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

Technical Report #448, Statistics Dept., University of Michigan

Jake Bowers and Ben B. Hansen[1]

October 16, 2006

Abstract

In a landmark study of political participation, A. Gerber and D. Green (2000) experimentally compared the effectiveness of various get-out-the-vote interventions. The study was well-powered, conducted not in a lab but under field conditions, in the midst of a Congressional campaign; it used random assignment, in a field where randomization had been rare. As Fisher (1935) showed long ago, inferences from randomized designs can be essentially assumption-free, making them uniquely suited to settle scientific debates. This study, however, prompted a contentious new debate after Imai (2005) tested and rejected the randomization model for Gerber and Green's data. His alternate methodology reaches substantive conclusions contradicting those of Gerber and Green.

It has since become clear that the experiment's apparent lapses can be ascribed to clustered treatment assignment, rather than failures of randomization; it had randomized households, not individuals. What remains to be clarified is how this structure could have been accommodated by an analysis as sparing with assumptions as Fisher's. The present paper adapts recent advances in randomization inference to this purpose, furnishing new theory to accommodate clustering and stratification in both small- and large-sample inference for attributable effects. Since the method estimates the number of votes attributable to treatment, rather than its coefficient in a maintained proportional odds model, it is well-suited to the assessment of get-out-the-vote studies; but it also applies more broadly, to most experiments and observational studies with binary treatments and binary outcomes.

Key words: asymptotic separability, attributable effect, group randomized trial, instrumental variable, randomization inference, voter turnout

[1] Jake Bowers is Assistant Professor, Department of Political Science and Center for Political Studies, University of Michigan, Ann Arbor MI 48106-1248 (`jwbowers@umich.edu`). Ben Hansen is Assistant Professor, Statistics Department, University of Michigan, Ann Arbor, MI 48109-1092 (`ben.b.hansen@umich.edu`). Parts of this work were presented at the 2004 Royal Statistical Society Conference, the 2005 meetings of the Midwest Politicial Science Association and the Political Methodology Section, APSA, and the 55th Session of the International Statistical Institute, as well to the Political Science Department of the University of Illinois, Urbana-Champaign, the Department of Medicine at Case Western Reserve University, and the Yale University Biostatistics Department. The authors thank participants in these seminars, as well as Wendy Tam Cho, Donald P. Green, and Kosuke Imai, for helpful comments.

1 Introduction

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, randomly varying the assignment of interventions in accordance with a full factorial design. The study was well-powered, conducted not in a lab but under field conditions, during the run-up to the 1998 Congressional elections in New Haven, Connecticut; it used recent techniques to account for non-compliance with minimal assumptions; and the design was based on random assignment, in a field where randomization was rare. As Fisher (1935) showed long ago, such a design supports randomization-based inferences about its interventions' efficacy, inference that is essentially model-free and ought in principle to be above reproach. This study's inference, however, prompted a contentious debate — in the flagship journal of the American Political Science Association — after Imai (2005) tested and rejected the randomization model for Gerber and Green's data. His alternate methodology, which avoids assuming that randomization was carried out as planned, delivers substantive conclusions that contradict Gerber's and Green's. They had found that impersonal appeals delivered by telephone did not mobilize voters while in-person appeals did; Imai's analysis attached statistically and materially significant benefits even to the telephone intervention. These incompatible conclusions have contradictory ramifications for both the theory and practice of voter mobilization (Gerber and Green 2000, 2005a; Imai 2005).

As it happens, the Vote '98 study's apparent anomalies did not arise from a failure of randomization. The design had randomized households, not individuals, a complication noted but not addressed in Gerber and Green's original report (2000). This clustered assignment induced treatment-control comparisons that by metrics appropriate to individual randomization would seem quite biased, although metrics appropriate to the design remove the appearance of bias. This is apparent in Table 1, which compares on selected baseline characteristics subjects to whom personal appeals were and were not attempted, first without and then with appropriate adjustments for clustering. The tests that accompany the descriptive comparisons are performed as follows. Let the study subjects be numbered $1, \ldots, n$ and let x be one of the baseline variables; let $I \subseteq \{1, \ldots, n\}$ identify the in-person intervention group; and let $\mathbb{I}$ consist of all subsets of $\{1, \ldots, n\}$ which, according to the maintained description of the design, could have been selected as the intervention group. (The precise composition of $\mathbb{I}$ depends on whether treatment was assigned to clusters or individuals, on whether it was assigned

Standardized Differences in Several Covariates (as % of a pooled s.d.)				
	Assumes Assignment by Household?			
Covariate	No		Yes	
1- vs. 2-voter household	2		2	
Voted in previous election	1		1	
was registered, didn't vote	−2		−4	
Member of a major party	0		−5	
Age: B-spline 1	0		−2	
$\vdots$				
B-spline 6	−2		−2	
Ward 2	0		−0	
Ward 3	−5	***	−12	**
$\vdots$				
Ward 30	1		1	
overall χ^2/d.f.:	58/38	*	40/38	

Table 1: Standardized differences on baseline measures between subjects to whom in-person appeals were and were not attempted, first ignoring and then accounting for household-level randomization. The standardized difference consists of the difference of intervention- and control-group means, either individual means or means of household totals, as a percentage of the variable's s.d. (as pooled across intervention and control groups). The age measure, an important predictor of voting, has been decomposed into natural cubic splines with knots at sextiles of the sample age distribution, generating 6 loadings onto a B-spline basis. Wards are contiguous regions of New Haven in which subjects were registered. Results of permutation tests for imbalance are indicated as follows: no flag, $p > .1$; ".", $p \leq .1$; . . . ; "***", $p \leq .001$.

within strata, and on n, in a fashion to be discussed presently.) Then the hypothesis of balance is rejected, at level α, if $\sum_I x_i$ falls outside the central $(1-\alpha)100\%$ of $\{\sum_{j \in J} x_j : J \in \mathbb{I}\}$. We perform these tests for each variable x, giving some 40 comparisons in each column, only a subset of which are shown in the table. The χ^2 statistics given at bottom summarize these comparisons (Hansen 2006a). When individual-level assignment is assumed, the hypothesis of well-functioning randomization is rejected ($p = .02$); but under the correct assumption of assignment by household, that hypothesis is sustained ($p = .4$). The experiment is vindicated.

The structure of the set $\mathbb{I}$ of possible treatment assignments, and thus the substance of tests in Fisher's style, depends subtly but importantly on the role of clustering in assignment to treatment. For the tests assuming individual assignment, $J \in \mathbb{I}$ if $\#J \cap S = \#I \cap S$, for each of the four subclasses S delineated by whether subjects were or were not assigned to the remaining treatments, mail and telephone GOTV. For tests assuming household assignment, $J \in \mathbb{I}$ if: (i) for all subjects i, j from the same household, either $i, j \in J$ or $i, j \notin J$; and (ii) for each subclass S of assignments to the remaining treatments, the number of households represented in $J \cap S$ is the same as in $I \cap S$. Table 1 proves that this is a distinction with a difference: the tests ignoring clustering declare that treatment had an effect on baseline variables, whereas the test accounting for clustering avoids this absurd conclusion. This is so despite the clusters' being no larger than two — had they been smaller, they would cease to be clusters — and their being relatively well-balanced across treatment groups, as shown by the first row of the table. It is a distinction, clearly, to which the analysis should carefully attend. Analytic methods accounting for clustered treatment assignment and binary outcomes, albeit from a model-based perspective, include those of Raudenbush (1997), Murray (2001), and Thompson et al. (2004); Braun and Feng's approach (2001) is randomization-based, but not readily adaptable to estimation of attributable effects.

