Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign

Ben B. Hansen and Jake Bowers[1]

August 18, 2009

Abstract

Early in the twentieth century, Fisher and Neyman demonstrated how to infer effects of agricultural interventions using only the very weakest of assumptions, by randomly varying which plots were to be manipulated. Although the methods permitted uncontrolled variation between experimental units, they required strict control over assignment of interventions; this hindered their application to field studies with human subjects, who could not ordinarily be compelled to comply with experimenters' instructions. In 1996, however, Angrist, Imbens and Rubin showed that inferences from randomized studies could accommodate non-compliance without significant strengthening of assumptions.

Political scientists A. Gerber and D. Green responded quickly, fielding a randomized study of voter turnout campaigns in the November 1998 general election. Non-contacts and refusals were frequent, but Gerber and Green analyzed their data in the style of Angrist *et al.*, avoiding having to model non-response. They did use models for other purposes: to address complexities of the randomization scheme; to permit heterogeneity among voters and campaigners; to account for deviations from experimental protocol; and to take advantage of highly informative covariates. Although the added assumptions seemed straightforward and unassailable, a later analysis found them to be at odds with Gerber and Green's data. Using a different model, it reaches the very opposite of Gerber and Green's central conclusion about getting out the vote.

This paper shows that neither of the models are necessary, addressing all of the complications of Gerber and Green's study using methods in the tradition of Fisher and Neyman. To do this, it merges recent developments in randomization-based inference for comparative studies with somewhat older developments in design-based analysis of sample surveys. The method involves regression, but large-sample analysis

[1] Ben Hansen is Assistant Professor, Statistics Department, University of Michigan, Ann Arbor (ben.b.hansen@umich.edu). Jake Bowers is Assistant Professor, Department of Political Science and NCSA, University of Illinois at Urbana-Champaign (jwbowers@illinois.edu). They are grateful for helpful discussions arising from several seminars in which they presented parts of this work. They also thank Michael Elliott, Donald Green, Kosuke Imai, Andrew Gelman, Roderick Little, David Nickerson, Shawn Treier and other discussants at the Society for Political Methodology 2005 conference, two anonymous referees, an anonymous associate editor and the editor, David Banks, for helpful comments and suggestions. Any errors or shortcomings are their own. Supported in part by The Robert Wood Johnson Foundation and by the NSF (# DMS-0102056). A compendium for reproducing all of the results reported in this paper may be downloaded from http://hdl.handle.net/1902.1/12174.

and simulations demonstrate its lack of dependence on regression assumptions. Its substantive results have consequences both for the design of campaigns to increase voter participation and for theories of political behavior more generally.

Key words: cluster randomization, group randomized trial, instrumental variable, model-assisted, randomization inference, voter turnout

1 Randomization in field studies of political participation

In a landmark study of political participation, A. Gerber and D. Green (2000) experimentally assessed effectiveness of get-out-the-vote (GOTV) appeals delivered over the telephone, by mail, and through personal contact. Their "Vote 98" study was large and well powered, conducted not in a lab but under field conditions in New Haven, Connecticut, prior to the 1998 congressional election; it used random assignment of interventions, in a discipline where randomization was rare. Random assignment had been used previously in the study of getting out the vote (Gosnell 1927; Eldersveld 1956; Adams and Smith 1980; Miller et al. 1981), but the design had limited appeal because potential voters assigned to intervention could never consistently be contacted, with the result that the eventual statistical analysis seemed to require assumptions going beyond randomization. Angrist, Imbens, and Rubin (1996) had recently established, however, that this was not so, that by treating random assignment as an instrumental variable one could address unintended non-receipt of treatment with few additional assumptions. The Vote 98 study was the first to marshal this advance for the study of political participation. By showing that the inevitability of non-contact could so elegantly be addressed in this context, it appears to have sparked a small renaissance in randomized studies of getting out the vote (Michelson 2003; Smith et al. 2003; Clinton and Lapinski 2004; Arceneaux 2005; Wong 2005; McNulty 2005; Nickerson et al. 2006; Niven 2006; Nickerson 2006).

Comparing different modes of getting out the vote in the same election and on the same population, Gerber and Green's study remains unique and of substantive interest, particularly given its notable conclusion that paid phone banks, a method of choice for many modern campaigns, were far inferior to personal contact. This conclusion has been called into question by Imai (2005), who also established that instrumental variables were not in themselves enough to address the various complications of Gerber and Green's data. Subjects assigned to treatment less resembled controls than should have been the case had they been a simple random sample of the overall experimental pool. Implementation, particularly of the telephone intervention, had been inconsistent, leading to ambiguity as to who precisely should be regarded as the treatment group. All this led Imai to question the study's randomization and ultimately reject it as "failed" (p. 285, 291). His alternate analysis sets aside assignment to treatment and control, instead propensity-matching to controls only subjects actually contacted by the campaign. Contra Gerber and Green, but consonant with common assumptions of political practice, it finds statistically and materially significant benefits for the telephone intervention. His and Gerber and Green's incompatible conclusions have contradictory ramifications for both the theory and practice of voter mobilization (Gerber and Green 2000, 2005; Imai 2005).

Methodological as well as substantive concerns are at stake in this debate. An analysis like Imai's requires the assumption that, by adjusting for available covariates, contacted voters can be rendered equivalent to non-contacted ones, so far as their eventual voting is concerned — an assumption about voting, not just about the manner of assignment of interventions. To be sure, in many studies there is little hope of progress without such substantive assumptions; but a central attraction of randomized studies is the possibility of doing without them, instead relying only upon the randomization itself as the "reasoned basis for inference" (Fisher 1935; see also Neyman 1990). If, once the inevitable complications of implementation have all been accounted for, analysis of the Vote 98 study requires meaningful assumptions about political behavior, then perhaps the benefits of randomization for field studies are more limited than experimentalists have come to think.

To illustrate that this is not the case, and to illuminate the substantive disagreement between Imai and Gerber and Green, this paper applies randomization-based inference to the Vote 98 study. We demonstrate presently that inference of this type is capable of assessing the magnitude as well as the statistical significance of the treatment effect, and (in § 2) that it can address *all* the lapses and inconsistencies known to have occurred in the New Haven Vote 98 experiment, without requiring special assumptions to do so. To be valid, inferences about treatment effects must be attentive to the manner in which randomization was carried out, respecting such features as stratification and cluster-level assignment; to be powerful, they should draw assistance from the several available covariates that potently predict voting. Similar challenges arise in survey sampling, and in Section 3, we adapt to this setting randomization-based methods of survey analysis. Section 4 addresses substantive questions around which debate about the experiment has centered and, in a demonstration of the power of this approach, brings into focus our understanding of how certain subgroups' voting was affected. Discussion appears in Section 5.

1.1 Votes attributable to treatment in a simple randomized turnout experiment

In 1978 Marion Barry became Mayor of Washington, D.C., leaving the city with a vacant seat on its city council. Before a special election to fill Barry's seat, Adams and Smith (1980) arranged to call $n = 1325$ subjects, soliciting their votes on behalf of one of the candidates, John Ray. These subjects had been randomly selected from a pool U of $N = 2650$ potential voters, none sharing a household, for whom turnout would later be determined from public records. Because the experiment is smaller and simpler than Gerber and Green's, and because it gives evidence that in its day, at least, brief messages from paid callers effectively got out the vote, we use it to illustrate the basis of our approach.

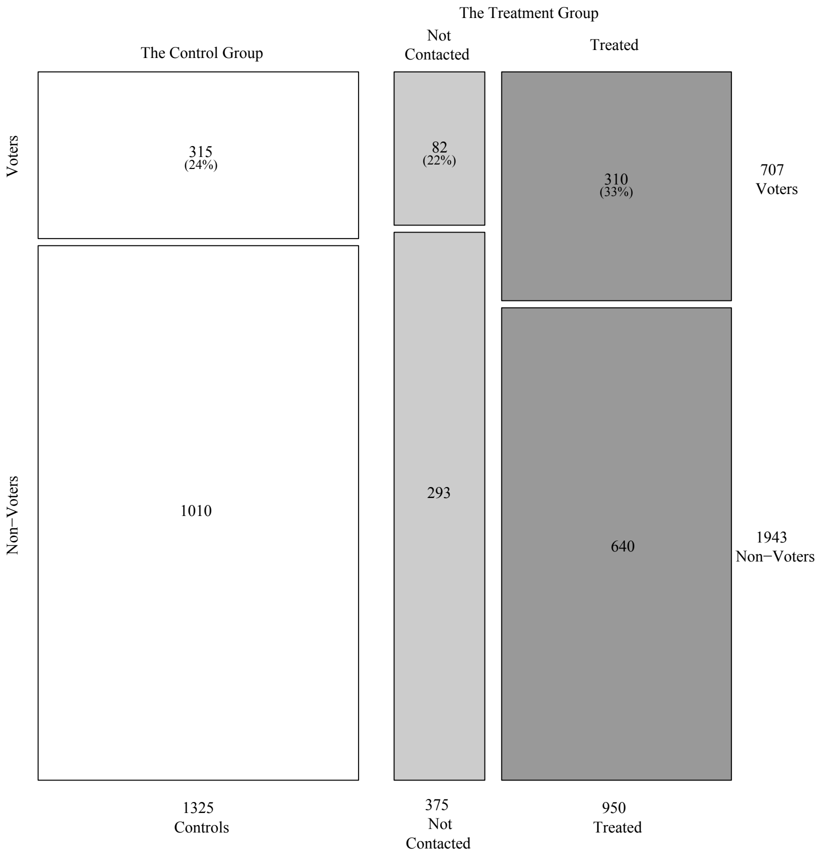

Figure 1: Assignment, compliance and voting for the Adams and Smith(1980) telephone field experiment. The columns labeled "Not Contacted" and "Treated" contain those subjects who were assigned to treatment but who either did not answer the telephone or did answer the call, respectively. Relative sizes of tiles reflect shares of the experimental pool (Hartigan and Kleiner 1984; Friendly 1994). For example, 315/1325≈24% of controls voted, and controls constituted 50% of experimental subjects, so the tile representing voting controls occupies 315/2650≈12% of the total area of the plot.

