Fixing Broken Experiments: How to Bolster the Case for Ignorability with Full Matching.*

Jake Bowers and Ben Hansen
Political Science and Statistics
jwbowers@umich.edu and bbh@umich.edu
University of Michigan

August 22, 2007

Abstract

When an treatment has been randomly assigned to a set of subjects, inference about the effect of assignment is straightforward and well-grounded. When some of the subjects refuse to comply with a randomly assigned treatment one may be tempted to think that the experiment has been broken — and that thus, techniques developed for analyzing completely randomized data may no longer be used — and that, inferences about the effect of treatment on the treated are no longer easily possible. In fact, it is well-known that a well-done random assignment allows an analyst to use indicators of assignment as an instrument which does provide consistent and confident estimate of the effect of treatment on those who complied with treatment — even if those who comply with treatment do so systematically.

What should an analyst do, however, if the assignment itself is detectably not random? Does this then mean that use of assignment indicators as an instrument in order to estimate effects of treatment on the treated is no longer appropriate? In this paper we argue that it is not necessary to ignore knowledge about assignment if random assignment was attempted but somehow failed to produce balance in a given experiment. Instead, we show how analysts can gently re-balance their data using subclassification and matching in order to do an instrumental variables analysis based on assignment indicators. This technique is especially useful in cases where treating the attempted experiment like an observational study and attempting to deal with potential confounds through direct adjustment founders on a lack of data. The case that we use to demonstrate our points is the New Haven Vote 98 field experiment (Gerber and Green, 2000).

1 The Importance of Ignorability: Ignorability puts the Gold into Gold Standard

A Google search for the words "experiment" and "gold standard" returns over 480,000 results.[1] Why is random assignment in controlled experiments seen as such a powerful way to assess whether a given treatment produced some effect? The answer lies in part in the fact that, given random assignment of treatments, in a

*Draft 11.1. We are grateful to participants in workshops at the Center for Political Studies at the University of Michigan, at the annual meetings of the Midwest Political Science Association, April 2005, the Royal Statistical Society, September 2004 and at the Department of Political Science at the University of Illinois, July 2004 for helpful comments on much earlier versions of this work.

[1] Google search as of Feb 4, 2006.

certain sense, one may ignore the effects of any observed or unobserved potential confounding covariate when calculating treatment effects.[2]

As an example, consider a study of Get Out The Vote (GOTV) efforts where some citizens are visited in-person and urged to vote, and others are not, and the goal is to find out how much these visits enhanced (or impeded) turnout. If before receiving a knock on the door people were all equally likely to vote, then we could compare the number of voters who received the visit with the number of voters who did not receive such a visit — and we could interpret the difference between these two numbers as being due only to a visit from a campaign activist. The idea here is that we imagine that each person has a fixed potential to vote if she receives a get-out-the-vote (GOTV) visit, and a fixed potential to vote if she does not receive such a visit. Perhaps these potential responses to a GOTV treatment are different among different types of people — say, people with high education might be more likely to vote in reaction to a GOTV visit than people with low education. However, if our treatment group has just as many people with high education as our control group, and education is the only attribute of people that changes their potential to respond to a GOTV appeal, then comparing a summary of vote turnout (like an average or a sum) between this treatment group and this control group will tell us something about differences caused by the treatment, not differences caused by education since the distribution of education is held constant between the two groups and equal. That is, in this simplified case we can *ignore* education as a potential confound by virtue of our research design.

It turns out that the reason why "experiment" is so associated with "gold standard" is because social-science experiments tend to use random assignment to ensure that members of treated groups are, on average, indistinguishable from members of control groups in terms of potential confounders.[3] That is, experiments are great because of random assignment. But random assignment is great, the common wisdom goes, because it ensures "ignorability". If the assumption of "strong ignorability" holds then people who receive the treatment are *a priori* no more likely to respond positively to it than people who do not receive the treatment. A formal way to say this, which introduces notation that will be useful for us in this paper, is:

$$\Pr(Z = 1|\mathbf{X}, r_t, r_c) \equiv \Pr(Z = 1), \tag{1}$$

where, Z is 1 if the person was assigned to treatment and 0 if she was assigned to control, r_t and r_c records whether the person would vote if that person were assigned treatment (a visit) or control (no visit) respectively, and $\mathbf{X}$ is a matrix of potentially confounding pre-treatment covariates — covariates that might make people differ in their odds of voting even before receiving a get-out-the-vote (GOTV) visit. Equation 1 says that the probability of a visit does not depend on covariates or potential responses. If this is true, we can look at the effect of treatment on responses without worrying about confounding. Thus, if asked, "Why do people call experiments the gold standard for identifying causal effects?" the most brief (and perhaps cryptic) answer is "Because of ignorability."

[2] Another part of the answer is that manipulation tends to be a more compelling basis for establishing causation than association. For a thoughtful review and exposition of different theories of causality, see Brady and Seawright (2004).

[3] In the physical sciences, isolation and strong theory can play the same role in justifying ignorability claims as random assignment does in social science Holland (1986*a*).

In order to make a case that we can observe a causal effect we need to argue that we have ignorability. People who do random assignment make this case by checking to see if the distribution of pre-treatment covariates that might matter for the outcome is the same in the treated group and the control group. If these distributions are the same (which they ought to be, in large experiments or across repeated replications of the same small experiment), then the experimentalist may feel that the assumption in (1) is a reasonable basis for simple comparisons between treated and control groups as causal inference — because random assignment provides similarity, or balance, in distributions of *both* unobserved and observed confounders, and the balance assessment of observed confounders provided no strong evidence of imbalance. Students of observational data must also make the same kind of balance assessment, but small departures from balance are more cause for worry — and yet more worry ought to exist in their hearts because they have no *a priori* protection against confounds due to unobserved factors. As a result of these two kinds of worries, analysts of observational studies tend to need to some how make adjustments to the organization of their data such that at least there are no clear differences in distributions of observed pre-treatment covariates. When these adjustments succeed, they add confidence to claims of *ignorability given covariates* $\mathbf{X}$:

$$\Pr(Z = 1|\mathbf{X}, r_t, r_c) \equiv \Pr(Z = 1|\mathbf{X}) \tag{2}$$

Thus, in an observational study we cannot ignore our covariates *a priori*, but we can try to condition on them. If, say, age of an individual were a potential confound in an observational study, we might assess treatment effects only among people who were the same age. By holding age constant in this way, we can ignore its effects.

Because of ignorability, when an treatment has been randomly assigned to a set of subjects, inference about the effect of assignment is straightforward and well-grounded. But, when some of the subjects refuse to comply with a randomly assigned treatment one may be tempted to think that the experiment has been broken — and that, thus, techniques developed for analyzing completely randomized data may no longer be used — and that, inferences about the effect of treatment on the treated are no longer easily possible. In fact, it is well-known that a well-done random assingment allows an analyst to use indicators of assignment as an instrument which does provide consistent and confident estimate of the effect of treatment on those who complied with treatment — even if those who comply with treatment do so systematically (Angrist, Imbens and Rubin, 1996*b*).

What should an analyst do, however, if the assignment itself is detectably not random? Does this then mean that use of assignment indicators as an instrument in order to estimate effects of treatment on the treated is no longer appropriate? In this paper we argue that it is not necessary to ignore knowlegde about assignment if random assignment was attempted but somehow failed to allow a strong case to be made for ignorability in a given experiment. Instead, we show how analysts can gently rebalance their data using subclassification and matching in order to do an instrumental variables analysis based on assignment indicators. This technique is especially useful in cases where treating the attempted experiment like an observational study and attempting to deal with potential confounds through direct adjustment founders on a lack of data. The case that we use to demonstrate our points is the New Haven Vote 98 field experiment (Gerber and Green, 2000).

1.1 The Plan

First, we present a more formal understanding of a causal effect — following the formulation of Rubin (1974) and Holland (1986*b*) that Brady and Seawright (2004) have called the Neyman-Rubin model of causality. We start with such an abstract discussion because we want to make sure that our modeling decisions hew as closely as possible to our understanding about what it means when we state one of our conclusions: "Personal contacts from campaigns caused between 66 and 239 more people to vote than would have occurred if no neighborhood canvassing had occurred in this campaign. Our best estimate of the number of people who voted because of in-person canvassing is 149."

Second, we propose a probability model for treatment assignment that enables us to test a null hypothesis of no treatment effects. Our framework for hypothesis testing and confidence intervals is often called "randomization inference" and it has a long history of development and use in statistics and biological applications starting at least from Fisher (1935). This means that instead of regressions or ANOVAs we will use Fisher's exact test for 2×2 tables and the Mantel-Haenzel test for $S \times 2 \times 2$ stratified tables. The use of tables and stratification is a key element in our approach here.[4] The use of stratification and tables as our analytic metaphor rather than smooth lines on a scatter plot will help us feel confident that we are not adding extraneous assumptions to our efforts to adjust an attempted random assignment that didn't produce the kind of balance one would have hoped for. Another major benefit of this approach is that appeals to asymptotic properties of sampling distributions of test statistics are always about one particular distribution, not an infinite class of distributions. This means that any asymptotic approximations made are open to validation and can also be replaced by exact calculations when necessary or desired. Imbens and Rosenbaum (2005) have recently shown that this property of randomization inference results in superior performance for instrumental variables estimation. We don't repeat their arguments here, but we will illustrate some of the advantages of randomization inference throughout this paper.[5] It is important to note, however, that the techniques that we use to bolster the case for ignorability do not require the use of any particular mode of probabilistic inference. We chose randomization inference in this particular case because it has particularly atttractive properties in comparison to Bayesian or repeated sampling based inference in the particular case of an experiment in which random assignment was attempted.

Third, we estimate the number of voters attributable to two different GOTV treatments: in-person visits and telephone calls.

2 The Neyman-Rubin Approach to Thinking about Causality

A causal effect changes a potential outcome. Our micro-foundations for the effect of various GOTV treatments on turnout are very simple. Following Holland (1986*b*) and Rosenbaum (2002*a*) we define a treatment effect τ_i as the difference between the outcome that we would observe for a person i who receives a treatment r_{ti}

[4] One can think of matching as a very finely grained form of stratification or a table with many many strata.