Clustering-aware balance tests vindicate the Vote 98 experiment's randomization, but they do not adjudicate between Gerber and Green's and Imai's contradictory inferences, each of which is supported by its own statistical model. Their methods, two-stage least squares and related techniques (Gerber and Green) and propensity-score matching (Imai), are both well-received and widely used. The methods' assumptions — Gerber and Green's, about potential response surfaces; Imai's, about conditional probabilities governing receipt of treatment — differ in character and perhaps also in degree, but resemble one another in that neither is entailed by established fact or theory. What the debate now requires is an analysis from first, Fisherian principles, eschewing speculation of either of type.

Fisher's randomization analysis culminates in tests of whether treatment had an effect — any effect, large or small. More recent techniques are needed to infer the number of events, votes for example, caused by a treatment; performing such inferences with clustered and stratified designs requires extension even of these methods. The remainder of the introduction reviews Rosenbaum's (2001) method of attributable effects using an experiment from the voter mobilization literature which has a simpler research design than the Vote 98 experiment. § 2 extends Rosenbaum's work to accommodate clustering. Section 3 applies this method to unmatched studies with stratification. New methodology also appears in § 4, which elaborates our randomization-based inferences so as to leverage covariate information for improved precision. Section 5 studies the potential for these methods to over- or understate confidence coefficients in small samples. Section 6 concludes.

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 that calls be placed to $n = 1325$ subjects, soliciting their votes on behalf of one of the candidates, John Ray. These subjects had been randomly selected from a pool of $N = 2650$ potential voters, no two of which shared a household, for whom turnout would later be determined from public records. Because the experiment is smaller and simpler than Gerber and Green's, we use it to illustrate the basis of our approach. The form of analysis sketched in this section is due to Rosenbaum (2001) (but see also Copas 1973).

Thirty percent of treatment group members voted in the special election, whereas only 24% of the control group voted. Could this difference be due to chance? Consider the hypothesis that it was, that treatment was inert. If this is so, then the labeling of one half-sample as treatment and another as the control group is in effect arbitrary, so far as their eventual voting, y, was concerned. From basic theory of simple random sampling, $\mathbf{E}_{J\in\mathbb{I}}(\sum_J y_j) = n\bar{y}$ and $\mathrm{Var}_{J\in\mathbb{I}}(\sum_J y_j) = n(1-n/N)s^2(y)$, where $\mathbb{I} = \{J \subseteq \{1,\ldots,N\} : \#J = 1325\}$. By these formulas, 353.5 ± 11.4 votes are expected for the treatment group. From tables of the hypergeometric distribution, if the treatment had no effect, 95% of possible samples would have tallied between 331 to 376 votes. Yet Adams and Smith recorded 392 votes from their intervention group. While not logically incompatible with our hypothesis, these data are at odds with it, as less than .1% of half-samples assemble so disproportionate a share of the 707 total votes. Fisher's test sets aside such improbabilities, encouraging us to conclude instead that this treatment was not inert.

Granting that treatment had an effect, let us probe this effect's likely magnitude. For concreteness, consider the hypothesis that treatment caused 50 votes. The analysis just given no longer simply applies, since an excess in $\sum_I y_i$ as compared to its permutation distribution can be explained by this hypothesized treatment effect, without supposing a treatment group improbably predisposed toward voting. To avoid this obstacle, begin by removing the hypothesized treatment effect of 50 votes. The hypothesis entails that $392 - 50$ intervention group members would have voted in the absence of treatment; there is no change to the number of voting controls (315). Those subjects' potential and actual responses to the control condition can be represented, under this hypothesis, with a binary variable y_c taking 1 as a value 342 times on I, 315 times on the complement of I, and otherwise 0. We therefore compare 342, not 392, to the distribution of $\sum_{j \in \mathcal{I}} y_{cj}$, not $\sum_{j \in \mathcal{I}} y_j$, for each uniform random draw $\mathcal{I}$ from $\mathbb{I}$. The result is a two-sided p-value of .21. The hypothesis attributing 50 votes to treatment, denoted $[A = 50]$, is sustained.

In like fashion p-values attach to each of $[A = 0]$, ..., $[A = 392]$. Inverting such hypothesis tests gives confidence intervals and point estimates. For Adams and Smith's experiment, the 95% confidence interval (CI) is $[33, 119]$ votes, or an increase in turnout of $\frac{392}{392-33} - 1 = 9\%$ to $\frac{392}{392-119} - 1 = 44\%$. Interpreted in terms of the proportion of the treatment group that voted because of treatment, the interval becomes $\frac{33}{1325} = 2.5\%$ up to $\frac{119}{1325} = 9.0\%$. Mimicking Hodges and Lehmann's (1963) technique for models with additive effects, a point estimate may be taken as the midpoint of the smallest nonempty $1 - \alpha$ CI. In this study, that would be the 3% CI, which includes 76,77, and 78; the point estimate is 77 votes, or a 24% turnout boost.

1.2 Three causal assumptions: noninterference, exclusion and nonnegative effects

No interference between units (Cox 1958, §2.4), or the stable unit treatment value assumption (Rubin 1986), states that only subject i's treatment assignment can affect subject i's response. Some version of this assumption is needed to justify the notation y_{ci} for subject i's potential response to control, by excluding the possibility of other subjects' treatment assignments influencing i's response to control. We have assumed noninterference outright for Adams and Smith's study, but the New Haven study requires a weaker assumption, since its cohabiting voters can be expected to influence one anothers' voting decisions (Stoker and Jennings 1995). Assuming noninterference between clusters (households), write y_i for subject i's observed response, y_{ci} for his potential response if his household were assigned to control, and τ_i for $y_i - y_{ci}$, the effect of treatment on subject i.

The exclusion restriction (Angrist et al. 1996) says assignment to the treatment group affects outcomes only via administration of the treatment. In GOTV intervention studies, this is the reasonable premise that only the voting of contacted subjects can have been influenced by the intervention: $\tau_i = 0$ unless $i \in C$, the set of treatment group members who complied with treatment (Rosenbaum 1996; Greevy et al. 2004). Hamilton (1979) assumes treatment may increase the response but never reduces it, in symbols $\tau_i \geq 0$ for all i; call this nonnegativity. Following Rosenbaum (2002a), a detailed hypothesis as to how each subject would have voted in the absence of treatment,

Households containing:	2 subjects			1 subject		Total no. of…	
Votes from household:	2	1	0	1	0	votes	subjects
Treatment	43	176	223	130	311	392	1325
Control	25	160	257	105	336	315	1325

Table 2: Adam and Smith's treatment and control groups, as imagined to have been assigned to treatment as households, each containing one or two experimental subjects.

$[y_c = \tilde{y}_c]$, is called *compatible* if it is consistent both with the exclusion restriction and with nonnegativity. Our analysis of the New Haven data will consider all and only the compatible hypotheses. By considering *all* the compatible hypotheses, we avoid making any assumptions about homogeneity of the treatment effect. This is in contrast with many other permutation-based approaches, including Braun and Feng's (2001) and Rosenbaum's (2002b).

2 Attributing effects by cluster

Adam and Smith's study placed calls to a simple random sample of individuals, whereas Gerber and Green's involved calling a random sample of households, some containing more than one subject. We now extend the method of § 1.1 to handle this complication. To illustrate the extension, this section adds fictitious clusters to Adams and Smith's data.