In the half-sample randomized to control, 315 subjects, or 23.8%, voted in the special election (Figure 1). Treating the control group, C, as a sample from U, the experimental universe, one estimates unbiasedly that 23.8% of subjects in U would have voted had none of them been called. It so happens, however, that 29.6% of the treatment group voted, so that in all 26.7% of U voted. Does the difference indicate the treatment had an effect, or could it be due to chance? For any $B \subseteq U$ denote by $\bar{r}_B$ the mean of r's in B, $|B|^{-1} \sum_{i \in B} r_i$, so that the proportion of controls voting was $\bar{r}_C = .238$. (Here "$|B|$" indicates the number of elements in B.) Let $\mathbb{C}$ be the set of all samples from U of size $n = |C|$, $\mathcal{C}$ a random subset of U drawn with uniform probability from $\mathbb{C}$. Elementary theory of survey sampling (Kish 1965, § 2.2–3; Cochran 1977, § 2.4–7; Lohr 1999, § 2.7) yields: $\mathbf{E}(\bar{r}_{\mathcal{C}}) = \bar{r}$, $V(\bar{r}_{\mathcal{C}}) = (\text{fpc}) * s^2[\mathbf{r}]/n$, and $\mathbf{E}s^2[(r_i : i \in \mathcal{C})] = s^2[\mathbf{r}]$, where $N = |U| = 2650$, $\mathbf{r} = (r_i : i \in U)$, (fpc) is the finite population correction $(1 - n/N)$, and $s^2[(r_1, \ldots, r_J)] = (J-1)^{-1} \sum_1^J (r_j - \bar{r})^2$; furthermore $\hat{V}(\bar{r}_C) = (1 - n/N)s^2[(r_i : i \in C)]/n$ is the natural estimate of $V(\bar{r}_C)$. With the finite-population central limit theorem (Hájek 1960), these facts suggest $\bar{r}_C \pm 1.96\hat{V}^{1/2}(\bar{r}_C) = .238 \pm 1.96(.0083) = [.222, .254]$ as an approximate 95% confidence interval for the overall proportion of subjects who would have voted even if none of the calls had been placed.

Evidently, sampling variability alone does not explain the difference in voting between Adams and Smith's treatment and control groups, as U's 26.7% turnout rate falls well outside of this confidence interval. At least some portion of the difference must be attributed to Adams and Smith's intervention. How much? If $2650 * [.222, .254]$, or from 587 to 673 of U's 2650 members, would have voted absent the GOTV calls, whereas in fact 707 of them voted, *then it follows that at least 34* (=707-673) *and as many as 119* (707-587) *of those votes can be attributed to treatment.* This is a 95% confidence interval for A, the *attributable effect* (Rosenbaum 2001). A point estimate is $707 - .238 * 2650 = 77$ votes. In other words, Adams and Smith's turnout campaign raised turnout by something between $34/1325 = 2.6\%$ and $119/1325 = 9.0\%$, with 95% confidence.

These statements make no claim about the efficacy of GOTV calls in general. They attribute effects to a particular intervention, Adams and Smith's 1978 turnout campaign; to a particular experimental universe, Adams and Smith's 2650 study subjects; and to a particular treatment group, those 1325 subjects the experiment selected for GOTV. This attributable effect is inherently an in-sample quantity. It relates closely, however, to more familiar targets of causal inference. The quantity $A/1325$ is equal in expectation to the "intention-to-treat effect," (ITT) parameter for Adams and Smith's 2650 subjects (and arguably for superpopulations of which they are representative). Together with data on the number of treated subjects, subjects who both were assigned to treatment and later received it, our inferences about A also speak to the effect of treatment *per se.* It follows that the ratio of votes spurred by treatment, A, to the number of subjects treated, O, lies between $34/950 = .036$ and $119/950 = .125$ — between 3.6% and 12.5% of experimental contacts effected a vote. The closely related parameter $\mathbf{E}A/\mathbf{E}O$ is sometimes called the effect of treatment on the treated, or "ETT" (Heckman 1997; see also Rosenbaum and Rubin 1985). However the result is presented, it appears that brief, scripted GOTV calls produced benefits of both statistical and material significance — at least in one special election in 1978.

Note carefully that the form of analysis just given relies only on the integrity of Adams and Smith's data, and on their having faithfully executed their maintained experimental design — no statistical model of the response variable is assumed, nor are non-contacted treatment-group subjects assumed exchangeable with controls. In both of these respects it differs from Adams and Smith's analysis. Their analysis compared to the control group only subjects to whom calls were successfully placed — *the treated*, a proper subset of *the treatment group,* the larger collection of subjects experimenters intended to contact by telephone (Figure 1). This type of comparison would be misleading, despite the randomization, had subjects who would have voted even if not called by the campaign been easier to reach than their non-voting counterparts. Consistent with the "intention-to-treat" principle (Lee et al. 1991), our alternate approach ensures parity by comparing treatment and control groups as randomized, irrespective of whether contact with treatment subjects was made.

Now Adams and Smith's analysis suggested a much greater turnout benefit than ours, a boost of nearly 40%. The discrepancy between these and our randomization-based results suggests that those subjects who would have voted whether reminded to or not took the campaign's calls in greater proportions than voters who needed reminding, a circumstance that would bias Adams and Smith's analysis but not ours. Imai's analysis of the Vote 98 experiment is protected against such bias to some extent, because it propensity-matched treated subjects to controls; but since within matched sets it compares the treated to controls, it remains vulnerable to a bias related to Adams and Smith's, in the event that conditioning on measured covariates does not suffice to make treated Vote 98 subjects — subjects who not only were assigned to intervention but also received it — exchangeable with Vote 98 controls.

1.2 Adapting design-based survey methods to experiments

In order to attribute effects to treatment, the only quantity about which one must draw statistical inferences is $\bar{y}_{cU}$, the average (over all of U) of outcomes that would have resulted had each study subject received the control condition. It, or rather the multiple $N\bar{y}_{cU} = \sum_U y_{ci}$ of it, is compared to $\sum_U y_i$, a quantity that is fully observed. When $\mathcal{C}$ is a probability sample from U, methods from survey sampling become available for estimation of $\sum_U y_{ci}$. Such complications as random assignment of groups rather than individuals and assignment within blocks map to common features of sample surveys, cluster-level selection and selection within strata, the consequences of which are well understood. When there are covariates, a mature literature establishes that randomization-based inference can borrow from model-driven covariate adjustment to improve precision (Isaki and Fuller 1982; Särndal et al. 1991; Firth and Bennett 1998). We shall bring both of these benefits to bear on the Vote 98 controversy.

Might something be lost by moving from methods designed for experiments to methods designed for surveys? One concern is that permutation-based inference for experiments can often be done exactly, whereas design- or randomization-based inference in surveys is generally not exact. The analysis of § 1.1, for example, involves two layers of approximation, neither of which would be invoked by an exact calculation:

L1. The distribution of the sample mean $\bar{y}_{\mathcal{C}}$ is approximated as Normal; and

L2. $V(\bar{y}_{\mathcal{C}}) = (1-\frac{n}{N})s^2[(r_i : i \in U)]/n$ is estimated by $\hat{V}(\bar{y}_{\mathcal{C}}) = (1-\frac{n}{N})s^2[(r_i : i \in \mathcal{C})]/n$.

Covariate adjustment will necessitate a further layer of large-sample approximation, to be discussed in § 3.

We studied the performance of these approximations in some detail. The results, many of which are to be given in this paper, support a methodological hypothesis to the effect that for simply- or block-randomized experiments like Gerber and Green's (2000), the combined approximation error is negligible. This hypothesis, call it H_M, carries the provisions that: (a) the experiment be relatively large, in terms of the number of units it independently assigned to treatment; (b) that if the outcome is binary then, in the absence of the treatment, it should be neither overwhelmingly common nor overwhelmingly rare; and (c) that the fraction of units assigned to control not be overly small, so that the control group is made large enough to be informative about both means and variances. Proviso (a) addresses L1, whereas provisos (b) and (c) address L2, by heading off known shortcomings of Wald-type variance approximations with small samples (Zheng and Little 2005, § 4, Elliott 2008, § 4.1) and with binary data (Brown et al. 2001).

The analysis of § 1.1 depends on L1 and L2, and as such offers a first test for H_M. Let us determine and evaluate the exact coverage probabilities of § 1.1's asymptotic 95% confidence interval. Write r_{ci} for subject i's potential response to the control condition; then $\bar{r}_C$ estimates $\bar{r}_{cU}$, a parameter that takes one of the values $\{397/2650, 398/2650, \ldots, 1347/2650\}$. In asserting this we assume an *exclusion restriction* (Angrist et al. 1996; Rosenbaum 1996), that r_i can differ from r_{ci} only for contacted subjects i. Since 397 votes were cast by the 1700 controls and treatment-group non-contacts who did not receive a GOTV call, our exclusion restriction entails that at least 397 and no more than $397+(|U|-1700) = 1347$ of the $|U| = 2650$ subjects would have voted absent the intervention. Some algebra shows that

$$\bar{r}_{cU} \in \bar{r}_{\mathcal{C}} \pm z_* \hat{V}^{1/2}(\bar{r}_C) \Leftrightarrow \bar{r}_{\mathcal{C}} \in \frac{\bar{r}_{cU} + c/2}{1+c} \pm c^{1/2} \frac{\sqrt{\bar{r}_{cU} - \bar{r}_{cU}^2 + c/4}}{1+c},$$

where $c = z_*^2 n^{-1}(1-n/N)N(N-1)^{-1}$. By evaluating the hypergeometric probability mass associated with this range of $\bar{r}_{\mathcal{C}}$s, we determined the *a priori* probability that $\bar{r}_{cU} \in \bar{r}_{\mathcal{C}} \pm 1.96\hat{V}^{1/2}(\bar{r}_C)$ for each value of $\bar{r}_{cU}$ not excluded by the data and the exclusion restriction. As the parameter $\bar{r}_{cU}$ varies across its feasible range, coverage probabilities fluctuated about a median value of .950, from as low as .944 (for $\bar{r}_{cU} = 632/2650$) to as high as .955 (for $\bar{r}_{cU} = 634/2650$) — a result that supports H_M. Further corroboration appear in § 3.

2 The randomization basis for analysis of Vote 98

The Vote 98 experiment was more complex than Adams and Smith's, with a much larger sample, multiple interventions and randomization that involved both stratification and clustering, not to mention unintended shortcomings of implementation. This section reviews the design and implementation of the Vote 98 experiment, exploring whether and how complications like those occurring in it can be addressed with randomization-based modes of inference.

2.1 Design of the Vote 98 study

From official records, Gerber and Green assembled a complete list of registered voters in New Haven as of September 1998. To isolate the non-student population, they excluded voters from the ward containing Yale University and many of its students, as well as those at addresses listing three or more registered voters and those without a street address; the remaining 31,100 subjects, residing in 22,450 households within the 29 remaining wards, constitute U, the universe of the Vote 98 experiment. (Our description is based on "2005 release" data posted to D. Green's Web site, which differ from earlier releases of the data in incorporating household identifiers, subjects dropped from the rolls after November 1998, and additional data cleaning, as described by [Gerber and Green 2005].) Postcards containing GOTV messages were randomly sent to half of the households, with the number of mailings varied at random between 1, 2 and 3. One-tenth of those households that were not sent a mailer were randomly selected to also be targeted for GOTV by telephone. Among households to which a mailer was sent, telephone contact was also attempted, but at a higher rate, with 40% randomized to telephone GOTV. Viewed unto itself, the telephone sub-experiment is randomized within blocks but not simply randomized, with mailed and unmailed blocks; likewise, mail was in effect block-randomized, with blocks defined by whether telephone GOTV calls were and were not attempted. A third form of intervention, in-person entreatment at potential voters' doors, was randomly assigned to 1/5 of the same pool, but this randomization was independent of the other two. A household could have been slated for no intervention or for any combination of interventions, up to and including mailers, multiple attempts at telephone contact over the three days up to and including the election, and a weekend personal visit during the month prior to the election; all of these combinations of experimental assignments occurred. The overall situation is depicted in Figure 2.1, which also speaks to compliance with assigned treatment.

