{V}_c}{\hat{\alpha}} = \frac{\pi_6 - \pi_7}{\pi_5 + \pi_6 + \pi_7 + \pi_8} \quad (5)$$

provides a consistent estimate of the average treatment effect on the treated.

An altogether different approach is to estimate the treatment-on-treated effect by comparing the voting rate among the treated members of the assigned treatment group (z_i =1, x_i =1) to the (untreated) members of the control group (z_i =0). This approach is roughly analogous to what observational researchers do when analyzing survey data in order to estimate the effects of voter mobilization activity.[3] The danger of this approach is the possibility that reachable people have different ex ante probabilities of voting than unreachable people. In order to address this concern, observational researchers turn to matching in an attempt to make the ex ante voting propensities of the treated and matched untreated as similar as possible. The core assumption is ignorability, the idea that

[3] Actually, the approach used by observational researchers is typically worse than the one we describe here. Without random assignment, subjects in the control group would be targeted in the real world, and consequently, analysts end up simply comparing treated subjects to untreated subjects.

conditional on covariates, the treated have the same ex ante voting propensity as the untreated.[4]

The role of the ignorability assumption becomes clear when we look more closely at the proposed matching estimator. Suppose, for simplicity of exposition, that the members of the treatment and control group have exactly matching background attributes such as age and past voting behavior.[5] The proposed estimator of the treatment-on-treated effect is the difference between the observed rate of voting among those who are reached by phone ($\hat{V}_r$) and the observed voting rate in the control group ($\hat{V}_c$). Referring back to Table 1, this estimator converges in probability to

$$\operatorname*{plim}_{N\to\infty}\left(\hat{V}_r - \hat{V}_c\right) = \frac{\pi_6 + \pi_8}{\pi_5 + \pi_6 + \pi_7 + \pi_8} - \left(\pi_3 + \pi_4 + \pi_7 + \pi_8\right). \tag{6}$$

This estimator provides a consistent estimate of the treatment-on-treated effect (see equation 1) when reachable people have the same voting rate in the absence of a phone call as the control group:

$$\frac{\pi_7 + \pi_8}{\pi_5 + \pi_6 + \pi_7 + \pi_8} = \left(\pi_3 + \pi_4 + \pi_7 + \pi_8\right). \tag{7}$$

---

[4] Following Wooldridge (2002, 607), we "assume ignorability in a conditional mean independence sense." Rather than assume that the potential outcomes of *Y* are independent of treatment conditional on covariates, we need only assume that the expected value of these potential outcomes is the same for the treated and untreated, conditional on covariates. Heckman, Ichimura, Smith, and Todd (1998) make the same assumption.

[5] Equation (6) below suppresses the notation for covariate strata. For the case of exact matching described in section 4, observations are grouped into strata comprised of those observations with the same covariate values and the difference in average outcome for the treated and the untreated is calculated for the strata. The population treatment on treated estimate is then obtained by calculating the weighted average of these strata differences, with weights equal to each stratum's proportion of the treated cases. In the notation we have employed in the paper, this calculation would provide an estimate of

$$\int_S [\mathrm{E}(Y_{i1} \mid X_i = 1, s) - \mathrm{E}(Y_{i0} \mid s)]\mathrm{p}(s)\, \partial x,$$

where S is the set of strata, $s \in$ S, and p(s) is the proportion of treated in s.

In effect, matching presupposes that, holding background attributes constant, reachable people have the same underlying voting propensities as those in the control group.