2.1 Clusters as units of analysis and assignment

Suppose in this section that Adams and Smith's treatment group had consisted of a simple random sample of one- or two-potential-voter households. Specifically, imagine that the vote totals presented in § 1.1 summarize the more detailed arrangement in Table 2. What modification to § 1.1's hypothesis tests would this require?

Let $y_1, \ldots, y_M$ be indicators of the $M = 2650$ subjects' actual voting and let $y_{c1}, \ldots, y_{cM}$ represent how they would have voted had none of them been called. Let the "cluster" function $\mathbf{clr} : \{1, \ldots, M\} \to \{1, \ldots, N\}$ map indices of subjects to indices of their clusters (households), write I for the indices of clusters assigned to treatment, and $C \subseteq I$ for the clusters in which someone received treatment. Write A for $\sum\{\tau_i : \mathbf{clr}(i) \in I\} = \sum_{\mathbf{clr}(i)\in I} y_i - y_{ci}$, the sum of effects attributable to treatment. Let $\mathbb{I}$ contain all possible treatment groups and let $\mathcal{I}$ be a random set distributed uniformly on $\mathbb{I}$. $\sum_{\mathbf{clr}(j)\in\mathcal{I}} y_{cj}$ is again the sum of a simple random sample, not of $m = 1325$ subjects' y_c values but of $n = 883$ from $N = 1766$ households' totals t_c of y_c values,

$t_{ck} = \sum_{\mathbf{clr}(i)=k} y_{ci}$. Its distribution has moments

$$\mathbf{E}(\sum_{j\in\mathcal{I}} t_{cj}) = n\bar{t}_c, \;\; \mathrm{Var}(\sum_{j\in\mathcal{I}} t_{cj}) = n(1-\frac{n}{N})s^2(t_c)$$

and is approximately Normal for large N (and n/N not close to 0 or 1), by the CLT for simple random samples (Erdős and Rényi 1959). The test that rejects if $\sum_{\mathbf{clr}(l)\in I} \tilde{y}_{cl} = \sum_I \tilde{t}_{ci}$, the treatment group's vote total net of votes hypothetically attributed to treatment, falls outside $\mathbf{E}(\sum_{j\in\mathcal{I}} \tilde{t}_{cj}) \pm z_{\alpha/2}\mathrm{Var}(\sum_{j\in\mathcal{I}} \tilde{t}_{cj})^{1/2}$, is asymptotically of level α.

Tested in this way the strict null hypothesis, which says $t_{ci} = t_i$ for all i, gives $\bar{t}_c = .4003$, $s^2(t_c) = (N/(N-1))(\overline{t_c^2} - (\bar{t}_c)^2) = (.4773 - .4003^2) = .3173$, and acceptance regions of form $353.5 \pm z_{\alpha/2}11.8$. Accounting for assignment by clusters has increased these regions' half-width slightly, from $11.4|z_{\alpha/2}|$ to $11.8|z_{\alpha/2}|$; accordingly the p-value for the strict null increases slightly, to .001.

Testing hypotheses asserting an effect now requires attention to where the effects are placed. Let two hypotheses, $H = [t_c = \tilde{t}_c]$ and $H^* = [t_c = \tilde{t}^*_c]$, satisfy $\sum_i (t_i - \tilde{t}_{ci}) = \sum_i (t_i - \tilde{t}^*_{ci}) = 2$. Then $\bar{\tilde{t}}_c = \bar{\tilde{t}}^*_c = .4003 - 2/1766$, so that the two hypotheses entail the same first moment for the test statistic; but $s^2(\tilde{t}_c)$ need not equal $s^2(\tilde{t}^*_c)$, so that $\mathrm{Var}(\sum_{j\in\mathcal{I}} \tilde{t}_{cj})$ and $\mathrm{Var}(\sum_{j\in\mathcal{I}} \tilde{t}^*_{cj})$ may differ. If H attributes its 2 votes to a single two-subject household, then $\overline{(\tilde{t}_c)^2} = .4773 - 2^2/1766$, whereas if $\tilde{t}^*_c$ attributes its to two separate one-subject households, then $\overline{(\tilde{t}^*_c)^2} = .4773 - 2 \cdot (1/1766)$. The implied difference in variances is small, 139.4 as opposed to 139.9, but the spread among such differences increases as hypothesized effect size increases, and cannot generally be ignored. Suppose now that $H = [t_c = \tilde{t}_c]$ has $\sum_{i\in I} t_i - \tilde{t}_{ci} = 31$, with $\tilde{t}_k = 2$ and $\tilde{t}_{ck} = 1$ for precisely 31 households k. Then $\sum_{i\in I} \tilde{t}_{ci}$ falls $2.08 \cdot \mathrm{Var}(\sum_{j\in\mathcal{I}} \tilde{t}_{cj})^{1/2}$ above $\mathbf{E}(\sum_{j\in\mathcal{I}} \tilde{t}_{cj})$, suggesting that at level $\alpha = .05$ the hypothesis $[A = 31]$ should be rejected. That composite hypothesis, however, contains other simple hypotheses. For instance, a hypothesis $H^* = [t_c = \tilde{t}^*_c]$ with $\sum_{i\in I} t_i - \tilde{t}^*_{ci} = 31$ but $t_i = 1$ and $\tilde{t}^*_{ci} = 0$ for 31 one-subject households i has $\sum_{i\in I} \tilde{t}^*_{ci} = \mathbf{E}(\sum_{j\in\mathcal{I}} \tilde{t}^*_{cj}) + 1.955 \cdot \mathrm{Var}(\sum_{j\in\mathcal{I}} \tilde{t}^*_{cj})^{1/2}$, and is narrowly sustained. In consequence, $[A = 31]$ is sustained, despite the rejection of H.

Both H and H^* issue the same test statistic, $\sum_{i\in I} \tilde{t}_{ci} = \sum_{i\in I} \tilde{t}^*_{ci} = 392 - 31$, and null expectation, $\mathbf{E}(\sum_{j\in\mathcal{I}} \tilde{t}_{cj}) = \mathbf{E}(\sum_{j\in\mathcal{I}} \tilde{t}^*_{cj}) = 883(.4003 - 31/1766)$, so the difference in z-statistics is due entirely to differences in induced variances. The test of a simple hypothesis $[t_c = \tilde{t}_c]$ is also a test of the composite hypothesis $[A = a]$, $\sum_{i\in I} t_i - \tilde{t}_{ci} = a$, if and only if $[t_c = \tilde{t}_c]$ maximizes $\mathrm{Var}(\sum_{j\in\mathcal{I}} \tilde{t}_{cj})$ among compatible $[t_c = \tilde{t}^*_c]$ such that

$\sum_{i \in I} t_i - \tilde{t}^*_{ci} = a$, since the composite is rejected only if each simple hypothesis falling under it is, and since among hypotheses giving the same test statistic and null expectation the variance-maximizing hypothesis is the most difficult to reject. (This holds for all two-sided tests, and for one-sided tests provided that $\alpha < \frac{1}{2}$.) Proposition 2.1 describes the variance-maximizing simple hypotheses within a composite $[A = a]$.