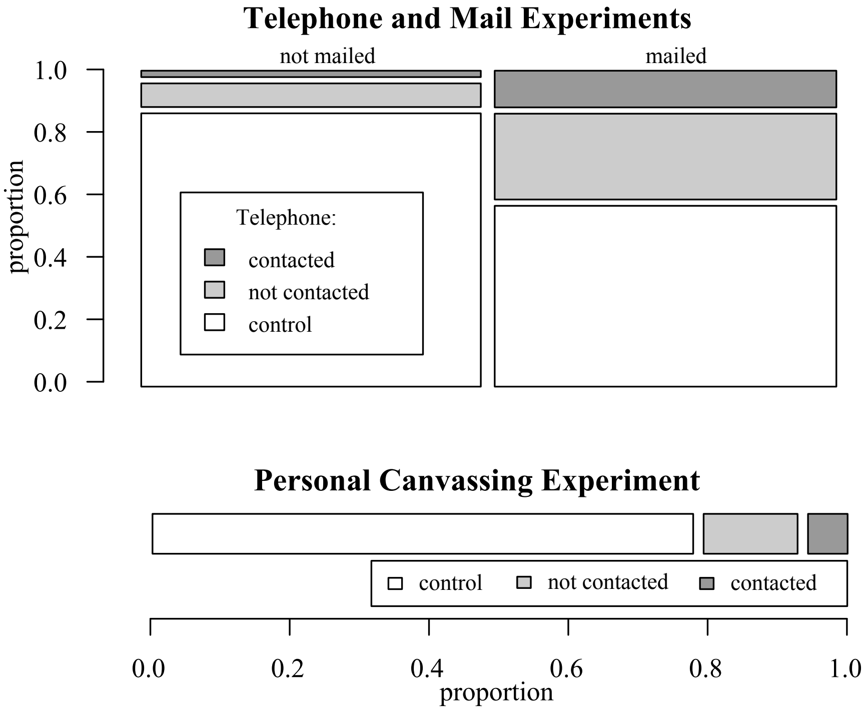

Figure 2: Assignment and compliance for mail, telephone and personal canvassing experiments. Relative sizes of tiles reflect proportions of households in the sample.

Compliance with telephone and in-person assignments was measured at the household level, with a household treated as complying if contact was made with any one of its members. Telephone GOTV calls were placed successfully to 28% of households randomized to the telephone condition, while personal contact was successful for 30% of households randomized to it. About 10% of those who could not be reached at their doors had leaflets left for them by canvassers, and roughly 15% of them instead were mailed a refrigerator magnet with the election date printed on it, a subsidiary intervention that for the purpose of inference about treatment effects must be regarded as part of the in-person appeal. These intervention supplements complicate interpretation of intervention effects, but since they were withheld from the group not randomized to personal canvassing, they are no threat to inference in the style of § 1.1 on intervention effects' presence and magnitude. Likewise, there was an irregularity in administering the telephone message, such that 10% of households assigned to telephone persuasion never were called with a GOTV message. (They were called, but with a script urging participation in a blood drive.) Whereas Imai's and Gerber and Green's analyses both treat these subjects as controls, ours regards them as non-complying intervention group members, hewing to the design. No measure of compliance is available for the mail intervention.

Some 5% of subjects drawn into the Vote 98 study pool from pre-election registration lists appeared neither as voters nor non-voters in official records of the 1998 election. Missing outcomes of this type are typical of voting data, as registrars may only infer when a voter has moved or passed away from repeated non-voting. Our analysis interprets these as non-votes, treating them the same as subjects coded as non-voters in 1998 election records.

2.2 Baseline comparability of treatment groups

In appraising experimental assignments to treatment or control, one seeks assurance that subjects slated for the two conditions are similar, or at least as similar as can be expected given the form of randomization used Raab and Butcher (2001, *e.g.*). The analogous question in surveys is whether a sample is representative of all units appearing in the sampling frame.

The Vote 98 study's covariates include voting in the prior election, registration at the time of the prior election, registration with either of the two major parties, whether the voter lives in a 1- or 2-voter household (households with 3 or more voters having been excluded), voter age, and which of the 29 non-student wards the voter resides in. Age information was available for more than 99% of voters, and other variables were always available; we handled missing ages by median imputation. Because the age variable was quite skewed, with one potential voter as old as 106, and because of its great value as a predictor of voting (Highton and Wolfinger 2001; Wolfinger and Rosenstone 1980), we decomposed it using a natural cubic spline with knots at quintiles of the age distribution, comparing the "sample," C, to the "sampling frame," U, in terms of the B-spline basis for this decomposition, rather than in terms of age itself. In a limited sense, compliance information can also be regarded as a covariate. Since the completion or non-completion of attempted telephone contacts is not plausibly influenced by independently-assigned personal interventions, having received the telephone intervention is presumptively a covariate, a variable not influenced by assignment to treatment conditions, from the perspective of the personal canvassing sub-experiment — although from the perspective of the telephone GOTV sub-experiment it can certainly be influenced by treatment assignment. Likewise, having made a personal GOTV appeal is a covariate for the telephone- and mail-GOTV sub-experiments, but not for the personal canvassing experiment.

To compare covariates in the sub-experiments' control groups to those of the experimental universe as a whole, we use the same method as was used in § 1.1, but this time to estimate $\sum_U x_i$, for various covariates x. In light of the treatments' assignment by household, we take U to be the experimental universe of households, not individuals; individual-level covariate measurements x_{ij} are summarized by household totals, $x_i = \sum_j x_{ij}$, in these calculations. In the case of the in-person experiment, then, we estimate covariate totals $\sum_{i \in U,j} x_{i,j} = \sum_U x_i$ by $N\bar{x}_C \pm z_* N\hat{V}^{1/2}(\bar{x}_{\mathcal{C}})$, $\hat{V}(\bar{x}_{\mathcal{C}}) = (1 - n/N)s^2[(x_i : i \in C)]/n$, with $s^2[.]$ as defined in § 1.1. In light of the telephone experiment having been randomized in blocks, totals of x are estimated separately in each block B and then added across blocks, as are the associated variance estimates $|B|^2\hat{V}(\bar{x}_{\mathcal{C}\cap B})$, to estimate the overall total and its error of estimation. Estimates of subject-level means in x, as shown in Figure 3, result from rescaling these estimated totals by the reciprocal of M, the number of subjects in the experiment.

This method accounts for the fact that randomization was performed at the household level, and so can be expected to be somewhat less effective at balancing the groups than individual-level randomization would have been. In contrast, Imai's conclusion that the Vote 98 study's randomization had failed followed from checks of group comparability that did not account for household-level randomization. (The household identifiers that we use here were not publicly available when his analysis was conducted.) Were we to do the same, the centering points of our confidence intervals would not have been substantially affected, but the intervals would have been too narrow. Extrapolating to the experimental universe from subjects the telephone sub-experiment assigned to control, 2 of the 39 interval estimates of baseline means in x would fail to cover their targets; in the extrapolation from subjects not assigned to in-person GOTV, 4 of 39 such 95% confidence intervals would fail to cover their targets. *Prima facie,* such results would suggest a problem with the randomization, but in truth they would only show that it had been held to an inappropriate standard. See Hansen and Bowers (2008) for more discussion of baseline comparability in cluster-randomized experiments.

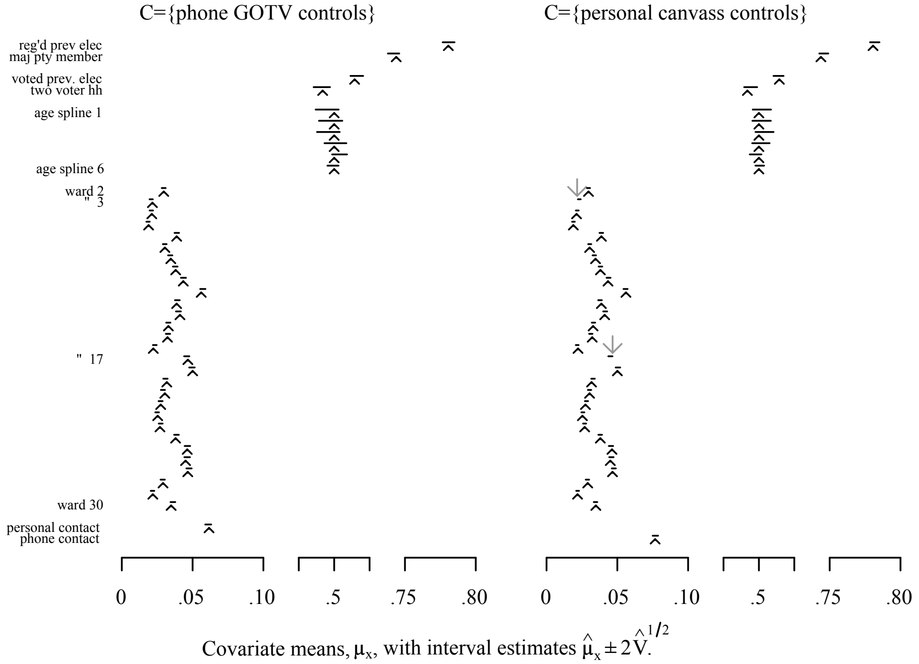

Figure 3: Control groups' representativeness of the experimental universe, in the telephone GOTV and personal canvassing sub-experiments. Arrowheads represent means over all of U, with the horizontal bars they point to giving intervals $\hat{\mu}_x \pm 2\hat{V}^{1/2}(\hat{\mu}_x)$ calculated from C. The larger, downward-pointing arrows indicate means not covered by corresponding interval estimates. Age spline loadings have been centered and re-scaled; for all other variables, scale is indicated on the lower horizontal axis. The 80 interval estimates that result should carry 95% confidence; consistent with this, all but 2 of them contain their targets.

As Figure 3 suggests, analysis that does account for randomization at the household level gives a different and more favorable picture than such an examination at the individual level. The figure compares covariate averages over the experimental universe to interval estimates of those averages arising by the application of our method to the telephone GOTV control group and to personal canvassing controls. With only 2 exceptions, extrapolations $\hat{\mu}_x \pm 2\hat{V}^{1/2}(\hat{\mu}_x)$ from the sample include their targets μ_x. These misses occurred for relatively skewed, binary variables, residence in wards 3 and 17, and they may reflect the known difficulty of our Wald-type confidence procedures with such variables, even in quite large samples (Brown et al. 2001). (This possibility motivates our proviso (b) in § 1.2, that the main estimand not be a binary variable with mean close to 0 or 1.) In any case, given that the figure shows some eighty 95% confidence intervals, it is to be expected that a few would exclude their estimands. Overall, the results cast no aspersions on the Vote 98 study's randomization, nor on the comparability of experimental and control groups it produced.

