Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference*

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

April 4, 2005

Abstract

Statistical analysis requires a probability model: commonly, a model for the dependence of outcomes Y on confounders X and a potentially causal variable Z. When the goal of the analysis is to infer Z's effects on Y, this requirement introduces an element of circularity: in order to decide how Z affects Y, the analyst first determines, speculatively, the manner of Y's dependence on Z and other variables. This paper takes a statistical perspective that avoids such circles, permitting analysis of Z's effects on Y even as the statistician remains entirely agnostic about the conditional distribution of Y given X and Z, or perhaps even denies that such a distribution exists. Our assumptions instead pertain to the conditional distribution $Z|X$, and the role of speculation in settling them is reduced by the use of such techniques as propensity scores, poststratification, testing for overt bias before accepting a poststratification, and optimal full matching. Such beginnings pave the way for "randomization inference", an approach which, despite a long history in the analysis of designed experiments, is relatively new to political science and to other fields in which experimental data are rarely available.

The approach applies to both experiments and observational studies. We illustrate this by applying it to analyze A. Gerber and D. Green's New Haven Vote 98 campaign. Conceived as both a get-out-the-vote campaign and a field experiment in political participation, the campaign as it turned out was in some ways more similar to an observational study than to a randomized experiment. Our analysis uses the strengths of the design of their study while adjusting for irregularities ignored by the original analysis. We estimate the number of voters who would not have voted had the campaign not prompted them to — that is, the total number of votes attributable to the interventions of the campaigners. Both our statistical inferences about these attributable effects and the stratification and matching that precede them rely on quite recent developments from statistics; our matching, in particular, has novel features of potentially wide applicability. Our broad findings resemble those of the original analysis by Gerber and Green (2000), although in the small, the method offers additional information as to the campaign's effects upon interestingly different subgroups, such as older voters or those who have not voted in a previous election.

*We are grateful to participants in workshops at the annual meetings of 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 Introduction and Overview

How many more people would vote if campaigns spent more money on neighborhood canvassing and less on television commercials?

In observational studies or experiments aimed at answering questions like this one, analysts must estimate the effect of some treatment (e.g. a visit from a campaign worker) on some binary response (e.g a record indicating whether a person turned out to vote or not). Since usually the types of people who answer their doors are different in politically consequential ways from the types of people who don't answer their doors, a simple comparison between them may reflect their types and not the effects of treatment. Thus, analysts seeking a treatment effect must also adjust for this difference in types. The combination of binary dependent variables and non-random compliance with treatment has tended to lead analysts to use a two-stage estimator to produce estimates of the increase in probability of voting associated with receipt of an in-person get-out-the-vote (GOTV) contact (See, e.g Green and Gerber, 2004; Gerber and Green, 2000). In this paper we present a mode of analysis which directly estimates the number of additional voters attributable to treatment, and which requires fewer assumptions from the analyst than the currently predominant approach. We use data from Adams and Smith (1980) and Gerber and Green (2000) on vote turnout throughout this paper in order to provide examples of the application of this method. Although these datasets both employ field experiments, the general structure of our analyses can also be applied to laboratory experiments or observational studies.

1.1 The Plan

First, we present our understanding of a causal effect — following the formulation of Rubin (1974) and Holland (1986) 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 inferential framework is often called "randomization inference" and it has a long history of development and use in statistics and biological applications. One major benefit of this approach is that its 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.

Third, we show how to use the framework of randomization inference to test hypotheses about the number of people who voted due to treatment. Finally, we estimate the number of voters attributable to two different GOTV treatments: in-person visits and telephone calls.

We develop and demonstrate these techniques first for a simple 2×2 table from Adams and Smith (1980) on the effects of phone calls on turnout in a city council race in Washington, D.C. in 1979. Then, we examine a $3 \times 2 \times 2$ table representing the in-person treatment from the Gerber and Green (2000) data. Finally, we create an optimal full matching for the telephone treatment from the Gerber and Green data.

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

$$\tau_i = r_{ti} - r_{ci} \tag{1}$$

The treatment effect for person i is the difference in potential responses for that person. The "fundamental problem of causal inference" (Holland, 1986) 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,

$$n^{-1}(\sum_{1}^{n} \tau_i) = n^{-1}(\sum_{1}^{n} r_{ti}) - n^{-1}(\sum_{1}^{n} r_{ci}), \tag{2}$$

which is defined as the difference between the average response across people exposed to treatment and the average response across people exposed to control. (Because our responses are binary, the treatment effects we estimate differ somewhat from this, as will be explained in due course.)

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 (2) would either fail to make sense or fail to express 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.

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

A Google search for the words "experiment" and "gold standard" returns over 92,000 results. Why is random assignment in controlled experiments seen as such a powerful way to assess whether a given treatment produced some effect? The answer is lies in part in the fact that, given random assignment of treatments, one may ignore the effects of any observed or unobserved potential confounding covariate when calculating treatment effects.[1] That is, the probability of a given unit receiving a treatment in a randomized experiment does not depend on its potential responses. If before receiving a knock on the door people were all equally likely to vote, then we can compare the number of voters who received the GOTV visit with the number of voters who did not receive such a visit — and we can interpret the difference between these two numbers as being due only to a visit from a campaign activist. If the assumption of "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:

$$\Pr(Z = 1|\mathbf{X}, r_t, r_c) \equiv \Pr(Z = 1) \tag{3}$$

where, Z is 1 if the person was assigned to treatment, and 0 if she was assigned to control and $\mathbf{X}$ is a matrix of potentially confounding covariates. Equation 3 says that, under random assignment, we don't need to pay attention to covariates or potential responses when testing for treatment effects — the only thing we need to know is how Z was generated. Thus, if asked, "Why do people call experiments the gold standard for identifying causal effects?" then most brief (and perhaps cryptic) answer is "Because of ignorability."

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 covariates that might matter for the outcome is the same in the treated group and the control group. And,

[1] 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).

people who do observational studies can do the same thing — but must do something else to adjust for potential confounds, $\mathbf{X}$, to the extent possible. When these adjustments succeed, they bring about *ignorability given covariates* $\mathbf{X}$:

$$\Pr(Z = 1|\mathbf{X}, r_t, r_c) \equiv \Pr(Z = 1|\mathbf{X}) \tag{4}$$

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. In the particular application that we pursue in this paper, we try to rely on the existence of random assignment as much as possible to strengthen our argument for ignoring covariates, but we will also have to make a few adjustments to make our case even stronger.

3 Randomization Inference

Could the results of a given study be merely due to chance? Is a given causal effect believable? In order to answer such questions, statistical inference always requires a probability model. In this paper we use the approach of randomization inference because it is the simplest way to specify a probability model based on our causal micro-foundations.

First, randomization inference enables the analyst to separate substantive theory from statistical specification (i.e. such that people are not encouraged to think in terms of regression models, for example, but in terms of the science itself). And it can enable a more direct and simple representation of a simple theory such as that depicted in (1) than, say, a likelihood function or a posterior distribution. Of course, where theory is strong and well developed in terms of probability distributions, then other approaches may offer benefits in terms of direct representation of scientific theory. It is our informal sense, however, that much political science theory does not have this character. If the formal substantive theory is simple, then you don't need the extra machinery and ontological commitments required of either the Bayesian or likelihood approaches.