Proposition 2.1 *Let $\mathcal{I}$ be uniform on $\mathbb{I}$. Let $[t_c = \tilde{t}_c]$ be a compatible hypothesis, and let $a = \sum_{i \in I} t_i - \tilde{t}_{ci}$. If $[t_c = \tilde{t}_c]$ maximizes $\mathrm{Var}(\sum_{j \in \mathcal{I}} \tilde{t}_{ci})$, in the sense that for all compatible $[t_c = \tilde{t}^*_c]$ such that $\sum_{i \in I} t_i - \tilde{t}^*_{ci} = a$, $\mathrm{Var}(\sum_{j \in \mathcal{I}} \tilde{t}_{cj}) \geq \mathrm{Var}(\sum_{j \in \mathcal{I}} \tilde{t}^*_{cj})$, then:*

(i) There exists an integer $\gamma_0 \geq 1$ such that if $t_k < \gamma_0$, then $\tilde{t}_k = 0$; there is at most one cluster k such that $t_k = \gamma_0$ and yet $0 < \tilde{t}_k < \gamma_0$, so that for other clusters l, if $t_l = \gamma_0$ then $\tilde{t}_l = 0$ or γ_0; and if $t_l > \gamma_0$ then $\tilde{t}_l = t_l$.

(ii) The γ_0 of (i) is the largest γ such that $\sum_{k \in C, t_k < \gamma} t_k \leq a$.

Writing r_0 for $a - \sum_{k \in C, t_k < \gamma_0} t_k$ and N_0 for $\#\{k : t_k = \gamma_0\}$, the maximum variance among compatible hypotheses $[t_c = \tilde{t}^*_c]$ such that $\sum_I t_i - \tilde{t}^*_{ci} = a$ is given by $\mathrm{Var}(t; a) = n[1 - n/N] s^2(t; a)$, where $s^2(t; a) =$

$$\left[\sum_{\substack{k: k \notin C, \\ t_k < \gamma_0}} t_k^2 + (N_0 - 1 - \lfloor r_0/\gamma_0 \rfloor)\gamma_0^2 + (\gamma_0 - (r_0 - \gamma_0 \lfloor r_0/\gamma_0 \rfloor))^2 + \right.$$
$$\left. \sum_{\substack{k: \\ \gamma_0 < t_k}} t_k^2 \right] / (N - 1) - \frac{N}{N - 1} \left(\frac{\sum_k t_k - a}{N} \right)^2. \tag{1}$$

For a proof, see the Appendix.

2.2 Addressing assignment by household in the Vote 98 experiment

We use Proposition 2.1 to infer the effect of in-person GOTV on New Haven voter turnout.

Of 23,450 households randomized to treatment conditions, 4.7% were dropped from the rolls in 1998; individuals in these households are excluded from our analysis. One in five of the remaining households were selected for personal canvassing, and canvassers succeeded in making contact with 30% of these. Households to which personal appeals had been directed enjoyed a higher turnout rate in the 1998 election, 47.4%, than than did their counterparts assigned to control, 44.7%; this difference is either a rare event or evidence against the null ($p = .007$, two-sided). Null hypotheses attributing some of the treatment group's votes to personal canvassing fare better, with p-values increasing to .331 at $A = 92$ and .340 at $A = 93$. (More specifically, compatible hypotheses $[t_c = \tilde{t}_c]$ with $\sum t_i - \tilde{t}_c = 92$ give two-sided p-values ranging from .329 up to .331, entailing a p-value of .331 for the composite hypothesis $[A = 92]$; and p-values for compatible asserting $\sum t_i - t_c = 93$ range from .338 to .340, so the p-value of $[A = 93]$ is .340.) These p-values continue to increase with increasing A, up to a maximum of .997 for $[A = 144]$, until declining to cross 1/3 again at $A = 196$. A 2/3 confidence interval runs from 93 up to 195, or 5.0 to 10.4% of members of contacted households, and the best single estimate of its effect is 144 votes, or 7.7% of such subjects.

3 Stratified randomization

Consider now treatments assigned to simple random samples of individuals or clusters of individuals from each of S strata. (In voter mobilization studies, strata might be precincts or districts, while clusters might be households or city blocks contained in the strata.) To appraise hypotheses $[A = a]$, about totals of treatment effects, we must appraise all compatible *stratum attributions*, stratum-wise specifications $[A_1 = a_1, \ldots, A_S = a_S]$ of treatment effects describing at least one compatible hypothesis $[t_c = \tilde{t}_c]$, such that $\sum_s a_s = a$, sustaining $[A = a]$ at level α if any stratum attribution consistent with it is sustained at level α. Writing U_s for the indices of clusters in the sth stratum, the test statistic for $[t_c = \tilde{t}_c]$ becomes $\sum_s \sum_{I \cap U_s} \tilde{t}_{ci}$, with null mean $\sum_s \mathbf{E}\left(\sum_{j \in (\mathcal{I} \cap U_s)} \tilde{t}_{cj}\right)$ and variance $\sum_s \mathrm{Var}\left(\sum_{j \in (\mathcal{I} \cap U_s)} \tilde{t}_{cj}\right)$. The statistic itself and its expectation are functions of y and of $(a_1, \ldots, a_S) = \left(\sum_{I \cap U_s} t_i - \tilde{t}_{ci} : s\right)$. We write the difference of the two as

$$d(t, I; \mathbf{a}) = d(t, I) - \sum_s (1 - p_s) a_s,$$

where $d(t, I) = \sum_s \sum_{I \cap U_s} t_{ci} - \sum_s \mathbf{E}\left(\sum_{\mathcal{I} \cap U_s} t_{cj}\right)$ is the difference of observed and expected under the strict null of no effect, $[y_c = y]$, and $p_s = n_s/N_s$ is the chance of assignment to treatment in stratum s.

Although y and $(a_s : s \leq S)$ do not suffice to determine the null variance, Proposition 2.1 applies separately to each stratum s to identify the largest $\mathrm{Var}\left(\sum_{\mathcal{I} \cap U_s} \tilde{t}_{cj}\right)$ that is compatible with $a_s = \sum_{I \cap U_s} t_i - \tilde{t}_{ci}$, as well as a compatible configuration

$(\tilde{t}_{ci} : i \in U_s)$ that attains this maximum. Concatenating these variances gives a *single* specification $(\tilde{t}_{ci} : i \in U = \cup_s U_s)$, $\sum_{I \cap U_s} t_i - \tilde{t}_{ci} = a_s$ for every s, with the properties that $[t_c = \tilde{t}_c]$ is compatible and is rejected at level α iff *every* compatible $[t_c = \tilde{t}^*_c]$ such that $\sum_{I \cap U_s} t_i - \tilde{t}^*_{ci} = a_s$, all $s \leq S$, is also rejected. Finding this $[t_c = \tilde{t}_c]$ and testing it is thus a recipe for testing the stratum attribution $[A_1 = a_1, \ldots, A_S = a_S]$ (more briefly $[\mathbf{A} = \mathbf{a}]$). Write $\mathrm{Var}(t; \mathbf{a})$ for $\mathrm{Var}\left(\sum_{j \in \mathcal{I}} \tilde{t}_{cj}\right)$.

