Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign: An Application of Randomization Inference Using Full Matching*

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

July 18, 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 existence of random assignment of Z in a field experiment as well as by 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 study assigned households to treatment and desired to estimate the effect of treatment on the individuals nested within the households. 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 — while taking into account the

*Authors listed in alphabetical order. Early versions of this paper were presented at the Department of Political Science at the University of Illinois, July 2004, and at meetings of the Royal Statistical Society, September 2004, of the Midwestern Political Science Association, April 2005, and of the International Statistical Institute, April 2005. We are grateful to participants of those meetings for their helpful comments.

non-independence of observations within households, non-random compliance, and missing responses. 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).

1 Introduction

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 New Haven 1998 Vote Turnout Experiment

A GOTV campaign in New Haven, the Vote '98 campaign, reached out to voters in three ways that are common to such campaigns: direct mail, telephone calls, and appeals delivered in person (Gerber and Green 2000). Vote '98 differed from most such campaigns, however, in that it used random assignment to determine by which of these means, if any, the campaigners would attempt to contact each voter. As a field experiment, Vote '98 was quite ambitious in scope, randomizing all three forms of intervention according to a three-way factorial design. For illustration, treatment assignments are shown in Table 1. The table shows, for example, that 200 people were assigned to receive a phone call and a visit from a canvasser but no mailings, while 2500 people were assigned no in-person visits and no phone calls, but 3 mailings.

Another novelty of the paper has to do with its allowances for the fact that only a fraction, a potentially unrepresentative fraction, of voters slated for personal visits or telephone entreatments could be reached by canvassers; their use of instrumental variable techniques permits them to validly estimate treatment effects despite the voters' "non-compliance" with the treatment to which they were assigned (Angrist et al. 1996a). They 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 ($\pm 2 \times 2.7 = 5.4$).

	0	1	2	3
In Person,Phone	200	400	400	400
In Person,No Phone	2900	600	700	600
No In Person,Phone	800	1600	1600	1600
No In Person,No Phone	11600	2600	2700	2500

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

In a later article, Imai (2005) argues that Vote '98's GOTV interventions can not have been properly assigned, because the subgroups it assigned to treatment and to control conditions were not as similar, in terms of pre-treatment covariates, as randomization should have made them. If this is the case, then the estimation procedure Gerber and Green used is not be valid. Imai's alternative approach, which rejects instrumental variable techniques in favor of matching and propensity scores, reaches a substantively different conclusion from Gerber and Green's — that telephone-based appeals substantially enhanced turnout. This paper revisits both the substantive and methodological debates among Gerber, Green, and Imai. Substantively, we will assess hypotheses about causal effects of GOTV appeals as delivered in person or via the telephone; and methodologically, we update both sides' methods so as to account for the fact that randomization was performed at the household rather than the individual without introducing additional assumptions; and we show that propensity scores and matching are available in combination with instrumental variable techniques.

1.2 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 (stating say, that a given treatment has an effect) 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.[1]

1.3 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 2 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 2: 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. If that is so, then since half of subjects fell in the treatment group, about half, or 354, of all votes by treatment and control group subjects should have been cast by members of the treatment group. A few more than 354 among treatment group voters might be explained by chance, but substantially more casts doubt on the null hypothesis. Since 392 seems a good bit more than 354, one is tempted to conclude that treatment had an effect.

Fisher's exact test formalizes this argument. Let $y_i = 1$ if subject i did turn out to vote, 0 otherwise. For sake of argument, it grants that the outcome y_{ci} subject i would have exhibited had he not been given a reminder call is the same as what was in fact observed whether or not he did receive a reminder call, *i.e.* y_i. These variables are taken as fixed quantities, not random variables. Subject i's treatment assignment, however, is a random variable Z_i, taking the value of 0 or 1 depending on whether i was assigned to treatment. The test measures association between $\mathbf{Z}$ and $\mathbf{y}$, rejecting the hypothesis that treatment was without effect if the association

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

is too large to be due to chance.

In a randomized experiment like Adams and Smith's, treatment is attempted on a simple random sample of eligible subjects. Association between treatment assignment and outcomes is assessed by comparing $\mathbf{Z}^t\mathbf{y}$ — the number of subjects in the treatment group who voted — with its reference distribution under the hypothesis of no effect. If treatment is without an effect, but treatment was assigned to a random sample of size m of a total of n subjects, then $\mathbf{Z}^t\mathbf{y}$ can be expected to be somewhere around m/n times the total number of votes cast by both treatment and control subjects — that is, $\mathrm{E}\mathbf{Z}^t\mathbf{y} = \frac{m}{n}\sum_i y_i$. Other moments, indeed exact probabilities of $\mathbf{Z}^t\mathbf{y}$ taking particular values, can be calculated under assumptions of randomization and of no effect. The Fisherian argument makes formal and precise what is meant when one asks, "could this relationship merely be due to chance?"

The distribution taken by $\mathbf{Z}^t\mathbf{y}$ under randomization and the hypothesis of no effect is known as the *hypergeometric distribution*. 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.

As Fisher originally presented it, his test did not simply extend to assessments of how large a treatment effect might have been (as opposed to whether there was a treatment effect). Recent developments in Statistics fill that gap. We now turn to explaining them.

1.4 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 $\log[(p_1(\mathbf{x})/(1-p_1(\mathbf{x})))/(p_0(\mathbf{x})/(1-p_0(\mathbf{x})))]$. Differences between log-odds may be a fallible or unreliable guide to differences between probabilities, with or without conditioning (Greenland 1987). Additionally, none of these parametric structures sits well with the Neyman-Rubin model, which assigns to unit i a potential response to treatment, y_{ti}, and a response that would obtain in the absence of treatment, y_{ci}, but not probabilities to respond one way or another. (In this framework, for each subject at most one of y_{ti} and y_{ci} are observed, and for statistical purposes unobserved potential responses are treated as missing data. See *e.g.* Holland (1986); Brady and Seawright (2004).) 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,

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

The attributable effect is never directly observed, since y_{ti} and y_{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 2002b). The following considerations recommend attributable effects.

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. In Bowers and Hansen (2005) we showed that 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 a book addressed to 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. This passage typifies this discussion:

> 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.

1.5 Attributable Effects: Some Formalities

We want to estimate the number of voters who would not have voted were it not for their exposure to the treatment. Let $D = 1$ for subjects who received the treatment, that is subjects who (i) fell in the experimental group and (ii) answered the GOTV call or visit from experimenters. (By design of the experiment, $D_i = 1$ only if $Z_i = 1$, but because not everyone answers the phone or the door sometimes $D_i = 0$ even though $Z_i = 1$.) Since the treatment effect for person i is $y_{ti} - y_{ci}$, the difference in that person's potential responses, the attributable effect is $A = \sum_{i=1}^{n} D_i(y_{ti} - y_{ci})$, the sum of the treatment effects among the treated. This A is no more available to direct observation than the individual effects $(y_{ti} - y_{ci})$. How can we estimate A if we can't observe it?

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

Table 3: Treatment Group by Potential Responses

Consider tables with the form of Table 3. Section 1.3 discussed such a table, with $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)^t$ equal to the vector of observed turnout outcomes (1=yes, 0=no) for all subjects of their experiment. More generally, tables of this form result from setting $\tilde{y}_i = y_{ci}$ for subjects for whom y_{ci} was observed, *i.e.* those subjects Adams and Smith either could not reach or did not attempt to reach with their GOTV call, and filling in the rest of $\tilde{\mathbf{y}}$ in such a way as to represent a detailed speculation about the y_{ci} values of remaining subjects. Table 2 from Section 1.3 may also be regarded as a table of this type – one in which the detailed speculation about the unobserved values of y_{ci} is that they are precisely what was observed in the presence of treatment, i.e. $y_i = y_{ti}$. That particular speculation has the helpful property that it is detailed enough to determine each of the four cells of Table 3; this makes it possible to test it via Fisher's method. Fortunately, the hypothesis of no effect is not the only one with this property.

Call any hypothesis that specifies each value of y_{ci}, even those not observed, an *atomic* hypothesis of effect. Only hypotheses $H : \mathbf{y}_c = \tilde{\mathbf{y}}$ with the property that for subjects i that did not receive the treatment ($d_i = 0$), $\tilde{y}_i = y_i$ (the observed response), can be credible, and we restrict attention to these. For simplicity, and because it suits the present application[2], we also restrict attention to atomic hypotheses with the property that for subjects i who did receive the treatment, $\tilde{y}_i \leq y_i$ — treatment can only have increased turnout, not decreased it. Atomic hypotheses with these properties are called *compatible* with observed data (Rosenbaum 2002c). We claim that any such atomic hypothesis with the property that $\sum_i y_i - \tilde{y}_i = a$ induces the same 2×2 table (Table 3), and that the entries in this table can be determined on the basis of patterns of observed data, *i.e.* $\mathbf{z}$ and $\mathbf{y}$, plus the value of a. Because of this, and because the information in this table gives both the observed value of $\mathbf{z}^t\tilde{\mathbf{y}}$ and the probability distribution of $\mathbf{Z}^t\tilde{\mathbf{y}}$ under hypothetical repetitions of the random assignment, Fisher's test can be applied to evaluate such an atomic hypothesis.

To verify that a, in combination with $(\mathbf{y}, \mathbf{z})$, determines the cells of this table, reason as follows. The observed response for person i, y_i, is written in terms of potential responses as $y_i = d_i y_{ti} + (1 - d_i) y_{ci}$. Since we are considering only atomic hypotheses in which $\tilde{y}_i$ and y_i may differ only among subjects i for which $d_i = 1$, and since by design d_i can be one only if also $z_i = 1$, $\sum_i y_i - \tilde{y}_i = \sum_i z_i(y_i - \tilde{y}_i)$. That is, $\sum_i z_i(y_i - \tilde{y}_i) = a$. Let t be the number of positive responses observed among the treatment group, $\sum_i z_i y_i$. Then

$$\begin{aligned} \sum_i z_i \tilde{y}_i &= \sum_i z_i y_i + \sum_i z_i(\tilde{y}_i - y_i) \\ &= t - a \end{aligned} \tag{1}$$

This shows that a and $(\mathbf{y}, \mathbf{z})$ suffice to determine the upper left cell of Table 3. In addition, comparing Tables 3 and 4 reveals that $\sum_i \tilde{y}_i = \sum_i y_i - \sum_i (y_i - \tilde{y}_i) = \sum_i y_i - a$, and $\sum_i (1 - \tilde{y}_i) = a + \sum_i (1 - y_i)$, so that the lower marginal totals are determined by a and $(\mathbf{y}, \mathbf{z})$. Totals in the

[2]This latter restriction is not strictly needed to estimate attributable effects.

right margins are the same for the observed table and for the $\tilde{\mathbf{y}}$-table. Jointly, the four marginal totals determine $\mathcal{L}[\mathbf{Z}^t\tilde{y}]$. Therefore, any atomic hypothesis $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$ with the property that, as the data happened to turn out, $\mathbf{z}^t(\mathbf{y} - \tilde{\mathbf{y}}) = a$, can be assessed in the manner of Section 1.3, by comparing $\mathbf{z}^t\tilde{y}$ to the exact distribution $\mathcal{L}[\mathbf{Z}^t\tilde{y}]$ or a Normal approximation to it.