4. EXPERIMENTAL DESIGN

Sample. In the fall of 2004, we obtained the voter file for the entire state of Illinois, which contained contact information, basic demographics, and voting history for 7,062,035 individuals. After removing large households with more than five registered voters, we constructed our target population by randomly selecting one individual to represent each of the households that had a listed phone number in the file, resulting in 2,779,315 experimental subjects. We then randomly assigned subjects to two different treatment groups: 16,047 subjects were slated to receive a phone call encouraging them to vote on Election Day, and a different 16,037 subjects were slated to receive a phone call encouraging them to buckle up their seatbelts while driving. The remaining 2,747,231 subjects were assigned to the control group and did not receive a phone call. These randomly assigned groups are essentially uncorrelated with subjects' background attributes. Using multinomial logit, we tested the null hypothesis that the covariates listed in Table 2 jointly fail to predict treatment assignment. As expected, the test statistic is statistically insignificant ($\chi^2[22] = 19.11$, $p = 0.639$). Of the 11 covariates we examined for balance, only one (whether age was available for the subject) was significantly correlated with experimental assignment at $p < 0.05$.

Analysis of the whole dataset generates results that are very similar to what we present below. (Data and replication programs are available on request.) For ease of exposition, we focus our attention in this essay on a subset of the data. Using the covariates listed in Table 2, we identified all combinations of background attributes shared by at least 1 treatment or placebo observation and at least 40 control observations for each treatment/placebo observation. Call each of these combinations a "matching stratum." Within each matching stratum, we randomly select without replacement 40

control observations separately for each treatment and placebo observation, so that all "matching blocks" contain exactly 41 observations. Although this procedure discards some data, it has no material effect on our results and simplifies our presentation because all of the matching blocks are of equal size. This restriction of the sample leaves 6,569 observations in the get-out-the-vote (GOTV) message group exactly matched to 262,760 control group observations, and 6,577 observations in the placebo group who are exactly matched to 263,080 control group observations.

By focusing on the exactly matched subset of the data, we sidestep the nettlesome question of how to define an acceptable inexact match, which accompanies the use of propensity score matching and related matching methods (Fröhlich 2004; Baser 2006). This question has attracted a great deal of attention in debates about the properties of matching, but there is no need to resort to propensity scores or other approximate methods in this application, because we can match directly on the covariates. Whereas matching applications are usually semi-parametric insofar as they rely on some form of regression to estimate propensity scores, ours is entirely nonparametric.

[Table 2 about here]

Because the treatment and placebo groups are exactly matched to their controls, the comparison groups share exactly the same descriptive statistics. Table 2 shows the means for covariates. Most important among them are voting behavior in each of the four years prior to the 2004 election. As one would expect, past voting behavior is a strong predictor of current voting, and controlling for past turnout also helps control for time-invariant favors that affect the decision to vote. Another well-known predictor of voter turnout is age. In addition, we include a dummy variable for registering with one of

the two major parties, for being female, and for being newly registered (i.e., registered since the last election). Dummy variables are also included to mark missing data in age, registration date, and gender.

As noted above, match strata are formed for each unique sequence of covariates. In these data, there are 2,716 matching strata or distinct configurations of covariates for the GOTV group and 2,704 for the placebo group. These match strata control for a good deal of variation in baseline voting rates. The cross-strata variability is illustrated in Figure 1, which presents a histogram of the average voting rates among control group respondents for each match stratum. These covariates also predict quite a bit of variability in propensity to be contacted by phone. Figure 2 shows the predicted values from a probit regression of contact on covariates, among the subjects assigned to the treatment and placebo groups. The dispersion of predicted values suggests that the covariates account for considerable heterogeneity in the probability of receiving the treatment.

[Figures 1 and 2 about here]

Researchers rarely know ex ante if observational data meet the ignorability assumption. Nonetheless, analysts are generally willing to proceed as if the assumption is met when their data fit accepted guidelines. As discussed above, Smith and Todd (2005, p.309) specify three criteria that observational data must meet in order to be suitable for matching. Our data arguably comport with all three. Treatment and control groups – and thus, treated and non-treated groups – are drawn from identical sources and outcomes were measured in the same way in both the treatment and comparison groups (i.e., using the Illinois voting file); they reside in the same geographical location (i.e.,

counties within Illinois)[6]; and we possess a "rich set" of covariates that predict both treatment and the outcome.