3.1 Example with two strata

Imagine now that in addition to the clustering described in Table 2, Adams and Smith's cohort had been randomized separately within two strata. Suppose stratum 1 contains 250, 50 and 20 households i for which $t_i =$ 0,1 or 2, respectively, with 100, 25 and 10 of these belonging to the treatment group, with remaining households in stratum 2. For simplicity, assume all treatment-group households complied with treatment. As before, a hypothesis $[A = a]$ determines the test statistic, $\sum_I \tilde{t}_{ci} = \sum_I t_i - a$; but now the statistic's null expectation as well as its variance vary among the simple hypotheses of which $[A = a]$ is comprised. To fix $\mathbf{E} \sum_{j \in \mathcal{I}} \tilde{t}_{cj}$, partition $[A = a]$ into $a + 1$ hypotheses $[A_1 = 0, A_2 = a], \ldots, [A_1 = a, A_2 = 0]$. Among hypotheses $[t_c = \tilde{t}_c]$ within any one of these, $\mathbf{E} \sum_{j \in \mathcal{I}} \tilde{t}_{cj}$ takes a single value. Still $\mathrm{Var}\left(\sum_{j \in \mathcal{I}} \tilde{t}_{cj}\right) = \mathrm{Var}\left(\sum_{j \in (\mathcal{I} \cap U_1)} \tilde{t}_{cj}\right) + \mathrm{Var}\left(\sum_{j \in (\mathcal{I} \cap U_2)} \tilde{t}_{cj}\right)$ may take multiple values, but its maximum,$\mathrm{Var}(t; \mathbf{a})$, is given by applying (1) separately in each of the two strata, and to test a composite $[\mathbf{A} = \mathbf{a}]$ one needs only to test the embedded simple hypothesis having the largest variance. For $[A = 20]$, one gets z-values increasing from 2.00 up to 2.15 for $[A_1 = 20, A_2 = 0]$, $\ldots$,$[A_1 = 0, A_2 = 20]$, giving an overall p-value of $2(1 - \Phi(2.00)) = .045$. In similar fashion, $[A = 21]$ gets a p-value of .051. $[A = 119]$ decomposes into $[A_1 = 45, A_2 = 74]$, $\ldots$,$[A_1 = 0, A_2 = 119]$ (stratum 1 has only 45 treatment-group subjects who voted, which limits the number of possible attributions to it) and z-values again increase, from -2.42 up to -1.97, entailing a two-sided p-value of $2\Phi(-1.97) = .049$ for $[A = 119]$. Likewise, the p-value for $[A = 118]$ is .054; a 95% CI is $[21, 118]$.

3.2 Stratified analysis via separability

When $S = 2$, an total effect hypothesis $[A = a]$ comprises no more than a stratum attributions $[\mathbf{A} = \mathbf{a}]$, but the number of stratum attributions falling under $[A = a]$ grows quickly with the sample size and the number of strata. With, say, five strata containing 500 or more compliers each, more than 2.7 billion stratum attributions fall under $[A = 500]$. A means of deciding whether any of these are sustained at a given level without explicitly testing each of them is desirable.

Rosenbaum (2002a), building on work of Gastwirth et al. (2000), proposes an easily implemented algorithm, the separability algorithm, to test $[A = a]$ in matched observational studies — studies with many strata. We extend this approach to stratified but not matched designs — studies with few strata. To translate it to this setting, suppose that a one-sided level-α test of $[A = a]$ against the alternative $[A > a]$, $\alpha < \frac{1}{2}$, is required, so that rejection of $[A = a]$ requires rejection of all compatible hypotheses $[\mathbf{A} = \mathbf{a}]$ such that $\sum_s a_s = a$, when tested against $[\mathbf{A} > \mathbf{a}]$; *i.e.* $[A = a]$ is rejected only if $z(\mathbf{a}) = \left(\sum_I \tilde{t}_{ci} - \mathbf{E}\sum_{j\in\mathcal{I}} \tilde{t}_{cj}\right) /\mathrm{Var}(t;\mathbf{a})^{1/2} > z_{1-\alpha}$ for all such $\mathbf{a}$ (where $\mathrm{Var}(t;\mathbf{a})$ is as defined in Proposition 2.1). The separability algorithm aims to identify a particular $\mathbf{a}$ whose z-score $z(\mathbf{a})$ approximates $m(a) = \min\{z(\mathbf{a}) : [\mathbf{A} = \mathbf{a}] \text{ is compatible, } \mathbf{a}'\mathbf{1} = a\}$ well enough to decide acceptance of $[A = a]$ by whether $z(\mathbf{a}) \leq z_{1-\alpha}$. Write $\mathcal{A}_s = \#(U_s \cap C)$ and $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_S)$, so that $m(a) = \min\{z(\mathbf{a}) : 0 \leq \mathbf{a} \leq \mathcal{A}, \mathbf{a}'\mathbf{1} = a\}$. (For two-sided tests of level α, apply two one-sided tests of level $\alpha/2$.)

The separability algorithm first narrows attention to those $\mathbf{a} = (a_1, \ldots, a_S)$ which minimize $d(t, I; \mathbf{a})$ subject to $\mathbf{a}'\mathbf{1} = a$ and $0 \leq \mathbf{a} \leq \mathcal{A}$. Since $d(t, I; \mathbf{a}) = d(t, I) - \sum_s (1 - p_s)a_s$, this only means setting $a_s = \mathcal{A}_s$, for s such that p_s exceeds a threshold, and $a_t = 0$ for p_t falling under the same threshold. If the set of stratum attributions minimizing $d(t, I; \cdot)$ is a singleton, $\{\mathbf{a}\}$, then its member is the separability solution.

There are multiple minimizers of $d(t, I; \mathbf{a})$ only if two or more strata $s_1, \ldots s_k$ have $p_{s_1} = \cdots = p_{s_k}$, while a falls between $\sum(\mathcal{A}_t : p_t < p_{s_1})$ and $\sum(\mathcal{A}_t : p_t \leq p_{s_1})$. To break such a tie, the separable algorithm selects $a_{s_1}, \ldots, a_{s_k}$ so as to maximize the contributions to $\tilde{\mathrm{V}}\mathrm{ar}(t; \mathbf{a})$ from strata $s_1, \ldots, s_k$, subject to $0 \leq a_{s_i} \leq \mathcal{A}_{s_i}$, $i = 1, \ldots, k$, and $\sum_i a_{s_i} = a - \sum(\mathcal{A}_t : p_t < p_{s_1})$. Thus the selected $\mathbf{a}$ minimizes the magnitude of $z(\mathbf{a})$ among $\mathbf{a}^*$ minimizing $d(t, I; \mathbf{a}^*)$ subject to constraints. The separability algorithm then rejects $[A = a]$ at level α if $z(\mathbf{a}) > z_{1-\alpha}$.

This separable optimization is simpler than direct minimization of $z(\mathbf{a})$ subject to $0 \leq \mathbf{a} \leq \mathcal{A}$ and $\mathbf{a}'\mathbf{1} = a$, or "joint optimization" (Gastwirth et al. 2000), and unlike joint optimization it is always computationally feasible. Ideally the separable and joint optima, $z(\mathbf{a}_*)$ and $m(a)$, coincide, or differ by very little, but there are cases in which they meaningfully differ; in such cases, tests based on separable optimization may exceed their nominal levels. Gastwirth et al.'s Proposition 1 protects separable optimization from this shortcoming in *matched* designs with large samples; what of our stratified but unmatched design?