[5] We have also described these techniques in more depth in Bowers and Hansen 2005a, 2005b and Bowers and Hansen 2006.

and the outcome that we would observe for that same person if he did not receive a treatment, r_{ci}. That is,

$$\tau_i = r_{ti} - r_{ci} \tag{3}$$

The treatment effect for person i is the difference in potential responses for that person. The "fundamental problem of causal inference" (Holland, 1986*b*) is that we can never observe τ_i since we only observe the response of person i either under treatment or control. In general, when scholars talk about an estimate of a treatment effect, they are in fact referring to an estimate of an average treatment effect,

$$\frac{(\sum_1^n \tau_i)}{n} = \frac{(\sum_1^n r_{ti})}{n} - \frac{(\sum_1^n r_{ci})}{n}, \tag{4}$$

which is defined as the difference between the average response across people exposed to treatment $\left(\frac{(\sum_1^n r_{ti})}{n}\right)$ and the average response across people exposed to control $\left(\frac{(\sum_1^n r_{ci})}{n}\right)$. Because our responses are binary, the treatment effects we estimate differ somewhat from this, as we will explain in § 3.5.

This individual-level theoretical model for the effect of a treatment is attractive because it defines a causal effect in terms of the result of some manipulation — and, this effect is understood to act at the level of the unit, not on aggregates. If we define the individual-level model in this way, the assumptions required for inference become more clear, and we can gauge these assumptions directly with the data in hand.

An example of an assumption the Neyman-Rubin model makes easier to understand is the assumption of no interference between units (Cox, 1958, §2.4), also known as the stable unit treatment value assumption (SUTVA) (Rubin, 1986). SUTVA states that the response of each unit depends on the treatment assigned to it, but not on treatments assigned to other units. To appreciate the meaning of SUTVA, suppose a treatment given to one person, i, were to affect another person j's outcome; this represents a failure of SUTVA. Then the effect of treating i would not be captured by a difference of form $r_{ti} - r_{ci}$, as such a difference misses the treatment's effect on j; consequently, expression such as (4) would either fail to make sense or fail adequately to express what we mean by "treatment effects". Without the Neyman-Rubin model, it can be difficult to recognize when a SUTVA assumption has been made.

This paper assumes SUTVA throughout: while effects of a GOTV campaign might in principle be felt by those not campaigned upon, we believe these effects to be quite small in the relatively brief, dispersed campaigns like the one we study. Another assumption, ignorability of treatment assignment, is extremely important in guiding our decisions in this paper, and so we give it special treatment. In fact, the point of this paper is that research design is way to make arguments that are stronger (or weaker) in favor of the ignorability assumption. Random assignment allows quite strong arguments. And when something has gone wrong with random assignment, the point of data analysis is to strengthen the case for ignorability. We believe that in cases like the Vote 98 field experiment, we can make the best argument for ignorability if we use information about random assignment — even if it was somewhat flawed — than if we threw out that information and treated the study merely as an observational study.

3 The New Haven 1998 Vote Turnout Experiment

Just before the 1998 Congressional election, Gerber and Green fielded a GOTV experiment in New Haven (Gerber and Green, 2000). In this study they tested a variety of potentially turnout-enhancing interventions: door-to-door canvassing, phone calls, and mailings (from 1 to 3 mailings were sent). In addition, they tested the effectiveness of different messages. Table 1 shows their design omitting the different appeals made for simplicity. For example, we can see that 288 people were assigned to receive a phone call and a visit from a canvasser but no mailings, while 2315 people were assigned no in-person visits and no phone calls, but 3 mailings.

	0	1	2	3
In Person,Phone	288	399	363	394
In Person,No Phone	2615	505	614	616
No In Person,Phone	1176	1506	1550	1582
No In Person,No Phone	10582	2351	2524	2315

Table 1: Treatment Assignments in Gerber and Green (2000): In-Person by Phone by Number of Mailings

In this paper we will be assessing hypotheses about two simple treatments — the in-person canvassing and the phone calling. Over all 5794 New Haven residents were assigned to be visited in-person by a canvasser, and 7258 people were assigned to be called on the phone.[6]

In their article (and later revisions) Gerber and Green estimated that direct face-to-face contact increased turnout in New Haven by roughly 9 percentage points (with 95% confidence interval (CI) $\pm 2 \times 2.6 = 5.2$), and phone calls decreased turnout by around 5 percentage points (with 95% CI $\pm 2 \times 2.7 = 5.4$). They estimated these effects using random assignment as an instrumental variable — which allowed them to produce consistent estimates of the effect of treatment on those who did comply, even though compliance with their treatment assignment was demonstrably not-random.

In a later article, Imai (2005) pointed out that the random assignment in the New Haven study might not have produced treatment and control groups as comparable as might have been expected. Imai suggested that, since Angrist, Imbens and Rubin (1996*a*) require random assignment as one of their assumptions before instrumental variables can work, the New Haven study ought to be analyzed as a pure observational study. In a response to Imai, Gerber and Green have argued that their random assignment is strong enough to support their use of instrumental variables Gerber and Green (2005).

In what follows, we will assess the success of the New Haven experiment in producing strongly ignorable treatment assignment. And then we will test the null hypothesis of no treatment effects and estimate the number of voters who did so because of the treatment. It turns out that the case for ignorability for the in-person treatment can be made fairly simply, but the case for phone treatment requires more work. This distinction between the treatments provides us a good opportunity to show how the argument for ignorability

[6] Our numbers for phone treatment differ from those published in Gerber and Green (2000) because they were corrected due to an error in the administration of the experiment discovered by the authors after publication of their article.

can in one case be easily bolstered (with a simple table), and in another case requires a much more sophisticated, non-parametric adjustment (using optimal full matching).

3.1 Assessing and Enhancing the Case for Ignorability of Treatment Assignment

Our first task when testing for the presence of causal effects is to assess the case for ignorability of treatment assignment. A simple test would allow us to decide whether random assignment alone ought to allow us to ignore potential confounds in our analysis. Note carefully that we test ignorability of treatment *assignment*, not ignorability of the distinction between *both* having been assigned to treatment *and* having received it, on the one hand, and *either* not having been assigned to treatment *or* having been assigned to treatment but then not having been willing to receive it or (not present to receive it). Experimentalists are in general powerless to enforce ignorability of the latter distinction, as it depends partly on experimenters but also partly on the movements and decisions of experimental subjects. Assignment to treatment conditions, on the other hand, may fall within within experimentalists' control. When it does, properly exercising such control usually makes treatment assignment ignorable.

Our first assessment of ignorability of in-person treatment involved an analysis of deviance test comparing the fit of a model of assignment to in-person treatment using only indicators of complementary treatment assignment and the fit of that model including also the covariates measured in this data set.[7] If random assignment to treatment will allow us to ignore potentially confounding covariates, then adding such covariates to the model should not improve the fit. In fact, our test casts doubt on the hypothesis of ignorability ($p = .012$) .[8]

We then used the Mantel-Haenszel test to gauge the differences between treated and controls on each covariate in the dataset before and after holding constant other potential confounds.[9] We made this test on 38 different covariates, and so, even if one desires to accept the null falsely 5% of the time, we'd expect to see our test reject the null for one or two covariates even when the random assignment was done perfectly, merely due to chance.[10] In this case we found 5 covariates with p-values less than .05. Table 2 shows the standardized biases for these covariates and p-values associated with a Fisher test. The standardized bias for a given variable x is:

$$\frac{\bar{x}_t - \bar{x}_c}{\text{pooled standard deviation}} \tag{5}$$

This measure of lack of ignorability allows us to compare the differences between members of the treated and

[7] The covariates in this model were age, age^2, voting behavior in 1996 (not registered, registered but didn't vote, registered and voted), ward of residence (29 wards), major party partisanship (yes or no), and the number of people living in the household (1 or 2).

[8] Although we did not describe randomization inference at length in this paper, notice that using a logistic regression here is consistent with the randomization inference in that we restrict our probability model to $Z|X$ since the binomial likelihood function here is only specified in terms of $Z|X$ and not $Y|X, Z$.

[9] We used the 5 covariates available in the dataset broken into indicators for each category: voting behavior in 1996 (not registered, registered but didn't vote, registered and voted), ward of residence (29 wards), major party partisanship (yes or no), and the number of people living in the household (1 or 2). We included age, the sole continuous covariate, not as one covariate but five, in the form of a natural cubic spline with five equally spaced inflection points. This created a model matrix with 38 different columns.

[10] We show all of the covariates in Table 5.

control groups across different variables, measured on different scales. In this case, the standardization is not crucial for interpretation, however, since the five covariates here are all different wards of New Haven. The most severely unbalanced ward is #3, where there were many more households assigned to control (598) than to treatment (91).[11]

	Standardized Bias	$p\,(< \lvert t(Z,r)\rvert)$
wardF3	−0.10	0.00
wardF8	0.04	0.07
wardF12	0.05	0.04
wardF15	0.04	0.05
wardF17	0.07	0.00
wardF19	−0.05	0.04

Table 2: Covariates Showing the Largest Differences between Treated and Control Subjects in the In-Person Condition

In medical trials it is common to assign subjects in clusters but to omit mention of this clustering in the published reports. We suspected that assignment by cluster might be the cause of the imbalance shown above. It is possible the treatments assigned to one person in a two person household might also affect the member who is not assigned treatment (say, if the non-assigned household member watched the in-person canvassing effort occur — or if the treated household member talked about the treatment with the control member), we decided to assess the ignorability assumption produced by random assignment separately by whether the household had 1 or 2 members. Our desire is to reduce the number of covariates showing imbalance here to no more than 1.[12]

We started again with an analysis of deviance test identical to that reported on page 7, only this time we did two tests, one for people who live alone and another for people who live in two person households. Our intuition about the source of weakness in the ignorability assumption was borne out here. Adding covariates did not appreciably improve the fit of the model of treatment assignment for people living in one person households (p =.44), but it did improve the fit substantially for people living in two person households (p =.0003).

We also used the Mantel-Haenzel test to discover which covariates were contributing to differences between the treated and control groups. For people living alone, we can see than none of the covariates differ across treated and control respondents more than .062 of a standard deviation. Table 3 shows that even the largest differences are not strongly distinguishable from zero by the Mantel-Haenzel test at α = .05.