2.3 Assumptions

Likely heterogeneity of treatment effects; exclusion restriction. The many callers and field workers contributing to a political campaign may do so with varying effectiveness, given differences in their experience and motivation as well as differences among potential voters. While it is appropriate that speculation about these factors should inform the experimental protocol — the Vote 98 campaign, for example, attempted to match the race of its canvassers to the neighborhoods in which they would be working, perhaps enhancing the quantity or quality of voter contacts — they may be difficult to parameterize reliably at the stage of analysis. Accordingly, our analysis seeks to minimize assumptions about intervention effects. It does, however, impose the exclusion restriction, here interpreted as the requirement that intervention effects are experienced only within households that received the intervention, so that $r_{ij} = r_{cij}$ unless i was an intervention household (Rosenbaum 1996).

No interference between households. The Vote 98 campaign randomized households rather than persons. Accordingly, we shall assume that intact households, but not individuals considered in isolation from their households, have stable unit treatment values (Rubin 1986), in that their outcomes may be determined by experimental interventions they receive but not by what interventions are delivered to other households. The analysis will allow cohabiting subjects' voting decisions to be correlated in arbitrary ways, with or without the treatment, a possibility that Gerber and Green's (2000) and Imai's (2005) models (if not Gerber and Green's [2005]) would deny.

Stability of non-focal interventions across possible assignments of the focal intervention. When there are other experiments in the same field, randomization-based assessments of an intervention's effects require neither that its intervention subjects nor its controls be protected from the other interventions. Instead, they require assignment of the focal intervention — not receipt, only assignment — to have been independent of both assignment and receipt of the other interventions. A GOTV effect observed against a backdrop of spirited campaigning may merit a different substantive interpretation than an effect of similar interventions observed in quiet political season, but from the randomization perspective the two inferential problems are the same. Likewise, viewing the New Haven Vote 98 study as a union of sub-experiments on GOTV by mail, by telephone and in person, our randomization analysis of each experiment conditions on the realized treatment assignments of the others. For analysis of the mail experiment, for instance, this means conditioning on assignments to the telephone intervention, which define the two blocks within which mail can be regarded as simply randomized.

A random variable as estimand. If, as is true of each of the Vote 98 sub-experiments, no subjects randomized to control received the intervention, then $a = \sum_{i \notin C} r_i - r_{ci}$. As $\mathbf{r}$ depends on which subjects receive the intervention, one could also write $a = \sum_{i \notin C} r_i(C) - r_{ci}$ — a representation emphasizing that a is the value of a random variable, $A = \sum_{i \notin \mathcal{C}} r_i(\mathcal{C}) - r_{ci}$, not a parameter. Since its value is determined by observed data in conjunction with the parameter $\sum_i r_{ci}$, however, inference about it is logically equivalent to inference about $\sum_i r_{ci}$, and can be made by conventional means.

Comparison with assumptions of other methods for cluster-randomized data. Other ways of accounting for clustered treatment assignment and binary outcomes include the empirical Bayes methods of Raudenbush (1997) and Murray (2001), the Bayesian approach of Thompson et al. (2004), which commit to models for the response as a function of covariates, and the randomization-based method of Braun and Feng (2001), who model the treatment effect as constant on a log-odds scale. Their setups all require modeling the effect of assignment to treatment, or intention-to-treat effect. In contrast, the present method supposes subjects to be characterized by deterministic indicators r_{cij} of whether they would have voted had the experiment not occurred, and adopts the limited goal of inferring the magnitude of $a = \sum_{i \in U} r_i - r_{ci}$, the sample-aggregate increase in voting attributable to treatment. It does not culminate in odds ratios, which can be difficult to relate to more readily interpretable parameters (Greenland 1987); nor make assumptions, other than the exclusion restriction, about intention-to-treat effects; nor require homogeneity of intervention effects across groups or subgroups of individuals.

It remains to be seen whether the method can retain these advantages while utilizing covariates to improve precision. Section 3 accomplishes this using standard regression techniques. Perhaps surprisingly, it also avoids the modeling assumptions that regression ordinarily requires.

3 Large-sample methods for experiments with covariates

Provided that households, rather than individuals, are taken as the unit of analysis, the method by which § 1.1 attributed votes to Adams and Smith's telephone intervention now applies directly to experiments like Gerber and Green's. Denote household i's observed turnout by r_i, and denote by r_{ci} its turnout had treatment been withheld (so that $r_i = r_{ci}$ for all $i \in C$, but r_i may differ from r_{ci} if $i \notin C$). We can estimate each intervention's effect on turnout as the difference between the total observed turnout, $\sum_U r_i$, and the estimate of total turnout one would extrapolate from its control group. As the in-person intervention was directed to a simple random sample of households, for it $\bar{r}_{\mathcal{C}}$ estimates the average votes per household in the absence of intervention, $\bar{r}_{cU}$, with variance approximately $\hat{V}(\bar{r}_{\mathcal{C}}) = (1 - n/N)s^2[(r_i : i \in \mathcal{C})]/n$, making $\sum_U r_i - N\bar{r}_C \pm N z_{\alpha/2}\hat{V}^{1/2}(\bar{r}_{\mathcal{C}})$ an approximate $(1-\alpha)*100\%$ confidence interval for the number votes won by the personal canvassing campaign. While the mail and telephone intervention groups are not simple random samples from U, they are unions of simple random samples, from blocks contained in U; the method applicable directly to the in-person experiment can be applied to separately to each block, after which both vote attributions and associated variances can simply be added across blocks.

As noted by Imai (2005), covariates in Vote 98 study were quite rich; age and prior voting, for example, are each important predictors of voting. The present section develops a method of extracting additional precision from them, modeled on the "design-based, model-assisted" approach to survey analysis. It uses regression adjustment, although the inferences it yields continue to flow from the strict logic of randomization alone, not regression modeling assumptions (Särndal et al. 1991, § 6.7). The approach is related to methods of regression adjustment for comparative studies discussed by Rosenbaum (2002), but differs from those in depending on large-sample approximations and in being somewhat simpler to implement. Our exposition of it is progressively more methodological than substantive in focus; readers interested primarily in our conclusions about voting can skip to § 4 from any point in § 3.

3.1 Known regression coefficients

Let $\mathcal{C}$ represent a simple random sample from U, and let $\hat{\mathbf{r}}_c(\cdot)$ be a function mapping regression parameters $\beta \in \Re^K$ to vectors of predictions $(\hat{r}_{ci}(\beta) : i \in U)$. Covariates $\mathbf{x}$ may play a role in determining $\hat{\mathbf{r}}_c(\beta)$, although this is suppressed in the notation. For example, in the analysis to follow $\hat{r}_{ci}(\beta), i \in U$, is defined by $\text{logit}(\hat{r}_{cij}) = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_K x_{Kij}$, each j in cluster i, and $\hat{r}_{ci}(\beta) = \sum_j \hat{r}_{cij}$. For this section only, peg β to a fixed position in regression-parameter space, the same position whatever $\mathcal{C} \subseteq U$ is chosen as the control group.

Writing $e_i(\beta) = r_{ci} - \hat{r}_{ci}(\beta)$, we simply regard $(e_i(\beta) : i \in \mathcal{C})$ as a sample from $(e_i(\beta) : i \in U)$, estimating $\bar{e}_U(\beta)$ with $\bar{e}_{\mathcal{C}}(\beta)$. Just as in § 1.1, a large-sample 95% confidence interval for $\bar{e}_U(\beta)$ is $\bar{e}_{\mathcal{C}}(\beta) \pm z_* \hat{V}^{1/2}(\bar{e}_{\mathcal{C}}(\beta))$, where $\hat{V}(\bar{e}_{\mathcal{C}}(\beta)) = (1 - n/N)s^2[(e_i(\beta) : i \in \mathcal{C})]/n$. The aim is to estimate $\mu_c = M^{-1} \sum_U r_{ci}$, the fraction of all M study subjects who would have voted absent the intervention, not $\bar{e}_U(\beta)$; but since $\bar{r}_{cU} = \bar{\hat{r}}_{cU}(\beta) + \bar{e}_U(\beta)$, the estimator

$$\hat{\mu}_c(\beta) = \frac{N}{M}(\bar{\hat{r}}_{cU}(\beta) + \bar{e}_{\mathcal{C}}(\beta)) \tag{1}$$

follows directly. Here $\bar{\hat{r}}_{cU}(\beta)$ is the average of $(\hat{r}_{ci}(\beta) : i \in U)$, a nonrandom quantity, so that the standard error of $\hat{\mu}_c(\beta)$ is N/M times the standard error of $\bar{e}_{\mathcal{C}}(\beta)$.

Observe that the argument just given avoids assuming that the "true" or "correct" regression of $\mathbf{r}_c$ on $\mathbf{x}$ is the inverse logit of $\mathbf{x}\beta$. Nor is there any need that the predictions $\hat{\mathbf{r}}_c(\beta)$ address correlations of response within a cluster; these issues have been addressed by aggregating residuals and predictions to the cluster level before estimating $\hat{\mu}_c(\beta)$ or its error.

3.2 Estimated regression surface

Although β can be chosen arbitrarily, it is advantageous to select it so as to maximize the quality of predictions of $\mathbf{r}_c$. This intuitive claim may be justified by observing that $V(\hat{\mu}_c) \propto s^2[(e_i(\beta) : i \in U)]$, where $e_i(\beta) = r_{ci} - \hat{r}_{ci}(\beta)$, and that $s^2[(e_i(\beta) : i \in U)]$ directly reflects how well $\hat{\mathbf{r}}_c(\beta)$ tracks $\mathbf{r}_c$. The β best describing r_c's relationship to $x$es within U —the logistic regression of $(r_{cij} : i \in U)$ on covariates $(\vec{x}_{ij} : i \in U)$ — would minimize $V(\hat{\mu}_c(\beta))$, at least approximately, and might be taken as the ideal value of β. We propose to estimate this β, written $\beta^{(0)}$, via a logistic regression restricted to the control group. (The restriction to controls allows us to avoid committing to a model relating $\mathbf{r}_t$ and $\mathbf{r}_c$.) Writing $\hat{\beta}$ for the result of this regression, our interval estimate for the attributable effect is

$$\begin{aligned} &\sum_U r_i - M\hat{\mu}_c(\hat{\beta}) \pm z_{\alpha/2} M \hat{V}^{1/2}\left[\hat{\mu}_c(\beta)\right]_{\beta=\hat{\beta}} \\ &\quad = \sum_U r_i - \sum_U \hat{r}_{ci}(\hat{\beta}) - N\bar{e}_{\mathcal{C}}(\hat{\beta}) \pm z_{\alpha/2} N \hat{V}^{1/2}\left[\bar{e}_{\mathcal{C}}(\beta)\right]_{\beta=\hat{\beta}} \qquad (2) \\ &\quad = \sum_U r_i - \sum_U \hat{r}_{ci}(\hat{\beta}) \pm z_{\alpha/2} N \hat{V}^{1/2}\left[\bar{e}_{\mathcal{C}}(\beta)\right]_{\beta=\hat{\beta}}\,. \qquad (3) \end{aligned}$$

(3) assumes the logistic regression to have been fit with an intercept, in which case the sum $N\bar{e}_{\mathcal{C}}(\hat{\beta})$ of its residuals must be zero.