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

Table 4: Adjusted Table of Observed Responses as Functions of Potential Responses

Call hypotheses that refer to the *number* of positive responses that are attributable to treatment, rather than specifying specific units' responses, *macroscopic*, so as to distinguish them from atomic hypotheses. In the discussion of this section, pertaining to studies with a single stratum, there was little difference between testing macroscopic and atomic hypotheses, since each test of an atomic hypothesis amounted to a test of a macroscopic one. That is, in testing $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$, one also tests any other compatible hypothesis $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}'$ with the property that $\sum_i d_i(y_i - \tilde{y}_i) = \sum_i d_i(y_i - \tilde{y}'_i)$; thus, a test of $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$,where $\sum_i d_i(y_i - \tilde{y}_i) = a$, is in fact a test of the macroscopic, composite hypothesis $H_0 : A = a$, asserting that a positive responses were caused by the treatment. On the other hand, in sections to follow, covering stratified designs, tests of a macroscopic hypothesis will be composed of many atomic ones. In both cases, confidence intervals for the attributable effect are delineated by repeated testing of macroscopic hypotheses about the value of A. In neither case are assumptions about large samples or repeated sampling from a superpopulation required.

1.6 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 2 and add it to our (Treated, Didn't Vote) cell to produce a table reflecting independence, as shown in Table 5. 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 5: Adjusted Responses for Voting by Telephone Treatment Assignment from Adams and Smith (1980)

The new test statistic is $\mathbf{z}^t\tilde{\mathbf{y}} = t - a = 392 - 50 = 342$. If the hypothesis $H_0 : A = 50$ is correct,then this new table reflects independence, and 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 confidence interval by doing this test for a whole range of values. We tested each null hypothesis of $A = 0$ to $A = 200$ using the Normal approximation of the Fisher exact test. The following plot shows p-values for this test. 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). The most probable hypothesis in both cases was 77 votes attributable to treatment.

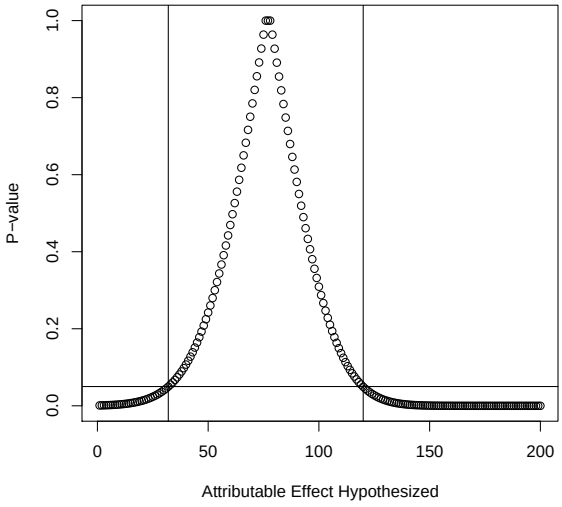

Figure 1: Attributable effects for Adams and Smith (1980).

2 Randomization tests for balance and for the absence of treatment effects

Take any pre-treatment characteristic of study subjects, measured or not, and code it as a numeric statistical variable $\mathbf{q} = (q_1, \ldots, q_n)'$. If m subjects are assigned to treatment at random, then the mean values of q in the treatment and the control groups are random variables even though $q_1, \ldots, q_n$ are not. While these random variables, $m^{-1}\mathbf{Z}'\mathbf{q}$ and $(n-m)^{-1}(\mathbf{1}-\mathbf{Z})'\mathbf{q}$, generally take different values, their expectations — evaluated along hypothetical, repeated random assignments — must coincide. This is the sense in which randomization "balances the distribution of covariates between the treatment and control groups," as it is sometimes put. If the randomization is performed within strata, with varying frequencies of treatment assignment

by stratum s, then the sense in which covariate distributions are balanced is that any weighted average of within-stratum differences of means,

$$\sum_s w_s[m_s^{-1}\mathbf{Z}_s'\mathbf{q}_s - (n_s - m_s)^{-1}(\mathbf{1}_s - \mathbf{Z}_s)'\mathbf{q}_s]$$

($w_s \geq 0$), has expectation zero (where w_s is a stratum specific weight, m_s is the number of subjects assigned to treatment in stratum s, n_s is the total number of subjects in stratum s, and $\mathbf{Z}_s$ indicates assignment to treatment within stratum).

Randomization balances the distribution of covariates (in the sense just explained) between groups assigned to treatment and to control conditions for both individual-level random assignment, in which a simple or stratified random sample of *individuals* is assigned to treatment, and for clustered random assignment, in which a simple or stratified random sample of *clusters of individuals* is assigned to treatment. Yet while their expectations may be the same, overall probability distributions for averages or for stratum-weighted averages in treatment and in control groups may differ enough between designs with and without clustering that to conflate the two, treating a clustered design as if it were not clustered or the reverse, can lead to substantial distortion. Reasoning, perhaps, that their experiment's clusters had been small enough this distortion should be negligible, Gerber and Green (2000) offered an analysis that ignored the clustered nature of the design. But as analyses given by Imai (2005) demonstrate, when evaluated against reference distributions that obtain in the absence of clustering, covariate balance in the Vote '98 experiment seems so improbably poor as to support a forceful rejection of the premise that treatment assignment really was random to begin with. Gerber and Green's (2005) rebuttal maintains that properly accounting for clustering dispels the appearance of an anomaly. The balance obtained in their experiment compares favorably, they argue, to balances obtained in simulated repetitions of the randomization that involved the same clustering as had the actual experiment's random assignment.

Gerber and Green's revised estimation of treatment effects addresses clustering of treatment assignments, but only by introducing new modeling assumptions. Clustering is better addressed, we believe, by abandoning both Gerber and Green's and Imai's analytic techniques in favor of randomization inference in the style of Fisher. Appropriate randomization-based methods account for clustering without additional assumptions, relieve the analyst of the need to conduct simulation experiments in order to approximate randomization distributions, and extend readily from randomization-checking to the estimation of treatment effects.

2.1 Descriptive indicators of covariate balance are insensitive to the presence of clustering

If treatment is assigned to a simple random sample of size m from among n individuals, then the *standardized bias* along covariate q is

$$b_q = \frac{(\mathbf{Z}'\mathbf{q})/m - ((\mathbf{1}-\mathbf{Z})'\mathbf{q})/(n-m)}{s_q}, \tag{2}$$

where s_q is the pooled s.d. of $\mathbf{q}$ in treatment and control groups. This statistic describes the degree of success of randomization at imposing balance on covariate distributions, with standardization by s_q insuring that each covariate's between-group variation is assessed relative to its within-group variation. Since $\mathrm{E}(\mathbf{Z}) = (m/n)\mathbf{1}$, it is straightforward to verify that the numerator of (2) has expectation zero.

If instead treatment were assigned to a simple random sample of $m^{(c)}$ clusters out of $n^{(c)}$ clusters in total, then balance along q might be assessed by applying (2) with $\mathbf{Z}^{(c)}$, a $n^{(c)} \times 1$ indicator of which clusters were assigned to treatment, in place of the $n \times 1$ matrix $\mathbf{Z}$, and with t_q, the cluster totals in q, in place of q. (Because this alternate expression still scales by the individual-level s.d. s_q rather than an s.d. of t_q, it differs slightly from (2) applied to t_q instead of q.) Expressed in terms of individual-level variables, this is

$$b_q^{(c)} = \frac{(\mathbf{Z}'\mathbf{q})/m^{(c)} - ((\mathbf{1}-\mathbf{Z})'\mathbf{q})/(n^{(c)}-m^{(c)})}{s_q}, \tag{3}$$

which differs from (2) only in terms of the denominators used to take averages in each treatment group, m versus $m^{(c)}$ and $n-m$ versus $n^{(c)}-m^{(c)}$. Because each cluster is assigned to treatment with probability $m^{(c)}/n^{(c)}$, the same is true of each individual, and again the numerator of the standardized difference has expectation zero.

Expression (3) gives a standardized bias that correctly takes clustering into account. If clustering were present and it were ignored for the determination of standardized biases, however, then then qualitatively similar patterns of biases among covariates $q^{(1)}, q^{(2)}, \ldots, q^{(k)}$ would result. To see this, observe that with clusters the number of individuals assigned to treatment is a random variable, M, which when treated as if it were the fixed number of individuals to be selected for treatment gives the expression

$$b_{qi} = \frac{(\mathbf{Z}'\mathbf{q})/M - ((\mathbf{1}-\mathbf{Z})'\mathbf{q})/(n-M)}{s_q}. \tag{4}$$

This random variable has $\mathrm{E}(M) = (m^{(c)}/n^{(c)})n$ and $\mathrm{E}(n-M) = [(n^{(c)}-m^{(c)})/n^{(c)}]n$, so if $M \approx \mathrm{E}(M)$ then $b_{qi} \approx (n^{(c)}/n)b_q^{(c)}$. The standardized differences that ignore clustering can be expected to be roughly a common multiple, $n^{(c)}/n$, of the standardized differences that properly take clustering into account.