Table 4 shows the results for subjects living in 2 person households. In this case, while most of the bias is small and comparable to the 1 person households, there are five wards where there were more treated (or

[11] Although we do not describe them here, the Fisher test and the Mantel-Haenszel test both allow an exact distribution for our test statistic, and are stated in terms of $Z|X$ — that is,they are randomization based tests, not sampling or likelihood based tests.

[12] In Bowers and Hansen (2005*a*), we directly model this clustering using randomization inference. In this paper, however, we pretend that we do not know that clustered treatment assignment is the culprit of the balance problems or at least that we don't know which respondents belong to which clusters. This makes our current analysis more useful to researchers who don't have the luxury of the full dataset as we have here. The estimates that we produce here are, in fact, the same as the ones that come from the clustered analysis — showing that our techniques for bolstering the instrument do work (at least in the case where the case for ignorability was prima facie relatively strong).

	Standardized Bias	$p(< \|t(Z,r)\|)$
wardF2	0.06	0.07
wardF3	−0.06	0.08
wardF10	−0.06	0.06
wardF13	−0.06	0.07
ns(AGE, 5)4	0.05	0.10

Table 3: One Person Households: Covariates Showing the Largest Differences between Treated and Control Subjects in the In-Person Condition

control) respondents than would be expected were the random assignment to have worked perfectly. The large differences for Ward 3 and 17 reflect the fact that the ratio of assigned to treatment to assigned to control for those two wards was 24/230 (many fewer treated than controls than would be expected under the null) and 204/627 (many more treated than controls than would be expected under the null) respectively.

	Standardized Bias	$p(< \|t(Z,r)\|)$
wardF2	−0.08	0.02
wardF3	−0.15	0.00
wardF16	0.07	0.04
wardF17	0.12	0.00
wardF19	−0.06	0.06

Table 4: Two Person Households: Covariates Showing the Largest Differences between Treated and Control Subjects in the In-Person Condition

In an effort to make the 2 person households in the treated group more like those in the control group, we grouped them into two strata using a median split on a propensity score. (Imai and van Dyk (2004) suggest an approach similar to this using different data.) Our propensity score predicts assignment to in-person canvassing using all of the relevant covariates listed in footnote 9 on page 7; we selected the form of the prediction equation using a forward-backward stepwise regression penalized by the AIC (Akaike, 1973). The final model involved all these covariates plus an indicator for assignment to phone and/or mailing treatment and several interactions.[13]

Table 5 shows that upon grouping the 2 person household respondents into two strata, one for propensity scores below the median and another for propensities above it, the treatment-control group imbalances seen in Tables 2 and 4 all but disappear. That is, by calculating biases separately in these two strata and in the single voter household stratum, then taking an appropriately weighted average of the three, the bias along each covariate is made substantially smaller than what it was prior to stratification. To assess the success of the adjustment, we compared these new standardized biases to what theory would have us expect had treatments been assigned at random within the three strata. For none except one of the covariates were remaining standardized differences significant at the .05 level.[14]

[13] Specifically, interactions between age terms and the major party membership-indicator, and between past voting behavior and major party membership.

[14] The outlier was residence in Ward # 3, with standardized bias .08 after stratification. As the bias on it was not large in absolute

	Pre Stratification Bias	Sig	Post Stratification Bias	Sig
votein96novote	−0.01		−0.01	
votein96notreg	0.01		0.01	
votein96voted	0.00		0.00	
wardF2	0.00		0.02	
wardF3	−0.10	***	−0.08	**
wardF4	−0.03		−0.01	
wardF5	−0.01		0.00	
wardF6	0.00		−0.01	
wardF7	−0.00		0.00	
wardF8	0.04	.	0.03	
wardF9	0.01		0.01	
wardF10	−0.01		−0.03	
wardF11	0.01		0.02	
wardF12	0.05	*	0.03	
wardF13	−0.02		−0.03	
wardF14	−0.01		0.00	
wardF15	0.05	*	0.03	
wardF16	0.03		0.01	
wardF17	0.07	**	0.03	
wardF18	−0.02		−0.01	
wardF19	−0.05	*	−0.02	
wardF20	0.01		0.01	
wardF21	−0.03		−0.01	
wardF22	−0.02		−0.01	
wardF23	−0.00		0.01	
wardF24	0.00		−0.01	
wardF25	0.02		0.02	
wardF26	−0.03		−0.00	
wardF27	−0.04		−0.02	
wardF28	0.00		−0.01	
wardF29	0.01		0.02	
wardF30	0.01		−0.01	
MAJORPTY	−0.04		−0.03	
ns(AGE, 5)1	−0.01		−0.02	
ns(AGE, 5)2	−0.00		−0.01	
ns(AGE, 5)3	−0.02		−0.00	
ns(AGE, 5)4	0.01		0.02	
ns(AGE, 5)5	−0.01		0.00	

Table 5: Balance for Three Strata in the In-Person Condition

The three-stratum subclassification with which we emerge with is shown in Table 6. Subsequent analysis for the effect of in-person treatments on voting assumes only that within each of the three strata taken separately, treatment assignment is ignorable — and now this assumption can be supported more strongly because of the way that we have now reorganized our data. That is, we showed that the case for ignorability was not as strong as it could have been, but then, by this relatively simple re-grouping of the data, we now have made a much

terms, and since of 37 true null hypotheses 1.8 are expected to be rejected in tests at the .05 level, we chose to accept this much bias. As it happens, with these data and with this propensity score, but splitting the two-subject households into four groups rather than two, we could have avoided having even one covariate bias that is significant at the .05 level. In the interests of simplicity, we opted for the two-stratum solution.

stronger case for ignorability. It turns out that, because we have chosen to use randomization inference for estimation, ignorability and knowledge of the research design are the only two ingredients necessary for us to obtain estimates of causal effects in which we have confidence.

	2 Persons:Lo	2 Persons:Hi	1 Person
Treated,Voted	686	809	1240
Treated,Did not vote	533	810	1716
Control,Voted	3064	2856	4639
Control,Did not vote	3020	2828	7179

Table 6: In-Person Treatment and Response by Household Size and Propensity Score Strata

3.2 A tempting simplification, and why the temptation is to be resisted

Focusing on effects of telephone solicitations, rather than in-person appeals, Imai (2005) analyzes the Vote 98 campaign by matching two groups to each other: (1) subjects who were assigned to treatment and then received it to (2) controls who either refused treatment, were not present to receive it, or were never assigned to get it. The matches were made without regard to which of these three reasons led to a matched control's not receiving the treatment, although Imai took pains to ensure comparability of treated and control units on the other covariates. His approach appears simpler than both Gerber and Green's original two-stage least squares analyses and the method we are presenting; why not use it?

The danger of drawing a comparison group from all three pools of untreated subjects is that doing so could introduce bias. If the sort of person who answers the door for a canvasser is also the sort of person who is more apt to vote, then this sort of person is overrepresented among the set of treatment-group members who complied and underrepresented among treatment group members who did not receive the treatment. By extension, they would also be underrepresented in the combined pool of untreated subjects.

Whether a subject would be available and willing to speak with a canvasser is known only for those with whom canvassing was attempted, not for the group initially assigned to control: coding this information as values of an observed variable would lead to missing values for each person assigned to control. Variables associated with accessibility to canvassers are available for the sample, however, and the relevant subgroups do not appear to be similar on them. Figure 1, to begin with, compares age distributions by subgroup. The left panel shows the distribution of age between people who were assigned to in-person treatment (in solid grey), and those who were not (the black line). Although the groups *assigned* to treatment and to control appear to be comparable, the age distribution among those who *actually* answered the door to receive treatment (in grey) is different from the distribution of those who were assigned the treatment but who did not answer the door (in black). Those who answered the door were systematically older than those who didn't. Since older people are systematically more likely to vote than younger people — see Verba, Schlozman and Brady (1995); Nie, Junn and Stehlik-Berry (1996); Rosenstone and Hansen (1993); Highton and Wolfinger (2001); Wolfinger and Rosenstone (1980), among many others — imbalance on age is particularly troubling.

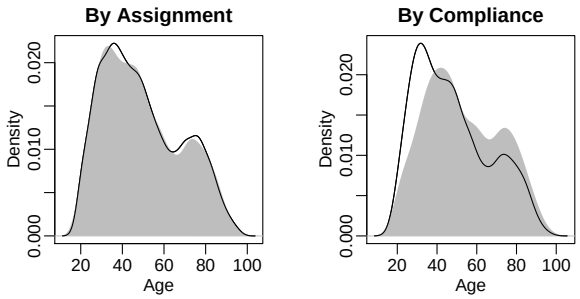

Figure 1: Age Distributions by In-Person Treatment Assignment versus Compliance with In-Person Treatment

	Standardized Bias	$p(< \|t(Z, r)\|)$
votein96novote	−0.15	0.00
votein96notreg	−0.09	0.03
votein96voted	0.20	0.00
wardF2	−0.13	0.00
wardF5	−0.07	0.10
wardF6	−0.16	0.00
wardF9	−0.13	0.00
wardF11	0.22	0.00
wardF13	−0.15	0.00
wardF14	0.10	0.01
wardF15	−0.12	0.00
wardF19	−0.15	0.00
wardF20	0.18	0.00
wardF21	0.12	0.00
wardF25	0.12	0.00
wardF26	−0.15	0.00
wardF28	0.11	0.00
wardF29	0.14	0.00
MAJORPTY	0.09	0.03
ns(AGE, 5)2	0.16	0.00
ns(AGE, 5)3	0.33	0.00
ns(AGE, 5)4	−0.23	0.00
ns(AGE, 5)5	0.33	0.00

Table 7: Covariates Showing the Largest Differences between People Who were Contacted and and Not-Contacted in the In-Person Condition Post-Stratification

More broadly, tests like those of section 3.1 showed that not only are compliers systematically older than non-compliers, but they tend to live in different neighborhoods and have different past voting behavior. Table 7 presents the largest standardized differences seen in comparisons of compliers to non-compliers among those assigned to treatment. Many of the standardized biases are quite large, and in addition, 22 of 36 null hypotheses asserting similarity between compliers and non-compliers are rejected (given conventional significance levels).