With respect to the last criterion, it is important to note that we do not interpret a "rich set" of covariates to mean the complete set of relevant covariates. Although the ignorability assumption requires that we match on all relevant covariates, there is no way of knowing a priori whether one possesses all relevant covariates. Consequently, we interpret a rich set of covariates to be a set that a judicious researcher would conclude is sufficient enough to satisfy the ignorability assumption. As we show in Table 2, the Illinois voter file includes demographic information and the voting history of each subject. The collective wisdom from nearly 40 years of empirical work on determinants of individual-level turnout indicates that age, gender, party registration, and past voting are strong predictors of turnout (cf. Verba, Brady, and Scholzman 1995; Wolfinger and Rosenstone 1980). Past voting history is an especially important covariate since it mediates the other attitudinal and demographic determinants of voting for which we do not possess data (Plutzer 2002). Once we control for vote history, demographics, and location, the only observable variable that distinguishes these individuals is that treated people were exposed to a GOTV phone call and untreated people were not.

Treatments. We hired a large and well-known political consulting firm and its commercial phone bank to call the individuals in our treatment groups. The firm was paid $0.95 for each completed call in the GOTV group and $0.55 for each completed call in the buckle-up group. Phone bank callers were instructed to deliver the script only to

[6] The criterion that treated and comparison group subjects reside in the same geographic area seems to be a study-specific consideration. It makes sense in the context of job training programs (Smith and Todd's example) and it makes sense in the context of GOTV drives, because there is a strong geographic component to voter turnout (Wolfinger and Rosenstone 1980). However, this criterion may make less sense in other contexts. We would like to thank an anonymous reviewer for making this point.

the listed person and to discontinue the call and call back if the listed person was unavailable. Calls were made during the last five days preceding the 2004 general election. The phone bank allocated calls and repeated call-backs across the five days according to the availability of its staff. GOTV and placebo calls were made each day but at different rates and by different callers, which accounts for the small but significantly different contact rates for the two scripts.

The voter mobilization script read as follows:

Hello, may I speak with [RESPONDENT NAME].

This is [CALLER NAME] calling on behalf of Vote Illinois 2004. We are not concerned if you're a Democrat or a Republican. What we want to do is make sure that every registered voter goes to the polls and votes. A lot of people think that this election is one of the most significant in years. Issues such as National Security, Taxes, and Healthcare will all be affected by this election, and we want every voter in Illinois to stand up and be counted. Can we count on you to vote Tuesday, November 2nd? [If YES: Great! We feel it's crucial that all Illinois voters go to the polls this year.] [If UNDECIDED/NO: Well, we want to remind you that this is a very important election – there's a lot at stake – and you can voice your opinion by going to the polls and voting.] Thanks for your time. Goodbye.

The buckle-up script was shorter but similar in structure:

Hello, may I speak with _____.
This is _____ calling on behalf of Buckle Up Illinois 2004.
We're heading in to the big holiday travel season and we wanted to remind you to buckle up whenever you're on the road. And if you have kids, be sure to buckle them up, too. Can we count on you to buckle up in the coming weeks?
Thank you for your time.

Subjects were coded as contacted if they answered the phone, even if they did not listen to the message in its entirety. This definition of contact is consistent with the model presented above, in which we assume that there can be no effect of the treatment unless contact occurs. As shown in Table 3, the phone bank contacted 51.85% of subjects

who were assigned to the GOTV treatment group and 56.27% of the subjects assigned to the buckle-up message group. The fact that contact rates differ between treatment and placebo groups is immaterial for our purposes, as we do not compare those who are contacted in the two groups. (For studies that make direct use of treatment, placebo, and baseline groups, see Gerber et al. 2010).