2.2 Statistical significance of covariate imbalance is sensitive to the presence of clustering

To evaluate the hypothesis that imbalances between treatment and control groups are due only to chance, the numerators of (2), (3), or (4) are compared to their reference distributions. Rearranging so as to simplify the necessary calculations, one has

$$s_q b_q = \frac{n}{m(n-m)}\mathbf{Z}'\mathbf{q} - \frac{1}{n-m}\mathbf{1}'\mathbf{q} \text{ and} \tag{5}$$

$$s_q b_q^{(c)} = \frac{n^{(c)}}{m^{(c)}(n^{(c)}-m^{(c)})}\mathbf{Z}^{(c)\prime}\mathbf{t_q} - \frac{1}{n^{(c)}-m^{(c)}}\mathbf{1}'\mathbf{t_q}. \tag{6}$$

As argued in Section 2.1, these have expectation zero (under simple random sampling of individuals and of clusters, respectively). Their variances follow from the following proposition.

proposition 2.1 *Let x, y be variables defined for units $1, \ldots, n(\geq 2)$, $m \geq 1$ of which are chosen in a simple random sample indicated by $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ (so that $\mathbf{Z}'\mathbf{1} = m$ and $Z_i \in \{0, 1\}$, $i = 1, \ldots, n$). Then*

$$\text{cov}(\mathbf{Z}'\mathbf{x}, \mathbf{Z}'\mathbf{y}) = \frac{m}{n}\frac{n-m}{n-1}\sum_1^n (x_i - \bar{x})(y_i - \bar{y}). \tag{7}$$

Proof of Proposition 2.1. Since $\mathbf{Z}'\bar{x}\mathbf{1} = \frac{m}{n}\bar{x}$ and $\mathbf{Z}'\bar{y}\mathbf{1} = \frac{m}{n}\bar{y}$ do not vary with $\mathbf{Z}$, $\text{cov}(\mathbf{Z}'\bar{x}\mathbf{1}, \mathbf{Z}'\mathbf{y}) = \text{cov}(\mathbf{Z}'(\mathbf{x}-\bar{x}\mathbf{1}), \mathbf{Z}'\bar{y}\mathbf{1}) = 0$. Thus it suffices to show that $\text{cov}(\mathbf{Z}'(\mathbf{x}-\bar{x}\mathbf{1}), \mathbf{Z}'(\mathbf{y}-\bar{y}\mathbf{1}))$ equals the right hand side of (7).

By exchangeability, $\text{var}(Z_i) = \text{var}(Z_1)$ and $\text{cov}(Z_i, Z_j) = \text{cov}(Z_1, Z_2)$ for all $i \neq j$. Thus

$$\begin{aligned}
&\text{cov}(Z_i(x_i - \bar{x}), \sum_j Z_j(y_j - \bar{y})) \\
&= (\text{var}(Z_1) - \text{cov}(Z_1, Z_2))(x_i - \bar{x})(y_i - \bar{y}) + \text{cov}(Z_1, Z_2)(x_i - \bar{x})\underbrace{\sum_{j=1}^n (y_i - \bar{y})}_{=0} \\
&= (\text{var}(Z_1) - \text{cov}(Z_1, Z_2))(x_i - \bar{x})(y_i - \bar{y}).
\end{aligned} \tag{8}$$

Calculating $\text{var}(Z_1) = \frac{m}{n}(1 - \frac{m}{n})$ and $\text{cov}(Z_1, Z_2) = \frac{m}{n}[(m-1)/(n-1) - \frac{m}{n}]$, then summing (8) over $i = 1, \ldots, n$, Proposition 2.1 follows.□

The proposition entails that for individual- and cluster-level randomization, respectively,

$$\text{var}(s_q b_q) = n^{-1}[m(n-m)/n^2]^{-1}\frac{\sum_{i=1}^n (q_i - \bar{q})^2}{n-1}, \text{ and} \tag{9}$$

$$\text{var}(s_q b_q^{(c)}) = n^{(c)-1}[m^{(c)}(n^{(c)} - m^{(c)})/(n^{(c)})^2]^{-1}\frac{\sum_{i=1}^n (t_{qi} - \bar{t}_q)^2}{n^{(c)}-1}. \tag{10}$$

Applied to the same data, these formulas can give quite different results. Under cluster sampling, $M(n-M)/n^2$ is likely to be close to $m^{(c)}(n^{(c)}-m^{(c)})/(n^{(c)})^2$, but the other terms on the left of (10) tend to exceed their counterparts in (9): the first because $n \geq n^{(c)}$; and the third because variability among cluster totals is likely to exceed variability among individuals, especially when the clusters are of varying size. Thus, treating a cluster randomization as if it were individual level randomization tends to understate the variability of standardized biases, inducing rejection of the null hypothesis of random assignment at rates exceeding nominal levels.

2.3 Standardized differences with clustered, stratified treatment assignments

Even (3) and (10), which do account for clustering, reflect designs that are simpler than that of the Vote'98 experiment. That experiment assigned three treatments which, considered one at a time, were given to stratified samples of households, with the probability of a household's assignment to treatment constant within strata but varying between them. For example, telephone GOTV calls were attempted for about 7% of households in the no-mailings, no-personal-canvass stratum, 6% of no-mailings, personal canvass households, 38% of mailing-but-not-personal-canvass households, and 39% of the mailing-plus-canvass households. This is not equivalent to attempting calls to a simple random sample of households.

To accommodate this difference, the standardized biases may be defined for stratified designs as weighted averages of the standardized biases within each stratum. Whatever the weighting scheme, this gives an overall standardized bias that has expectation zero. However, if each contribution from a stratum s is weighted in proportion to $m_s^{(c)}(n_s^{(c)}-m_s^{(c)})/n_s^{(c)}$, then by (6) one has

$$s_q b_q^{(c)} \propto \mathbf{Z}^{(c)\prime}\mathbf{t_q} - \pi'\mathbf{t_q},$$

where $\pi \equiv (\Pr(Z_1^{(c)}=1),\ldots,\Pr(Z_n^{(c)}=1))'$ is the vector of household-level probabilities of receiving the treatment. Besides yielding a relatively neat expression for the overall standardized difference, this weighting scheme has the property that under certain conditions it maximizes the likelihood that the derived standardized difference will differ significantly from zero; see Kalton (1968).

Writing $K = \sum_s m_s^{(c)}(n_s^{(c)}-m_s^{(c)})/n_s^{(c)}$ and weighting stratum contributions in proportion to $m_s^{(c)}(n_s^{(c)}-m_s^{(c)})/n_s^{(c)}$, Proposition 2.1 applied to each stratum separately yields

$$\mathrm{var}(s_q b_q^{(c)}) = K^{-2}\sum_s \frac{m_s^{(c)}(n_s^{(c)}-m_s^{(c)})}{n_s^{(c)}}\left[\sum_{i=1}^{n_s^{(c)}}(t_{qsi}-\bar{t}_{qs})^2\right]/(n_s-1) \qquad (11)$$

(*cf.* (10)).

In moderate and large samples, $s_q b_q^{(c)}$ is distributed roughly as $\mathcal{N}(0,\mathrm{var}(s_q b_q^{(c)}))$. We use this fact to assess statistical significance of standardized biases along covariates and various transformations of covariates, for each of the three treatments and corresponding treatment assignments. Note carefully that (11) gives the variance of $b_q^{(c)}$ *exactly*, not an estimate or

approximation to the variance. This improves the quality of the Normal approximation relative to those that are common in point estimation, which are based on z-scores of the form $(\widehat{\text{var}}(\hat{\theta}))^{-1/2}(\hat{\theta} - \text{E}_0\hat{\theta})$.

2.4 Tests of balance and of strictly no effect for the Vote '98 experiment

Table 6 gives standardized biases for assignment to in-person canvassing, which we have selected because among the three treatment conditions it gave the strongest suggestion of imbalance. One covariate, Residence in Ward 3, is biased away from the treatment condition to an extent that is statistically significant at the .01 level, and other standardized differences were significant at the .05 and .10 levels. (In each of the other two treatment assignments, one variable was significantly imbalanced at the .10 level and no other imbalances were significant.) For purposes of balance assessment multinomial variates are split into separate binary indicators and the one continuous covariate, age, is decomposed according to a natural spline with knots at the five quintiles of the age distribution, with balance assessed not on age directly but on the resulting B-spline basis variables (Hastie et al. 2001).

On the basis of Table 6 alone, it is unclear whether covariate imbalances in the Vote '98 experiment are consonant with treatment having been assigned at random. The many covariates on which treatment and control groups do not significantly differ speak in favor, but the imbalances along Ward 3 residence and other variables speak against (so far as the in-person experiment is concerned, at least). Were Ward 3 residence the only covariate, we would certainly conclude that something had been wrong; but with many covariates we would expect that a small fraction might exhibit some imbalance, at least by chance.

Postpone this issue, at least until Section 2.5, and assume for the moment that there is no reason to question the randomization. Under properly functioning randomization, all covariates are balanced; in particular, potential responses $y_{c1}, \ldots, y_{cn}$ and $y_{t1}, \ldots, y_{tn}$ are balanced, even if they incompletely observed. Because of this, if the hypothesis of properly functioning randomization is accepted, then we are in a position to test whether treatment had an effect. The test is similar to the test of no effect in the Adams and Smith experiment (§ 1.3); the main difference is that each Vote '98 experiment involves stratification, which the test has to take into account.

If treatment had no effect, then $y_{ci} = y_{ti} = y_i$ for all i. If treatment had no effect, then the bias $s_y b_{cy}$ should be small enough in magnitude as to be statistically insignificant. By comparing this statistic to its reference distribution — that is, by effecting the same calculations that were performed in Table 6 for each covariate, but this time on the observed response — we are led to a test of the hypothesis that treatment was without an effect. Applying this test, calibrated to have type I error rate .10, hypotheses of no effect or mail or telephone stand, whereas the hypothesis of no in-person effect is rejected (and would have been even at the .01 level). If randomization worked as it should, then we may conclude that in-person entreatments had an effect, some effect, whereas telephone and mail appeals may not have.

Covariate	Standardized Bias	
persons1	.015	
persons2	-.015	
v96.abst	-.025	
v96.vote	.005	
majpty	-.031	
age.Bspline1	-.012	
age.Bspline2	-.011	
age.Bspline3	-.026	
age.Bspline4	-.013	
age.Bspline5	.023	
age.Bspline6	-.011	
ward2	.000	
ward3	-.072	**
ward4	-.023	
ward5	-.005	
ward6	.001	
ward7	-.005	
ward8	.032	
ward9	.000	
ward10	-.010	
ward11	.012	
ward12	.035	
ward13	-.013	
ward14	-.012	
ward15	.035	
ward16	.026	
ward17	.052	*
ward18	-.014	
ward19	-.042	.
ward20	.001	
ward21	-.032	
ward22	-.013	
ward23	.002	
ward24	.002	
ward25	.024	
ward26	-.026	
ward27	-.029	
ward28	.007	
ward29	-.003	
ward30	.006	

Table 6: Standardized biases for assignment to in-person canvass

2.5 A χ^2 test of the randomization null

To assess imbalance along all covariates at once, rather than separately, note that Proposition 2.1 gives each stratum's contribution to the covariance matrix of $(s_{q^1} b_{cq^1}, \ldots, s_{q^n} b_{cq^n})$ and

by extension, since these contributions are independent, the covariance C of $(s_{q^1} b_{cq^1}, \ldots, s_{q^n} b_{cq^n})$. By the multivariate central limit theorem, under the null hypothesis this sum is distributed roughly as $\mathcal{N}(\mathbf{0}, C)$, provided sample size is large enough. If C^- is a generalized inverse of C, then in large samples

$$(s_{q^1} b_{cq^1}, \ldots, s_{q^n} b_{cq^n}) C^- (s_{q^1} b_{cq^1}, \ldots, s_{q^n} b_{cq^n})'$$

has the χ^2 distribution on $\text{rank}(C)$ degrees of freedom. To our knowledge, this global test of balance is new to this paper.