In analyses of the type we are considering, then, the threat of bias from observed covariates is real (in addition to the ever present danger of bias from hidden covariates). Were the treated and not-treated groups to be compared without prior benefit of matching, the justifying assumption required would be that the treated/not-

treated distinction is ignorable *simpliciter*; however, by using matching, Imai (2005) relaxes this requirement somewhat. He need only assume that, given covariates, the distinction between treated and not-treated subjects is ignorable. He can concede that subjects not treated because they did not comply with treatment may differ systematically from subjects not treated because they were initially assigned to the control group. He must insist, however, that these systematic differences are entirely captured by the observed covariates.

In other words, his strategy asks us to believe that the information in the dataset accounts for differences in compliance. Perhaps this is so, but it is unclear to us why it should be so. In broad strokes his approach is similar to the as-treated analysis in clinical trials, a method which has fared poorly in empirical assessments of it (see Lee et al., 1991, for an example of such an assessment). We prefer to proceed on weaker assumptions. By using an instrumental variable, we avoid the need to model compliance.

3.3 The method of instrumental variables

If subjects who received the treatment differ in unmeasured ways from those who declined it, how can bias be avoided in the assessment of treatment effects? One important approach makes use of a variable, an "instrument," that is not related directly to the response, but that influences who actually receives the treatment. In an instrumental variables analysis with a binary IV and binary treatment, subjects whose instrumental variable-value is one are compared with subjects whose instrumental variable-value is zero, regardless of whether the instrument is of substantive interest in itself. The point of the comparison is to shed light indirectly on effects of certain other variables with which the instrument happens to be associated. A recent exposition and explanation of this approach by Angrist, Imbens and Rubin (1996*a*) shows that when treatment assignment is ignorable (as it is in most randomized studies), then treatment assignment is the perfect instrumental variable. Should ignorability and other the assumptions hold, then instrumental variables estimates of treatment effects are not undermined by systematic differences between subjects who would and would not comply with treatment; this is why Gerber and Green (2000) selected the approach in their original analysis.

What assumptions does an instrumental variable require? The first is ignorability, which with an IV can be put as follows: the instrument should not be correlated with other potential confounds. A second assumption, known as the exclusion restriction, states that the instrument (treatment assignment) only affects voting via the treatment itself (answering the door to receive a GOTV message). A third requirement of common IV estimators is that treatment assignment must increase the probability that a person will actually receive the treatment. There are a few more minor assumptions that IV estimators typically require (Angrist, Imbens and Rubin, 1996*a*), but as Imbens and Rosenbaum (2005) have pointed out (and will be illustrated in the course of this paper) only the first two, ignorability and the exclusion restriction, are required to use instrumental variables with randomization inference.

To say that IV analyses assume ignorability somewhat overstates their requirements. As seen in § 3.1, assignment to in-person canvassing appears, strictly speaking, not to have been ignorable; so the IV approach would seem not apply. Yet it does, granting that treatment assignment is ignorable conditionally upon our three-level

stratification. It also turns out that randomization inference accommodates the stratification naturally via the Mantel-Haenzel test. In terms of assumptions, it will require only the exclusion restriction and that treatment assignment is ignorable within strata.

Let us apply this assumption to a test of the hypothesis that in-person canvassing in the Vote'98 campaign was without any effect. The null hypothesis, that treatment had no effect, is automatically in accord with the assumption that treatment can have affected no others than those who received the treatment. To take into account that there are three strata rather than one, we use the Mantel-Haenszel test. The test statistic is the number of treatment-group members, across strata, who voted in the subsequent election, centered and scaled by the null expectation and variance of that number from Table 6. (The statistic itself is not influenced by the degree of compliance with intended treatment.) The number generated in this way is improbably large; its p-value is 0.00037. Were the null hypothesis true, so extreme a statistic would occur in no more than three comparable experiments in 10,000. We conclude that the null hypothesis is false; the Vote 98 campaign's in-person contacts did influence subsequent voting.

3.4 Summary

So far, we have mostly focused on hypothesis testing to check balance, and we have only done one small analysis (also a simple hypothesis test) of the null hypothesis of no treatment effects.. Most political scientists, however, would like to do more than test hypotheses about no treatment effects (even though the amount of asterisks and language about "significant coefficients" in published articles might be evidence against this statement). Can we represent and appraise hypotheses about the particular effect of a given treatment or are we stuck with observed differences and some number of stars from a test of no effect? The answer is yes. We will use a method first suggested in (Rosenbaum, 2001) called "attributable effects" to produce confidence intervals and point estimates of treatment effects using random assignment — suitably adjusted — as our instrument. We do not explain attributable effects here, but the interested reader can see our other explanations (Bowers and Hansen, 2005*a*,*b*). We have chosen this framework because it allows us to continue to do randomization inference while still using an instrumental variable and reporting confidence intervals and point estimates.

3.5 How Many People Voted Because of In Person Treatment

No more than 435 people who were assigned an in-person visit actually opened their doors who voted. By the exclusion restriction, this means that we can attribute no more than 435 votes to treatment (that is, $A \leq 435$, where A is "# of votes attributable to treatment" or "# of votes which would not have happened in the control condition, but which could have happened in the treatment condition"). We also assumed that opening one's door in the face of a GOTV appeal would not have a negative effect ($A \geq 0$). Since a confidence interval is defined as the set of null hypotheses that would be accepted at a given significance level, we can create a confidence interval for A by just doing all of the different tests for each value of A.

Following the logic explained more fully in (Bowers and Hansen, 2005*b*), in this case with only 3 strata, we

need to test all of the possible configurations of treatment assignment for any given hypothesized A. If we wanted to assess the potential for attributable effects from 0 to 435, this would require 11,982,429 tests — that is, there are nearly 12 million different ways that a binary treatment can be assigned such that between 0 and 435 votes are observed among the respondents who actually opened their doors. In order to narrow the scope of our search, we first tested just a few As spread across the range of possible values, and narrowed in on the range from 0 to 300 as containing the most probable values. We then tested each A in this range, amounting to 4,590,551 tests. To generate the potential $\boldsymbol{\tau}$s took about 45 seconds and to execute the tests took another 12 seconds on a 2 processor Unix workstation. Figure 2 shows the results of testing all of the attributable effects between 0 and 300. The curve peaks at 149, the Hodges-Lehmann point estimate (Rosenbaum, 2002*c*, ch. 2).

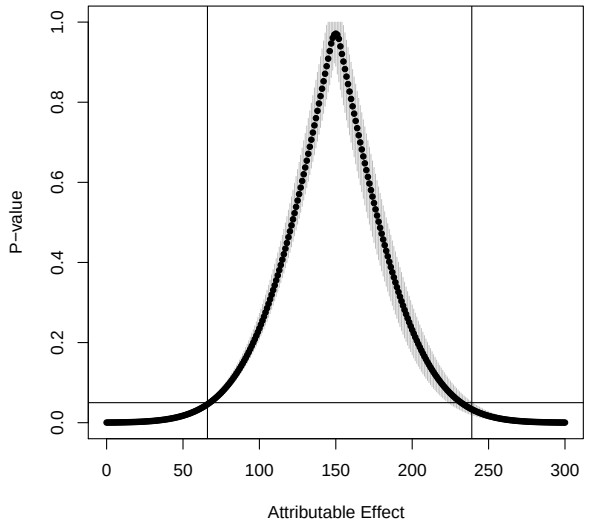

Figure 2: Attributable Effects for In-Person Treatment

Notes: The points represent the median probability associated with a null hypothesis (y axis) of each attributable effect (x axis). The gray line segments show the range of p-values returned for SPPRs that attribute that amount of voting to treatment. The horizontal line shows p=.05. The two vertical lines show the boundaries of a 95% confidence interval.

3.6 Summary

The job of enhancing the case for ignorability was made easy in the case of in-person treatment because random assignment of treatment worked rather well in that case. In fact, in order to justify the use of random assignment as an IV, we did not need to make any adjustments other than a very gentle stratification of the sample by type of household (1 and 2 person) and among 2 person households by propensity to be assigned treatment.

We now turn to the case of GOTV telephone calls. For this treatment, making the case for ignorability was more difficult than for the in person condition. For this reason we use optimal full matching rather than

stratification. The logic of calculating treatment effects, however, remains exactly the same.

4 Inference for the effect attributable to telephone solicitations using full matching

4.1 The need for propensity-score adjustments

Gerber and Green's Vote 98 experiment studied various GOTV encouragements, in design giving similar emphasis to telephone, mail, and face-to-face contact as methods of delivery of the appeal; but in practice there were important differences between these three parts of the experiment. Because of the differing costs of the interventions, for instance, more subjects received mailings than were called, while still fewer were canvassed in-person. More consequentially, by-mail appeals were distributed in a demonstrably random fashion, and assignment to in-person canvassing was nearly random, whereas the distribution of attempted telephone solicitations decidedly differed from distributions a properly functioning randomizer could be expected to produce. In the in-person part of the experiment, there was some imbalance between the subgroups with whom personal contact was and was not attempted; but we have shown that the imbalance was readily explained as a matter of having assigned pairs of voters who shared a household to treatment or to control together, as clusters of size two, rather than independently, and simple adjustments restored this balance. For the telephone intervention, there were again imbalances between groups assigned to treatment and to control. Unfortunately, these differences between treated and control groups cannot be explained as a by-product of having randomized two-voter households together, as they also affect subjects from households with only a single voter. For the experiment with telephone solicitations, a more studied adjustment is called for.

Were the assignment of telephone calls ignorable in the simple sense — ignorable given or not given covariates; ignorable in the sense ordinarily secured by random assignment — then model-based attempts to predict treatment assignment, Z, from covariates, $\mathbf{x}$, ought not enjoy discernibly more success than predictions of Z on the basis of no information at all. This is not the case, as Imai (2005) establishes in his critique of the experiment and its analysis. Imai presents a logistic regression model of Z as a function of pre-treatment variables including age, location of residence, and prior voting history, and in an analysis of deviance soundly rejects the hypothesis that improvement of fit of his model over one without covariates is due to chance. It follows that the ignorability assumption upon which Gerber and Green's original analysis depends must also be rejected. Noting this, Imai goes on to argue that a more appropriate analysis would exclude those subjects who received treatments other than telephone solicitations (mailings or in-person solicitations) and would ignore the instrumental variable, assignment to receive a telephone call, comparing only those who answered a telephone call (and no other treatments) to controls who were not assigned to any of the experimental treatments (2005).