Second, even if substantive theory is complex, randomization inference exchanges reliance on potentially dubious point-estimates based on asymptotics that are difficult to validate in finite samples, for a framework in which asymptotic simplifications are available but not necessary and are straightforward to validate within a particular dataset. This article is not meant to be a general primer on randomization inference, but as we describe our methods, we will not assume prior experience with this body of techniques.[2] So, we will explain as we go.

[2]For good basic exposition see Rosenbaum (2002*a*) and for more references Imbens and Rosenbaum (2005). See also Ho and Imai (2004) for an example of randomization inference in political science.

3.1 Randomization Inference in 2×2 Tables

The city of Washington, D.C. was left with a vacant seat on its city council after Marion Barry was elected Mayor in 1978. To fill the newly empty seat, the city held a special election on May 1, 1979. Just before the election Adams and Smith (1980) fielded a small experiment in which 1325 randomly selected registered voters were called on the phone and were given a message urging them to turnout to vote for John Ray (one of the candidates for that city council election). Another 1325 registered voters who were not called served as controls. After the election, public voting records were collected for all 2650 subjects. Table 1 shows that, of the 1325 people assigned to receive a phone call, 392 turned out to vote while 315 of the people who were not assigned to receive a phone call voted.

	Vote	No Vote
Treatment	392	933
Control	315	1010

Table 1: Voting by Telephone Treatment Assignment from Adams and Smith (1980)

We'd like to know whether assigning people to receive a phone call influenced their voting behavior. This question suggests a null hypothesis that the treatment had no effect on turnout. Fisher's exact test represents independence of treatment and responses by the idea that each of treatment and response can be seen as independent binomial random variables with the same probability of success. In a randomized experiment like this one, the treatment is physically generated as a binomial random variable. The idea that response is also like a coin flip captures neatly what we mean when we ask "could this relationship merely be due to chance?" In this case, the number of people in the (Treated, Voted) cell is the test statistic. A side benefit of this approach is that the quantity of substantive interest — the number of treated people who voted — is also the test statistic with a well defined probability distribution.

Fisher showed (and we demonstrate in Appendix A) that the distribution of the number of treated voters is hypergeometric under the null hypothesis of uncertainty. This distribution is the one used in the common Fisher's exact test of independence for 2×2 contingency tables. We can evaluate this null hypothesis exactly (p=.00042) or using a Normal approximation (p=.00036). Both versions of the test cast great doubt on the null hypothesis that we'd observe 392 or more events merely due to chance. It is not plausible that the phone calls had no effect on the vote turnout of people in the Adams and Smith example.

3.2 Randomization Inference with an IV in $S \times 2 \times 2$ Tables

If we can't rely entirely on random assignment for ignorability, then we need to use observed covariates to adjust for potential differences between treated and control units. The simplest way to do this adjustment is to group like units with like. If we did this, however, we would end up with one 2×2 table like Table 1 for each group of comparable subjects. We could calculate a different hypothesis test for each table, but this would not help us test an overall treatment effect. In this situation we need a test that applies to tables nested in multiple strata.

The Mantel-Haenszel test is a generalization of the Fisher exact test which allows us to test the null hypothesis of no treatment effect across 2×2 tables which are nested in S strata.[3] In our example, the MH test statistic (denoted here by $t(\mathbf{Z}, \mathbf{r})$) is the number of treated voters across all of the strata.

$$t(\mathbf{Z}, \mathbf{r}) \equiv \sum_{s=1}^{S} \sum_{i=1}^{n_s} Z_{si} r_{si} \tag{5}$$

Here, Z_{si} indicates the treatment status of each person i in stratum s using a 1 for treated, and 0 for control; r_{si} is the response observed for person i in stratum s (1 if voted, 0 if didn't vote); Z is upper case because it is a random variable — that is, the assignment to treatment and control could have come out other than it was observed in a particular experiment. The response r is written in lower case because, under the null hypothesis of no effect, it is not a random variable — the hypothesis of no effect says it would have been the same whether or not the subject received treatment. Given ignorability, the only variation in this framework is due to Z.

We present table 2 as an example of such a multi-way table. This $3 \times 2 \times 2$ table contains data from the Gerber and Green (2000) experiment. Each column represents a single 2×2 table containing the number of subjects who were assigned to receive an in-person GOTV treatment and the number of people who voted. We will explain the construction of this table in depth in §4.1. For now, suffice it to say that within these three strata, the subjects who were assigned treatment are indistinguishable on observed covariates from those who were assigned control. We observe 686, 809 and 1240 treated voters in strata 1, 2, and 3 respectively.

Under the null hypothesis of no effect, $t(\mathbf{Z}, \mathbf{r})$ in (5) is distributed as the sum of S independent hypergeometric random variables (random because of Z). We could calculate the probability in the tail of this distribution exactly, however a Normal approximation speeds computation and is easily available in two different situations. We'll sketch the rough intuition for this here.

If the number of strata is fixed (say, at 3), but the number of observations in each strata grows large, then the number of treated voters in each table can be seen as a simple random sample from

[3] See (Agresti, 2002, page 231) and (Rosenbaum, 2002*a*, Chapter 2) for more information about the Mantel-Haenszel exact test. For more technical details and information about the development of this test see also (Drake, 1993; Hauck, 1979; Somes, 1986; Mantel, 1963; Breslow, 1996).

	Strata		
	1	2	3
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 2: In-Person Treatment and Response in 3 Strata

a bigger and bigger population. In this case the hypergeometric random variable from the 2×2 table acts very similarly to a binomial random variable — which we know rapidly approximates a Normal as the number of trials in $\binom{\#\text{ trials}}{\#\text{ successes}}$ grows large. Thus, for small S but large n_s, $t(\mathbf{Z}, \mathbf{r})$ can be thought of as a sum of a small number of Normally distributed random variables, and is thus approximately Normal.

If the number of strata grows large, each one generating a small random integer (i.e. n_s remains fixed), then the Normal approximation appeals to the fact that the test statistic is a sum of independent integers. Since the sum of independent integers approximates a Normal distribution, the null distribution of the MH test statistic also approximates a Normal distribution as the number of strata increases.

Under the null hypothesis, the mean and variance of $t(\mathbf{Z}, \mathbf{r})$ can be determined precisely. Its standardization is

$$\tilde{t}(\mathbf{Z}, \mathbf{r}) \approx \frac{\sum \left(t(\mathbf{Z}_s, \mathbf{r}_s) - E(t(\mathbf{Z}_s, \mathbf{r}_s)) \right)}{\sqrt{\text{var}(\sum t(\mathbf{Z}_s, \mathbf{r}_s))}}, \tag{6}$$

which follows the standard Normal distribution (at least approximately). In sum, the MH test has an exact probability distribution representing what "no effect" means. But, if S and/or n_s is relatively large we can use an asymptotic approximation to speed computation — although we are not required to use this asymptotic approximation, and can always check it with the exact calculation if we desire.

4 The New Haven 1998 Vote Turnout Experiment

Just before the 1998 Congressional election, Gerber and Green fielded a GOTV experiment in New Haven. They report on the results of this experiment in (Gerber and Green, 2000). In this experiment they tested a variety of 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 3 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.