Assessed by this method, the telephone, mail and in-person treatments yield χ^2 values of 35, 31 and 40, all on 38 degrees of freedom, none of which a yield a p-value less even than 1/3. This shows that the imbalances noted in Section 2.4 between subjects with whom personal contact was and was not attempted do not undercut the hypothesis that this treatment was randomly assigned, at least not when they are viewed against a backdrop of covariates that were well balanced.

3 Attributing effects to treatment in a stratified, cluster-randomized design

3.1 Attributable effects with strata and non-compliance under individual-level randomization

To produce estimates and confidence intervals for the number of votes attributable to, say, telephone entreatments to vote, the method of inference used in the introduction to determine the number of votes attributable to Adams and Smith's telephone entreatment might be used, but for two complications. First, Adams and Smith placed calls to a simple random sample of voters, whereas the Vote '98 experiment made calls to simple random samples from each of four strata. Second, Adams and Smith assigned individual voters to treatment or to control conditions, whereas in the Vote '98 experiment assignment was made at the household level. As noted above, Gerber and Green's original analysis of these data ignored household-level clustering in the analysis of treatment effects (Gerber and Green 2000, 2005; Imai 2005). For the present section only, we follow them in this practice. The purpose of this is to ease the exposition. In Section 3.2, we elaborate the analysis so as to account for both clustered and stratified aspects of the design.

For purposes of analyzing effects of telephone entreatments, there are four strata: the "M-plus-I" stratum, consisting of those subjects assigned also to receive mailers and in-person appeals; the "M-plus-not I" stratum of subjects to whom mailers were sent but with whom in-person appeals were not attempted; and analogous "not M-plus-I" and "not M-plus-not I" strata. Corresponding to each of the four strata is a two-by-two table classifying subjects according to treatment assignment and outcomes, which we combine and regard as a single $2 \times 2 \times 4$, treatment by outcome by stratum, table. As in the unstratified analysis of § 1.3, hypotheses attributing effects to treatment can be represented as modifications of the table. For example, Table 7 represents hypotheses to the effect that: (i) in the M-plus-I stratum,

a votes are attributable to the telephone treatment; but (ii) in the three remaining strata, the telephone intervention generated no votes that would not have been made in its absence. Let $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$ be an atomic hypothesis subsumed by this one, so that $\sum_{M+I}(y_i - \tilde{y}_i) = a$ and $y_i = \tilde{y}_i$ in strata other than M-plus-I. Then Table 7 is a sufficient statistic for the test of $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$. The test consists simply of computing $s_{\tilde{y}} b_{\tilde{y}}$, by formula (4) or (5), and its null variance, applying Proposition 2.1 to each stratum, then comparing $s_{\tilde{y}} b_{\tilde{y}} / \sqrt{\mathrm{var}(s_{\tilde{y}} b_{\tilde{y}})}$ to the standard Normal distribution. If the test results in a rejection, then the attribution of a votes in the M-plus-I stratum, and none anywhere else, is deemed untenable. Before considering the

M-plus-I		
	vote	no vote
called	538	621
	$-a$	$+a$
not called	831	957

not M-plus-I		
	vote	no vote
called	89	96
not called	1331	1427

M-plus-not I		
	vote	no vote
called	2043	2512
not called	3382	4058

not M-plus-not I		
	vote	no vote
called	354	461
not called	4894	6217

Table 7: Attributing a of the M plus I stratum's votes to telephone entreatment.

role of such tests in assessing the total number of votes, irrespective of stratum, that a given treatment may have brought about, note that not every hypothesis of the form we are discussing — that a M-plus-I votes, and no others, are attributable to telephone entreatment — is worth evaluating by a statistical test. We assume that telephone GOTV calls do not prevent anyone from voting; thus negative a need not be considered. Certainly a is no larger than the upper left cell of the M-plus-I subtable of Table 7, 538, since only members of the treatment group, and subjects who actually voted, can be eligible to have their votes attributed to treatment. A related restriction on a follows from the fact that only a portion of those assigned treatment actually received it: writing d_{m+i} for the number M-plus-I subjects in households that received the telephone GOTV message, one has $a \leq d_{m+i}$. (The value of d_{m+i} is 241.) This is the manner in which the exclusion restriction (Angrist et al. 1996b) expresses itself in this setting, as a restriction on which hypotheses of attribution need be evaluated.

How shall we estimate the total number votes attributable to the telephone intervention? The hypothesis that a M-plus-I votes and no others are attributable to calls is one of a large number of hypotheses contained in the composite hypothesis to the effect that a votes overall are attributable to treatment. In order to assess the plausibility of a votes' being due to treatment, one must assess, directly or implicitly, whether it is plausible that a_{m+i}, a_{m-i}, a_{-m+i}, and a_{-m-i} votes from the four strata, respectively, are due to treatment, for all four-tuples of natural numbers $(a_{m+i}, a_{m-i}, a_{-m+i}, a_{-m-i})$ summing to a. There are $(a+1)^3/6+(a+1)^2/2+(a+1)/3$ such four-tuples. Provided that a is no more than a few hundred, direct assessment of each such four-tuple is feasible with a modern computer; for instance $a = 200$ translates to about 1.4

million alternatives. In general the number of natural-number sequences adding to a number a is a polynomial of degree one minus the number of strata, so that the chore quickly becomes infeasible, even with modern computers, as the number of strata increases. Indirect methods that avoid considering each possibility separately will be discussed in Section 4.

Straightforward (if highly repetitive) calculations of this type lead to confidence intervals for attributable effects. We illustrate by sketching the calculations used to delimit effects attributable to telephone entreatments. Of 23,500 hypotheses with $a = a_{m+i} + a_{m-i} + a_{-m+i} + a_{-m-i} = 50$, 22,900 are compatible, and these give z-statistics ranging from -2.08 to -1.59, indicating somewhat less plausibility. Accept for the moment that the standard Normal distribution closely approximates each of these statistics' null distribution. Then the hypotheses attributing 50 votes to treatment give two-sided p-values ranging from .037 to .113, and the p-value attaching to the composite hypothesis that treatment is responsible for 50 votes is the largest of those, .113. The 95% confidence interval for A, the number of votes attributable to treatment, consists of those a not rejected at the .05 level, so (continuing to take on faith that each z-statistic's null distribution is adequately approximated as $\mathcal{N}(0,1)$) we conclude that 50 belongs inside the interval. Continuing in this fashion, the composite hypothesis that $a = 70$ barely escapes rejection, with p-values ranging from .008 to .051; whereas each compatible hypothesis with $a = a_{m+i} + a_{m-i} + a_{-m+i} + a_{-m-i} = 71$ is rejected at the .05 level. Our 95% confidence interval runs from 0 through 70 votes attributable to telephone calls (of 5,030 calls attempted and 1,620 completed).

Can calibrating these z-statistics according to the standard Normal distribution be counted upon to give accurate p-values and confidence limits? No; but yes. No, because the z-statistics just given were calculated on the false assumption that randomization had been performed at the individual level. As a result, the variances calculated en route are bound to have been too small (§ 2.2). However, this fault is to be remedied presently, in Section 3.2; and once it is remedied the Central Limit Theorem for simple random samples (Erdős and Rényi 1959; Hájek 1960) ensures that the distribution of our test statistics is roughly Normal. But ordinarily this theorem is invoked to assure Normality of a single statistic, whereas the present method requires us to approximate the distribution of each of a large battery of statistics, and to do so with uniform standards of accuracy. So, even with z-statistics that appropriately account for the clustered design, an elaboration of the ordinary CLT argument is required.

For each compatible $\mathbf{a}$, let $\tilde{y}_\mathbf{a}$ represent a pattern of potential outcomes differing from the observed pattern y only in that in each stratum s, it records 1-responses (votes) for a_s fewer treated subjects. Write $F_\mathbf{a}$ for the distribution function of $s_{\tilde{y}_\mathbf{a}} b_{\tilde{y}_\mathbf{a}} / \sqrt{\mathrm{var}(s_{\tilde{y}_\mathbf{a}} b_{\tilde{y}_\mathbf{a} y})}$ under the hypothesis that outcomes $\tilde{y}_\mathbf{a}$ would have obtained had the GOTV calls not been made. With slight modifications, a theorem of Hoeglund (1978) shows[3] that even as $\mathbf{a}$ is permitted to vary freely, $\max_t |F_\mathbf{a}(t) - \Psi(t)|$ is bounded above by a universal constant that approaches zero as the sample size increases. In other words, the Central Limit Theorem applies uniformly. This warrants the

[3]Höglund's theorem, which pertains to simple random sampling rather than stratified random samples, asserts

use of the standard Normal distribution to evaluate each statistic $s_{\tilde{y}_\mathbf{a}} b_{\tilde{y}_\mathbf{a}} / \sqrt{\text{var}(s_{\tilde{y}_\mathbf{a}} b_{\tilde{y}_\mathbf{a} y})}$.

3.2 z-statistic profiling to account for strata and clusters

By ignoring clustering, in Section 3.1 we tested composite hypotheses to the effect that a votes resulted from treatment by decomposing them into simpler hypotheses that could separately be appraised by the method of Section 2.4, with each appraisal culminating in its own z-statistic. Such z-statistics are as readily calculated for clustered designs as for designs with treatment assignment at the individual level, so the presence of clusters is not an obstacle to such approach, at least in principle. Practically speaking, it is an obstacle, because the presence of clusters greatly expands the number of distinct atomic hypotheses that must be evaluated in order to test a the smallest macroscopic ones. A single formalism both clarifies the difficulty and aids in articulating a way around it.

Return for the moment to the setting of Section 3.1, in which clustering is ignored, and let $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$ be $\{0,1\}$-valued vectors of length n such that $\tilde{y}_i = \tilde{y}'_i = y_i$ for all i with $z_i = 0$, or with $z_i = 1$ but d_i (an indicator of whether treatment was received) equal to zero, and such that $y_i \leq \tilde{y}_i, \tilde{y}'_i \leq 1$ for all i. Then $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$ and $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}'$ are compatible atomic hypotheses (as discussed in the Introduction). Hypotheses like those considered early in Section 3.1, to the effect that a votes in the M-plus-I stratum (and no others) are attributable to telephone entreatments, are macroscopic composites of such atomic hypotheses — despite our earlier use of the term "macroscopic" only to refer to suppositions that a votes overall, irrespective of stratum, are attributable to treatment. To maintain this distinction, refer to $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$ and $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$ as *atomic* attributions of effect and to composites such as the hypotheses that ten, or that 50, M-plus-I votes are attributable to treatment, but no others are, as *molecular*.