In contrast to Imai, we think it important both to retain the subjects who received treatments complementary to the telephone treatment and to analyze the data in a way that makes use of the experimental randomization, perfect or not. But we share his conclusion that this experiment's random assignment to telephone treatment conditions must have been broken or somehow flawed. In order to accommodate variation in levels of the

complementary treatments, before testing for association between covariates and assignment to receive a telephone solicitation, we condition on assignment and receipt of complementary treatments: whether a subject was sent a mailing, how many mailings were sent, whether personal contact was attempted and whether it occurred (see Table 1 on page 6 for a depiction of the Green and Gerber research design). If treatment assignment is ignorable within levels of complementary treatment, then no such association should obtain. As the analysis reveals a distinct association, the hypothesis of ignorability must be rejected. Table 8 shows the results of this analysis of deviance. Clearly, assignment to phone treatment is not independent of covariates.

	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
Complementary Treatment Assignment	29368	26469.50				
...+Covariates	29334	26383.47	34	86.03	2.53	0.0000

Table 8: Assessing the Ignorability of Assignment to Phone Treatment

The same could have been said of assignment to personal contact: we showed on page 7 that its association with covariates was statistically significant. However, that significance disappeared if attention was restricted to subjects from single-voter households, supporting our surmise that the association was an artifact of clustered treatment assignment. With assignment to telephone appeal, restriction to single-voter households does *not* remove that association; rather, the association remains highly significant among one- and two-voter households, taken jointly or taken separately (not shown).

Like Imai, we use propensity-score methods to address the resulting imbalance of covariates. Our model specification was found by a forward-backward model search, beginning with the larger model of treatment assignment as a function of covariates given in the preceding display and limited above by the model containing it and all second-order interactions of its covariates. (Using a similar procedure, Rosenbaum and Rubin (1984) generated "considerably greater balance on the observed covariates ... than would have been expected from random assignment" (p. 517).) This procedure added to our model some eight interactions among original variables and enhanced its fit quite significantly ($p < 10^{-4}$). We differ from Imai in not modeling the propensity *both* to be assigned to treatment *and* to comply with it, $\mathbf{P}(Z = 1, C = 1|\mathbf{x})$, but simply the propensity to be assigned to treatment, $\mathbf{P}(Z = 1|\mathbf{x})$. To emphasize this distinction, we refer to our score as an *assignment score*. The propensity to be assigned to treatment, rather than the propensity to both be assigned to treatment and to comply with it, is the propensity appropriate to the IV method of estimation that we shall eventually use. In what follows, we describe some other techniques that we use to ensure maximal similarity of members of matched sets — and to thus to make the case for ignorability of assignment conditional on set as strong as possible.

4.2 Assignment-score calipers

An old technique to adjust for possible confounding from a single continuous variable is to match within a *caliper* on it (Cochran and Rubin, 1973), that is to match treatment and control units subject to the restriction that matched units differ by no more than a fixed constant c on the variable. The point of reducing bias is to

strengthen the case for ignorability conditional on the adjustments made. The constant c is then the half-width of the caliper. The caliper matching literature offers some guidance on the selection of c, recommending

caliper half-width	percent bias reduction if …		
	$2\sigma_t = \sigma_c$	$\sigma_t = \sigma_c$	$\sigma_t = 2\sigma_c$
0.2	99	99	98
0.4	96	95	93
0.6	91	89	86
0.8	86	82	77
1.0	79	74	69

Table 9: Guidelines for widths of calipers. Findings reported by Cochran and Rubin (1973) on bias reduction on a continuous covariate after matching within calipers on it, as a function of half-width of the calipers and of within-group s.d.s on the continuous covariate.

that it be selected with attention to the difference in s.d.s on the underlying continuous variable within the treatment and control groups and to the fraction by which the between-group discrepancy on that variable is to be reduced: see Table 9, adapted from Cochran and Rubin's paper. Calipers are not a requirement for the use of propensity scores or of optimal full matching, but they appeal to us because of the availability of rough guidelines on the choice of caliper width, and because their use automates parts of the full-matching process that can otherwise be tedious. Following recommendations of Rosenbaum and Rubin (1985), we impose our caliper not on the fitted probabilities of assignment to treatment, $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$, but on their logits, $\{\log(\hat{p}_i/(1-\hat{p}_i)) : i = 1, \ldots, n\}$; since we used logistic regression to fit our scores, these coincide with the linear predictor component of the model-fitting output.

A commonly used caliper half-width is a fifth of a pooled s.d. in the propensity score; this is the value chosen by Rosenbaum and Rubin (1985) for an epidemiological application on the basis of Cochran and Rubin's earlier results. After examining the data, we decided to impose a much narrower caliper. Our examination consisted of partitioning the data by the assignment score divided at equally spaced points, and within these bins testing for association between treatment assignment and covariates, or between treatment assignment and the assignment score.[15] We used all covariates that had been selected into the specification of the assignment score, representing categorical covariates with separate dummy variables for each level; this yielded 54 variables. Split into two, three, four, five, or six bins, most of these variables were not significantly associated with treatment assignment at the .05 level, but some were. Once the data were split into seven bins, the associations of treatment assignment with each of the pre-treatment variables became statistically insignificant, and they remained so for finer partitions as well. At this point the bin-width was slightly more than $s_p/2$, where s_p is

[15] Our tests were randomization-based: for each variable $\mathbf{v}$ we compared the overall sum of its values in the treatment group, $\mathbf{z}^t\mathbf{v}$, to the distribution of values of a random sum $\mathbf{Z}^t\mathbf{v}$. In these expressions, $\mathbf{z}^t = (z_1, \ldots, z_n)$ is a nonrandom sequence of 0s and 1s indicating which subjects were actually assigned to treatment, whereas $\mathbf{Z}^t = (Z_1, \ldots, Z_n)$ represents a sequence of $\{0,1\}$-valued random variables abstracting from the configuration of treatments in the achieved sample to configurations a random assignment mechanism might have generated. The random sequence $\mathbf{Z}$ is such that in each possible realization of it, each matched set M's number of subjects assigned to treatment, $\sum_{i\in M} Z_i$, equals the number of treated subjects it contained in the actual sample, $\sum_{i\in M} z_i$; the distribution of $\mathbf{Z}$ is defined to be the uniform distribution on $\{0,1\}$-valued sequences respecting this constraint. This fully specifies the distribution of $\mathbf{Z}$, and by extension it specifies the probability distribution of $\mathbf{Z}^t\mathbf{v}$ for each covariate $\mathbf{v}$, since we are treating the covariates as constants. Finally, when $\mathbf{z}^t\mathbf{v}$ is found to lie in either tail of the this distribution for $\mathbf{Z}^t\mathbf{v}$, the hypothesis of treatment-control balance on $\mathbf{v}$ is rejected.

the pooled s.d. of assignment scores ($s_p = [(n_t + n_c - 1)^{-1}((n_t - 1)s_t^2 + (n_c - 1)s_c^2)]^{1/2}$).

With a caliper of $s_p/2$, treatment assignment remained significantly associated with the assignment score itself. (It was to be expected that this association would be more difficult to remove than association of treatment with any of the covariates individually, since by construction the assignment score is the linear combination of covariates that is maximally associated with treatment assignment.) To remove this association required a much finer partition. Only with 38 bins, of width $.1s_p$ each, was this association also removed. To ensure that our matching would compare subjects no farther on the assignment score than did this stratification, we set our caliper half-width to be half this, $s_p/20$.

Such a requirement is far stricter than any Cochran and Rubin (1973) had considered — in Table 9, which is taken from their paper, the narrowest half-width considered is $s_p/5$. One might fear that so restrictive a criterion would exclude many subjects from the analysis for lack of comparison subjects whose assignment scores are near enough to theirs. This was not so; all but 4/10 of a percent of the sample had *some* counterpart assigned to the opposite treatment condition whose assignment score was within this distance of theirs, although for most subjects, only a small subset of the total pool of potential matches met this standard of comparability. The small proportion of subjects without such a counterpart consisted of nine subjects assigned to treatment and 116 assigned to control; these subjects are excluded from further analysis. Of the remaining subjects, most have many potential comparisons and all are within caliper distance of at least one subject with whom they might be compared. Our full-matching routine will be able to find suitable matches for all of them.

4.3 A response-tailored assignment score

This section introduces a way of modifying propensity scores (or assignment scores) so as to better reduce bias in estimates of causal effects — which is yet another way that we attempt to strength the case for ignorability of treatment assignment. Whereas ordinary propensity-score adjustment aims to control bias on available pre-treatment covariates $\mathbf{x}_1, \ldots, \mathbf{x}_k$ and, by extension, potential responses $\mathbf{r}_c$ and $\mathbf{r}_t$, the modification aims more narrowly at bias on potential responses, particularly $\mathbf{r}_c$, and to be beneficial it requires some mild additional assumptions. Because of its use of additional assumptions, we chose not to base our calipers upon this modification, reserving it instead for a supporting role (to be explained in § 4.4). However, the additional assumptions required are mild, data-driven, and in this application quite tenable. Should they hold, our use of the modified assignment score gives us additional insurance against bias (i.e. violations of the ignorability assumption) in assessment of treatment effects.

The point of a propensity-score adjustment is to create groups within which the potential responses to treatment are not different from potential responses to control. If there is no relationship between an estimated propensity score and potential responses, we would not expect any adjustment to occur upon stratification or matching on the propensity score. Because potential responses are never observed jointly, one is never in a position to discern the non-existence of such a relationship with any certainty, but when such a relationship does exist, the data are likely to provide positive indications of its existence and perhaps some hints as to its form. In this spirit, we performed a non-parametric, local regression of responses, *i.e.* whether or not a subject eventually

voted, on fitted assignment scores, *restricting the fitting to the control group only*. The restriction reflects our assumption that observed responses among control group members sharing assignment score e^* are a simple random sample from $\{y_{ci} : e_i = e^*\}$, the set of potential responses in the absence of treatment among all study subjects, experimentals and controls, sharing estimated score e^*. (Were we in possession of a "true" propensity score, and certain of ignorability given covariates, then control group members' representativeness would be a fact, not an assumption.)