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

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

by a canvasser, and 7258 people were assigned to be called on the phone.[4]

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 $\pm 2 \times 2.6 = 5.2$), and phone calls decreased turnout by around 5 percentage points ($\pm 2 \times 2.7 = 5.4$). 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 (forthcoming) 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) 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 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 move from a simple application of randomization inference using only a simple stratification to adjust for slight imbalance in the in-person treatment to a more difficult application requiring optimal full matching.

[4]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.

4.1 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. Just as we assessed the independence of treatment and response in Table 1, we can assess the independence of treatment and covariates 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.[5] 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 ($p = .012$) . Notice that using a logistic regression here is consistent with our desire to restrict our probability model to $Z|X$ since the binomial likelihood function here is only specified in terms of $Z|X$.

We then used the MH test, to gauge the differences between treated and controls on each covariate in the dataset before and after holding constant other potential confounds.[6] 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.[7] In this case we found 5 covariates with p-values less than .05. Table 4 shows the standardized biases for these covariates and p-values associated with a Fisher test. The standardized bias for a given variable x is:

$$\frac{\bar{x}_t - \bar{x}_c}{\text{pooled standard deviation}} \tag{7}$$

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

[5]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).

[6] 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.

[7]We show all of the covariates in Table 7.

treated and 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).

	Standardized Bias	$p(<\|t(Z,r)\|)$
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 4: 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 stood around watching 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.

We started again with an analysis of deviance test identical to that reported on page 10, 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 MH test to examine exactly 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 5 shows the largest differences are not strongly distinguishable from zero by the MH test at $\alpha = .05$.

Table 6 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 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

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

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 6: 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, if with different data.) Our propensity score predicts assignment to in-person canvassing using all of the relevant covariates listed in footnote 6 on page 10; 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.[8]

Table 7 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 4 and 6 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.[9]

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

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

The three-stratum subclassification with which we emerge with is shown in Table 8. 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.

large in absolute 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.

	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 7: Balance for Three Strata in the In-Person Condition

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

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

Focusing on effects of telephone solicitations, rather than in-person appeals, Imai (forthcoming) analyzes the Vote '98 campaign by matching treated persons, subjects who were assigned to treatment and then received it, to 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 is arguably 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.

Now 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: were this coded as values of an observed variable, it would be missing 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 Barry (1996); Rosenstone and Hansen (1993); Highton and Wolfinger (2001); Wolfinger and Rosenstone (1980), among many others — imbalance on age is particularly troubling.

More broadly, tests like those of section 4.1 showed that not only are compliers systematically older than non-compliers, but they tend to live in different neighborhoods and have different past voting

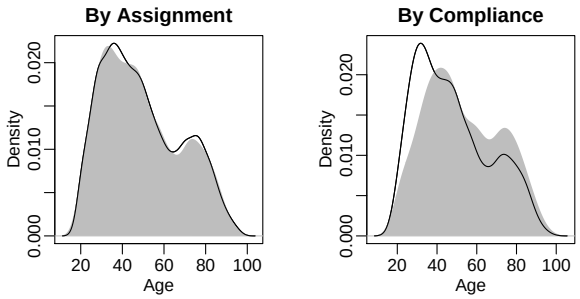

Figure 1: Age Distributions by In-Person Treatment Assignment versus Compliance with In-Person Treatment

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

behavior. Table 9 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 (forthcoming) 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.

4.3 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 receives the treatment compliance. In an instrumental variables analysis, 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) 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), 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.

Seen from within randomization inference, in fact, an instrumental-variables analysis differs scarcely at all from analyses without an IV. Section 3.1's analysis of Adams and Smith's (1980) experiment becomes an IV analysis once it takes into account the fact that since only 950 of the 1,325 individuals assigned to treatment were actually contacted, and since only 310 of them eventually voted, the treatment can have caused, at the very upper limit, no more than 310 votes. This is no impediment to the test of the hypothesis of no effect that we carried out in § 3.1, as the null hypothesis alleged that treatment caused no change in voting and the alternative said only that the telephone calls affected some subjects' voting. Nor will it impede the assessment of attributable effects more broadly, as we shall see.

To say that IV analyses assume ignorability somewhat overstates their requirements. As seen in § 4.1, assignment to in-person canvassing appears, strictly speaking, not to have been ignorable; so the IV approach would seem not apply. Yet it does, if it is granted that our three-level stratification makes treatment assignment ignorable conditionally upon it. Again, randomization inference accommodates the stratification naturally. 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 rather than Fisher's test (see § 3.2). 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 8. (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 $p = 0.00037$. 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.

4.4 Summary

We have shown how one can test a strict null hypothesis in both a simple 2×2 table using the Fisher exact test and in a more general way using the Mantel-Haenszel test. In executing these tests we made no more assumptions than were needed to believe in our basic model of causality defined for individual people in these studies. Further, we were protected from the inferential errors possible with 2SLS as explained by Imbens and Rosenbaum (2005).

These particular benefits of randomization inference is well known. 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?

5 Attributable Effects

In order to summarize the association between a binary explanatory variable Z and a binary outcome Y, it is both common and standard to posit two parameters, $p_1 = \Pr(Y|Z = 1)$ and $p_0 = \Pr(Y|Z = 0)$, and to use data to estimate a comparison of them: perhaps their difference, $p_1 - p_0$, or perhaps the log-odds ratio, $\log[(p_1/(1-p_1))/(p_0/(1-p_0))]$. If covariates are present, further parametrization will be required to define a conditional estimand, for instance $p_1(\mathbf{x}) - p_0(\mathbf{x})$ or $\log[(p_1(\mathbf{x})/(1-p_1(\mathbf{x})))/(p_0(\mathbf{x})/(1-p_0(\mathbf{x})))]$. This parametric structure sits uneasily with that of the Neyman-Rubin model, which at the unit level assigns non-random potential responses but not probabilities to respond one way or another. The two structures can be reconciled with some effort (Holland and Rubin, 1989), but it is simpler to chose one of the two and reject the other. By choosing Neyman's and Rubin's structure and abandoning comparisons of p_1 to p_0, one is led to attributable effects (Rosenbaum, 2001).

The effect attributable to treatment is simply the sum of treatment effects among treated subjects,

$$\sum_i Z_i \tau_i = \sum_{i:Z_i=1} \tau_i \equiv \sum_{i:Z_i=1} (r_{ti} - r_{ci}).$$

The attributable effect is never directly observed, since r_{ti} and r_{ci} are never observed jointly; but we are committed to its existence once we commit to the Neyman-Rubin model. In the strict sense of mathematical statistics, it is not a parameter, since its value is partly determined by $\mathbf{Z}$ and thus varies from sample to sample; in this it differs from "attributable risk" and "excess risk" in epidemiology (Walter, 1976). Still, common strategies for inference about statistical parameters are applicable to inference about attributable effects (Rosenbaum, 2002*b*). The following considerations recommend attributable effects.