The estimate $\hat{\beta}$ is a random variable, not a constant, so the argument of § 3.1 does not alone suffice for large-sample normality of $\hat{\mu}_c(\hat{\beta})$, nor for $\hat{V}(\bar{e}_{\mathcal{C}}(\hat{\beta}))$ to approximate its variance. This turns out not to be an impediment: under appropriate conditions, one can act as if $\hat{\beta}$ were $\beta^{(0)}$, without degrading the quality of inference.

Proposition 3.1 *Let $\hat{\mu}_c(\beta), e_i(\beta)$ be as defined in (1) and surrounding discussion, all $\beta \in \Re^K$. Suppose U and $\mathcal{C}$ to be embedded in sequences such that $N = |U| \uparrow \infty$ and $|\mathcal{C}| = n \uparrow \infty$; that $n\mathbf{E}(\hat{\beta} - \beta^{(0)})^2$ is asymptotically bounded, some $\beta^{(0)}$; that $s^2[(e_i(\beta^{(0)}) : i \in U)] \rightarrow$ some limit; that covariates x_{ijk} and cluster sizes are uniformly bounded; and that $n/N, M/N \rightarrow$ some limits. One then has the representation*

$$n^{1/2}(\hat{\mu}_c(\hat{\beta}) - \mu_c) = n^{1/2}(\hat{\mu}_c(\beta^{(0)}) - \mu_c) + \underbrace{n^{1/2}(\hat{\beta} - \beta^{(0)})^t T(\mathcal{C})}_{*} \qquad (4)$$

in which $n^{1/2}(\hat{\beta}-\beta^{(0)})$ is bounded in probability while $T(\mathcal{C}) \xrightarrow{P} 0$, so that $(*) \xrightarrow{P} 0$, where $T(\mathcal{C})$ is defined in Appendix A. Furthermore $s^2[(e_i(\hat{\beta}) : i \in \mathcal{C})] \xrightarrow{P} s^2[(e_i(\beta^{(0)}) : i \in U)]$, so that $(\hat{\mu}_c(\hat{\beta}) - \mu_c)\hat{V}^{-1/2}(\hat{\mu}_c(\beta))|_{\beta=\hat{\beta}} \xrightarrow{P} N(0,1)$.

Binder (1983, Appendix) gives natural, if rather technical, conditions on samples $\mathcal{C}$ from sampling frames U under which $\hat{\beta}$ has $o(n^{-1/2})$ bias and $O(n^{-1})$ variance, making $n\mathbf{E}(\hat{\beta} - \beta^{(0)})^2$ asymptotically bounded. The practical meaning of these and the conditions of Proposition 3.1 is that the control group should be sufficiently large and that, taken together, the data $((r_{cij}, \vec{x}_{ij}) : i \in U)$ and the model used to estimate $\hat{\beta}$ are such that few of $(e_i(\beta) : i \in U)$ are large relative to their standard error and few of $(\sum_j x_{kij}\hat{r}_{cij}(\beta)(1 - \hat{r}_{cij}(\beta)) : i \in U)$ are large relative to their standard error, for $k \leq K$ and β among the likely values of $\hat{\beta}$; see *e.g.* Scott and Wu (1981, p.101). Proposition 3.1 is proved in Appendix A.

3.3 Checking finite-sample performance and maximizing power

When inference is carried out using the procedure of § 3.2, one relies on three asymptotic approximations:

A1 The distribution of $\bar{e}_{\mathcal{C}}(\beta^{(0)})$ is approximated with a Normal distribution;

A2 the distribution of $\bar{\hat{r}}_c(\hat{\beta}) + \bar{e}_{\mathcal{C}}(\hat{\beta})$ is approximated with that of $\bar{\hat{r}}_c(\beta^{(0)}) + \bar{e}_{\mathcal{C}}(\beta^{(0)})$; and

A3 $s^2[\mathbf{e}(\beta^{(0)})]$ is approximated with $s^2[(e_i(\hat{\beta}) : i \in \mathcal{C})]$.

Assumption A1 is comparable to L1 of § 1.2. A3 strengthens L2, and A2 is new. A1 is relatively safe, at least in large samples with few outliers (Hájek 1960; Höglund 1978), but A2 and A3 are likely to err in predictable ways.

As noted following (3), when it holds, the fitted residuals $(e_{ij}(\hat{\beta}) : i \in \mathcal{C})$ necessarily sum to zero, unlike the corresponding deviations $(e_{ij}(\beta^{(0)}) : i \in \mathcal{C})$ from the population regression surface. Note that $\hat{\mu}_c(\hat{\beta}) \propto \bar{e}_{\mathcal{C}}(\hat{\beta}) + \bar{\hat{r}}_{cU}(\hat{\beta})$, wherein $\bar{\hat{r}}_{cU}(\hat{\beta})$ but not $\bar{e}_{\mathcal{C}}(\hat{\beta})$ is random, whereas $\hat{\mu}_c(\beta^{(0)}) \propto \bar{e}_{\mathcal{C}}(\beta^{(0)}) + \bar{\hat{r}}_{cU}(\beta^{(0)})$, wherein $\bar{e}_{\mathcal{C}}(\beta^{(0)})$ but not $\bar{\hat{r}}_{cU}(\beta^{(0)})$ is random. For finite n and N, one might expect variation of $\bar{\hat{r}}_{cU}(\hat{\beta})$ to be smaller than that of $\bar{e}_{\mathcal{C}}(\beta^{(0)}) = \bar{r}_{c\mathcal{C}} - \bar{\hat{r}}_{c\mathcal{C}}(\beta^{(0)})$, as $\bar{\hat{r}}_{cU}(\hat{\beta})$ is affected only indirectly by variation in $\mathcal{C}$. If so, this would undercut approximation A2, in such a way as to cause overestimation of $V(\hat{\mu}_c(\hat{\beta}))$.

As to A3, while $s^2[(e_i(\beta) : i \in \mathcal{C})]$ may be unbiased for $s^2[(e_i(\beta) : i \in U)]$ when β is fixed, it is well known that when coefficients $\hat{\beta}$ are estimated on one sample, C say, then the MSE of residuals, *i.e.* $s^2[(e_i(\hat{\beta}) : i \in C)]$, is often an "optimistic" or downwardly biased estimate of the error of predictions made using the same estimated coefficients $\hat{\beta}$ on a separate sample, such as $U \setminus C$ (*cf., e.g.,* Efron 1983). In the limit, as sample sizes increase towards infinity with the dimension of the regression model staying fixed, this bias shrinks to zero. In finite samples, however, it could in principle lead to appreciable under-estimation of $V(\hat{\mu}_c(\hat{\beta}))$. In summary, in finite samples the method of § 3.2 could either systematically overestimate or systematically underestimate its error of estimation.

Which of the two biases dominates is likely to be a function of the complexity of the regression surface fit to controls and then used for predictions $\hat{\mathbf{r}}_c$, with greater complexity contributing to under-estimation of $V(\hat{\mu}_c(\hat{\beta}))$. At the same time, underfitting of that regression surface should be avoided, as it would decrease precision of the estimate. To minimize errors of both types — Type I errors due to overfitting, Type II errors due to underfitting — we compared regression specifications of varying complexity in simulated repetitions of the experiment, performed on Vote 98 controls. This simulation study, details and results of which appear in Appendix B, found appreciable inflation of Type I errors for none of the sub-experiments or regression specifications considered, and suggested that a relatively saturated model ("F3", in which independent variables consume about 40 d.f.) would appreciably increase power relative to others considered.

4 Outcome Analysis

4.1 Overall effects of in-person, mail and telephone GOTV

Separately for each of the three interventions, we estimated the proportion μ_c of subjects who would have voted in its absence using the method of § 3.2. In the case of the telephone intervention, for example, this meant fitting a logistic regression surface to the subset of the control group that had not been sent mailers, and fitting another logistic regression surface to the remaining controls; extrapolating these fits to generate predictions $\hat{r}_{cij}(\hat{\beta})$ for all $i \in U$ and all j; and calculating $\hat{r}_{ci}(\hat{\beta}) = \sum_j \hat{r}_{cij}(\hat{\beta})$, each $i \in U$. (Specifications for these regressions, and our method of settling on them, are described in Appendix B.) Our estimate of the total number of votes that would have been cast had none of the telephone GOTV calls been made is $\sum_{i \in U} \hat{r}_{ci}(\hat{\beta})$. Our estimate of the number of votes attributable to telephone appeals, then, is simply $\sum_U r_i - \sum_U \hat{r}_{ci}(\hat{\beta})$, with standard error equal to that of $\sum_U \hat{r}_{ci}(\hat{\beta})$.

Since, by the assumed exclusion restriction, only subjects contacted by telephone can have been either prompted or dissuaded from voting by the telephone intervention, we checked that the resulting confidence intervals did not extend above the total number of subjects contacted by telephone who eventually voted in the 1998 election, nor below -1 times the total number of contacted subjects who did not vote in that election. (They fell within these limits; had they not, we would have truncated them.) We then divided these figures by the total number of subjects who had been contacted by telephone, so as to estimate the number of votes generated per contact. Parallel calculations were made for the mail and telephone interventions.

Personal canvassing appeared to produce 9 votes per 100 contacts (95% CI= [5, 13]), the best of the three interventions studied. Mailers were also demonstrably better than control, generating 14 votes per 1000 households mailed (95% CI= [1,27]). Although the votes-per-household-mailed estimate is relatively small, political campaigns should balance this small effect against the greater ease of mailing a large number of households. In our analysis, the study does not give evidence of a benefit for telephone appeals. The point estimate is negative, -3 votes per 100 completed calls, with a 95% confidence interval of -7 up to 1 votes per 100 telephone contacts. Although the results stop short of showing GOTV calls to have reduced turnout in the aggregate, they do exclude substantial telephone GOTV benefits.

4.2 Subgroup effects

These methods also apply to the estimation of subgroup effects. To see this, suppose $G \subseteq \{(i,j) : i \in U\}$ is a subgroup of individuals that can be specified in terms of their covariate values $\vec{x}_{ij}$. Then the attributable effect within G is $\sum_G r_{ij} - r_{cij} = \sum_{i\in U;j} r_{(G)ij} - \sum_{i\in U;j} r_{(G)cij}$, where $(r_{(G)ij}, r_{(G)cij}) = (r_{ij}, r_{cij})$ if $(i,j) \in G$, $(0,0)$ otherwise. Define $\hat{r}_{(G)cij}(\beta) = \hat{r}_{cij}(\beta)$ if $(i,j) \in G$, 0 otherwise, and for each i write $r_{(G)i} = \sum_j r_{(G)ij}$, $r_{(G)ci} = \sum_j r_{(G)cij}$, and $\hat{r}_{(G)ci}(\beta) = \sum_j \hat{r}_{(G)cij}(\beta)$. Then (2) applies to estimation of $\sum_G r_{ij} - r_{cij}$, once $r_{(G)i}$ and $\hat{r}_{(G)ci}(\hat{\beta})$ have been substituted for r_i and $\hat{r}_{ci}(\hat{\beta})$ and $\bar{e}_{\mathcal{C}}(\hat{\beta})$ has been interpreted as $n^{-1}\sum_{\mathcal{C}}(r_{(G)i} - \hat{r}_{(G)ci}(\hat{\beta}))$. If the indicator of G is a linear combination of the covariate, then $\sum_{i\in\mathcal{C};j}(r_{(G)ij} - \hat{r}_{(G)cij}(\hat{\beta})) = 0$ and the simpler form (3) applies.