3.3 Large-sample theory for stratified designs

New theory is needed for unmatched samples with a limited number of strata. The proposition to follow covers this case as well as Gastwirth et al.'s, showing that in sufficiently large samples the separable optimum z-score coincides with the joint optimum. We invoke a triangular array of assignment units, here clusters. For studies $\kappa = 1, 2, \ldots$, let the subjects be arranged in N_κ clusters, growing in number without limit but uniformly bounded in size. These clusters sit in strata $U_{\kappa 1}, \ldots, U_{\kappa S}$, within the sth of which $n_{\kappa s}$ of $N_{\kappa s}$ clusters are assigned to the treatment group, I_κ, of which a subset C_κ complies with treatment. The vectors $t_\kappa^{(s)}$ record cluster totals of responses in stratum s of study κ, and their concatenation is t_κ. The largest possible total of effect attributions in $U_{\kappa s}$, $\sum_{U_{\kappa s} \cap C_\kappa} t_{\kappa i}^{(s)}$, is denoted $\mathcal{A}_{\kappa s}$, and $\mathcal{A}_\kappa$ stands for $(\mathcal{A}_{\kappa 1}, \ldots, \mathcal{A}_{\kappa S})$. Write $p_{\kappa s} = n_{\kappa s}/N_{\kappa s}$, $\Psi = \{p_{\kappa s} : \kappa = 1, 2, \ldots, 1 \leq s \leq S_\kappa\}$, $\psi_l = \inf \Psi$, $\psi_u = \sup \Psi$; $\Delta = \{p_{\kappa s} - p_{\kappa t} : p_{\kappa s} > p_{\kappa t}, \kappa = 1, 2, \ldots, 1 \leq s, t \leq S_\kappa\}$, $\delta = \inf \Delta$; and $\tilde{\Sigma} = \{\tilde{s}^2(y_\kappa^{(s)}; a) : \kappa = 1, 2, \ldots, s = 1, \ldots, S_\kappa, 0 \leq a \leq \mathcal{A}_{\kappa s}\}$, and $\tilde{\sigma}^2 = \inf \tilde{\Sigma}$.

Proposition 3.1 *Assume $\delta, \tilde{\sigma} > 0$; $0 < \psi_l, \psi_u < 1$. Suppose $0 < \alpha < 1/2$, and level-α tests of hypotheses $[\mathbf{A} = \mathbf{a}]$ against $[\mathbf{A} > \mathbf{a}]$ (or against $[\mathbf{A} < \mathbf{a}]$) have acceptance regions of form $d(t, I; \mathbf{a})/\mathrm{Var}(t; \mathbf{a})^{1/2} \leq z_{1-\alpha}$ (respectively, $d(t, I; \mathbf{a})/\mathrm{Var}(t; \mathbf{a})^{1/2} \geq z_\alpha$). Then there exists κ_0 such that for all $\kappa > \kappa_0$ and compatible $[A = a]$, any separable optimizer $\mathbf{a}_*$ of $[A = a]$ against $[A > a]$ (respectively, $[A < a]$) is such that $[\mathbf{A} = \mathbf{a}_*]$ is rejected at level α if and only if all compatible $[\mathbf{A} = \mathbf{a}]$ such that $\sum_s a_s = a$ are rejected at level α.*

A proof of Proposition 3.1 is given in the Appendix. For the Adams and Smith study, as recast in § 3.1, $(p_1, p_2) = (.42, .52)$, and the joint optimizers found in §3.1 are the same stratum attributions that separable optimization would have produced. Is this also true of the New Haven experiment?

3.4 Telephone and Mail GOTV effects via separable optimization

To test hypotheses about effects of telephone calls, we consider the sample as stratified by assignment to in-person GOTV, yes or no, and by the number of direct mailings sent to a household, 0, 1, 2, or 3; this gives 8 strata. For hypotheses about the effects of mail, we use the 2×2 stratification in terms of (attempted) in-person and telephone GOTV. Testing hypotheses that no votes are attributable to these treatments requires testing only one simple hypothesis for each treatment; for neither of these treatments can this hypothesis be rejected at conventional levels ($p = .64$ and $.37$, respectively, two-sided). Tests of hypotheses $[A = a]$, $a > 0$, require separable or joint optimization.

In order to get two-sided hypothesis tests from one-sided tests, as in Proposition 3.1, say that $[A = a]$ is rejected at level α if for all compatible $[\mathbf{A} = \mathbf{a}]$ such that $\mathbf{a}'\mathbf{1} = a$, $[\mathbf{A} = \mathbf{a}]$ is rejected when tested at level $\alpha/2$ against $[\mathbf{A} \geq \mathbf{a}]$, or if all such $[\mathbf{A} = \mathbf{a}]$ are rejected in level $\alpha/2$ tests against $[\mathbf{A} \leq \mathbf{a}]$. Tested in this way at the 2/3 level, hypotheses attributing $A = 1, 2, \ldots, 35$ votes to telephone intervention are sustained, as for each of them the separable optimization gives at least one stratum attribution whose z-statistic falls above $z_{1/6} = -.96742$ (as well as many that fall below $z_{5/6}$). For $[A = 36]$, the largest z-statistic the routine locates is $z = -.96744$, just below $z_{1/6}$. Assuming the sample is large enough that 3.1 applies, every stratum attribution falling under $[A = 36]$ has a z-statistic less than $z_{1/6}$, entailing rejection of the composite hypothesis. The 2/3-confidence interval extends from zero up to only 35 votes; with 2/3 confidence, fewer than 2.2% of GOTV calls generated a vote.

For the mail intervention, $[A = 0]$ is also within the 2/3 confidence interval. The upper end of the confidence interval is much larger, 652 votes; since 11,200 households were sent a mailer, this translates to an upper limit of 5.8% of mailed households' having someone who voted because of the mailing. Again, this statement holds with 2/3 confidence, and also assumes that the sample is large enough for Proposition 3.1 to apply.

4 Adding covariates for precision

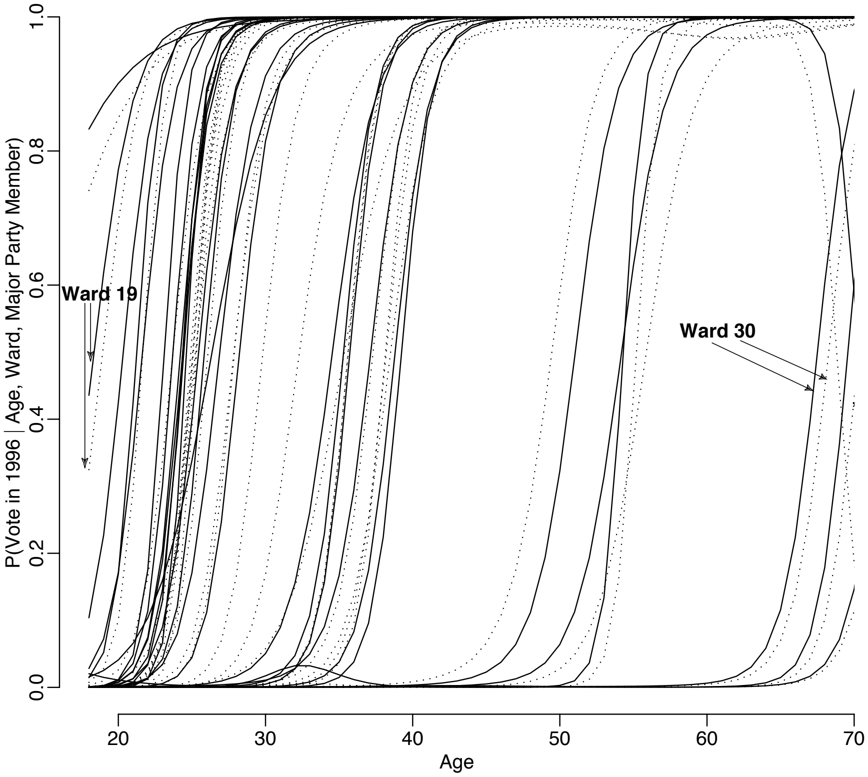
So far we have estimated GOTV effects for the in-person treatment using only treatment assignment, compliance and outcome data, but ignoring potentially quite informative covariates. Besides outcome and intervention data, Gerber and Green collected demographic information from voter rolls, specifically voters' ages, wards of residence, and whether they were members of a major political party, along with their registration status and voting in the November election two years before. These data are powerful predictors of future voting, and the estimation procedure shouldn't ignore them.