Attributable effects adapt the counterfactual view of causality to binary dependent variables without additional complications. When potential outcomes are continuous, the Neyman-Rubin approach defines treatment effects as differences in the response a person would exhibit were he treated, r_{ti}, and the response he would have shown were he not treated, r_{ci}. In the interest of simplicity, these differences are often supposed to be constant across all people: $r_{ti} = r_{ci} + \tau$, where i varies from person to person, but the treatment effect τ does not. To take a trivial example, let "treatment" be asking a subject to hold a ten-pound weight, and let the response be his weight as measured by a scale. Then $\tau = 10$ pounds. With a binary dependent variable, r_{ti} and r_{ci} can only be 0 or 1, not anything in between. If treatment has no effect, so that $\mathbf{r}_t = \mathbf{r}_c$, or if yes-responses are observed only in the presence of treatment but never otherwise, so that $\mathbf{r}_c \equiv 0$ and $\mathbf{r}_t \equiv 1$, then the constant effect model is appropriate; but for every intermediate possibility, it necessarily fails. Once cannot have $\tau = .1$. One can, however, have a situation in which a tenth of all yes-responses in the treated group are attributable to treatment, while the remaining 90% of them would have occurred even in

the absence of treatment. Attributable effects relax the constant effects model, yet they introduce no additional structure.

Attributable effects pertain to subjects studied, not to hypothetical superpopulations. To assert that some number of votes can be attributed to a given GOTV campaign is to say something narrower than that the intervention increased the probability of voting by Δp, the quotient of the same number of votes and the total number of voters contacted. The assertion about probabilities of voting describes a superpopulation of voters that might have been contacted, alleging that a fraction Δp of them would vote if intervened upon but not otherwise. It would require, therefore, that we hold the circumstances in which the intervention was studied to be precisely representative of those in which it might apply to the superpopulation — or that we imagine a hypothetical superpopulation, figuratively constructed for the express purpose of giving the realized sample a population to represent. In contrast, to attribute the corresponding number of votes to treatment is to make a statement only about the sample at hand — indeed, only about the subset of the sample that happened to receive the treatment.

As compared to alternative estimands, attributable effects impose fewer incidental assumptions. Models for association between an outcome and an explanatory variable bring with them mathematical structure, as noted at the beginning of this section. At a minimum, likelihood-based approaches introduce latent variables $\Pr(Y = 1|Z = 1)$ and $\Pr(Y = 1|Z = 0)$, and commonly an entire latent distribution, that of $Y|Z, X$; this in turn introduces a link function, to translate the linear predictor to the probability scale, and a functional form for the regression of Y on Z and X. In the framework of attributable effects, each person is assumed to have one latent variable δ_i which is either 0 if they responded because of the treatment and 1 if their response did not occur because of the treatment. Probability only enters into our story as we combine these individual level δ_is over subpopulations.

Attributable effects allow us to use matched and stratified data. When random assignment is weaker than we'd like, we can gently re-balance the data by matching or stratification. Although such problems of imbalance can also be solved by regression if the functional form of the imbalance is known exactly, or by instrumental variables, if random assignment provides a strong enough instrument, matching and/or stratification provide ultra-simple ways to do this, too — and by allowing for checks of balance make it easier on the analyst to do this adjustment.

To speak of the number of additional people voting due to treatment is more intuitive to non-technical audiences than predicted probabilities or coefficients. For example, in their book meant for non-technical audiences Green and Gerber (2004) speak directly to concerns about the cost of an additional vote, and the concerns of practical campaigns interested in turning out the vote. Consider this early passage:

> How many votes would you realistically expect to generate as a result of [a variety of treatments]? By the time you finish this book you will understand why the answer is approximately 200 (p. 22).

The assumptions required to estimate attributable effects are the same as those for testing the null of no treatment effects. One needs a probability distribution for $\mathbf{Z}$; within strata, this distribution must be blind to potential responses (ignorability); SUTVA must hold; and assignment to receive the intervention is assumed to affect outcomes only via its influence on whether the intervention occurs. Although not strictly necessary, we add the plausible assumption that these interventions either encouraged voting or did not affect it, but never prevented voting by someone who without the intervention would have voted[10].

5.1 Definition and Explanation of an Attributable Effect

We want to estimate the number of voters who would not have voted were it not for their exposure to the treatment: call this number A. Recall that the treatment effect for person i is $\tau_i = r_{ti} - r_{ci}$, the difference in that person's potential responses. We've also made an assumption that $r_{ti} \geq r_{ci}$ and that any response to treatment (potential or observed) must be either a 0 or a 1. Therefore our estimand A is the same as $\sum_{i=1}^{n} Z_i \tau_i$, the sum of the treatment effects among the treated. This A is no more available to direct observation than the individual effects τ_i. How can we estimate A if we can't observe it?

What can we observe? We can see R_i, the observed response for person i. Written in terms of potential responses, $R_i = Z_i r_{ti} + (1 - Z_i) r_{ci}$. The number of positive responses among treated subjects, T, is also observed, and we can also write this in terms of first observed responses and second potential responses: $T = \sum_i^n Z_i R_i = \sum_i^n Z_i r_{ti}$, since the responses among treated people are a reflection of how they would respond under treatment, not under control.

[10] This last assumption will be relaxed somewhat for the analysis of effects attributable to the telephone intervention (§ 6, below).

Notice that:

$$\begin{aligned} T - A &= \sum_{i}^{n} Z_i r_{ti} - \sum_{i=1}^{n} Z_i \tau_i \\ &= \sum_{i=1}^{n} Z_i (r_{ti} - \tau_i) \\ &= \sum_{i=1}^{n} Z_i (r_{ti} - (r_{ti} - r_{ci})) \\ &= \sum_{i=1}^{n} Z_i r_{ci} \end{aligned} \tag{8}$$

By subtracting the hypothesized attributable effect from the observed number of events among the treated, we get the number of events that would have been observed among people who received the no treatment. Under the null hypothesis of no treatment effect, this number has a hypergeometric probability distribution. The following tables shows how this works in the 2×2 case.

First consider the simplest case, the hypothesis that $A = 0$. If this null hypothesis is true, then the results of the experiment are as represented in Table 10.

	Voted	Didn't Vote	Total
Treated	$\sum Z_i r_{ci}$	$n_t - \sum Z_i r_{ci}$	n_t
Control	$\sum (1 - Z_i) r_{ci}$	$n_c - \sum (1 - Z_i) r_{ci}$	n_c

Table 10: Potential Responses in the Control Condition

By ignorability, $\sum Z_i r_{ci}$ has a hypergeometric distribution, which was the basis for section 3.1's test of the hypothesis of no effect. The same test can be applied to the hypothesis that $A = 0$.

To mount an analogous test of the hypothesis that A is some other number, one applies the same reasoning to a table that is adjusted, as in equation (A1) to reflect the attribution of A events to treatment. This is represented schematically in Table 11.

	Voted	Didn't Vote	Total
Treated	$\sum Z_i r_{ti} - A = \sum Z_i r_{ci}$	$n_t - [\sum Z_i r_{ti} - A] = n_c - \sum Z_i r_{ci}$	n_t
Control	$\sum (1 - Z_i) r_{ci}$	$n_c - \sum (1 - Z_i) r_{ci}$	n_c

Table 11: Adjusted Table of Observed Responses as Functions of Potential Responses

What this means is that we can test any given hypothesis $H_0 : A = A_0$ by subtracting A_0 from the (Treated, Voted) cell of the observed table of responses, and adding A_0 to the (Treated, Didn't Vote) cell. By testing different values for A we can delineate a confidence interval. That is, the set

of values for A_0 not rejected by a test at the α level is the set of A_0 within the $1-\alpha$ confidence interval. Notice that we didn't make any assumptions about large samples or repeated sampling. The only thing we needed was knowledge about the random assignment.