[Table 3 about here]

4. RESULTS

4.1 Intent-to-treat results

Table 3 reports the experimental outcomes for the matched data. We find a 64.41% voting rate among those assigned to be encouraged to vote (regardless of whether they are in fact contacted) and a 63.38% voting rate in the matched control group; we find a 63.21% voting rate among those assigned to be encouraged to buckle-up, and a 62.67% voting rate among the matched control group. The contrast between turnout in the voter mobilization condition and the control condition is marginally statistically significant using a one-tailed test ($z = 1.70$, $p = 0.045$), while, as expected, the contrast between the buckle-up group and the control group is not significant ($z = 0.88$, $p = 0.378$, two-tailed test).

4.2 Treatment-on-treated effects

Table 4 reports two-stage least squares estimates of the effect of phone contact on voter turnout for the GOTV and buckle-up treatments. We estimate the average treatment-on-treated effect (ATT) by regressing whether the subjected voted in the 2004 election (0=subject abstained, 1=subject voted) on an indicator for phone contact

(0=subject not contacted, 1=subject contacted), using random assignment to the treatment as an instrumental variable (0=subject assigned to control group, 1=subject assigned the treatment group). The average treatment effect of a GOTV phone call is estimated to be 1.98 percentage-points with a standard error of 1.16 percentage-points. The 95% confidence interval therefore spans from -0.30 to 4.25 percentage-points. The corresponding estimated effect of the buckle-up message is 0.95 percentage-points with a 1.07 standard error. As expected, the 95% confidence interval for the buckle-up message (-1.16, 3.05) includes zero.

[Table 4 about here]

In contrast, the observational analyses displayed in Table 4 show large, statistically significant effects for both the GOTV and the buckle-up phone scripts. For the GOTV call, the matching estimate is 7.53 percentage-points (SE = 0.91). These estimates are nowhere near the 1.98 percentage-point estimate supplied by our experiment. Figure 3a shows the kernel density of the difference between the matching estimate and experimental benchmark drawn from 1,000 bootstrap sample simulations. The matching estimate is considerably larger than the experimental benchmark in every simulation. The same point holds for the estimated effect of our placebo treatment. The matching estimate suggests that encouraging people to buckle-up for safety has powerful effects on voter turnout, raising turnout by 5.36 percentage-points (SE = 0.89). Bootstrap sample simulations tell a similar story in Figure 3b. The matching estimate is larger than the experimental benchmark in each of the 1,000 simulated samples.

[Figure 3 about here]

Readers may wonder how the matching estimates can diverge so markedly from the experimental benchmark. Taking advantage of the simplified format of our data, Table 5 traces the bias by subdividing the total sample into its contacted and uncontacted components. The contacted matching blocks give the same large, positive matching estimates of the ATT as reported in Table 4. Because the contacted and uncontacted components must add up to the ITT results for the total sample, it follows that the uncontacted component shows strong, significant *negative* treatment effects. In other words, had we used identical matching procedures to measure the causal effect of calling and not reaching someone, we would have come to the nonsensical conclusion that not speaking to someone sharply reduces their turnout rate. This absurd pattern arises because reachable people are more likely to vote than unreachable people, even after controlling for a long list of background variables.

[Table 5 about here]

5. DIAGNOSTIC ANALYSIS

Could a researcher working only with the observational component of our data set have detected sensitivity of the results to unobserved biases? There are a number of standard tests that help researchers gauge the sensitivity of their statistical inferences to hidden bias. Nevertheless, a look at the social science literature that employs matching methods suggests that sensitivity tests are underused (or underreported). We identified 298 articles published between January of 2006 and May of 2008 and indexed in the ISI social science research database that included the keyword "propensity score." We randomly sampled 40 of these articles and noted the use of sensitivity tests in matching

analyses.[7] The results are shown in Table 6. Of the 40 randomly sampled articles, only two feature formal tests for the presence of unobserved heterogeneity (5 percent of the sample). Rather than conducting formal sensitivity tests, most researchers (18 or 45 percent of the article sample) accompany their matching estimates with supplementary results, estimating different propensity score models, comparing different estimation approaches, or checking the robustness of the findings across different sample definitions. The next most popular approach is to do nothing or, at the very least, to report nothing by way of sensitivity tests (13 or 32.5 percent of the article sample). It is also instructive that of the five articles that do not fall into one of these three categories, three are review pieces designed to educate researchers about matching methods and their proper application. Only one of these review pieces mentions the need to conduct sensitivity tests (Rubin and Stuart 2006), but the authors of the piece do not go into much detail about the sorts of sensitivity tests researchers should use.