We used this recipe to analyze treatment effects by subgroups defined in terms of age, receipt of complementary treatments, and prior voting. For age, we split the sample at quartiles; the resulting four subgroups were not precisely representable as linear combinations of the covariate, so formula (2) had to be used. “Complementary treatment” refers, in (for instance) the telephone sub-experiment, to whether a subject was assigned to in-person GOTV, and if so whether they had been contacted; alternately, it may be taken to mean whether mailers were sent to the subject, and if so how many. We divided the sample in these two ways separately, conducting two sets of subgroup analyses for treatment complementary to telephone GOTV, as well as two each for in-person GOTV and mailers. In each of these cases, the relevant dummy variables had been among the covariates used for prediction, so the simpler formula (3) could be used. For prior voting, we simply split the sample according to whether subjects had voted in New Haven in the previous election, as slightly more than half of them had done; again formula (3) applied. We do not present specific results of age and complementary treatment subgroup analyses, as they did not suggest interactions with the treatment, or did so only very weakly.

Figure 4 displays estimates of treatment effects overall and by voting in the previous election. While the effectiveness of personal canvassing appears to have been roughly similar for voters and nonvoters in the previous election, the results suggest that both mail and telephone GOTV differed in their effects on those who had and

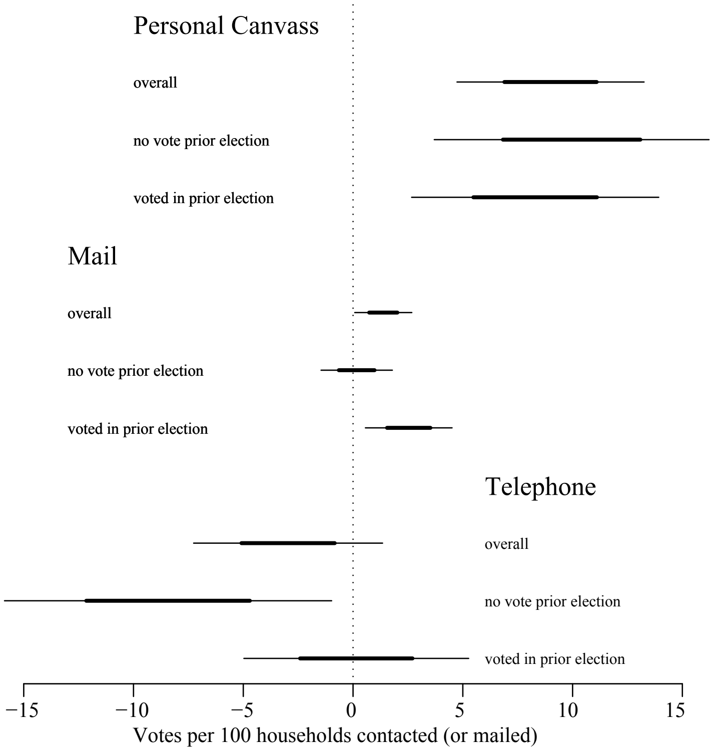

Figure 4: Effectiveness of the three modes of GOTV message delivery, overall and by voting in the previous election. Thick lines denote 2/3 CIs (Mosteller and Tukey 1977); thin lines, 95% CIs.

had not voted two years before. The suggestion is strongest in the case of telephone GOTV, a form of intervention that may have *dissuaded* voting, according to these results. Without attention to multiple comparisons, the hypothesis that telephone GOTV was neutral or beneficial for non-voters in the prior election receives a p-value of .01, one-sided, although a correction for multiplicity would render it nonsignificant. In the case of mail, the intervention does not appear to have been harmful, but there is the suggestion that its benefits were concentrated among prior voters.

5 Discussion

5.1 Methodology

Analysis of the Vote 98 experiment presents a number of important challenges. Although assignment to treatment was randomized, non-contact rates were high, execution was somewhat inconsistent, and effectiveness of the treatment could be expected to vary even when treatment was properly delivered; subjects were assigned to treatment with varying probabilities and in clusters; and the data included covariates of rich prognostic value, raising the question of how best to leverage them to enhance precision. Similar challenges can be expected to arise in other high-quality field experiments. The randomization-based method here adapted from survey sampling methodology addresses each of them, and in addition produces confidence statements attributing total numbers of votes, rather than changes to the log-odds of voting, to intervention, thus summarizing the effectiveness of the intervention on the same scale on which elections are decided. Its only requirements about intervention effects are that they could be experienced only by members of contacted households, that a GOTV appeal directed to one household could not in itself affect other households, and that the random assignment of each experimental intervention be independent of other interventions that may have affected voting (§ 2.3). It makes use of the covariates, borrowing strength from regression techniques, but it has no need for regression models' assumptions (§ 3).

High non-contact rates put special demands on the methods of analysis. They increase the risk inherent to "as-treated" analyses, which compare only subjects receiving the treatment to control, by magnifying the impact on effect estimates of the difficulty of isolating controls who, like the treatment group members who actually received treatment, could have been contacted had they been randomized to intervention. One avoids this risk with instrumental variable (IV) methods; but common model-based IV methods struggle with high rates of non-contact or non-compliance, even in very large experiments (Bound et al. 1995). Randomization-based methods do not share this difficulty, yielding tests, confidence intervals and point estimates which remain valid with arbitrarily weak instruments, a property that seems unique to these methods (Imbens and Rosenbaum 2005). This seems particularly relevant to political participation field experiments, where message delivery rates can be quite low. (In one recent experiment targeting young voters, only 8% of voters slated for in-person appeals could be contacted (Nickerson et al. 2006).)

Our choice of randomization-based methods more typical of survey analysis than experiments has the benefits of making available simple uses of regression in combination with straightforward adjustment for cluster-level assignment. Its drawback is that it invokes additional layers of asymptotic approximation. In studies with small samples, with very rare or very common binary outcomes, or with very small control groups, our variance estimators cannot be expected to perform as well as in this application. In studies adjusting for a covariate of high dimension or with outliers or heavy tails, treating an estimated regression coefficient as if it had been fixed *a priori* may not be as innocuous as it was found to be here. These exclusions leave a large class of experiments, including most get-out-the-vote experiments, for which present methods can be expected to perform well. In ambiguous cases the bootstrap method of Section 3.3 and Appendix B is available to check finite-sample performance.

An aspect of our formulation that may be limiting in some contexts is that it leads to inferences addressing uncertainty in our knowledge about the treatment effect A achieved in the experiment, a random variable, or about the random variable A/O, the number of votes per contact, but not specifically about such parameters as $\mathbf{E}A$ or $\mathbf{E}A/O$. That is, sampling variability in A and O is not addressed by the inference statement. This may be a limitation if A and/or O is felt to be drawn from a distribution shared with other contexts of substantive interest. A benefit is that by attending strictly to internal validity, greater precision of estimation may be possible, a point made by Abadie and Imbens (2006) in their discussion of sample-average and population-average treatment effects. This may explain why our analysis was able to distinguish the benefit of mailed GOTV appeals from zero even when clustering was properly addressed, whereas Gerber and Green's model-based analyses either ignored clustering (2000) or failed to discern mailer effects (2005).

5.2 Getting out the vote

We have estimated treatment effects for the Vote 98 experiment with quite minimal assumptions. Our analysis requires certain sample-size and other data conditions in order that its large-sample approximations apply; it depends on the data representing what they claim to represent; and it requires treatment assignment to have been blind to who would have voted in the absence of treatment. As regards the first of these, in Sections 1.2, 2.2, and 3.3 above we subjected the applicability of our large-sample approximations to rather extensive tests, confirming their applicability to the Vote 98 data. Regarding the second, we have taken Gerber and Green's most recently edited version of the data (Gerber and Green 2005, p.301–02), the only version to include cluster identifiers, at face value. Although their explanation of its other differences with earlier versions of the data satisfied us, Imai (2005, pp.288–89) regards some of the changes as suspicious. Interested readers should compare his and Gerber and Green's discussions and judge this for themselves. If these two requirements are granted, then only independence of treatment assignment and potential outcomes remains; but this flows naturally from experimental randomization. To protect this implication, we analyzed comparison groups quite strictly as they had been randomized. This may be contrasted with Gerber and Green (2000, 2005), who moved to the control group treatment group subjects who mistakenly had been given a placebo message, and it is in marked contrast with Imai (2005), whose as-treated analysis compared to control only the treated, the subset of the treatment group who had actually received the treatment.

Our overall results accord with those originally presented by Gerber and Green (2000): personal canvassing had clear and positive effects; mail GOTV had statistically significant but smaller benefits; and there was no evidence of a benefit for brief, scripted calls from an out-of-state professional calling firm. One caveat is that the positive effect of personal canvassing may be partially attributable to impersonal reminders left for subjects randomized to be canvassed but not contacted in person (§ 2). (Results of Nickerson et al. (2006) suggest that this is unlikely.) Another is that the Vote 98 experiment's mistaken delivery of a placebo message to part of the telephone intervention group would have reduced its power to detect a telephone benefit.

It would also have reduced power to detect a telephone GOTV *detriment*, a possibility that is at least as consistent with these data as is that of a GOTV benefit. Although our result on telephone GOTV differs from Imai's (2005), it accords with those of a separate experiment reported by Arceneaux, Gerber, and Green (2006), which also failed to find benefits for brief, mechanically delivered calls placed to voters in Iowa and Michigan before the 2002 elections. In recent years, telephone GOTV benefits have been seen in experiments, but only in especially favorable settings. Nickerson (2006) finds an overall average benefit of GOTV in a meta-analysis of eight randomized telephone campaigns with volunteer callers, but the overall benefit appears to have been driven by one particularly efficacious campaign. Wong (2005) also finds benefits for GOTV calls placed by volunteers, but the campaign had targeted Asian immigrant voters, many of them non-native English speakers, and the callers were coethnics and near-coethnics who often could address voters in their native tongues. Nickerson (2007) found positive effects from calls made by contractors, but the callers, already professionals, had been given special training and instruction in making "conversational" appeals, along with an irregular incentive structure to encourage "high-quality" interactions. That professional GOTV calls made without such special measures could backfire with some voters is consistent with these findings.