With individual-level assignment and binary outcomes, $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}$ and $H_0 : \mathbf{y}_c = \tilde{\mathbf{y}}'$ fall under the same molecular attribution if they allocate the same number of votes to treatment in each stratum: for each S, $\sum_{i \in S} y_i - \tilde{y}_i = \sum_{i \in S} y_i - \tilde{y}'_i$. When this is so, testing the one in the manner of Section 3.1 invariably gives the same result as testing the other using the same procedure, since both test statistics and their null distributions are functions of sufficient statistics that coincide. This can be seen by review the procedures with which the test statistic and its moments were determined. For a formal expression, let $\mathbf{a}$ be an integer vector of length $|S|$, let $\mathbf{c\text{-}tab}(\mathbf{z}, \mathbf{d}, \mathbf{y}, \mathbf{s})$ be the treatment assignment- by treatment received- by observed

---

of (centered and scaled) sample sums X that

$$\max_{-\infty < t < \infty} |F_X(t) - \Psi(t)| \leq C \left[\frac{n}{N}\frac{N-n}{N}\right]^{-1/2} N^{-1/2} \left[N^{-1}\sum_1^N |x_k - \bar{x}|^3\right] \left[N^{-1}\sum_1^N (x_k - \bar{x})^2\right]^{-3/2},$$

where n and N are the sample and population sizes, $\Psi(\cdot)$ is the standard Normal cdf, and C is a universal constant (not specified in the paper). His method of proof is to bound the difference between X's characteristic function and that of the standard Normal distribution, then invoke Esseen's smoothing lemma (Feller 1971, ch. 16) to translate that bound into a bound on $\max_t |F_X(t) - \Psi(t)|$. Now the sample sum Y of a stratified random sample is the sum of sample sums $Y_1, \ldots, Y_k$ of simple random samples. The difference between the c.f.'s of $Y_1, \ldots, Y_k$ and those of (appropriated centered and scaled) Normal deviates can be bounded in Höglund's fashion, which again lifts via Esseen's lemma to a bound on $\max_t |F_Y(t) - \Psi(t)|$.

response- by stratum-cross-tabulation, and choose $\tilde{\mathbf{y}}(\mathbf{a})$ arbitrarily from among compatible atomic hypotheses with $\mathbf{a} = (\sum_{i\in S} y_i - \tilde{y}_i(\mathbf{a}) : i \in S)$; then any such hypothesis is appraised by comparing

$$\begin{aligned} z(\mathbf{a}; \textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{y}, \mathbf{s})) &= \frac{s_{\tilde{y}} b_{\tilde{y}}}{\sqrt{\text{var}(s_{\tilde{y}} b_{\tilde{y}})}} \\ &= \frac{\mathbf{z}^t \tilde{\mathbf{y}}(\mathbf{a}) - \sum_s (\text{E}Z_{s1}) \bar{\tilde{y}}_s(\mathbf{a})}{\sqrt{\text{var}(\mathbf{Z}^t \tilde{\mathbf{y}}(\mathbf{a}))}} \end{aligned} \tag{12}$$

$$\tag{13}$$

to the same (roughly Normal) reference distribution.

With clusters, the combination of $\textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{y}, \mathbf{s})$ and $\mathbf{a} = (\sum_{i\in S} y_i - \tilde{y}_i : i \in S)$ is no longer sufficient to determine the moments of $\mathbf{z}^t\tilde{\mathbf{y}}$. With clustering by household, the contribution to $\text{var}(\mathbf{Z}^t\tilde{\mathbf{y}})$ from stratum s is determined households' totals of votes under $H_0 : \tilde{\mathbf{y}}$, i.e. by the numbers $(t_{\tilde{y}i} = \sum\{\tilde{y}_{cj} : j \text{ in cluster } i\})$, rather than by individuals' votes; the formula is

$$\begin{aligned} &\frac{m_s^{(c)}(n_s^{(c)} - m_s^{(c)})}{n_s^{(c)}} \frac{\sum_1^{n_s^{(c)}} (t_{\tilde{y}i} - \bar{t}_{\tilde{y}})^2}{(n_s^{(c)} - 1)} \\ &= \frac{m_s^{(c)}(n_s^{(c)} - m_s^{(c)})}{n_s^{(c)}(n_s^{(c)} - 1)} \sum_{T=0}^{2} |\{i : t_{\tilde{y}i} = T\}| (T - \frac{\sum_i t_{\tilde{y}i}}{n_s^{(c)}})^2, \end{aligned} \tag{14}$$

where $m_s^{(c)}$ and $n_s^{(c)}$ are the number of treatment-group clusters and the total number of clusters in s, respectively. Inspection of (14) confirms that s may contribute differently to $\text{var}(\mathbf{Z}^t\tilde{\mathbf{y}})$ and to $\text{var}(\mathbf{Z}^t\tilde{\mathbf{y}}')$ provided that $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$ differ in the way they assign votes by household — even if they hypothesize the same number of votes in total.

Nor is $\textbf{c-tab}(\mathbf{z}, \mathbf{d}, \tilde{\mathbf{y}}, \mathbf{s})$, nor $\textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{y}, \mathbf{s})$ in combination with $\mathbf{a} = (a_1, \ldots, a_s)$, sufficient for the test statistic. Instead, atomic hypotheses $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$ give rise to the same z-statistic when for all strata s, and for $T = 0$, 1 or 2,

$$|\{i : i \in s, t_{\tilde{y}i} = T\}| = |\{i : i \in s, t_{\tilde{y}'i} = T\}|.$$

To accommodate this, identify molecular hypotheses with $2 \times 2 \times S$ tables α of natural numbers, where $\alpha[i, j, s]$ indicates the number of households in stratum i from which in actuality i members voted but from which, according to the hypothesis, only $i - j$ members would have voted had the household received no GOTV call. This information suffices to calculate

$$z(\alpha; \textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{t}_y, \mathbf{s})) = \frac{\mathbf{z}^t \mathbf{t}_{\tilde{y}}(\alpha) - \sum_s (\text{E}Z_{s1})(n_s^{(c)})^{-1} \sum_1^{n_s^{(c)}} t_{\tilde{y}si}(\alpha)}{\text{var}(\mathbf{Z}^t \mathbf{t}_{\tilde{y}}(\alpha))} \tag{15}$$

where $\tilde{\mathbf{y}}(\alpha)$ is any of the atomic hypothesis of which α is comprised. A hypothesis α attributes $a(\alpha) = \sum_s \alpha[2, 2, s] * 2 + \alpha[2, 1, s] + \alpha[1, 1, s]$ votes to treatment.

For positive integers a of moderate size, the number of α such that $a(\alpha) = a$ well exceeds the number of by-stratum attributions $\mathbf{a}$ that sum to a. To skirt the resulting computational difficulty, define $\mathbf{a}(\alpha) = (\alpha[2,2,s] * 2 + \alpha[2,1,s] + \alpha[1,1,s] : s = 1,\ldots,S)$ and note that the numerator of (15), if not its denominator, depends on α only though information contained in $\mathbf{a}(\alpha)$. Rather than calculating each member of

$$\{z(\alpha; \textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{t}_y, \mathbf{s})) : a(\alpha) = a\}, \tag{16}$$

for each $\mathbf{a} = (a_1, \ldots, a_s)$ such that $a_1 + \cdots + a_s = a$ we calculate $z(\alpha^*; \textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{t}_y, \mathbf{s}))$, where

$$\alpha^* = \arg \min_{\mathbf{a}(\alpha)=\mathbf{a}} |z(\alpha; \textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{t}_y, \mathbf{s}))|. \tag{17}$$

In other words, we calculate a profile of (16), selecting from each

$$\{z(\alpha; \textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{t}_y, \mathbf{s})) : \mathbf{a}(\alpha) = \mathbf{a}\} \tag{18}$$

for which $\mathbf{a}^t\mathbf{1} = a$ that element that is closest to zero. Because the denominator in (15) is a function of $\mathbf{a}(\alpha)$ only, the sets (18) being profiled are always uniform in sign. Thus if a set of form (18) contains element z that is larger than a negative acceptance limit $-z^*$, then its $z(\alpha^*; \textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{t}_y, \mathbf{s}))$ exceeds that limit and the hypothesis $H_0 : A = a(\alpha)$ is not rejected; or if it contains some element z that is smaller than a positive acceptance limit z^* then its $z(\alpha^*; \textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{t}_y, \mathbf{s}))$ falls below the same limit and again $H_0 : A = a(\alpha)$ is not rejected. The profile method permits assessment of both one and two-sided hypotheses.

Given that $\mathbf{a}(\alpha) = \mathbf{a}$, the magnitude of (15) is minimized by selecting α so as to maximize $\text{var}(\mathbf{Z}^t t_{\tilde{y}}(\alpha))$. With clusters of size 1 or 2, this maximum is achieved by any α that for each s maximizes $\alpha[1,1,s] + (1/2)\alpha[2,2,s]$[4]. By selecting such an α for each $\mathbf{a}$ summing to a, we reduce the number of z-statistics that need be calculated to assess a hypothesis $H_0 : A = a(\alpha)$ to the same number as was required in the absence of clustering.

3.3 Accommodations for drop-out

If treatment assignment is ignorable, either because researchers controlled it or because adjustment for measured covariates removed hidden bias, and if a certain plausible assumption holds, then the 5% of subjects in the Vote '98 experiment whose voting in 1998 could not be determined from voter rolls pose no threat to the validity or to the implementation of randomization inference for the attributable effect. The simple and intuitive procedure of imputing a fixed value, such as zero, to each missing-response household as the household total number of votes, then suffices for valid inference. The current section, which readers may skip without loss of continuity, argues this point and suggests a slight elaboration on the imputation procedure that

[4]This can be verified by writing each term $|\{i : t_{\tilde{y}i} = T\}|$ in (14) as a sum of entries in $\textbf{c-tab}(\mathbf{z}, \mathbf{d}, \mathbf{t}_y, \mathbf{s})$ and entries in α or their opposites and then rearranging terms to get a quadratic in $\frac{\sum_i t_{\tilde{y}i}(\alpha)}{n_s^{(c)}}$ (which is determined by $\mathbf{a}(\alpha)$).

we use in our analysis.