[Table 6 about here]

The lack of reported sensitivity tests is unfortunate, as they may provide valuable diagnostic information. In our empirical example, for instance, multiple lines of investigation call the accuracy of the matching estimates into question. First, we consider the vulnerability of the estimates to errors in the key assumptions underlying the application of matching in this context. Using the notation presented in section 3 and equation (7), and again suppressing notation for the measured covariates, the bias in the matching estimates of the average treatment on treated effect shown in Table 4 can be expressed as:

[7] A table of these articles is available at [replication website].

$$Bias = \frac{\pi_7 + \pi_8}{\pi_5 + \pi_6 + \pi_7 + \pi_8} - (\pi_3 + \pi_4 + \pi_7 + \pi_8). \quad (8)$$

This expression can be rewritten as:

$$(1-\alpha)[\frac{\pi_7 + \pi_8}{\pi_5 + \pi_6 + \pi_7 + \pi_8} - \frac{\pi_3 + \pi_4}{\pi_1 + \pi_2 + \pi_3 + \pi_4}]. \quad (9)$$

Substituting in expected values, the bias expression is:

$$(1-\alpha)[E(y \mid x = 1, z = 0) - E(y \mid x = 0, z = 0)], \quad (10)$$

where α is the portion of the population that is reachable by phone (and therefore treatable), y is 1 if the subject votes, 0 otherwise, x=1 if the subject is reachable, 0 otherwise, and z=1 if the subject is treated, 0 otherwise.

[Table 7 about here]

Equation 10 permits an intuitive decomposition of the bias. Matching produces bias when the turnout rate for the matched control group (e.g. E (y | x = 0, z = 0)) does not provide an accurate answer to the counter-factual question: what would the turnout rate for the treated individuals have been had they been left untreated (e.g., E (y | x = 1, z = 0)? The proportion α of the observations in the control group are reachable subjects (x = 1). This fraction of the control group is interchangeable with the actually treated subjects and so can provide unbiased measurement of the counterfactual turnout rate. Since this portion of the untreated subjects does not contribute to bias, this explains the (1-α) term in the bias expression.

In Table 7 we use Equation 10 to explore the sensitivity of the matching estimates to violations of the assumption that, conditional on matching covariates, Pr(Vote | reachable, not treated) = Pr(Vote | not reachable). Table 7 shows the estimates that would be observed if there were in fact no real treatment effect from phone calls, under biases of

different magnitudes. From Table 7 we observe that if the voting rates among those not reachable by phone was 10 percentage points lower than those subjects who would have been reachable had an attempt been made, this would produce an erroneous treatment effect estimate of 5 percentage points. The matching estimate for the GOTV treatment could be completely explained by hidden bias if the voting rate of reachable subjects were 15 percentage points higher than the voting rate of non-reachable subjects. For the placebo condition, reachable subjects need only be 10 percentage points more likely to vote than non-reachable subjects to account for the observed effect estimate.

The interpretation of these sensitivity results depends on the degree of hidden bias that one believes to exist. Observational researchers can never be sure about the level of hidden bias; they must rely on educated guesses. As a first step, researchers can formulate an opinion on the basis of their understanding of selection bias. In this study, for instance, one would reflect on unmeasured attributes that might cause reachable and non-reachable people to vote at different rates. Reachable people might be more likely to vote than non-reachable people if the type of person who picks up a phone when called is also more likely to take an interest in politics, controlling for past voting behavior, age, gender, party, and registration date. Another conjecture is that reachable people are more likely to be in town during Election Day and therefore more likely to cast a ballot. Not all of the conjectures about hidden bias go in the same direction, however. Reachable people might be less likely to be full-time workers, and their lower socioeconomic status may make them less likely to vote.