The fitted regression, with approximate confidence bars and with a "rug" below it to indicate the sample distribution of assignment scores, is shown in Figure 3. We used Loader's "locfit" routines (Loader, 1999), with default settings, to fit the curve. The figure suggests, without definitively confirming, a weak relationship between assignment score and potential responses (i.e. probability of voting in the absence of treatment). This appearance supports our decision to adjust for assignment scores rather than to ignore the treatment-control imbalances described in § 4.1: in the event that the suggestion is correct, then the analysis is prone to bias unless it adjusts for assignment scores.

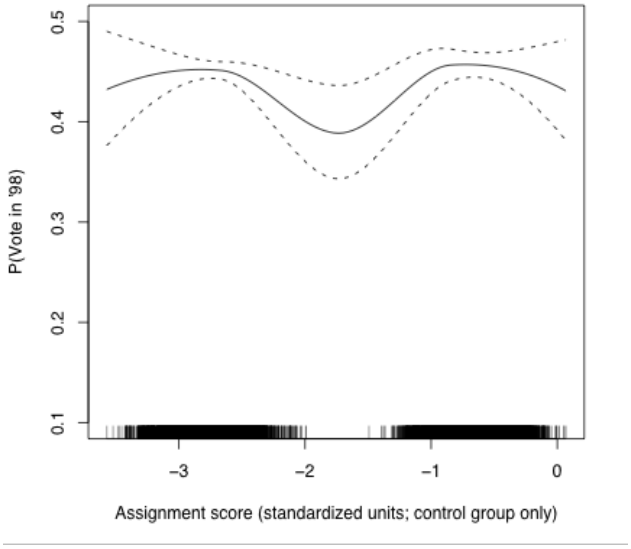

Figure 3: Probability of voting and propensity to be assigned to treatment. A local regression of response on fitted assignment score, control group only, with the distribution of assignment scores among the controls indicated by the marks at bottom. Dashed lines indicate pointwise 95% CIs.

The pattern shown in Figure 3 offers guidance on another issue. According to its original definition, the propensity score is either the probability of assignment to treatment (conditional on covariates) or any monotone transformation of that probability (See, for example Rosenbaum and Rubin, 1983). The theory leaves open just which transformation of that probability is to be preferred. Propensity-score specifications are validated by checking that within propensity-score strata, no covariate shows association with treatment status; and in case studies this is often achieved by stratifying on quantiles of the estimated propensity score (Rosenbaum and Rubin, 1984), which are unchanged by monotone transformations of it. Despite the

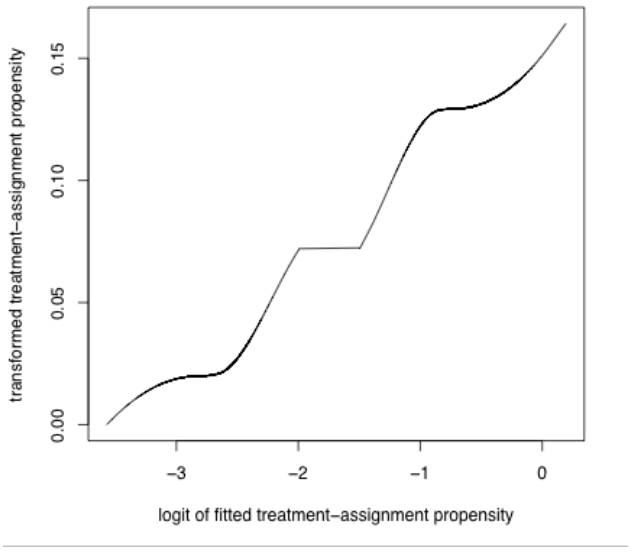

Figure 4: Response-tailored assignment score as a transformation of the simple propensity score. If the curve shown in Figure 3 faithfully reflects associations between the (logit of the) probability of assignment to treatment and the likelihood of voting in the absence of treatment, then bias will be minimized if comparisons are restricted to units similar on the y-coordinate of this figure, the response-tailored assignment score.

importance of covariate balance in their validation, the end goal of propensity-score adjustment is to remove associations of treatment status with *potential outcomes*, not covariates, and the curve of Figure 3 suggests a transformation that, for these data, is adapted specifically to this purpose.

To represent assignment scores on such a scale that bias due to treating unequal propensity scores as if they were the same is reflected in the magnitude of the difference between them, we use the total variation of the curve shown in Figure 3. If f is the curve in that figure and l is the least of the x-coordinates (*i.e.* assignment scores prior to transformation) represented in it, its total variation is the function $t(x) = \int_l^x |f'(t)|dt$. It is an increasing function but it increases more in those parts of the curve where assignment score and probability of voting in the absence of treatment covary, less so where they do not. One can approximate $t(x)$ by arranging the values of f in increasing order $a_{(1)}, \ldots, a_{(n)}$ of the raw assignment score, calculating differences $\{f(a_{(i)}) - f(a_{(i-1)})\}$ between f-values of adjacent units, and summing the magnitudes of the differences to get $t(a_{(i)}) \approx \sum_{j=2}^{i} |f(a_{(j)}) - f(a_{(j-1)}|$ before returning the data to their original order. To distinguish it from the logits of fitted assignment probabilities from which it was derived, call the result the *response-tailored assignment score.* Figure 4 plots response-tailored assignment scores against logits of fitted assignment probabilities. If matching is performed so as to promote closeness of matches on the response-tailored assignment score, then differences on the assignment score will be kept small in just those cases where large differences on it would lead to bias. And, this additional effort to decrease bias leads to a stronger case for ignorability.

4.4 Optimal full matching with calipers

The ability of any adjustment to reduce bias depends on making appropriate comparisons. So far we've discussed two ways to measure the potentially confounding distance between treated and controls: calipers based on the assignment score, and the response-tailored assignment score. Now, we want to use these assessments of distance to create appropriate matched sets.

When restrictions as to who can be matched with whom, such as those issuing from the imposition of one or more calipers, are to be observed when matching treated subjects to subjects in a pool of controls, optimal full matching is the one technique that bears an important guarantee. The guarantee is that it will match each subject capable of being matched to at least one suitable counterpart, and to no unsuitable counterparts (Rosenbaum, 1991). The importance of the guarantee is that it gives license to the data analyst to design, and then enforce, precise requirements as to which units may be considered similar enough to be compared. In designing her requirements, she will have to balance competing aims, because too stringent a standard of comparability makes it impossible to find matches for anyone, while too loose a standard permits poor matches and may introduce bias. While some such trade-off is unavoidable, with optimal full matching the analyst can manage it in the knowledge that, because of the guarantee, her compromise solution will be precisely adhered to, with no ineligible matchings tolerated and no eligible candidates excluded from the matching to be produced.

The analyst first decides on the standards of comparability in her dataset — essentially creating strata containing only subjects who are, in her eyes, comparable. Usually, such standards are guided by a desire to make a strong case for ignorability (as our standards of comparability are here). Once such post-stratifications are in place, full matching removes subjects from the analysis only for lack of suitable comparison units, never because of operational limitations of the matching routine itself. With-replacement matching offers no comparable guarantee because it doesn't create poststrata; while without-replacement routines other than full matching can (i) fail to place everyone with a suitable counterpart into some matched set, (ii) match some subjects to counterparts with whom they are not reasonably comparable, or (iii) both. Full matching's improvement upon older forms of with-replacement matching was documented with simulation studies by Gu and Rosenbaum (1993), and with data by Hansen (2004, §2).

The improvement is achieved by way of greater flexibility in the configuration of matched sets. In traditional forms of matching, such as those implemented in the `matchit` add-on package (Ho et al., 2004) for the R statistical software program, matched sets consist of a single treatment unit and either one or a fixed number of controls. With full matching, by contrast, a matched set may contain one treated subject and a positive, not-necessarily predesignated number of controls; one control subject and a positive, variable number of treated ones; or anything in between; and one poststratification produced with full matching may contain matched sets of each of these types. In a small comparison with results produced by `matchit`, this flexibility permitted `optmatch`, an R add-on package focused more squarely on full matching, to generate a poststratification giving a 30% narrower confidence interval than the one that would have been generated had the matching been performed using `matchit` (Hansen and Klopfer, 2005, §5). (The two packages cooperate to a degree,

with `matchit` making use `optmatch` in order to produce matches that are best-possible among matches of the traditional type; but as of this writing the greater flexibility of full matching is available only through direct use of `optmatch`.)

The simplest application of full matching to the New Haven telephone experiment matching problem illustrates the flexibility of full matching, and is also suggestive of some of the hazards that attend to it. Here, the sole input to the function that produces full matches is a matrix containing one row for each treatment unit and a column for each control and filled with numbers indicating the suitability and desirability of each potential match. In this matrix, the cell in the *i*th row and *j*th column contains a finite number only if treated subject *i* and control *j* are deemed comparable; if not, the cell should be empty or filled with `Inf`, the R representation of infinity. Propensity-score calipers may be represented in this distance matrix by emptying each cell of the matrix corresponding to a pair of subjects separated by more than a caliper on the propensity score. Within the remaining cells, we convey relative desirability of matches to the matching routine by assigning large positive numbers, "match discrepancies," to pairs that are less suitable for matching, while reserving smaller, but still nonnegative, match discrepancies for cells corresponding to more desirable matches. (The software appraises candidate poststratifications on the basis of the sum of discrepancies of matched units, returning one that achieves the least-possible sum of discrepancies or something very close to it.) To match subjects of the telephone matching experiment in such a way as to respect the requirements laid out in section 4.2, it would suffice to put any finite number in each cell corresponding to a permissible match, with `Inf` in the others. Again following Rosenbaum and Rubin (1985), we rated the permissible matches using a Mahalanobis distance on the age variable and on the response-tailored assignment score. Because we already were inspecting the phone treatment stratified by the other experimental treatments, we preceded our full matching with a stratification along these complementary treatments. Only treatment and control group members who both had been slated for in-person canvassing or both had not, who had both been reached by the canvassers or both had not, and who both had been sent the same number of mailings (0, 1, 2, or 3), would be eligible to be matched to one another. Matching in this way also secures the important benefit of excluding conclusively the possibility that the telephone-treatment effect be confounded with the effects of personal canvassing or direct mail.[16] The subjects fell into twelve strata, and were matched (subject to caliper constraints and to the Mahalanobis distance) only within these strata.[17] This could in principle be achieved by applying the function `fullmatch` twelve times, to each of the strata, but the software has options that obviate the need for such repetition, at least on the part of the user. Then we generated a full matching.