A convenient model to relate voting, V, to demographic characteristics, D, and household (H) is

$$\operatorname{logit}(\mathbf{P}(V|D, H)) = l + D\beta + \gamma_H, \tag{2}$$

where l is an election-specific intercept and $\gamma_H \sim N(0, \sigma^2)$ is a household-specific random effect. To make use of data from several elections, one could add individual

Figure 1: Fitted 1996 voting probabilities, conditional on ward (separate lines), age, and membership in a major political party (solid lines for members, dotted for nonmembers). The marked differences between curves reflect the covariates' high prognostic value. For example, people living in Ward 19 were predicted to vote with very high probability, both young and old, whereas subjects in Ward 30 were generally unlikely to vote. Ward 19 is roughly coterminous with the affluent Yale faculty neighborhood of Prospect Hill, while Ward 30 sits in the poorer West Rock neighborhood, where nearly half of households earned less than $10,000 as of Census 2000.



random effects to the model and fit it to the available elections simultaneously; Gerber and Green collected data on just one prior election, however, so we do not pursue this here. Fit to voting and demographic data from a prior election or elections, the model produces fitted probabilities $\hat{\mathbf{P}}(V|D,H)$ that smooth subjects' binary voting indicators, borrowing information from demographically similar subjects to appraise the certainty that their voting behavior would turn out as it did. Figure 1 plots age against fitted 1996 voting probabilities given ward of residence, age, and major party membership, demonstrating pronounced geographic and generational trends.

To estimate (2), we restrict our sample to the 24,300 subjects who were registered in New Haven as of the previous election. To create D, we expanded the age variable into natural cubic splines with knots at quintiles of the age distribution, included indicator variables for the 29 wards represented in the study, and added major party membership as another indicator variable; then we included also first-order interactions of these. This expanded covariate basis had a few hundred elements, less than a hundredth of the overall number of study subjects. The mixed logistic regression model accommodated overdispersion and was fit by the Laplace method, using the `lmer` function from Bates and Maechler's "Matrix" package (2006) for R. The fit yielded covariate coefficients and, for each household with a voter registered for the previous election, a random deviation from the overall intercept. To obtain 1996 voting probabilities for all subjects on the rolls at the 1998 election, households without a voter registered in New Haven in 1996 were assigned a deviation of zero.

Because overall turnout varies systematically between presidential and midterm elections (Rosenstone and Hansen 1993, p.57), it would be incorrect to use these as probabilities of voting in an upcoming election; but if (2) is generally correct, then the sum $D\beta + \gamma_H$ is a sufficient statistic with which to predict voting in an upcoming cycle, a prognosis score (Hansen 2006b). On the other hand, were we to misspecify the prognostic model, or otherwise poorly estimate its score, we would introduce no marginal bias, nor jeopardize the legitimacy of randomization-based tests: the potential penalties are conditional bias, and deficits of efficiency relative to inference based upon better estimated scores.

These individual-level prognosis scores were used to subclassify the sample of households. After splitting on household size, into one- and two-voter households, we partitioned the sample of one-voter households at the quintiles of its prognosis scores, and partitioned two-voter households first at teriles of household mean prognosis scores, then within each terile at the median of within-household ranges of scores. The resulting 11 prognostic subclasses were then crossed with the complementary treatment subclassifications, leaving in the case of the telephone experiment, for example, an $11 \times 2 \times 4$-way cross-tabulation, prognosis score by in-person assignment (treatment or control) by number of mailings sent (0,1,2, or 3). The in-person experiment is also given a $11 \times 2 \times 4$-way subclassification, prognosis on telephone on mailings, while the mail experiment was given a $11 \times 2 \times 2$-way prognosis by telephone by in-person assignment subclassification. We then proceed with inference for each experiment as if its treatment had been assigned to simple random samples within each of the resulting subclasses, rather than to random samples within the more coarse subclassification along complementary treatments. This amounts to narrowing $\mathbb{I}$, the set of potential treatment assignments to which the actual treatment assignment I is to be compared, to a class of assignments relatively similar to I in terms of the prognostic comparability of their treatment and control groups – a step consonant with the conditionality principle (see *e.g.* Barndorff-Nielsen and Cox 1994, ch. 2).

Type of GOTV	Point Estimate	CIs 2/3	CIs 95%
phone	0	0 to 2	0 to 5
mail	2	0 to 7	0 to 9
in-person	9	6 to 13	3 to 16

Table 3: Votes attributable to GOTV interventions, per 100 contacts. These inferences stratify on prognosis scores and complementary treatments.

Table 3 gives 2/3 and 95% confidence intervals derived by this method, in terms of votes per contacted household. While the confidence intervals overlap, the results clearly suggest an ordering of effectiveness of the interventions, with personal canvassing the most and telephone GOTV the least effective.

A comparison with confidence intervals that would have been obtained without the additional subclassification demonstrates the benefit of prognosis scoring. Without it, 2/3 and 95% interval estimates of the in-person benefits would have been 12 and 16% wider, respectively. Before comparing widths of intervals for mail and telephone effects, for intervals that meet 0 we substitute twice the upper half-width, or distance from the point estimate to the confidence interval's upper limit, for their lengths, recognizing that the intervals have been limited *a priori* to nonnegative values. By this measure, prognosis scoring improves 2/3 and 95% intervals for the telephone effect, and 2/3 and 95% intervals for mailer effects, by 3%, 10%, 17%, and 2%, respectively.

5 Validating the result of separable optimization

In marked contrast with both Imai's and Gerber and Green's inferences, ours have assumed little other than that the households were properly randomized. However, since we have relied on assumptions about large sample sizes , one might worry that our analysis has traded uncertainty about assumptions for uncertainty as to whether asymptotics apply. To remove this remaining uncertainty, this section explains a way to check whether the separable and joint optimizers coincide, and to bound the discrepancy between them if they do not, without relying on a large-sample justification. It is more technical than previous sections, and readers not concerned with this issue should skip it.

5.1 Testing $[A = a]$ as a convex minimization problem

Among compatible hypotheses $[t_c = \tilde{t}_c]$ falling under the composite hypothesis $[\mathbf{A} = \mathbf{a}]$, the supremum of $\mathrm{Var}\left(\sum_{j \in \mathcal{I}} t_j\right)$ is $\mathrm{Var}\,(t; \mathbf{a}) = \sum_s n_s(1 - n_s/N_s)s^2(t^{(s)}; a_s)$, where $t^{(s)} = (t_i : i \in U_s)$ and $s^2(t; a)$ is as in (1). Since all such $[t_c = \tilde{t}_c]$ share the same value $d(t, I; \mathbf{a})$ of $\sum_I \tilde{t}_{ci} - \mathbf{E} \sum_{i \in \mathcal{I}} \tilde{t}_{ci}$, we accept $[\mathbf{A} = \mathbf{a}]$ in a level-α test against $[\mathbf{A} \geq \mathbf{a}]$, $\alpha < \frac{1}{2}$, if and only if

$$g(t; \mathbf{a}) = \max(0, d(t, I; \mathbf{a}))^2 - z_{1-\alpha}^2 \mathrm{Var}(t; \mathbf{a}), \tag{3}$$

where $\mathrm{Var}(t; \mathbf{a}) = \sum_s n_s(1 - p_s)s^2(t^{(s)}; a_s)$, falls at or below 0. By extension, $[A = a]$ is accepted at level α if and only if the minimum of $g(t; \cdot)$, constrained by $\sum_s a_s = a$ and $0 \leq \mathbf{a} \leq \mathcal{A}$, falls at or below 0. What sort of function is $g(t; \cdot)$?