As a next step, analysts should attempt to construct empirical tests to detect hidden bias.[8] If one possesses data about past outcomes, it is possible to employ the same placebo test that Heckman and Hotz (1989) and Smith and Todd (2005) use. If hidden bias is present, there will likely be a difference between treated and non-treated groups in terms of their past behavior. To illustrate, we use matching to estimate the correlation between *previous* voting behavior and phone contact. Because phone contact cannot affect how voters behaved in the past, any correlation between phone contact and past voting behavior (controlling for the remaining covariates) suggests the presence of bias.

In our example, we treat the 2002 midterm (i.e., the last federal election held before the 2004 contest) as the dependent variable and redefine the matching strata so that they perfectly balance the treated and control groups on the demographics included in the main analysis and voting behavior in the 2000 and 2001 elections, thereby excluding two covariates included in the main analysis (i.e., voting behavior in the 2002 and 2003 elections). With fewer covariates, this approach will necessarily increase the number of exact matches available to each treatment and placebo group observation. We identify exact matches using the same procedure employed in the main analysis.

Table 8 reports the results. We find some hints of a problematic correlation between phone contact and past voting. Relative to the matched control group, those who were contacted with the GOTV message were 0.75 percentage-points more likely to vote in the 2002 election (SE = 0.70), and those contacted with the buckle-up message were 1.63 percentage-points more likely to vote in the previous federal election (SE = 0.68).

[8] We also estimate the amount of hidden bias necessary to reverse the direction of the treatment effect using the method proposed by Rosenbaum (1995), which is increasingly employed by social scientists (e.g., Harding 2003). The results, which are available upon request, show that always voters need only be 1.75 times as likely to be treated in the GOTV group to reverse the sign on the treatment effect, and always voters need only be 1.5 times as likely to be treated in the buckle-up group to reverse the sign of its effect.

Here we confront a basic conundrum that emerges whenever sensitivity tests are used: are these correlations large enough to call our results into question? Because observational researchers cannot directly estimate hidden bias, the most conservative approach is to interpret any correlation in a placebo test as an indication that problematic amounts of hidden bias are present. At the very least, researchers should present the results from both diagnostic tests and let readers decide for themselves.

Experimental design enables us to gauge the amount of hidden bias with unusual accuracy. Matching estimates are found to depart wildly from the experimental benchmark, and a comparison of voting rates of those reached and not reached in the placebo group gives us some indication about the size of the bias in our particular sample. Treated placebo group subjects are 20 percentage points more likely to vote than untreated placebo group subjects (see Table 5), which is well above the 0.10 mark identified in Table 7.

[Table 8 about here]

6. CONCLUSION

Because we have access to an experimental benchmark in this application, we see in hindsight the potential for bias in the observational approaches. In retrospect, we infer that the people who answer phone calls from commercial phone banks tend to have elevated voting propensities, even after controlling for their past voting habits and their demographic attributes. Imagine the plight of the observational researcher who lacks the luxury of an experimental benchmark. The extensive list of covariates and the availability of exact matches may encourage this researcher to grossly misestimate the effectiveness and cost-efficiency of GOTV phone calls.

The fact that observational researchers so seldom have recourse to experimental benchmarks means that tremendous weight is placed on the substantive assumptions that they bring to bear when asserting the adequacy of their estimation strategy. In comparison to linear regression, exact matching may weaken these assumptions slightly by allowing the covariates to influence the outcome in nonlinear and non-additive ways (Winship and Morgan 1999, p.673-674; Ho et al. 2007). Still, the nagging problem of unobserved heterogeneity remains, and matching only addresses the heterogeneity in observed variables. The application presented here shows that one can have an extensive array of observables at one's disposal and still miss the experimental benchmark by a wide margin.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. 1996. "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association,* 91, 444-455.