Fix attention on n households in a given stratum, within which ignorability is assumed. Let $\mathbf{T}_t, \mathbf{T}_c, \mathbf{D}_t$, and $\mathbf{Z}$ be n-vectors of random variables representing households' potential responses to treatment and control conditions ($\mathbf{T}_t, \mathbf{T}_c = 0$, 1, or 2 votes), whether the subjects would receive the treatment ($D_t = 1$) or not ($D_t = 0$) if assigned to it, and assignment to treatment ($Z = 1$) or control ($Z = 0$) conditions. Under ignorability and without missing data, the likelihood contribution for this stratum takes the form

$$p(\mathbf{d}_t|\mathbf{t}_t, \mathbf{t}_c)p(\mathbf{t}_t|\mathbf{t}_c)p(\mathbf{t}_c)p(\mathbf{z}).$$

As a result, $(\mathbf{T}_t, \mathbf{T}_c, \mathbf{D}_t) \perp \mathbf{Z}$, and for non-random functions f

$$\mathcal{L}[\mathbf{Z}^t f(\mathbf{T}_t, \mathbf{T}_c, \mathbf{D}_t)|(\mathbf{T}_t, \mathbf{T}_c, \mathbf{D}_t) = (\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t); \mathbf{Z}^t\mathbf{1} = m] = \mathcal{L}[\mathbf{Z}^t f(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t)|\mathbf{Z}^t\mathbf{1} = m]. \quad (19)$$

(Whereas the probability law at left may in general depend on the joint distribution of all of $\mathbf{T}_t, \mathbf{T}_c, \mathbf{D}_t$, and $\mathbf{Z}$, the law at right depends only on that of $\mathbf{Z}$.) An hypothesis about the attributable effect fixes a suitable function f, the value of which can be calculated from observed data: $f(t_t, t_c, d_t) = (1 - d_t)t_c + d_t(t_t - \delta)$, where δ is given by the hypothesis. In the event that the hypothesis is true, then $\mathbf{Z}^t f(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t) = \mathbf{Z}^t\mathbf{t}_c$, a random variable the distribution of which can be determined precisely.

To accommodate missing data, add a random vector $\mathbf{O}$, with $O_i = 1$ indicating that the ith household's response is observed and $O_i = 0$ that it is missing, and assume that a stratum's likelihood factors as

$$p(\mathbf{d}_t|\mathbf{t}_t, \mathbf{t}_c, \mathbf{o})p(\mathbf{t}_t|\mathbf{t}_c, \mathbf{o})p(\mathbf{t}_c, \mathbf{o})p(\mathbf{z}). \quad (20)$$

By the same reasoning as above, for any fixed function g of $\mathbf{T}_t, \mathbf{T}_c, \mathbf{D}_t$, and $\mathbf{O}$ the distribution $\mathcal{L}[\mathbf{Z}^t g(\mathbf{T}_t, \mathbf{T}_c, \mathbf{D}_t, \mathbf{O})|\mathbf{T}_t, \mathbf{T}_c, \mathbf{D}_t, \mathbf{O}; \mathbf{Z}^t\mathbf{1} = m]$ is determined by $p(z)$ from (20) only. If missing responses are to be treated as no-votes, to evaluate a hypothesis attributing votes δ one would define

$$g(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t, \mathbf{o})_i = \begin{cases} t_{ci}, & o_i = 1, d_{ti} = 0; \\ t_{ti} - \delta_i, & o_i = 1, d_{ti} = 1; \\ 0, & o_i = 0. \end{cases} \quad (21)$$

Whether or not the hypothesis is true, $g(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t, \mathbf{o})$ can be calculated from observed data. Assuming that δ correctly identifies those subjects who voted only because they received the treatment, then the distribution of $\mathbf{Z}^t g(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t, \mathbf{o}) = \mathbf{Z}^t(\mathbf{t}_c \cdot \mathbf{o})$ can be determined precisely and used to calibrate p-values and acceptance regions.

The variation on this approach that we prefer varies this by instead defining

$$g(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t, \mathbf{o})_i = \begin{cases} t_{ci}, & o_i = 1, d_{ti} = 0; \\ t_{ti} - \delta_i, & o_i = 1, d_{ti} = 1; \\ \frac{\sum_i o_i[(1-d_{ti})t_{ci} + d_{ti}(t_{ti} - \delta_i)]}{\sum_i o_i}, & o_i = 0. \end{cases} \quad (22)$$

One source of appeal of this approach is that its centered test statistic,

$$\mathbf{Z}^t g(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t, \mathbf{o}) - (\mathrm{E}Z_1) n^{-1} \sum_i g(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t, \mathbf{o}))_i,$$

works out to be the same as would have resulted from deleting missing-outcome clusters, pretending they did not exist, and using as test statistic $\mathbf{z}^t f(\mathbf{t}_t, \mathbf{t}_c, \mathbf{d}_t)$ (where $f(t_t, t_c, d_t) = (1 - d_t)t_c + d_t(t_t - \delta)$). While the centered test statistics that result from that approach and from ours are the same, the casewise-deletion approach generally errs in its calculation of the variance of its test statistic, whereas retaining the missing-outcome cases, but using the g from (22) in order to mimic the casewise deletion test statistic, protects us from that mistake.

A second attraction is that our approach amounts to a form of mean imputation of missing responses, whereas that represented in (21) imputes a value that is blind to available data. We expect that our data-driven approach should be more efficient. An exception occurs when a stratum contains very few observed responses, so that the mean of them would be subject to great sampling variability. With large strata, containing many observed responses, there is little threat of this, although in matched analyses, where each stratum (matched set) may contain only a few observations to begin with, it is a greater concern. In the matched analyses to follow in Section 4, we address this issue by using a hybrid of (21) and (22): when the matched set contains only one observed response, (21) is used; otherwise we use (22).

3.4 Effects attributable to treatment

Our confidence intervals for the attributable effect, reported to two significant digits, are as follows. With 95% confidence, between 40 and 250 votes may be attributed to in-person canvassing. With 2/3 confidence[5], between 90 and 200 votes may be attributed to Vote '98's in-person canvass. For direct mail appeals to vote, 95% and 2/3 confidence intervals run from 0 to 1000 and from 0 to 670 votes attributable to treatment. For the telephone experiment, 95% and 2/3 confidence intervals are $[0, 100]$ and $[0, 30]$ votes, respectively. Table 8 expresses these results in terms of the numbers of contacts and attempted contacts, for each GOTV intervention.

Treatment	level	up to Votes overall	Number of votes per 100 up to contacts	attempted contacts
In-person	2/3	90 up to 200	4.7 up to 10	1.5 up to 3.3
	95%	40 up to 250	2.1 up to 13	.7 up to 4.1
Telephone	2/3	0 up to 30	0 up to 1.3	0 up to .4
	95%	0 up to 100	0 up to 4.2	0 up to 1.4

Table 8: Confidence intervals for effects attributable to Vote '98 interventions.

[5]Since the 2/3 acceptance region runs from $-.97$ to .97, or approximately from -1 to $+1$, the 2/3 confidence interval is an appropriate analogue to the interval $\hat{\mu} \pm \text{s.e.}(\hat{\mu})$ for inferences not based on point estimates.

4 Flexible matching to handle missing

To attribute effects to treatment in the in-person and phone conditions we relied on the fact that treatment was randomly assigned to clusters of people (households) to justify our statistical inference. However, we made one further assumption in that analysis that ought to be examined: we assumed that the potential of finding missing data on vote in 1998 (the outcome of interest) would be the same regardless of treatment assignment. That is, just as it is common to base inference on an assumption that the values of potential outcomes are independent of random assignment, so to, we assumed that potential missingness on potential outcomes is also independent of assignment. If this assumption is correct then our imputation procedure would enhance the precision of our stratified analysis without any concerns about bias. That is, the fact of random assignment ought to protect our inferences from bias due to missing responses.

However, what if something about treatment assignment itself caused people to have missing responses? In a drug trial this would be easy to imagine — the treatment could cause onerous side-effects which could cause people to drop out of the study. Without observation of the side-effects, we have an unobserved confound — and thus some adjustment would be necessary in order to make the case for ignorability for the IV assumptions to hold (and thus to obtain an estimate of the effect of the drug on those who took it). In the case of the Vote '98 experiment, we don't imagine that being assigned a phone call or in-person visit had a large effect on propensity to show up missing in the registrar of voters's records. However, one might imagine that some people, by assignment to the control group, did not receive the encouragement to turnout that others did — and that not turning out for them (people who had not voted in many previous elections) could have caused them to be purged from the voting records. In this case, the treatment could be seen as enabling people who were assigned treatment to stay on the voter records in addition to encouraging them to turnout to vote, and thus the potential for someone to have missing responses would not be independent of treatment assignment but in fact caused by it.

One way to address concerns about bias associated with observed confounds in observational studies is to reduce the heterogeneity between the groups of people assigned control and treatment (Rosenbaum 2005). Although this is an experiment, our concern about potential biases due to missingness can also be addressed by reducing the heterogeneity that obtains within the four strata that so far have organized our analyses. And, a natural way to make strata that are more homogeneous would be to break those four strata into smaller groups. Carried to an extreme, matching individual members of the treated to individual members of the control group allows the maximum heterogeneity reduction due to observed covariates within any given sample. Thus, our general goal is to create matched sets of people (treated matched to controls) who are as similar as possible in terms of their propensity to have missing outcomes. When we attribute effects to treatment we will (1) impute response values to those people in sets who have missing responses based on the responses of the other people within those sets and (2) exclude people with imputed responses from those who are eligible for attribution of treatment

effects.

So, one benefit of our proposed strategy is that it ought to dampen any effects of bias due to associations between treatment and missing responses. Another benefit is that this mode of analysis (that of reducing heterogeneity by matching or otherwise grouping subjects together based on the observed information in the dataset) demonstrates the applicability of attributable effects to observational data. That is, in an observational study an analyst cannot assume that treatment has been assigned in some way that is unconfounded with other attributes of units — and thus must "control for" those potential confounds by making comparisons among units that differ only on treatment and are similar in as many other ways as possible.

4.1 Scores

Any attempt to group or match units requires some metric which defines what it means for units to be "similar". In this application we combine three such metrics. We sought to create matched sets that are not only similar in terms of propensity to exhibit missing responses (in order to protect us against bias), but also which are similar in terms of propensity to be treated (mainly here to enhance precision since we already know that households — our unit of matching — are already quite balanced in terms of treatment assignment), and also in potential response to control (which we will explain in a bit, and which is a score meant to enhance the precision of our estimates).