5.2 Attributable Effects for a 2×2 Table

Let us turn back to the data from Adams and Smith for a moment. Consider the hypothesis that 50 people were moved to vote because of this treatment. If the true number of people who voted because of the treatment were 50, then we could just subtract 50 people from our (Treated,Voted) cell in Table 1 and add it to our (Treated, Didn't Vote) cell to produce a table reflecting independence, as shown in Table 12. Since our table reflects a situation where the treatment and the response ought to be independent, we can specify a distribution of the adjusted responses in the (Treated, Voted) cell just as we did before.

	Vote	No Vote
Treatment	342	983
Control	315	1010

Table 12: Adjusted Responses for Voting by Telephone Treatment Assignment from Adams and Smith (1980)

If our hypothesis about $A_0 = 50$ is correct,then this new table reflects independence, and therefore the new test statistic is 392-50=342. In this case, the probability of observing a value of 342 or greater if the treatment and turnout were independent is .11 (using the Normal approximation), and .12 (using the hypergeometric distribution). This suggests that it is plausible that 50 people voted because of the treatment.

We can create a Hodges-Lehmann point-estimate (the value of A for which the null-hypothesis is most plausible) and a 95% confidence interval by doing this test for a whole range of values. We tested each null hypothesis of $A_0 = 0$ to $A_0 = 200$ using the Normal approximation of the Fisher exact test. The following plot shows how the p-values for this test increase up to $A_0 = 77$. The horizontal line is drawn at the p-value of .05, and the vertical lines are drawn at the values of A for which the p-value is closest to .05. In this case, the 95% acceptance interval goes from 34 to 118 votes due to the treatment (i.e. between about 5% and about 17% of the voters in this study are estimated to have done so because of the treatment). We also assessed the accuracy of the approximation using the exact test around the boundaries of the confidence interval and near the Hodges-Lehmann estimate. The exact test accepts hypotheses of 33 and 119 — thereby widening the confidence interval very slightly. The most probable hypothesis in both cases was 77 votes attributable to treatment.

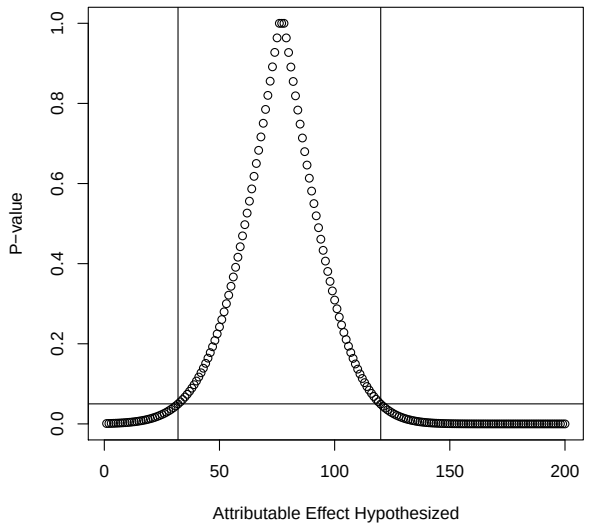

Figure 2: Attributable effects for Adams and Smith (1980).

5.3 Attributable Effects for an $S \times 2 \times 2$ table: Estimating the Number of People who Voted Because of In Person Treatment

Since our test of the null hypothesis of no treatment effects generalized easily from one table to S, estimation of attributable effects across multiple strata is also possible. In this case our potential response to treatment of person i in strata s is $\tau_{si} = r_{Tsi} - r_{Csi}$, and $A = \sum_{s=1}^{S} \sum_{i}^{n_s} Z_{si}\tau_{si}$. However, the introduction of multiple strata does add some conceptual and computational complexity.

We tend to write the null hypothesis about A in terms of a given set of τ_i's, such that we would write our null hypothesis as $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ and then $A_0 = \sum Z_i \tau_{0i}$. Now, remember that any given vector of τs is a record of the way an experiment might have turned out. Thus, there are a lot of these hypothetical vectors. Say we think that 20 people voted because of the treatment (i.e. they would not have voted were it not for the treatment). This means that we'd have 20 1s in our vector $\boldsymbol{\tau}$ and a bunch of 0s. Say we had 100 subjects in our experiment. There are then $\binom{100}{20} = 5.36 \times 10^{20}$ different vectors of length 100 with 20 1s. In fact, not all $\boldsymbol{\tau}$ are logically compatible with our data. For example, if we have a situation where a control subject has a 1 in their particular τ we know this is wrong — no control subject can vote because of the treatment since they didn't get the treatment. This is how our instrumental variable is represented in our test. We also have assumed that $r_{ti} \geq r_{ci}$, so we can't allow a treated subject who has $R_i = 0$ (i.e. observed not to vote) have a $\tau_i = 1$. Thus, even before we do a hypothesis test, we reject many impossible hypotheses represented by different $\boldsymbol{\tau}$s — each of which have 20 1s and, in this case, 80 0s.

For the in-person treatment in the Vote 98 experiment, we can begin to think about compatible τ's by finding out how many people answered their doors in response to a GOTV canvasser who

also voted. These are the only people i for whom it is possible that τ_i might be 1. Across the three strata that balanced the in-person treatment, no more than 238 people could have turned out to vote because of treatment in the 2 Person Household, Low Probability of Treatment group, no more than 280 could have turned out to vote because of treatment in the 2 Person Household, High Probability of Treatment group, and no more than 435 people in 1 Person households could have turned out to vote because of treatment. Thus, the maximum number of people who could have turned out to vote because of the treatment is 953 (238+280+435 people who answered their doors and voted). This means that $\boldsymbol{\tau}$ can have no more than 953 1s. The next question is about exactly which 953 positions in the 29,380 $\times$ 1 vector of $\boldsymbol{\tau}$ should get a 1 and which should get a 0 to represent exactly which hypothesis about A.

Say we hypothesize that $A = 10$ — meaning that 10 people would not have voted were it not for the treatment. There are 66 different ways to have 10 votes occur across the three strata. If $A = 4$, there are 15 different ways. For $A = 200$, there are 20301.

We can think of any given null hypothesis about an attributable effect, $H : A = A_0$, as implying a test of the set of all of the specified patterns of potential responses (SPPRs) that are compatible with that hypothesis. We could test every SPPR attributing A_0 votes to treatment, and if all of them lead us to reject the null, we would say that the hypothesis $H : A = A_0$ is rejected. If any of them is greater than our predetermined α, however, we would accept the null. This procedure ensures that our confidence interval is as wide as necessary. For an example, consider $A = 5$. This leads to 21 different configurations of $\boldsymbol{\tau}_0$. Testing each of these different configurations leads to p-values ranging from .00051 to .00054. We reject the hypothesis that $A = 5$.

In this case with only 3 strata, we need to test all of the possible configurations of $\boldsymbol{\tau}$ 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. 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.

Although A_0 could be any integer value from 0 to 935, a preliminary coarse grid search found that values over 300 were completely implausible. Figure 3 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*d*, ch. 2).