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods using a Large-Scale Voter Mobilization Experiment." *Political Analysis*, 14 (Winter): 37-62.

Barabas, J. 2004. "How Deliberation Affects Policy Opinions." *American Political Science Review,* 98, 687-702.

Baser, O. 2006. "Too Much Ado about Propensity Score Models? Comparing Methods of Propensity Score Matching," *Value in Health,* 9 (6), 377-385.

Campbell, D. T., and Stanley, J.C. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Houghton Mifflin Co.

Cook, T. D. and Campbell, D. T. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin Co.

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

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics,* 125: 355-64.

Della Vigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics, 122* (3): 1187-1234.

Freedman, David A. 2006. "Statistical models for causation: What inferential leverage do they provide?" *Evaluation Review* 30: 691–713.

Frölich, Markus. 2004. "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators." *The Review of Economics and Statistics* 86(1) 77-90.

Gentzkow, Matthew. 2006. "Televison and Voter Turnout." *Quarterly Journal of Economics,* 121 (3): 931-72.

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 (September): 653-64.

Gerber, Alan S., Donald P. Green, Edward H. Kaplan, and Holger L. Kern. 2010. "Baseline, Placebo, and Treatment: Efficient Estimation for Three-Group Experiments." *Political Analysis* (forthcoming)

Gilligan, Michael J., and Ernest J. Sergenti. 2008. "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference." *Quarterly Journal of Political Science*, 3 (2): 89-122.

Glick, R., Guo, X., and Hutchison M. 2006. "Currency Crises, Capital-Account Liberalization, and Selection Bias," *Review of Economics and Statistics,* 88, 698-714.

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

Green, Donald P., Alan S. Gerber, and David W. Nickerson. 2003. "Getting Out the Vote in Local Elections: Results from Six Door-to-Door Canvassing Experiments." *Journal of Politics*, 65 (November): 1083-96.

Green, K. M., and Ensminger, M. E. 2006. "Adult Social Behavioral Effects of Heavy Adolescent Marijuana Use among African Americans," *Developmental Psychology,* 42, 1168-1178.

Hahs-Vaughn, D. L., and Onwuegbuzie, A. J. 2006. "Estimating and Using Propensity Score Analysis with Complex Samples," *Journal of Experimental Education,* 75, 31-65.

Harding, David J. 2003. "Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping Out and Teenage Pregnancy." *American Journal of Sociology*, 109: 676-719.

Hastings, Justine S., Thomas J. Kane, Douglas O. Staiger, and Jeffrey M. Weinstein. 2007. "The Effect of Randomized School Admissions on Voter participation." *Journal of Public Economics,* 91 (5-6): 915-37.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association,* 84 (408): 862-74.

Heckman, J., Ichimura, H., and Todd, P. 1997. "Matching as and Econometric Evaluator Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies,* 64, 605-654.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1996. "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method." *Proceedings of the National Academy of Sciences of the United States of America* 93 (23): 13416-13420.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, 66 (5): 1017-98.

Ho, D., Imai, K., King, G., Stuart, E. A. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis* 15(3): 199-236.

Imai, K. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments," *American Political Science Review,* 99, 283-300.

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

Mithas, S., Almirall, D., and Krishnan, M.S. 2006. "Do CRM Systems Cause One-to-One Marketing Effectiveness?" *Statistical Science*, 21, 223-233.

Morgan, S. L., and Harding, D. J. 2006. "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice," *Sociological Methods and Research*, 35, 3-60.

Neyman, Jerzy. 1923 [1990]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4): 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed.

Nickerson, David W. 2006. "Volunteer Phone Calls Can Increase Turnout: Evidence from Eight Field Experiments." *American Politics Research,* 34 (May): 271-92.

Nickerson, David W. 2007. "Quality is Job One: Volunteer and Professional Phone Calls." *American Journal of Political Science* 51(2):269-282.