All three scores (assignment, missingness, and response to control) were created using a logistic regression model, the model specification of which was found by a forward-backward AIC based model search on a randomly sampled training dataset 1/5th the size of the entire individual level dataset. Our models for treatment assignment began as an additive function of complementary treatment assignment (12 categories), ward of residence (29 categories), number of people in the household (1 or 2), voting behavior in 1996 (abstain, registered but not voting, and voting), membership in a major party or not, whether age was missing, and natural cubic b-spline bases for age with 5 cutpoints dividing the sextiles. The model search was limited from above by the model containing it and all second-order interactions of its covariates (except for the interaction of "missing age" and the age b-spline bases). 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). We then applied the results of these training models to our full dataset. In order to assess the applicability of our training models to our full dataset we compared the mean-squared error (MSE) from the same model applied to both sets of data. In each case, the MSEs were quite close and thus indicated that our relatively unsupervised data mining exercise (Hastie et al. 2001) had produced useful results.

Assignment Scores For telephone treatment assignment this procedure did not add any new terms to the model above and beyond those already listed (i.e. it did not end up moving from the simple base model). For the in-person treatment one more element was added — an interaction of the number of persons in a household and voting behavior in 1996 (2 terms). The MSEs of the training and full datasets were quite close: .91 versus .90 for the phone treatment and .99 versus .98 in the in-person treatment.

In our stratified analysis, we used treatment assignment as a IV and we did the same, albeit with many more strata, in this matching analysis. Thus, we differ from Imai (2005) 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.

Missingness Scores We followed the same general procedure to predict the propensity to have missing responses. This time the selection procedure added a ward by 1996 voting behavior interaction — adding some 56 terms to the equation (2 included categories of 1996 voting behavior by 28 included ward terms). For the missingness score, the MSE on the full model and training set model were identical at 2 significant digits (.23).

Response Scores Among Controls We also estimated what we call a "response score," which is the probability of voting in 1998 (the outcome of the experiment) *among the control group* only. For the response score, the MSE on the full model versus training model was 1.05 versus 1.07 in the phone control group and 1.04 versus 1.06 in the in-person control groups. We do this in order to produce tighter confidence intervals around our attributed effects.

The justification for conditioning on these scores is as follows. Under the assumption of ignorability, we can factorize the joint density of potential responses (y_t, y_c), d_t for "dose" or compliance with the treatment, o (observation i.e. missingness of potential responses to either treatment or control), treatment assignment z, and covariates $\mathbf{x} = (x_1^t, \ldots x_k^t)^t$ so that we can write

$$p(y_t, y_c, d_t, o, z, x) = p(d_t|y_t, y_c, o, x)p(y_t|y_c)p(y_c, o|\mathbf{x})p(z|\mathbf{x})p(\mathbf{x}). \tag{23}$$

Writing our overall probability model in this way highlights the fact that, in order to understand something about effect of treatment — which we defined earlier by the relationship between the potential responses to treatment (y_t) and the potential responses to control (y_c) — that we can condition on quantities that are ancillary to $p(y_t|y_c)$, in the sense that they carry no information on parameters determining $p(y_t|y_c)$. Our goal is to diminish heterogeneity between treated and controls without introducing more bias. Equation 23 shows what we can use in pursuing this goal: data determined by, and reflecting on, $p(y_c, o|\mathbf{x})$, $p(z|\mathbf{x})$ and$p(\mathbf{x})$; but not data that reflect on $p(d_t|y_t, y_c, o, x)$ or $p(y_t|y_c)$.

Our assignment score is an estimate of $p(z|\mathbf{x})$, and our missingness and response scores are estimates of $p(y_c, o|\mathbf{x})$. Both are dimension-reducing functions of $\mathbf{x}$, so by conditioning on them we condition on $\mathbf{x}$, at least partly. We are prevented from conditioning on $\mathbf{x}$ in its entirety by the "curse of dimensionality": even if our covariates each only had 2 values, for 10 covariates we'd need 2^{10} different types of treated and control people in our dataset – and the odds of finding suitable matches would be slim. Although it is common to only reduce the dimension of $\mathbf{x}$ via an assignment score, we find no reason to not use different "slices" of $\mathbf{x}$ to aid our matching. However, by using three different scores, each of which is a projection of $\mathbf{x}$ onto theoretically relevant axes (treatment assignment, missingness, and potential response to the control condition), we will do a better job of isolating the variance in $\mathbf{x}$ that is important for our inferences.

4.2 Distance Matrices with Calipers on Missingness

After creating these scores (using the linear predictors from logit models), we collapsed them by taking the average scores within each household. Following Rosenbaum and Rubin (1985), we rated the permissible matches of households using a Mahalanobis distance on all three scores. We created these distances separately within complementary treatment strata, and also separately for people living in 1 person versus 2 person households (giving us 24 distance matrices, each posing its own, independent, matching problem).

Within the distance matrices we wanted to privilege missingness distance above the other scores because we worry about most about bias due to missingness and, because of random assignment, less about differences between treated and control households on potential responses or propensity to be assigned treatment. That is, we wanted to make sure that our matched sets contain people who are as similar as possible on the missingness dimension. 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 that 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 difference

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

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

in s.d.s on the underlying continuous variable within the treatment and control groups and to the fraction by which the between-group discrepancy on that variable is to be reduced: see Table 9, adapted from Cochran and Rubin's paper. Calipers are not a requirement for the use of scores like ours 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 in this case of using 3 scores that are all equally weighted for the Mahalanobis distance matrices, they allow us to add extra emphasis to the missingness score in the matching procedure. 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.

We selected our caliper based on our examination of the distance matrices. We desired to enforce the closest possible distance between treated and control households on the missingness score while simultaneously avoiding excluding too many people from the analysis. That is, if we chose to only allow matches within, say, .01 of a pooled sd on the missingness score, we might have labeled many pairs of respondents as unmatchable. After we inspected the matrix of distances between treated and control respondents on missingness scores, we realized that we could impose a caliper of .25sd and only lose 5 treated households out of 6131 in the in-person condition, and only 19 treated households out of 6996 in the phone condition. This caliper also caused us to exclude 279 control households from the in-person condition and 270 control households from the phone condition. Thus the total amount of data that we lost in exchange for the virtue of ensuring homogeneity of people in terms of missingness is roughly 1% in each condition.

4.3 Optimal Full Matching

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 during the matching of 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. Some such trade-off is unavoidable; but 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.

Full matching removes subjects from the analysis only for lack of suitable comparison units, never because of operational limitations of the matching routine itself. Other strains of without-

replacement 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 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.

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 (i.e. within our caliper); if not, the cell should be empty or filled with `Inf`, the R representation of infinity. 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 experiment in such a way as to respect the requirements laid out in section 4.2, it would suffice to put any finite number in each cell corresponding to a permissible match, with `Inf` in the others. 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 5275 telephone treatment households and a column for each of the 18175 controls (or the 4645 in-person treatment versus 18805 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 and in-person treatments stratified by the other experimental treatments, we preceded our full

matching with a stratification along these complementary treatments. We further stratified the matching by size of household (1 or 2 members), producing 24 strata containing independent distance matrices. 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), and who both lived in the same size of household (1 or 2 person), 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 treatment effects of telephone or personal canvassing be confounded with the effects of the other treatments.[6]

A simple call to `fullmatch`, however, does not produce the best set of matches (in the sense of minimizing mean-squared error of resulting estimates). For example, the poststratification obtained in this way for the telephone treatment contained matched sets with one treated subject and controls varying in number from one to 68, as well as matched sets with only one control but as many as five treated subjects. Such a variable set of matched-set configurations tends to increase the width of confidence intervals to be produced in the analysis (Hansen 2004). For the 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 `max.c` of controls to treated subjects in matched sets — perhaps a whole number k, indicating that only matched sets with one treatment and k or more controls are to be allowed, or perhaps a fraction $1/k$, indicating that sets of one control and k treated subjects are permissible. Without such constraints, full matching may generate matched sets with a treatment and any number of controls, or a control and any number of treated subjects, but with them the matched sets' ratios of controls to treatments are forced to fall between `min.c` and `max.c`. Not every such pair of constraints is possible to meet: if a data set contains three times as many controls as treated subjects, then clearly the requirement `max.c` $= 2$ is impossible to observe while placing all available subjects into matched sets. Indeed, if also some potential matches were forbidden by caliper requirements, then it might be necessary to set `max.c` to a number larger than three in order that the matching problem be feasible.

This section (§ 4) 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

[6] 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.

calipers, the most favorable restrictions consistent with the caliper can be determined by trial and error. The `optmatch` functions `minControlsCap()` and `maxControlsCap()` 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, `minControlsCap()` 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 `maxControlsCap()`. 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 24 separate strata, compound scores, and missingness calipers, in turn, first maximizing `min.c`, then passing this largest feasible `min.c` to `maxControlsCap()` in order to set `max.c`. This procedure regularizes the matched sets appreciably, as shown in Tables 10 and 11. Whereas the unrestricted matching for the telephone treatment produced at least one matched set with 68 controls, here, no matched set contains more than 24 controls. The in-person matching produced one set with 94 controls without restrictions (and seven sets with more than 25 controls). With restrictions the largest set had 50 controls, and the next largest had 27.

	Number of direct mailings sent			
	0	1	2	3
Personal canvass not attempted				
	6 to 24	$\frac{1}{2}$ to 3	$\frac{1}{3}$ to 2	1 to 1
Assigned to receive personal canvass				
Contact occurred	1 to 11	$\frac{1}{2}$ to 4	$\frac{1}{2}$ to 2	1 to 3
Not contacted	2 to 18	1 to 2	$\frac{1}{2}$ to 3	$\frac{1}{2}$ to 5

Table 10: Controls per treatment subject in matched sets (Telephone Treatment). Represented is the variation, across strata of direct mail and in-person treatments, in the number of controls per subject assigned to receive the telephone treatment. For instance, within the subgroup of subjects with whom a personal canvass was unsuccessfully attempted (last row of the table) and who received two encouragements to vote in the mail (third column), subjects with whom a telephone call was attempted are matched to as many as three 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.

	Number of direct mailings sent			
	0	1	2	3
Phone call not attempted				
	$\frac{1}{2}$ to 50	1 to 16	2 to 5	$\frac{1}{2}$ to 11
Assigned to receive phone call				
Contact occurred	1 to 8	1 to 12	1 to 5	1 to 6
Not contacted	1 to 8	1 to 11	1 to 9	1 to 8

Table 11: CONTROLS PER TREATMENT SUBJECT IN MATCHED SETS (IN PERSON 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 in-person treatment. For instance, within the subgroup of subjects with whom a telephone call 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 nine subjects not called. The same subgroup contains matched sets with as few as one attempted-telephone-entreatment subject per subject for whom no telephone treatment was attempted.