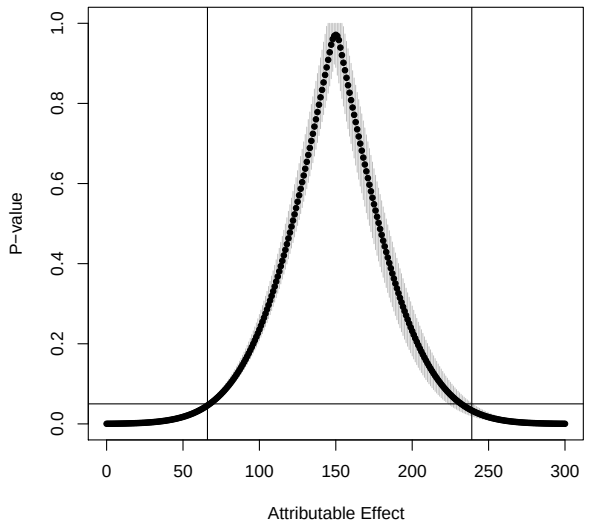

Figure 3: Attributable Effects for In-Person Treatment

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 boundaries of a 95% confidence interval.

5.4 Summary

After showing how one can test a null hypothesis using an instrumental variable, we explained how to invert this test to produce a confidence interval, and how to obtain a Hodges-Lehmann point estimate from this confidence interval.

Our estimand was tied directly to our substantive concerns (i.e. the number of people who would not have voted were it not for the treatment), and we made no extraneous assumptions about functional form or distributions or latent variables in order to produce our estimates.

Our job was made easier because the random assignment of treatment to individuals worked rather well in the case of in-person treatment. In fact, 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.

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

6.1 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 (forthcoming) 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 (forthcoming).

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 3 on page 9 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 13 shows the results of this analysis of deviance. Clearly, assignment to phone treatment is not independent of covariates.

```
Model 1: PHNTRT1 ~ interaction(MAILINGS, PERSNGRP, CONTACT)
Model 2: PHNTRT1 ~ MAJORPTY + PERSONS + AGE + I(AGE^2) + votein96 +
        ward + interaction(MAILINGS, PERSNGRP, CONTACT)
```

	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
Model 1	29368	26469.50				
Model 2	29334	26383.47	34	86.03	2.53	0.0000

Table 13: Assessing the Ignorability of Assignment to Phone Treatment

The same could have been said of assignment to personal contact: we showed on page 10 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.

6.2 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 constant c is then the half-width of the caliper. The caliper matching literature offers some guidance on the selection of c, recommending that it be selected with attention to the

caliper	percent bias reduction if ...		
half-width	$2\sigma_t = \sigma_c$	$\sigma_t = \sigma_c$	$\sigma_t = 2\sigma_c$
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 14: 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.

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 14, 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.[11] We used all covariates that

[11] 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.

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 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 14, 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.

6.3 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. 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 § 6.4). 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 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 4. 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 § 6.1: in the event that the suggestion is correct, then the analysis is prone to bias unless it adjusts for assignment scores.

The pattern shown in Figure 4 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 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 4 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 4. 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

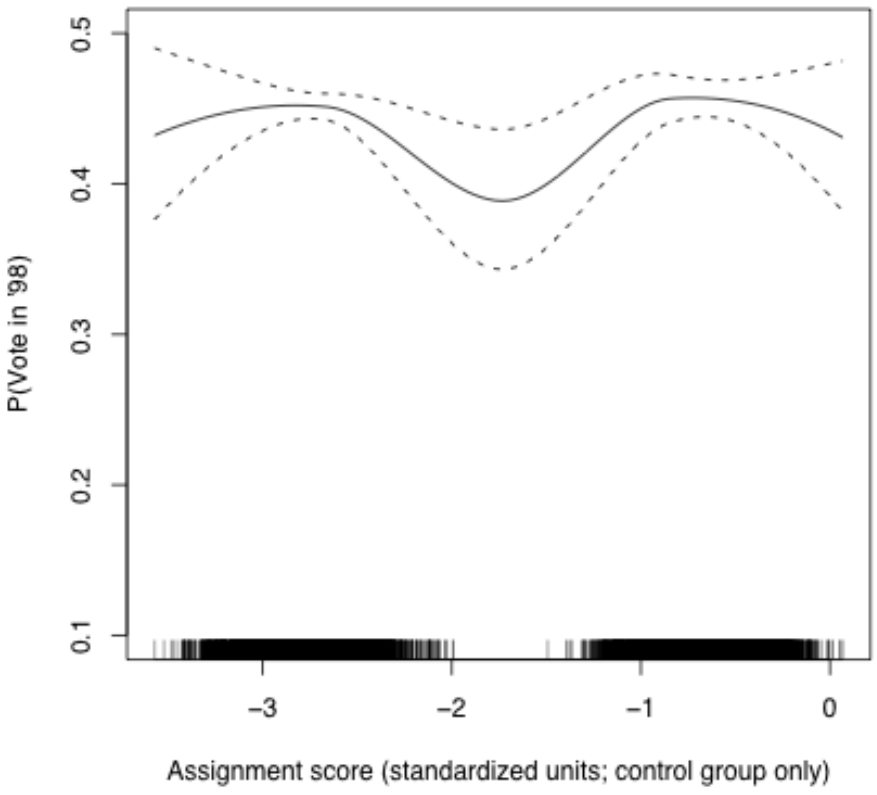

Figure 4: Probability of voting and propensity to be assigned to treatment. A local regression of response on fitted assignment score, control group only, with the distribution of assignment scores among the controls indicated by the marks at bottom.

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 5 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.

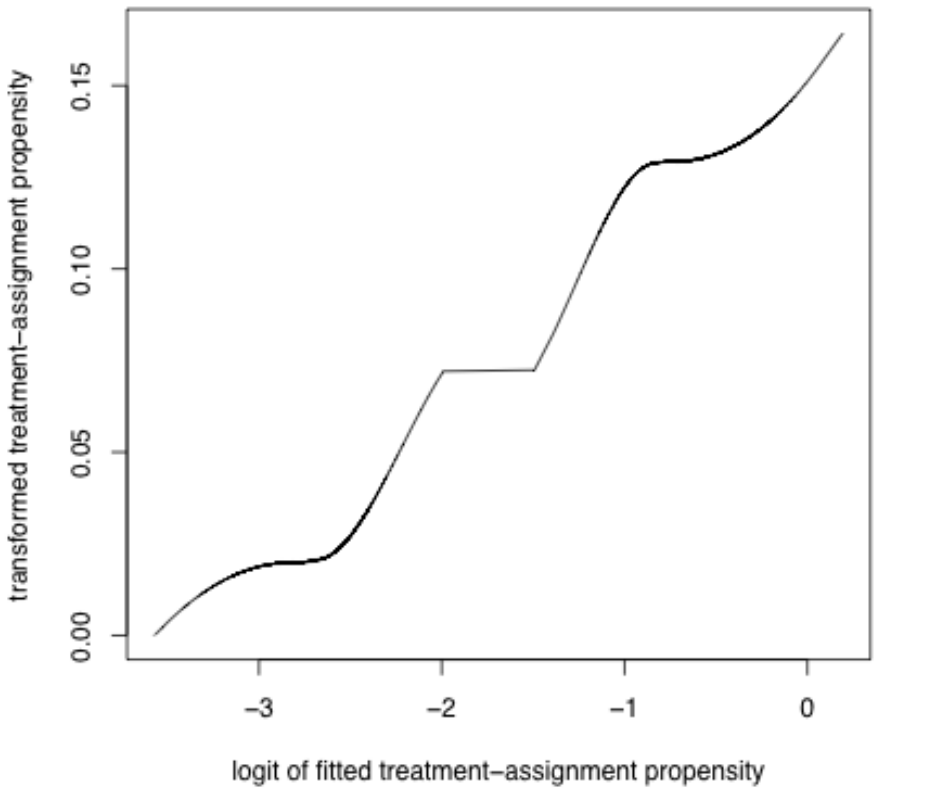

Figure 5: Response-tailored assignment score as a transformation of the simple propensity score. If the curve shown in Figure 4 faithfully reflects associations between the (logit of the) probability of assignment to treatment and the likelihood of voting in the absence of treatment, then bias will be minimized if comparisons are restricted to units similar on the y-coordinate of this figure, the response-tailored assignment score.