The poststratification obtained in this way contained matched sets with one treated subject and controls varying in number from one to 86, as well as matched sets with only one control but as many as seven treated subjects. So variable set of matched-set configurations tends to increase the width of confidence intervals to be

[16] A side benefit of this stratification is that it would allow us to examine differences in treatment effects by complementary treatment. In the interests of space, we do not pursue the idea that, say, phone calls plus in-person visits are especially effective at mobilizing voters. However, we could use the same results of the matching that we report below to answer questions about such interaction effects with no modification.

[17] A technical aside: when matching within strata and with calipers on a propensity score, we find it important to include stratum effects in the propensity score stratification, as this tends to maintain the balancing property of the propensity score while reducing the number of within-stratum potential matches that are forbidden by the caliper constraint.

produced in the analysis (Hansen, 2004). For sake of the precision of later statistical inferences, we decided to set limits on the matching procedure, using only as much flexibility as is needed. As is needed, that is, to permit that each subject with comparable counterparts be matched to one or more of them, but not to counterparts she is not comparable with. The full matching software offers the option to constrain (i) the maximum number `max.c` of controls to be matched to a single treated subject and (ii) the minimum ratio `min.c` of controls to treated subjects in matched sets — perhaps a whole number k, indicating that only matched sets with one treatment and k or more controls are to be allowed, or perhaps a fraction $1/k$, indicating that sets of one control and k treated subjects are permissible. Without such constraints, full matching may generate matched sets with a treatment and any number of controls, or a control and any number of treated subjects, but with them the matched sets' ratios of controls to treatments are forced to fall between `min.c` and `max.c`. Not every such pair of constraints is possible to meet: if a data set contains three times as many controls as treated subjects, then clearly the requirement `max.c = 2` is impossible to observe while placing all available subjects into matched sets. Indeed, if also some potential matches were forbidden by caliper requirements, then it might be necessary to set `max.c` to a number larger than three in order that the matching problem be feasible.

This section (§ 4.4) began by noting that full matching is the one approach to matching guaranteeing that all and only the subjects who can, considered in isolation, be well matched, will be placed into poststrata alongside well-matched counterparts. This is always so for full matching without restrictions, but whether it remains true once `min.c` and `max.c` are specified depends on the values to which they are set, and varies from one data set to the next. With calipers, the most favorable restrictions consistent with the caliper can be determined by trial and error. The `optmatch` functions `min.controls.cap()` and `max.controls.cap()` automate this task, performing line searches of the positive half-line $(0, \infty)$ to determine the largest value of `min.c`, or the smallest value of `max.c`, with which the restricted full matching problem permits a solution. Optionally, `min.controls.cap()` accepts an argument `max.c`, in which case it explores the feasibility of various `min.c` parameters subject to the `max.c` restriction it was given, and likewise for `max.controls.cap()`. This permits the user to apply the two functions in sequence, first finding the largest value for `min.c` and then selecting a small value for `max.c` that is consistent with it. We apply these functions to our data set, with its twelve separate strata and compound calipers, in turn, first maximizing `min.c`, then passing this largest feasible `min.c` to `max.controls.cap()` in order to set `max.c`. Since full matching is invoked repeatedly by these functions, this is the most time-consuming part of the analysis, requiring up to a day with modern computers; but it regularizes the matched sets appreciably, as shown in Table 10. Whereas the unrestricted matching produced at least one matched set with 84 controls, here, no matched set contains more than 23 controls.

To compare candidate matchings for the likely efficiency of causal effect estimates they would support, Hansen (2004) introduced *relative precision*, an approximation to a ratio of widths of confidence intervals that would result from using a set of proposed matchings in an analysis. Although relative precision applies most directly to studies with continuous, rather than binary, outcomes, an assessment of relative precision is instructive. Without restrictions on matched sets' treated to control ratios, the full matching that arises has precision

	Number of direct mailings sent			
	0	1	2	3
Personal canvass not attempted				
	7 to 14	1 to 3	1 to 3	1 to 3
Assigned to receive personal canvass				
Contact occurred	7 to 14	$\frac{1}{2}$ to 3	1 to 6	$\frac{1}{2}$ to 4
Not contacted	11 to 23	1 to 6	$\frac{1}{2}$ to 5	1 to 3

Table 10: Controls per treatment subject in matched sets (Telephone Treatment). Represented is the variation, across strata of direct mail and in-person treatments, in the number of controls per subject assigned to receive the telephone treatment. For instance, within the subgroup of subjects with whom a personal canvass was unsuccessfully attempted (last row of the table) and who received two encouragements to vote in the mail (third column), subjects with whom a telephone call was attempted are matched to as many as five subjects not called. The same subgroup contains matched sets with as few as one-half an attempted-telephone-entreatment subject per subject for whom no telephone treatment was attempted — *i.e.* matched sets in which two treatment subjects share a control — but no matched sets with a lesser ratio of controls to treated subjects than 1:2. These upper and lower limits coincide with the full-matching restrictions `min.c` and `max.c` found by the optmatch functions `min.controls.cap()` and `max.controls.cap()`.

.99 (an improvement of 1%) relative to matches that pair each treated subject with one control.[18] When the largest feasible values of `min.controls` are insisted upon, precision improves to .92. When in addition the smallest possible values of `max.controls` are specified, precision is still better, at .90.

As is usual with adjustments using propensity scores, our match sharply decreased covariate imbalances between treated and control groups. Prior to matching, our randomization tests found four of 37 first-order covariate standardized biases to be significantly different from zero.[19] Each of these biases was sharply reduced by matching, three to insignificance and one from strong ($p = 10^{-8}$) to mild ($p = .04$) statistical significance. Because we matched subjects only to others who had received the same treatments other than a telephone solicitation, treatment-control imbalances in the frequency with which complementary treatments were applied are completely eliminated by our matching. Our propensity adjustment also addressed second-order combinations of covariates and complementary-treatment variables, for example age^2 and the interaction of major-party membership and having received an in-person solicitation. Viewed together, the covariates generate 530 first and second-order terms (only a few of which contributed to the propensity score). Prior to matching, standardized biases differed significantly from zero in 199 of them; with matching only 32, or 6% of the biases, differed significantly from zero — quite close to the 5% that ignorability would entail, over repeated administrations of this treatment to the same group of people.

These benefits accrue from our having matched on the assignment score within strata defined in part by complementary treatment assignment. The post-stratification that we used, however, also took into account

[18] We use pair matching as reference point only because it is the best known form of matching. Despite the fact that pair matching makes use of many fewer controls than ours did, with these data it would not have been able to pair each treated subject to a unique comparison unit within caliper distance of it: either the caliper requirement would have had to be relaxed, or some treated subjects could not have been matched. Table 10 gives a hint of this, in showing that for three subclasses, the full matching could not have gone through without some treated subjects' sharing a matched control. Because of this, the comparison to pair matching somewhat understates the advantage of full matching over it.

[19] There are only five covariates, ward of residence, age, a descriptor of subjects' voting and eligibility in the previous election, and whether they belonged to a major political party; we get 37 standardized biases by treating the levels of categorical variables separately.

the age of the individuals in the experiment. Since voting is a habitual behavior (Plutzer, 2002), GOTV interventions can be expected to act differently on the young than on the old, who will have had many more opportunities to pick up the habit of voting. Since the relationship between age and turnout is well established, to compare treated and control units that differ greatly in age would be untoward. Yet unstratified analyses, and propensity-adjusted analyses that do not specifically address age, will make such undesirable comparisons.

This is an issue particularly for analyses that are propensity-matched but not age-matched, as the dispersion of age among subjects matched on propensity score but not age can be expected to be only slightly less than the overall age dispersion in the data set. While our matching reduced dispersion along each of the 37 first-order variables contributing to our propensity score, the reductions were mostly (32 of 37) on the order of ten to twenty percent. For a few more variables (four of 37), the pooled standard deviation across matched sets was between half and three-quarters of the overall s.d.; for age, however, matching reduced the s.d. from 19 to three years, only 17% of what it previously had been. On complementary treatments, of course, our matched sets are entirely without dispersion. Our adjustments for age and complementary treatment are significantly more exacting than adjustment by the propensity score alone.

4.5 Votes attributable to GOTV telephone calls

In broad terms, our inferences as to the effects of telephone solicitations to vote proceed in parallel with our earlier inferences (§ 3.5) about effects of personal canvassing. To assess the hypothesis that telephone solicitations brought about A_0 votes, list all possible sets of $A = A_0$ subjects who were called, who then answered the telephone, and who later turned out to vote; for each such set formulate an list of all of the possible treatment assignments that to the effect that precisely these A_0 subjects voted but would not have voted in the absence of treatment (We call this list of possibilities a Specified Pattern of Potential Responses (SPPR)); for each SPPR separately, determine the full set of potential responses in the absence of treatment that is implied by that list and by the pattern of observed responses; and finally, use the Mantel-Haenszel method to test each set of responses separately for lack of association with treatment. Only when all hypotheses on the list are rejected is $H_0 : A = A_0$ rejected. To get a confidence interval for the attributable effect, repeat the entire process so as to find the smallest and largest A_0 for which $H_0 : A = A_0$ is accepted; these delimit a confidence interval for the effect on voting of the telephone GOTV campaign.

Our analysis uses treatment assignment as an instrument. We have shown that we can assume that treatment assignment is ignorable within matched sets, and we assume the exclusion restriction. The ignorability assumption manifests itself in our use of the Mantel-Haenszel method, which takes our full matching into account; and the exclusion restriction enters into the analysis via limits on which SPPRs are considered. SPPRs are not permitted to specify treatment effects for treatment-group subjects who did not comply with treatment.