We found suggestive evidence of differences in GOTV effects among those who had and had not voted in the prior election. Telephone GOTV seems to have had little or no effect on those who voted in the previous election, but it appeared to de-mobilize prior election nonvoters more than it mobilized them. Mail benefits seem to have been concentrated among those who had voted in the last major election.

The evidence for a negative effect of phoning on prior election nonvoters is somewhat weaker than Figure 4 would suggest, since its error bars don't correct for the fact that several subgroup analyses were performed. However, it is natural to expect that a GOTV intervention's effectiveness might vary by likelihood of voting; had it been this possibility that prompted our analysis from the beginning, then no correction would be called for, and these negative conclusions would hold with full force. As matters stand, the evidence is less than conclusive, but it in any case suggests hypotheses which may merit further research. One is that GOTV mailings may help as reminders for those who intended to vote, but are less helpful for persuading those whose voting intentions were not yet formed; another is that scripted, impersonal GOTV calls made across social divides may tell *against* voting in the deliberations of less reliable voters.

References

Abadie, A. and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

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

Arceneaux, K. (2005), "Using cluster randomized field experiments to study voting behavior," *Annals of the Americal Academy of Political and Social Science*, 601, 169–179.

Arceneaux, K., Gerber, A. S., and Green, D. P. (2006), "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment," *Political Analysis*, 14, 37–62.

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review/Revue Internationale de Statistique*, 51, 279–292.

Bound, J., Jaeger, D., and Baker, R. (1995), "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association*, 90.

Braun, T. M. and Feng, Z. (2001), “Optimal permutation tests for the analysis of group randomized trials,” *Journal of the American Statistical Association*, 96, 1424–32.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001), “Interval estimation for a binomial proportion (with discussion),” *Statistical Science*, 16, 101–133.

Clinton, J. and Lapinski, J. (2004), “‘Targeted’ Advertising and Voter Turnout: An Experimental Study of the 2000 Presidential Election,” *Journal of Politics*, 66, 69–96.

Cochran, W. (1977), *Sampling Techniques*, Wiley, 3rd ed.

Efron, B. (1983), “Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation,” *Journal of the American Statistical Association*, 78, 316–331.

Eldersveld, S. J. (1956), “Experimental Propaganda Techniques and Voting Behavior,” *American Political Science Review*, 50, 154–165.

Elliott, M. R. (2008), “Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models,” *Journal of Official Statistics*, to appear.

Firth, D. and Bennett, K. E. (1998), “Robust models in probability sampling,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 3–21.

Fisher, R. A. (1935), *Design of Experiments*, Edinburgh: Oliver and Boyd.

Friendly, M. (1994), “Mosaic Displays for Multi-Way Contingency Tables,” *Journal of the American Statistical Association*, 89, 190–200.

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 to Gerber and Green (2000), replication of disputed findings, and reply to Imai (2005),” *American Political Science Review*, 99, 301–313.

Gosnell, H. F. (1927), *Getting Out the Vote: An Experiment in the Stimulation of Voting*, The University of Chicago press.

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

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. and Bowers, J. (2008), "Covariate balance in simple, stratified and clustered comparative studies," *Statistical Science*, 23, to appear.

Hartigan, J. and Kleiner, B. (1984), "A Mosaic of Television Ratings," *The American Statistician*, 38, 32–35.

Heckman, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely Used Estimator," *Journal of Human Resources*, 32, 441–462.

Highton, B. and Wolfinger, R. (2001), "The First Seven Years of the Political Life Cycle," *American Journal of Political Science*, 45, 202–209.

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

Imai, K. (2005), "Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments," *American Political Science Review*, 99, 283–300.

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, 109–126.

Isaki, C. T. and Fuller, W. A. (1982), "Survey design under the regression superpopulation model," *Journal of the American Statistical Association*, 77, 89–96.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

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.

Lohr, S. (1999), *Sampling: Design and Analysis*, Brooks/Cole.

McNulty, J. E. (2005), "Phone-Based GOTV–What's on the Line? Field Experiments with Varied Partisan Components, 2002-2003," *The Annals of the American Academy of Political and Social Science*, 601, 41.

Michelson, M. R. (2003), "Getting Out the Latino Vote: How Door-to-Door Canvassing Influences Voter Turnout in Rural Central California," *Political Behavior*, 25, 247–263.

Miller, R. E., Bositis, D. A., and Baer, D. L. (1981), "Stimulating Voter Turnout in a Primary: Field Experiment with a Precinct Committeeman," *International Political Science Review/ Revue internationale de science politique*, 2, 445.

Mosteller, F. and Tukey, J. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley Reading, MA.

Murray, D. M. (2001), "Statistical Models Appropriate for Designs Often Used in Group-randomized Trials," *Statistics in Medicine*, 20, 1373–1385.

Neyman, J. (1990), "On the application of probability theory to agricultural experiments. Essay on principles. Section 9," *Statistical Science*, 5, 463–480, transl. by D.M. Dabrowska and T.P. Speed from 1923 Polish original.

Nickerson, D. W. (2006), "Volunteer Phone Calls Can Increase Turnout: Evidence From Eight Field Experiments," *American Politics Research*, 34, 271.

— (2007), "Quality Is Job One: Professional and Volunteer Voter Mobilization Calls," *American Journal of Political Science*, 51, 269–282.

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

Niven, D. (2006), "A Field Experiment on the Effects of Negative Campaign Mail on Voter Turnout in a Municipal Election," *Political Research Quarterly*, 59, 203.

Raab, G. M. and Butcher, I. (2001), "Balance in Cluster Randomized Trials," *Statistics in Medicine*, 20, 351–365.

Raudenbush, S. W. (1997), "Statistical Analysis and Optimal Design for Cluster Randomized Trials," *Psychological Methods*, 2, 173–185.

Rosenbaum, P. R. (1996), "Identification of causal effects using instrumental variables: Comment," *Journal of the American Statistical Association*, 91, 465–468.

— (2001), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot," *Biometrika*, 88, 219–231.

— (2002), “Covariance adjustment in randomized experiments and observational studies,” *Statistical Science*, 17, 286–327.

Rosenbaum, P. R. and Rubin, D. (1985), “The bias due to incomplete matching,” *Biometrics*, 41, 103– 116.

Rubin, D. B. (1986), “Comments on “Statistics and Causal Inference”,” *Journal of the American Statistical Association*, 81, 961–962.

Särndal, C.-E., Swensson, B., and Wretman, J. (1991), *Model assisted survey sampling*, Springer-Verlag.

Scott, A. and Wu, C.-F. (1981), “On the Asymptotic Distribution of Ratio and Regression Estimators,” *Journal of the American Statistical Association*, 76, 98–102.

Smith, J., Gerber, A., and Orlich, A. (2003), “Self-Prophecy Effects and Voter Turnout: An Experimental Replication,” *Political Psychology*, 24, 593–604.

Thompson, S. G., Warn, D. E., and Turner, R. M. (2004), “Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale,” *Statistics in Medicine*, 23, 389–410.

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

Wong, J. (2005), “Mobilizing Asian American Voters: A Field Experiment,” *Annals of the American Academy of Political and Social Science*, 601, 102.

Zheng, H. and Little, R. J. A. (2005), “Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model,” *Journal of Official Statistics*, 21, 1–20.

Appendix A: Proof of Proposition 3.1.

Suppose a sequence of increasingly large experiments U_ν with simple random samples $\mathcal{C}_\nu \subseteq U_\nu$ ($|\mathcal{C}_\nu| = n_\nu, |U_\nu| = N_\nu$). Taylor approximation gives (4) with $T(\mathcal{C}) = \nabla_\beta \hat{\mu}_{\nu c}(\beta)|_{\beta=\mathbf{B}_\nu}$, where $\mathbf{B}_\nu$ is a vector bracketed by $\hat{\beta}_\nu$ and $\beta^{(0)}$ and

$$\nabla_\beta \hat{\mu}_{\nu c}(\beta) = N_\nu^{-1} \sum_{U_\nu} \nabla_\beta \hat{r}_i(\beta) - n_\nu^{-1} \sum_{\mathcal{C}_\nu} \nabla_\beta \hat{r}_i(\beta).$$

For each k and γ, $\mathbf{E}\partial/\partial\beta_k\hat{\mu}_{\nu c}(\beta)|_{\beta=\gamma} = 0$. Uniform boundedness of cluster sizes and covariates x_{kij} entails that the variances $s^2_{\nu k}(\gamma) = s^2[(\partial/\partial_k\hat{r}_i(\beta)|_{\beta=\gamma} : i \in U_\nu)]$ stay bounded as $\nu \uparrow \infty$, so that $V\left(\partial/\partial_k\hat{\mu}_{\nu c}(\beta)|_{\beta=\gamma}\right) = (1 - n_\nu/N_\nu)s^2_{\nu k}(\gamma)/n_\nu \to 0$ and $\partial/\partial\beta_k\hat{\mu}_{\nu c}(\beta)|_{\beta=\gamma} \xrightarrow{P} 0$. The uniform boundedness conditions also suffice to bound the Hessians $\nabla^t_\beta\nabla_\beta\hat{r}_i(\beta)|_{\beta=\gamma}$ uniformly in γ, i and ν, in which case $\nabla_\beta\hat{\mu}_{\nu c}(\beta)|_{\beta=\mathbf{B}_\nu} - \nabla_\beta\hat{\mu}_{\nu c}(\beta)|_{\beta=\beta_0} \to 0$ in probability provided that $\hat{\beta}_\nu \to \beta_0$ in probability. In particular, if $n_\nu\mathbf{E}(\hat{\beta}_\nu - \beta^{(0)})^2$ does not diverge then surely $\hat{\beta}_\nu \xrightarrow{P} \beta_0$, so that $T(\mathcal{C}) = \nabla_\beta\hat{\mu}_{\nu c}(\beta)|_{\beta=\mathbf{B}} \xrightarrow{P} 0$. (4) follows by an application of Slutsky's theorem.

For the second assertion of the Proposition, note that $|\hat{r}_{cij}(\beta) - \hat{r}_{cij}(\beta^{(0)})| \leq (1/4)|\vec{x}_{ij}(\beta - \beta^{(0)})|$, since the inverse logit function is increasing with maximum derivative 1/4. Thus

$$s^2[(\hat{r}_{ci}(\hat{\beta}) - \hat{r}_{ci}(\beta^{(0)}) : i \in \mathcal{C})] \leq \frac{1}{16}(\hat{\beta} - \beta^{(0)})^t\hat{\Sigma}_x(\mathcal{C})(\hat{\beta} - \beta^{(0)}) \xrightarrow{P} \mathbf{0}^t\Sigma_x\mathbf{0} = 0.$$

In consequence, differences between $s^2[(r_{ci} - \hat{r}_{ci}(\hat{\beta}) : i \in \mathcal{C})]$ and $s^2[(r_{ci} - \hat{r}_{ci}(\beta^{(0)}) : i \in \mathcal{C})]$ are asymptotically negligible, and consistency of the former follows from consistency of the latter. For the third assertion, Hájek's (1960) CLT says that $V^{-1/2}(\mu_c(\beta^{(0)}))(\mu_c(\beta^{(0)} - \mu_c)$ is $N(0,1)$, so the convergence follows from (4), $V(\mu_c(\beta^{(0)})/\hat{V}(\mu_c(\beta)|_{\beta=\hat{\beta}}$ - 1, and Slutsky's theorem.□

Appendix B: Details of simulation study

We simulate random assignment by protocols mirroring those of the Vote 98 randomization within bootstrap experimental universes U^* drawn from the Vote 98 control group. The reason to construct U^* by bootstrap sampling from the control group is that for controls but not other subjects, r_c is known, so that for such a U^* one can calculate a benchmark, $\mu^* = \bar{r}_{cU^*}$, against which to compare estimates $\hat{\mu}^*$. The relationship of r_{cij}s to $\vec{x}_{ij}$s in U^* should resemble their relationship in U, but no particular functional relationship is assumed of them in either the real or the contrived universes.