4.4 Attributing Effects with Missingness, IV and Cluster adjustments

In broad brush strokes, our procedure for attributing effects to treatment across the matched sets is the same as before. The main difference (other than the imputation of missing values) is that, previously we had poststratified the data into just four sets, whereas now there are many more. In the earlier case, the job of enumerating atomic hypotheses was simplified greatly by the fact that atomic hypotheses the same number of outcome events in each of the four strata could be handled in a single calculation. While the same principle is valid in the current analysis, it is of little use with the present 5,498 and 4,920 matched sets for the phone and in-person treatments respectively.

Fortunately, since we have many matched sets, another principle, that of asymptotic separability, is available (Gastwirth et al. 2000). When testing the hypothesis that A takes a given value A_0, it points the way to that atomic hypothesis among the many attributing A_0 events to treatment whose associated test statistic has the largest left-tailed, one-sided p-value, and to that atomic hypothesis 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 atomic hypotheses 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 atomic hypotheses 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 (2002c)

for more details as to how the principle is applied.

To illustrate the importance of reducing the number of atomic hypotheses to be evaluated, consider testing the hypothesis that $A = 100$ votes were produced by the telephone treatment. Of the 2,385 subjects (nested in 1,668 households) contacted by telephone as part of the VOTE'98 campaign, 1415 eventually voted, and for 56 the registrar of voters did not record whether they voted or not. Our matching placed these voters into 1655 separate matched sets. This makes for roughly $\binom{1400}{100}$, about 10^{155}, atomic hypotheses and associated null hypotheses subsumed under the composite hypothesis that $A = 100$: about 10^{155} separate ways to assign $y_{ti} - y_{ci} = 1$ to members of the treatment group who complied; or, put in terms of statistical computation, about 10^{155} ways to distribute 100 1's among 1415 positions in a vector of length 31,098, the size of our sample of individuals (including those with missing responses). If $A = 100$ is to be rejected, it is only because each of these 10^{155} 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^{155} test statistics would also have merited rejection.

Votes Attributable to Telephone Calls The hypothesis that $A = 100$ is, in fact, accepted at the .05 level, with the z-statistics for tests of the simple hypotheses subsumed under it delimited by -3.17 and -1.79. Applying the same analysis repeatedly with varying A_0 starting at 0 and going up to 1415 (the maximum number of votes attributable to the telephone treatment), produces a 95% confidence interval bounded above by 114 votes, and which includes 0. A 2/3 confidence interval is bounded above by 39 votes.Thus, while it is probable that the telephone treatment did not change the voting behavior of any experimental subject, as many as 4.8% of those contacted may have voted as a result of the call.

Votes Attributable to In-Person Visits Personal visits were assigned to 6131 people in 4645 households. Of those assigned, 1916 people (in 1377 households) were successfully contacted. Of those 1091 voted and 38 were missing information about whether they voted or not.[7] This means that the maximum attribution of votes to in-person treatment is 1091 (since we do not attribute effects to respondent whose outcomes are not available). When we assessed the evidence for the range of A_0 from 0 to 1091, we found that we could reject attributions of less than 48 or more than 325 votes given our 95% confidence interval. The 2/3 confidence interval ascribes as few as 41 (2.1% of those contacted) or as many as 270 (14% of those contacted) votes to the in-person treatment.

[7]Our matching assigned the people who answered their doors to 1373 sets.

5 Discussion

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 for a randomized experiment using simple stratification under the assumption that missingness of responses was independent of treatment assignment and also using an optimal full matching to protect our inference from the possibility that the missing data might be related to treatment assignment. We say "confident" because the crucial assumptions rest on a piece of the research design about which researchers often know most. 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 — "number of votes" versus "odds of voting". 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.

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

Although the bases for our inference are simple, the techniques that we deployed to make these inferences may not seem so — it has certainly taken us quite a few pages just to report a few confidence intervals! Focusing on effects of telephone solicitations, rather than in-person appeals, Imai (2005) 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.

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. For example, figure 2,

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 et al. (1995); Nie et al. (1996); Rosenstone and Hansen (1993); Highton and Wolfinger (2001); Wolfinger and Rosenstone (1980), among many others — imbalance on age is particularly troubling.

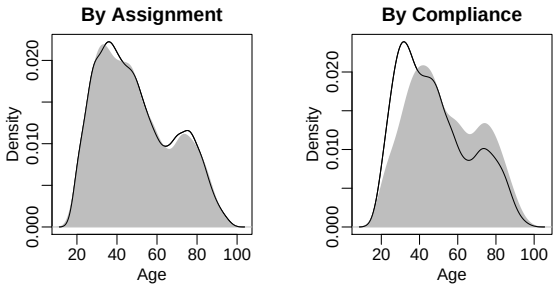

Figure 2: Age Distributions by In-Person Treatment Assignment versus Compliance with In-Person Treatment This figure shows Gaussian kernel density estimates of the distributions of age among subjects assigned to in-person treatment versus control (left panel) and among those assigned to personal canvassing, subjects who answered their doors versus those who did not (right panel). The default bandwidth selection algorithm for the `density()` function in R was used to select the different bandwidths of the different curves.

In analyses of the type we are considering, then, the threat of bias from observed covariates is real (in addition to the ever present danger of bias from hidden covariates). Were the treated and not-treated groups to be compared without prior benefit of matching, the justifying assumption required would be that the treated/not-treated distinction is ignorable *simpliciter*; however, by using matching, Imai (2005) relaxes this requirement somewhat. He need only assume that, given covariates, the distinction between treated and not-treated subjects is ignorable. He can concede that subjects not treated because they did not comply with treatment may differ systematically from subjects not treated because they were initially assigned to the control group. He must insist, however, that these systematic differences are entirely captured by the observed covariates.

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

to implement an IV analysis within the framework of attributable effects. Thus, we not only prefer the IV analysis on the grounds that its assumptions are more reasonable in this particular study than the assumptions required to trust entirely in covariance adjustment (no matter how non-parametric), but also because it is easy to do.

5.2 Assumptions

Throughout the paper we have attempted to highlight our 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.

- 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 attributions of effects using matched sets, we also assumed that there were sufficiently many matched sets for asymptotic separability. (This is a mild assumption; we had upwards of 4,900 matched sets, whereas Gastwirth et al. (2000) found asymptotic separability to hold in examples with as few as 13 matched sets.)

- We assumed that the lack of bias due to observed covariates supports our decision use random assignment to treatment as an instrumental variable.

- We made the stable unit treatment value assumption, SUTVA (Rubin 1986), with households as the units whose treatment values, t_{yt} and $_{yc}$, were assumed stable across hypothetical treatment assignments.

- 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. 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 two-stage least squares estimation, even when

demonstrably consistent, may be substantially biased in moderate, even moderately large, samples, whereas randomization inference with an instrumental variable is free of bias even in small samples. We don't repeat their arguments here, but we have illustrated some of the advantages of randomization inference throughout this paper.

Above and beyond randomization inference for instrumental variables and attributable effects, we also demonstrated that randomization inference can handle complex data analysis problems. For example, we accounted for the clustered administration of treatment and missing data in our estimation of effects. We also showed how to attribute effects within the framework of full matching. We used matching to assess whether the results from our stratified analyses might be biased due to some relationship between treatment assignment and propensity to have missing outcomes. This turned out not to be the case. However, by demonstrating how we can generalize our analysis from a small stratified table to a full matching, we showed how one might use these tools in the context of an observational study.

Why do we think it is a virtue that we made few assumptions and that we made them explicit? Because assumptions can fail, and often do. We believe that taking this possibility seriously has consequences for the substantive interpretation of research. For example, our substantive conclusions are quite close to those that emerged from the 2SLS analysis of Gerber and Green (2000), but, because ours relied on fewer assumptions, we have more confidence that the results they presented were not the result of problems such as those described by Imbens and Rosenbaum (2005). 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 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.

References

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

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996a), "Identification of Causal Effects Using Instrumental Variables (Disc: P456-472)," *Journal of the American Statistical Association*, 91, 444–455.

— (1996b), "Identification of causal effects using instrumental variables (Disc: p456-472)," *Journal of the American Statistical Association*, 91, 444–455.

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

Brady, H. and Seawright, J. (2004), "Framing Social Inquiry: From Models of Causation to Statistically Based Causal Inference," Working Paper.

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

Erdős, P. and Rényi, A. (1959), "On the central limit theorem for samples from a finite population," *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 4, 49–61.

Feller, W. (1971), *An introduction to probability theory and its applications. Vol. II.*, Second edition, New York: John Wiley & Sons Inc.

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

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

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

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

Greenland, S. (1987), "Interpretation and choice of effect measures in epidemiologic analyses," *American Journal of Epidemiology*, 125, 761–768.

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

Hájek, J. (1960), "Limiting distributions in simple random sampling from a finite population," *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5, 361–374.

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

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*, Springer-Verlag Inc.

Highton, B. and Wolfinger, R. E. (2001), "The first seven years of the political life cycle," *American Journal of Political Science*, 45.

Ho, D. and Imai, K. (2004), "Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election," Unpublished Manuscript.

Ho, D., Imai, K., King, G., and Stuart, E. A. (2004), MatchIt: *Matching Software for Causal Inference*.

Hoeglund, T. (1978), "Sampling from a finite population. A remainder term estimate," *Scandinavian Journal of Statistics*, 5, 69–71.

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

Holland, P. W. and Rubin, D. B. (1989), "Causal inference in retrospective studies," *Evaluation Review*, 12, 203–231.

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

Imbens, G. W. and Rosenbaum, P. R. (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+.

Kalton, G. (1968), "Standardization: A technique to control for extraneous variables," *Applied Statistics*, 17, 118–136.

Lee, Y. J., Ellenberg, J. H., Hirtz, D. G., and Nelson, K. B. (1991), "Analysis of clinical trials by treatment actually received: Is it really an option?" *Statistics in Medicine*, 10, 1595–1605.

Nie, N., Junn, J., and Barry, K. S. (1996), *Education and Democratic Citizenship in America*, Chicago: University of Chicago Press.

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

— (2002a), *Observational Studies*, New York: Springer-Verlag, 2nd ed.

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

— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

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

— (2002b), "Attributing Effects to Treatment in Matched Observational Studies," *Journal of the American Statistical Association*, 97, 183–192.

— (2002c), "Attributing effects to treatment in matched observational studies," *Journal of the American Statistical Association*, 97, 183–192.

— (2005), "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies," *The American Statistician*, 59, 147–152.

Rosenstone, S. and Hansen, J. M. (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.

Verba, S., Schlozman, K. L., and Brady, H. (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, R. and Rosenstone, S. (1980), *Who Votes? (Yale Fastback Series)*, Yale University Press.