There are several important differences between this and the analysis of effects of in-person canvassing. The first is a difference in the results: when assessing telephone effects, the hypothesis that $A = 0$ is not rejected. (It yields a standardized test statistic of $-.6$, with associated p-value .53.) We thus had to consider the possibility that treatment depressed voting; accordingly, we conduct two-sided hypotheses of both positive and negative

attributions of effect. (In a test of the hypothesis that $A = -100$, say, our SPPRs attribute *non*-votes, rather than votes, to treatment, asserting of some 100 members of the telephone treatment group who received treatment but did *not* eventually vote that they *would* have voted, but for the GOTV call that caused them to do otherwise.) A second difference with the earlier inferential setup is structural: for the analysis of in-person contact effects, we had poststratified the data into just three sets, whereas now there are many more. In the earlier case, the job of enumerating SPPRs was simplified greatly by the fact that two SPPRs are equivalent, so far as the Mantel-Haenszel statistic is concerned, if they attribute the same number of outcome events in each of the three strata. While the same principle is valid in the current analysis, it is of little use with the present 6,270 matched sets.

Fortunately, since we have many matched sets, another principle, that of asymptotic separability, is available (Gastwirth, Krieger and Rosenbaum, 2000). When testing the hypothesis that A takes a given value A_0, it points the way to that SPPR among the many attributing A_0 events to treatment whose associated Mantel-Haenszel test statistic has the largest left-tailed, one-sided p-value, and to that SPPR which within the same class has the largest right-tailed, one-sided p-value. The two test statistics that result represent the extremes among the test statistics that would result were all SPPRs attributing A_0 events to be tested. Under their null distributions, the standardized versions of each would be expected to fall close to zero. If both fail this test, and if they differ from zero in the same direction, then it follows that tests of the other SPPRs would have resulted in rejection as well, since these two lie at the extremes. If one or the other is within a two-sided acceptance region, then the hypothesis that $A = A_0$ is not rejected, because then the SPPR that generated it is an attribution of A_0 events to treatment that is compatible with the data. (In the event that both lie outside the two-sided acceptance region but they straddle the origin, we assume that at least one of the SPPRs between the two extremes would generate a test statistic within the acceptance region, and the hypothesis that $A = A_0$ is again not rejected.) The reader is referred to Rosenbaum (2002*b*) for the details of how the principle is applied.

To illustrate the importance of reducing the number of SPPRs to be evaluated, consider testing the hypothesis that $A = 100$ votes were produced by the treatment. Of the 2,182 subjects contacted by telephone as part of the Vote 98 campaign, 1314 eventually voted, and our matching placed these voters into 1305 separate matched sets (with nine separate pairs of them sharing a matched control). This makes for roughly $\binom{1300}{100}$, about 10^{152}, SPPRs and associated null hypotheses subsumed under the composite hypothesis that $A = 100$: about 10^{152} separate ways to assign $\tau_i = 1$ to members of the treatment group who complied; or, put in terms of statistical computation, about 10^{152} ways to distribute 100 1's among 1314 positions in a vector of length 29,380, the size of our sample. If $A = 100$ is to be rejected, it is only because each of these 10^{152} hypotheses separately is rejected. However, with the principle of asymptotic separability, it is necessary to calculate only two z-statistics, those the principle asserts to be largest and smallest. The composite hypothesis that $A = 100$ is rejected if both of these lie in the same tail of the standard Normal distribution, far enough from the center to merit rejection — for then it follows that all 10^{152} test statistics would also have merited rejection.

The hypothesis that $A = 100$ is, in fact, rejected at the .05 level, with the z-statistics for tests of the simple hypotheses subsumed under it delimited by -3.8 and -2.2. Applying the same analysis repeatedly with varying A_0, one gets a 95% confidence interval of -159 up to 84 votes. For as many as 7.3% of those contacted, the call

may have dissuaded their voting; while up to 3.8% may have voted as a result of the call. The 2/3 confidence interval ascribes as few as −98 (4.5% of those contacted) or as many as 23 (1.1%) votes to treatment.

5 Conclusion

Random assignment guarantees that treated and control groups have the same distributions of pre-treatment covariates *in expectation*. In any given application of random assignment, however, treated and control groups may well look different — especially if sample sizes or small, or if something went wrong in the administration of treatment itself. Given a manifest difference, or imbalance, in an experimental research design, experimenters have up until now faced three choices (1) ignore the problem and hope that the imbalance is small enough that it doesn't bias estimates of treatment effects (which is probably uncommon and fairly dangerous); (2) throw away the data from a given experiment and try again (which is quite common, and does not expose the researcher to bias, but does waste resources); or (3) treat the experiment as an observational study and hope to directly adjust for the differences in treated and control respondents based on observed covariate values (which never guarantees protection from unobserved biases). In this paper, we have suggested a fourth way that combines the benefits of adjustment with the benefits of using random assignment as an inferential tool (both for the creation of an instrument and as a basis for hypothesis testing, confidence intervals, and point estimates). We showed that one can effectively (and often easily) make enough adjustments to what might seem like a broken experiment to make it whole again. This approach turns out to be particularly important with field experiments, which are very expensive and difficult to replicate (compared to, say, throwing away data from 20 undergraduates in a computerized lab). However, given that one can accomplish this kind of adjustment in a few hours for a given study, it still might make sense to consider it even if an experimenter plans to do another study to replace data from a small and cheap study that was considered broken. Our approach is also quite useful in cases where there are very few covariates on which to adjust. For one thing, our matching procedure effectively squeezed a great deal of information from the few covariates available on the Vote 98 study. For another, we didn't have to entirely rely on our ability to get a propensity score correct, nor did we have to worry (too much) about unobserved confounds (which matching has no ability to avoid) since we used an instrumental variable to estimate the actual treatment effects.

In the end, we hope that analysts of experimental data realize that (1) random assignment is a "gold standard" because of ignorability (and, also because of randomization inference, but that is another story) and (2) that the case made for ignorability can be strengthened by adjustments based on what is observed in a given dataset. Putting these two things together means that resources can be saved, studies don't have to be thrown away, and knowledge can grow, relatively easily and with great confidence because at the very least, random assignment was attempted.

References

Akaike, Hirotugu. 1973. "Maximum likelihood identification of Gaussian autoregressive moving average models." *Biometrika* 60:255–265.

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

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996*b*. "Identification of causal effects using instrumental variables (Disc: p456-472)." *Journal of the American Statistical Association* 91:444–455.

Bowers, Jake and Ben B. Hansen. 2005*a*. "Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign: An Application of Randomization Inference Using Full Matching." Presented at annual meeting of the Political Methodology Section of the American Political Science Association.
URL: *http://www-personal.umich.edu/ jwbowers/PAPERS/bhPolmeth2005July.pdf*

Bowers, Jake and Ben B. Hansen. 2005*b*. "Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference." Prepared for presentation at the Annual Meeting of the Midwestern Political Science Association.

Brady, Henry and Jason Seawright. 2004. "Framing Social Inquiry: From Models of Causation to Statistically Based Causal Inference." Working Paper.

Cochran, William G. and D. B. Rubin. 1973. "Controlling Bias in Observational Studies: A Review." *Sankhyā, Series A, Indian Journal of Statistics* 35:417–446.

Cox, D.R. 1958. *The Planning of Experiments*. John Wiley.

Fisher, R.A. 1935. *The design of experiments. 1935*. Edinburgh: Oliver and Boyd.

Gastwirth, J.L., A.M. Krieger and P.R. Rosenbaum. 2000. "Asymptotic Separability in Sensitivity Analysis." *Journal of the Royal Statistical Society* 62:545–555.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94:653–663.

Gerber, Alan S. and Donald P. Green. 2005. "Correction go Gerber and Green (2000), Replication of Disputed Findings, and Reply to Imai (2005)." *American Political Science Review* 99(2):301–313.

Gu, X.S. and P.R. Rosenbaum. 1993. "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms." *Journal of Computational and Graphical Statistics* 2(4):405–420.

Hansen, Ben B. 2004. "Full matching in an observational study of coaching for the SAT." *Journal of the American Statistical Association* 99(467):609–618.

Hansen, Ben B. and Stephanie Olsen Klopfer. 2005. Optimal full matching and related designs via network flows. Technical Report 416 Statistics Department, University of Michigan.

Highton, Benjamin and Raymond E. Wolfinger. 2001. "The first seven years of the political life cycle." *American Journal of Political Science* 45.

Ho, Daniel, Kosuke Imai, Gary King and Elizabeth A. Stuart. 2004. *MatchIt: Matching Software for Causal Inference*.

Holland, P. W. 1986*a*. "Statistics and Causal Inference (with discussion)." *Journal of the American Statistical Association* 81:945–970.

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

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

Imai, Kosuke and David A. van Dyk. 2004. "Causal inference with Generalized Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99(467):854–866.

Imbens, Guido W. and Paul R. Rosenbaum. 2005. "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(1):109+.

Lee, Young Jack, Jonas H. Ellenberg, Deborah G. Hirtz and Karin B. Nelson. 1991. "Analysis of clinical trials by treatment actually received: Is it really an option?" *Statistics in Medicine* 10:1595–1605.

Loader, Clive. 1999. *Local Regression and Likelihood*. New York, NY: Springer.

Nie, Norman, Jane Junn and Kenneth S. Stehlik-Berry. 1996. *Education and Democratic Citizenship in America*. Chicago: University of Chicago Press.

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

Rosenbaum, Paul. 2002*a*. *Observational Studies*. 2nd ed. New York: Springer-Verlag.

Rosenbaum, Paul R. 2001. "Effects Attributable to Treatment: Inference in experiments and observational studies with a discrete pivot." *Biometrika* 88:219–231.

Rosenbaum, Paul R. 2002*b*. "Attributing effects to treatment in matched observational studies." *Journal of the American Statistical Association* 97(457):183–192.

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

Rosenbaum, P.R. 1991. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society* 53:597– 610.

Rosenbaum, P.R. 2002*c*. *Observational Studies*. Second ed. Springer-Verlag.

Rosenbaum, P.R. and D.B. Rubin. 1984. "Reducing Bias in Observational Studies using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516–524.

Rosenbaum, P.R. and D.B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *The American Statistician* 39:33–38.

Rosenstone, Steven and John M. Hansen. 1993. *Mobilization, Participation and Democracy in America*. MacMillan Publishing.

Rubin, D. B. 1986. "Comments on "Statistics and Causal Inference"" *Journal of the American Statistical Association* 81:961–962.

Rubin, D.B. 1974. "Estimating the Causal Effects of Treatments in Randomized and Nonrandomized Studies." *J. Educ. Psych.* 66:688–701.

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

Wolfinger, Raymond and Steven Rosenstone. 1980. *Who Votes? (Yale Fastback Series)*. Yale University Press.