6.4 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 actual 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 ith row and jth 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 6.2, 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. Then we generated a full matching by the command

```
fm1 <- fullmatch(discrepancy.matrix)
```

The result, `fm1`, is stored as a certain type of list, a so-called factor object, containing a unique identifier for the matched set into which a subject has been placed for each treated or control subject represented in `discrepancy.matrix`.

We have somewhat misrepresented the command that produced this poststratification. The Gerber and Green data set is so large that a matrix with a row for each of the 6300 treated units and a column for each of the 23000 controls is too large to be stored in memory, much less fed in one chunk to `fullmatch()`. For this reason, and 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.[12] The subjects fell into twelve strata, and were matched (subject to caliper constraints and to the Mahalanobis distance) only within these strata.[13] This could in principle be achieved by applying the function `fullmatch`

[12] 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.

[13] 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.

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.

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 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 (§ 6.4) 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 15. Whereas the unrestricted matching produced at least one matched set with 84 controls, here, no matched set contains more than 23 controls.

	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 15: 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()`.

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 .99 (an improvement of 1%) relative to matches that pair each treated subject with one control.[14] 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.[15] Each of these biases was sharply reduced by matching, three to insignificance and one from strong ($p = 10^{-8}$)

[14]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 15 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.

[15]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.

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 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.

6.5 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 (§ 5.3) 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 SPPR (see p. 24) to the effect that precisely these A_0 subjects voted but would not have voted in the absence of treatment; for each SPPR separately, determine the full set of potential responses in the absence of treatment that is implied by the SPPR 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 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 (See equation 6).) 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*c*) 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.

The process of generating these confidence intervals produces specific patterns of attribution, patterns which can be illuminating to examine. Consider the upper 95% confidence limit of 84 votes attributable to treatment, for example. In testing it, we found the largest and the smallest z-statistics associated with SPPRs attributing 84 votes, and in the process we determined the SPPRs generating them. Because 84 is the upper limit of the confidence region, most of these z-statistics fall outside the ± 1.96 acceptance region; but at least one of them, one of those we calculated, falls narrowly within it.

This SPPR represents the scenario most favorable to telephone calls as a GOTV tactic. As noted in § 6.4, our matching method took some pains to compare subjects of similar ages. What does the scenario most favorable to telephone solicitation say about its effect by age? The quartiles of the experiment's age distribution span from 18 to 34, 35 to 46, 47 to 64, and from 64 to 97, and the likelihood that a subject would take a call from a paid GOTV soliciter increased with age (from 23% to 27%, to 36%, on to 54% in the highest age group). When one restricts consideration to just those who answered calls, the calls appear to have been more effective with the old than with the young. In the scenario most favorable to telephone solicitations, less than 1% of 18–34 year old subjects contacted were induced by the contact to vote, while four or five percent of older voters were induced to vote. (The proportions were .008, .052, .042, and .043 in the four respective age

groups.) For a confidence level of 2/3, rather than .95, the most favorable acceptable SPPR suggests still more strongly that if telephone solicitation to vote was beneficial, then it was beneficial among subjects other than the very youngest; in this scenario, no votes among 18 to 34 year-old voters can be attributed to treatment, while among voters 35 years or older every hundred calls generated one or two votes.

Independently of age, *all* votes attributed by either of these (2/3 and 95%-confidence) most favorable scenarios went to subjects who had voted in the previous election. For those not in the habit of voting, the data appear to suggest, a call from a paid vote soliciter has little hope of success, although calls placed to people who had voted in the last election may have successfully prevented them from dropping the habit.[16]

7 Randomization Inference Encourages Simple and Confident Data Analysis

In this paper we have tried to present a new perspective on statistical inference that allows for principled decisions to be made about data analysis in simple and confident steps. We say "simple" because the analysis occurs in parts — and not all parts are necessary for all causal models. For example, we showed how to estimate attributable effects where ignorability was simple (akin to laboratory experiments) and where ignorability required more work (akin toobservational studies). We say "confident" because the crucial assumptions rest on a piece of the researchdesign about which researchers often know most. For example, in this paper our most important causal variable was a binary valued experimental manipulation. While this simple setup made exposition easy for us, it is not necessary (See Imbens and Rosenbaum, 2005, for an example analyzing the causal effect of years of education on earnings). A researcher may also be most confident in her interpretations of results from a randomization inference analysis because the estimands tend to reflect the theoretical model very directly. Finally, randomization inference provides both simplicity and confidence because it does not require the substantive researcher make decisions about, or interpret her results in terms of, entities outside of her particular sample and her particular research design. In so doing, it clearly delineates the important scientific process of generalization of conclusions as something separate from the statistical assessment of causal effects.

7.1 Assumptions

Throughout this paper we have been very careful to highlight the moments where we made assumptions. Let us list them all here (perhaps readers of this draft will help us by pointing out assumptions that we have overlooked):

- We assumed that our treatment variable Z can be treated as if it were generated at random. This seemed reasonable since, in fact, random assignment was attempted by both Adams and Smith and Gerber and Green.

[16] In their study of Oregon's switch to mail-in voting, Berinsky, Burns and Traugott (2001) also found that an attempt to decrease the costs of voting mainly operated to prevent habitual voters (who tend to be older, wealthier, and more educated) from skipping elections rather than bringing new voters into the electorate.

- At times we also assumed that the number of strata or the number of observations within strata were large enough that the Central Limit Theorem characterizes the distribution of our test statistic (under a null hypothesis of independence of response and treatment). This assumption is fully testable, and could be dispensed with if we were concerned about it.
- In our assessment of the telephone treatment, we also assumed that there were sufficiently many matched sets for asymptotic separability. (This is a mild assumption; we had some 6,000 matched sets, whereas Gastwirth, Krieger and Rosenbaum (2000) found asymptotic separability to hold in examples with as few as 13 matched sets.)
- We assumed that the lack of bias on observed covariates achieved by our matching allows us to use random assignment to treatment as an instrumental variable.[17]
- We made the stable unit treatment value assumption (SUTVA).
- We assumed the exclusion restriction, that treatment assignment can have affected the response only via the administration of treatment.

We avoided making other common assumptions. For example, we made no claims about the distribution of our response variable or the parameters that might govern its distribution, or about the functional form of the relationship between treatment, covariates, and response. Our interpretations of p-values were in terms of repeated administrations of the same research design to the same group of individuals (not in terms of a population of represented by these individuals or in terms of changes in our own willingness to bet on our hypotheses).