Nickerson, David W., Ryan D. Friedrichs, and David C. King. 2006. "Partisan Mobilization Campaigns in the Field: Results from a Statewide Turnout Experiment in Michigan." *Political Research Quarterly* 59 (1): 85-97.

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

Rosenbaum, P.R. 1995. *Observational Studies*. New York: Springer.

Rosenbaum, P.R., and Rubin D. B. 1985. "The Bias Due to Incomplete Matching," *Biometrics*, 41, 103-116.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 66 (5): 688-701.

Rubin, Donald B., and Elizabeth S. Stuart. 2006. "Affinely Invariant Matching Methods with Discriminant Mixtures of Proportional Eillipsoidally Symmetric Distributions." *Annals of Statistics*, 34 (4): 1814-26.

Shadish, William R., M. H. Clark, Peter M. Steiner. 2009. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments." *Journal of the American Statistical Association,* 103 (484): 1334-44.

Smith, J., and Todd, P. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics*, 125, 305-353.

Titus, M. A. 2007. "Detecting Selection Bias, Using Propensity Score Matching, and Estimating Treatment Effects: An Application to the Private Returns to a Master's Degree," *Research in Higher Education,* 48, 487-521.

VanderWelle, T. 2006. "The Use of Propensity Score Methods in Psychiatric Research," *International Journal of Methods in Psychiatric Research*, 15 (2), 95-103.

Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality : Civic Voluntarism in American Politics*. Cambridge, Mass. : Harvard University Press.

Winship, C., and Morgan, S. L. 1999. "The Estimation of Causal Effects from Observational Data," *Annual Review of Sociology,* 25, 659-706.

Wolfinger, Raymond, and Steven J. Rosenstone. 1980. *Who Votes?* New Haven, CT: Yale.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Table 1: Classification of Target Population

Group Number	Latent Type	Reachable by Phone? (x_i)	Votes if Reached by Phone (y_{i1})	Votes if Not Reached by Phone (y_{i0})	Share of the Population
1	Non-Reachable or "Never Taker"	No	No	No	π_1
2		No	Yes	No	π_2
3		No	No	Yes	π_3
4		No	Yes	Yes	π_4
5	Reachable or "Complier"	Yes	No	No	π_5
6		Yes	Yes	No	π_6
7		Yes	No	Yes	π_7
8		Yes	Yes	Yes	π_8

In the text, we denote No=0 and Yes=1. In this case, there are by construction no Always-takers or Defiers, because one cannot receive treatment phone call without being assigned to the treatment group.

Table 2. Summary Statistics for Covariates by Treatment Groups (Only for Observations that have Exact Matches in the Control Group)

	Message	
Covariate	**GOTV**	**Buckle Up**
Percent Vote 2000	61.24	60.83
Percent Vote 2001	16.69	16.27
Percent Vote 2002	53.89	52.82
Percent Vote 2003	13.49	13.55
Percent Missing Age	8.60	8.99
Percent Newly Registered	2.07	1.92
Percent Missing Registration Date	0.55	0.59
Percent Female	55.91	55.62
Percent Missing Gender	0.12	0.12
Percent Major Party Registration	48.07	46.21
Mean Age	41.19	40.72
(Standard Deviation of Age)	(19.22)	(19.37)
Matched Treatment Group N	6,569	6,577
[Matched Control Group N]	[262,760]	[263,080]

Note: Because we restrict the experimental sample to observations that match exactly on these covariates, the control group means are identical to those in the treatment groups. To avoid redundancy, we present only the treatment group means.

Table 3. Voting by Randomized Treatment Assignment

	GOTV Message		Buckle-up Message	
	Treatment Group	**Matched Control**	**Treatment Group**	**Matched Control**
Percent Voting in 2004	64.41	63.38	63.21	62.67
Number of Observations	6,569	262,760	6,577	263,080
Percent Contacted*	51.85	0.00	56.27	0.00

*A subject is coded as *contacted* if someone at the targeted phone number answered the phone bank's call.