A repetition of our bootstrap experiment consists of sampling such a U^* from the controls, and calculating and storing μ^*; randomly selecting a size-n subset of it as a pseudo-control group C^*; fitting a regression to the individual-level observations in the pseudo-control group to produce $\hat{\beta}^*$; calculating the mean and s.d. of $e_i(\hat{\beta}^*)$ over C^*, and the mean of the predicted responses $\hat{r}_{ci}(\hat{\beta}^*)$ over U^*, to produce $\hat{\mu}^* = \hat{\mu}^*_c(\hat{\beta}^*)$ and $\hat{V}(\hat{\mu}^*) = \hat{V}(\hat{\mu}^*_c(\hat{\beta}^*))$; then calculating and storing $z^* = (\hat{\mu}^* - \mu^*)\hat{V}^{-1/2}(\hat{\mu}^*)$. The last three of these steps (finding $\hat{\beta}^*$, $\hat{\mu}^*$ and $\hat{V}(\hat{\mu}^*)$, and z^*) were performed for each of three candidate specifications of the regression model. In order to compare efficiency of $\hat{\mu}^*(\hat{\beta}^*)$ under the alternate specifications of the regression surface, we also computed and stored s.d.s $\sigma(\beta^*)$ of $e_i(\hat{\beta}^*)$ over U^*, using them to approximate $\sigma(\beta^*) \propto V^{1/2}(\hat{\mu}(\beta^*)) \approx V^{1/2}(\hat{\mu}(\hat{\beta}^*))$. (Compared to $s[(e_i(\hat{\beta}^*) : i \in C^*)]$, $\sigma(\beta^*) \approx s[(e_i(\hat{\beta}^*) : i \in U^*)]$ has the advantage that it is not prone to optimism.) We applied the procedure separately for each of the three interventions; in the case of the block-randomized mail and telephone experiments, we applied it separately within each of the two assignment blocks. In total, we performed five bootstrap simulations, with 2000 replications for each.

Our most parsimonious specification ("F1") regressed individuals' voting in the *previous* election on covariates, using all of U^*, rather than C^* only, for fitting. Its predictions of the dependent variable were made from demographic and household-membership data, using a binomial mixed model with random effects for household and fixed effects for voting ward, age (expanded into cubic splines using 6 df), membership in a major political party, number of voters in the household (1 or 2), and first-order interactions of these. Analysis assisted by this model would require only the arguments of § 3.1; in particular, it would not rely on Proposition 3.1, since F1's coefficients are the same whatever $\mathcal{C}$ is selected. Alternately, this model could be seen as using demographic and household information to smooth subjects' voting in the prior election, exchanging a 0/1 variable for a vector of empirical-Bayes posterior predictive voting probabilities.

"F2" used these smoothed prior votes, along with ward, the spline expansion of age, and complementary treatment assignment and compliance, to predict voting in the control group. This prediction was done using ordinary logistic regression at the individual level. The third specification, "F3," also using ordinary logistic regression, had the same independent variables as F2, except that instead of smoothed prior votes it used as predictors indicators of having been registered in New Haven at the time of the prior election, and of having voted in it.

Results were quite favorable, as seen in Table B1. For none of the procedures or sub-experiments were Type 1 errors significantly inflated relative to their asymptotic levels, although for the mail experiment as applied to the subgroup assigned to telephone error rates approach significance. This was the only sub-experiment assigning a minority of households to control; see Figure 2. In the remaining conditions, variance overestimation due to approximation A2 (§ 3.3) appears to have swamped variance underestimation due to A3. On the basis of these results, we expect that any of the procedures tested in our bootstrap experiment would lead to somewhat conservative statistical inferences. Consistent with effects of "optimism" having been

	Fit	Type I error rates: $\alpha = .05$	$\alpha = .10$	Relative Efficiency
Personal Canvas	F1	0.05	0.10	1.10
	F2	0.05	0.10	1.60
	F3	0.04	0.10	1.67
Mail|No Phone	F1	0.04	0.09	1.08
	F2	0.04	0.10	1.57
	F3	0.04	0.10	1.64
Mail|Phone	F1	0.05	0.11	1.14
	F2	0.05	0.12	0.89
	F3	0.06	0.11	1.06
Phone|No Mail	F1	0.05	0.10	1.08
	F2	0.05	0.11	1.58
	F3	0.05	0.10	1.64
Phone|Mail	F1	0.05	0.10	1.13
	F2	0.06	0.11	1.64
	F3	0.06	0.11	1.71

Table B1: Bootstrap Type 1 error rates and efficiency relative to estimation without covariate adjustment, for three fitting strategies (F1, F2, F3) and five sub-experiments. All cases achieved error rates comparable to nominal levels. In "Mail | Phone," a minority of households were assigned to control, and the most parcimonious specification is the most efficient; in the remaining conditions, control groups were larger and F3, the richest specification, was most efficient.

modest whenever the control group was not too small, in 4 of the 5 sub-experiments power increased steadily with increasing complexity of the surface fit to the control group, with F3 being the clear winner.

Appendix C: Reproducing Results

This appendix is not included in the JASA paper. It is provided for researchers interested in reproducing the results or experimenting with the data. A basic familiarity with the `R` statistical environment is assumed. File paths are listed relative to the paper archive path.

The entire compendium can be downloaded as a compressed archive from `http://dvn.iq.harvard.edu/dvn/dv/jakebowers` or directly from `http://hdl.handle.net/1902.1/12174`.

C.1 Data and Analysis

This paper archive includes all the source data and analysis files used in constructing this document. While all the data has been supplied as part of the paper archive, you may wish to create it yourself or modify the analysis to test other hypotheses. For convenience, the included `Makefile` provides commands to simplify this process.

C.1.1 Gerber and Green Data

The paper archive includes a copy of the Vote98 data as published by Gerber and Green on their website. The data are from the 2005 revision of the data, which included several changes to improve data quality (see Section 2.1 for more information). The data can be found in the `data/NHrep_household.dta` and `data/NHrep_individual.dta` files. If you wish to fetch the data directly from Gerber and Green's website, you may use the following `make` targets:

```
$ make clean-gerber-green-data gerber-green-data
```

C.1.2 Vote98 Attribution of Effect

Section 4.2 considers the different subgroups in the Vote98 universe and provides analysis of treatment effects on each of the treatment groups. This analysis is preformed in the `data/data-src/v98-effect-attribution.Rnw` Sweave file. The results are cached in `data/v98-effect-attribution.rda`. Updates to the Sweave file will cause the data results (and related figures) to be rebuilt. To rebuild the data without making any changes, use the following:

```
$ rm data/v98-effect-attribution.rda
$ make data/v98-effect-attribution.rda
```

As an Sweave file, `data/data-src/v98-effect-attribution.Rnw` contains useful explanatory text in LaTeX format as well as `R` code. To "weave" the file into a PDF:

```
$ R CMD Sweave data/data-src/v98-effect-attribution.Rnw
$ pdflatex v98-effect-attribution.tex
```

C.1.3 Increasing precision through the use of covariates: The Bootstrap Experiment

Section 4.1 computes estimates of effect the different treament alternatives. These results and the bootstrap experiment discussed in Section 3.3 and further explained in Appendix B are contained in three various Sweave files in the `data/data-src` directory, one for each treatment group.

The experimental data is saved in the files `data/bootstrap-inperson.rda`, `data/bootstrap-ma` and `data/bootstrap-phone.rda`. These files are generated from the corresponding `inperson.Rnw`, `mail.Rnw`, and `phone.Rnw` files in the `data/data-src` directory. These Sweave files can be "tangled" into `R` files that run the simulation experiment for one subgroup of the Vote 98 control data. The `Makefile` includes targets to tangle and run the bootstrapping in a single command. For example, to generate the `data/bootstrap-inperson.rda` data, use the following:

```
$ rm data/bootstrap-inperson.rda
$ make data/bootstrap-inperson.rda
```

By default, the bootstrapping experiment will run with 2000 repetitions. You can run the experiments with fewer repetitions by setting a command line variable. Combined with the overall bootstrap target, you can regenerate all the bootstrap data with more or fewer repetitions:

```
$ make BOOTSTRAP_REPS=123 bootstrap-data
```

C.2 Figures and Tables

All of the figures and tables in this paper are reproducible from the source data. If data, or source files are updated, the figures will automatically be regenerated by the `make` process. Individual figures can also be built directly using the directions below. If you wish to remove all the figures and start from scratch, please use the following:

```
$ make clean-figures
```

Figure 1 depicts the design of the Adams and Smith 1980 GOTV experiment discussed in § 1.1. This figure is generated from the `figures/ASdesign.R` file. This file can either be loaded into an interactive R session or used to generate `build/ASdesign.pdf` using the included Makefile:

```
$ make build/ASdesign.pdf
```

Figure 2 depicts the different sub-groups in the Gerber and Green experimental design. This figure is generated from `figures/v98design.R`. This file can either be loaded into an interactive R session or used to generate `build/v98design.pdf` using the included Makefile:

```
$ make build/v98design.pdf
```

Figure 3 uses control-group data to "estimate" means of baseline variables over treatment and control groups taken together, comparing associated confidence intervals to the known full-sample means of these variables. This figure is created by "tangling" the `figures/balance.Rnw` file to generate `build/balance.R`. This file, in turn, is used generated `build/balance.pdf`. For convenience, both of these steps are automated using the following Makefile target:

```
$ make build/balance.pdf
```

For additional information, you can weave the source file into a PDF:

```
$ R CMD Sweave figures/balance.Rnw
$ pdflatex balance.tex
```

Figure 4 shows confidence intervals for the effectiveness of the three treatment methods in the Gerber and Green experiment, broken out by previous vote history. This file is generated from `figures/aeByPrVote.R`. This file can either be loaded into an interactive R session or used to generate `build/aeByPrVote.pdf` using the included Makefile:

```
$ make build/aeByPrVote.pdf
```

Table B1 summarizes results from the bootstrapping simulation experiment as discussed in Appendix B. This table is generated from `figures/bootstrap-results.R`. This file summarizes the `data/boostrap-*.rda` data.

To generate this table use:

```
$ make build/bootstrap-results.tex
```