Above and beyond randomization inference for instrumental variables and attributable effects, we also demonstrated several techniques for strengthening the case that causal effects may be assessed without fear of confounding from observed covariates. The simplest technique was the good old contingency table. However, we also showed how full matching can strengthen the case for ignorability without making assumptions about the functional form of the relationship between treatment, covariates, and response. The matching also allowed us to suggest some ways that attributing effects to treatment might shed light on the mechanisms behind the experimental manipulations being tested.

Why is it a virtue that we made few assumptions and that we were so explicit about it? In this particular paper it was a virtue to show that it is possible to do this at all. For example, our substantive conclusions are quite close to those that emerged from the 2SLS analysis of Gerber and Green (2000), but we feel ours rest on firmer statistical ground. In more applied work, we want to emphasize that assumptions are choices that scholars may, or may not, desire to make. That is, we believe that there is plenty of room in political science for likelihood functions, posterior distributions, and linear additive functional forms. But as our paper has demonstrated, such

[17] We considered a sensitivity analysis to check this but decided to postpone that work in the interests of sparing our readers another 10 pages of text!

assumptions are not always necessary. Analysts should be free to choose the inferential framework that best matches their substantive concerns; but we all bear the burden of justifying our choices.

Appendix A Deriving the Hypergeometric Distribution

Let X and Y be independent Binomial RVs with the same probability of success p, but different numbers of trials, say n and m. For example X could be assignment to treatment and Y could record whether a person voted.

We can depict the relationship between these two variables with the following table:

	Voted	Didn't Vote	
Treated	x	$n-x$	n
Control	$k-x$	$m-k-x$	m
	k	$m+n-k$	$m+n$

The probability of X being some number x given that $X+Y=k$ can be written using the rules of probability as:

$$Pr(X=x|X+Y=k) = \frac{Pr(X=x, X+Y=k)}{Pr(X+Y=k)} \tag{A1}$$

$$= \frac{Pr(X=x, x+Y=k)}{Pr(x+Y=k)} \tag{A2}$$

$$= \frac{Pr(X=x, Y=k-x)}{Pr(Y=k-x)} \tag{A3}$$

Now, let's make use of the fact that we've assumed that X and Y are independent Binomial RVs with same p.

$$Pr(X=x|X+Y=k) = \frac{Pr(X=x, Y=k-x)}{Pr(Y=k-x)} \tag{A4}$$

$$= \frac{\binom{n}{x}p^x(1-p)^{n-x} \cdot \binom{m}{k-x}p^{k-x}(1-p)^{m-(k-x)}}{\binom{m+n}{k}(1-p)^{m+n-k}p^k} \tag{A5}$$

$$= \frac{p^k(1-p)^{m+n-k}\binom{n}{x}\binom{m}{k-x}}{p^k(1-p)^{m+n-k}\binom{m+n}{k}} \tag{A6}$$

$$= \frac{\binom{n}{x}\binom{m}{k-x}}{\binom{m+n}{k}} \tag{A7}$$

Equation A7 is the hypergeometric probability mass function.

Here is another way to think about where the hypergeometric distribution comes from (following Hodges and Lehmann (1964, pages 154-158).)

Say we have 7 items in the population, we are taking samples of size 3 from this population, and this population contains 4 voters and 3 non-voters. We want to know the probability of observing, say, 2 voters out of any given sample of 3 people from this population.

There are $\binom{7}{3} = 35$ ways to take 3 items from 7. But, what proportion of these 35 outcomes have exactly 2 voters? One way to find this out is just to generate the samples. If persons numbers 4,5,6,7 are voters, then you discover that there are 18 sets that have exactly 2 voters in them.

{1,2,3}	{1,2,4}	{1,2,5}
{1,2,6}	{1,2,7}	{1,3,4}
{1,3,5}	{1,3,6}	{1,3,7}
{1,4,5}	{1,4,6}	{1,4,7}
{1,5,6}	{1,5,7}	{1,6,7}
{2,3,4}	{2,3,5}	{2,3,6}
{2,3,7}	{2,4,5}	{2,4,6}
{2,4,7}	{2,5,6}	{2,5,7}
{2,6,7}	{3,4,5}	{3,4,6}
{3,4,7}	{3,5,6}	{3,5,7}
{3,6,7}	{4,5,6}	{4,5,7}
{4,6,7}	{5,6,7}	

The number of ways you can select 2 voters from a population of 4 is $\binom{4}{2} = 6$; and we can get 1 non-voter out of the population of 3 $\binom{7-4}{3-2} = \binom{3}{1} = 3$ ways. So, we can get 2 voters $\binom{4}{2} \cdot \binom{7-4}{3-2} = 6 \cdot 3 = 18$ ways.

If each of the 35 samples were equally likely, then, in 100 draws from from this sample we'd expect to get 2 voters 18 out of 35 draws or 51% of the time. Thinking of probabilty as relative frequency we can say that the probability of observing 2 voters at random in a sample of 3 is:

$$\frac{\binom{4}{2} \cdot \binom{7-4}{3-2}}{\binom{7}{3}} = .51.$$

References

Adams, William C. and Dennis J. Smith. 1980. "Effects of Telephone Canvassing on Turnout and Preferences: A Field Experiment." *Public Opinion Quarterly* 44:389–395.

Agresti, Alan. 2002. *Categorical Data Analysis (Wiley Series in Probability and Statistics).* Wiley-Interscience.

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. "Identification of Causal Effects Using Instrumental Variables (Disc: P456-472)." *Journal of the American Statistical Association* 91:444–455.

Berinsky, Adam J., Nancy Burns and Michael W. Traugott. 2001. "Who Votes by Mail? A Dynamic Model of the Individual-Level Consequences of Voting-by-Mail." *Public Opinion Quarterly* 65:178–197.

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

Breslow, N. E. 1996. "Statistics in Epidemiology: The Case-Control Study." *Journal of the American Statistical Association* 91:14–28.

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.

Drake, Christiana. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect." *Biometrics* 49:1231–1236.

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.

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

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.

Hauck, Walter W. 1979. "The Large Sample Variance of the Mantel-Haenszel Estimator of a Common Odds Ratio." *Biometrics* 35:817–819.

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

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

Ho, Daniel and Kosuke Imai. 2004. "Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election.".

Hodges, Joseph Lawson and Erich L. Lehmann. 1964. *Basic Concepts of Probability and Statistics*. San Francisco: Holden-Day.

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

Holland, Paul W. and Donald B. Rubin. 1989. "Causal inference in retrospective studies." *Evaluation Review* 12(3):203–231.

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

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)* 168:109–126.

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.

Mantel, Nathan. 1963. "Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure." *Journal of the American Statistical Association* 58:690–700.

Nie, Norman, Jane Junn and Kenneth S. Barry. 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* 97:183–192.

Rosenbaum, Paul R. 2002*c*. "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*d*. *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.

Somes, Grant W. 1986. "The Generalized Mantel-Haenszel Statistic." *The American Statistician* 40:106–108.

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

Walter, S. D. 1976. "The estimation and interpretation of attributable risk in health research." *Biometrics* 32:829–849.

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