$\mathbf{a} \mapsto d(t, I; \mathbf{a})^2$ has a positive-definite Hessian, $2\{(1 - p_s)(1 - p_t) : s, t \leq S\}$, and is convex; and the set of $\mathbf{a}$ for which $d(t, I; \mathbf{a}) \leq 0$ is convex. Thus the first term of (3) is convex in $\mathbf{a}$. However, $a \mapsto s^2(t^{(s)}; a)$ is neither convex nor concave, with the result that $g(t; \cdot)$ is not generally convex. This complicates its minimization. We endeavor to replace g with a close, convex approximation.

The quantities $s^2(t^{(s)}; a_s)$ contributing to g are closely bounded above (since the third term of (1) has form $\gamma_0^2 x^2$, $|x| < 1$, and is $\leq \gamma_0^2 x$) by $\tilde{s}^2(t^{(s)}; a_s) =$

$$\frac{1}{(N-1)} \left[\sum_{\substack{k \in U_s : k \notin C, \\ t_k < \gamma_s}} t_k^2 + \left(N_s - \frac{r_s}{\gamma_s} \right) \gamma_s^2 + \sum_{\substack{k \in U_s : \\ \gamma_s < t_k}} t_k^2 \right] - \frac{N}{N-1} \left(\frac{\sum_{k \in U_s} t_k - a}{N} \right)^2,$$

with (γ_s, N_s, r_s) depending on $(t^{(s)}, a_s)$ as (γ_0, N_0, r_0) depend on (t, a) in (1). These functions $\tilde{s}^2(t^{(s)}; \cdot)$ are continuous, piecewise quadratic and concave. Consequently

$$\tilde{g}_\alpha(t, I; \mathbf{a}) = \max(0, d(t, I; \mathbf{a}))^2 - z_{1-\alpha}^2 \tilde{\mathrm{V}}\mathrm{ar}(t; \mathbf{a}), \tag{4}$$

where $\tilde{\mathrm{V}}\mathrm{ar}(t; \mathbf{a}) = \sum_s n_s(1 - p_s)\tilde{s}^2(t^{(s)}; a_s)$, is convex in $\mathbf{a}$. At the same time, $\tilde{g}_\alpha(t; \mathbf{a})$ bounds g from above, and should do so closely. (For tests of $[\mathbf{A} = \mathbf{a}]$ against $[\mathbf{A} \leq \mathbf{a}]$, substitute $\min(0, d(t, I; \mathbf{a}))$ for $\max(0, d(t, I; \mathbf{a}))$ in (3) and (4).)

Although in principle using $\tilde{g}$ rather than g to delineate confidence regions can lead to wider intervals, in the Vote 98 study it does not. Evaluated at the 12 separable optimizers $\mathbf{a}$ of total effect hypotheses $[A = a]$, for a just inside the boundaries of the 2/3 and 95% CIs given in Table 3 — at the most nearly rejected of those stratum attributions that we tested but did not reject at levels 1/3 and .05 — $\mathrm{Var}(t;\mathbf{a})$ and $\tilde{\mathrm{V}}\mathrm{ar}(t;\mathbf{a})$ very nearly coincide, with the consequence that $g_\alpha(t,I;\mathbf{a})$ and $\tilde{g}_\alpha(t,I;\mathbf{a})$ are for practical purposes the same. The differences are small enough as to produce differences no larger than 2×10^{-5} in Normal approximation p-values calculated from $z(\mathbf{a}) = \mathrm{Var}(t;\mathbf{a})^{-1/2}d(t,I;\mathbf{a})$ and from $\tilde{z}(\mathbf{a}) = \tilde{\mathrm{V}}\mathrm{ar}(t;\mathbf{a})^{-1/2}d(t,I;\mathbf{a})$, occasioning no change to Table 3.

In sum, a valid test of $[A = a]$ rejects if the minimum of $\tilde{g}_\alpha(t,I;\mathbf{a})$, constrained by $\sum_s a_s = a$ and $0 \leq \mathbf{a} \leq \mathcal{A}$, is positive. The test is similar to, if potentially more conservative than, tests based on whether unmodified z-statistics $\mathrm{Var}(t;\mathbf{a})^{-1/2}d(t,I;\mathbf{a})$ fall within level-α acceptance limits. Applied to the Vote 98 study, these tests give only negligibly different results.

5.2 Comparing separable and joint optima

Proposition 3.1 continues to hold if $\mathrm{Var}(t;\mathbf{a})$ is replaced with $\tilde{\mathrm{V}}\mathrm{ar}(t;\mathbf{a})$; in sufficiently large samples the modified z-statistic can also be separably optimized. The modification is advantageous when the sample is not known to be sufficiently large, because since $\tilde{g}_\alpha(t,I;\cdot)$, unlike $g_\alpha(t,I;\cdot)$, is convex, its joint optimization is tractable: since the constraint space is also convex, $\tilde{g}$'s local minima are also global minima.

Even with separable optimization, which may be more convenient than joint optimization, convexity of $\tilde{g}$ helps. By checking whether a separable optimizer $\mathbf{a}$ attains a local minimum of $\tilde{g}_{p(\mathbf{a})}(t,I;\cdot)$, where $p(\mathbf{a})$ is the p-value of $[\mathbf{A} = \mathbf{a}]$, one can determine conclusively whether it is a joint optimizer. If the separable optimizer is also a joint optimizer, then its acceptance or rejection decides that of $[A = a]$, $a = \sum_s a_s$; if not, then it can be used as a starting point for optimization of $\tilde{g}_\alpha(t,I;\cdot)$, or combined with derivatives in order to linearly approximate $\tilde{g}_\alpha(t,I;\cdot)$ from below. (The necessary derivatives are given by Lemma 6.1.)

Table 4 uses this refinement technique to bound from above the jointly optimized p-values of those a that fall just above 95% CI limits obtained by separable optimization. Only upper boundaries of the 95% CIs are shown because just outside of the lower limits of the separable 95% interval, and just outside either end of the separable 2/3 intervals, separably and jointly optimized p-values coincide. Of the CI limits reported in Table 3, refinement changes only the upper limit of the 95% interval for mail, extending it by 9 votes, or .1 votes per 100 households mailed.

In sum, confidence intervals produced by inversion of separably optimized hypothesis tests may be anti-conservative. With additional effort on the part of the analyst,

		p-values	
Intervention	$a =$	separable	upper bound
phone	96	.02415	.02426
mail	1060	.02483	.02686
in-person	227	.02391	.02432

Table 4: Separable p-values and their refinements, at the upper boundaries of the 95% (two-sided) confidence intervals reported in Table 3. The separable p-values are one-sided, assessing a hypothesis $[\mathbf{A} = \mathbf{a}_*]$ against $[\mathbf{A} \leq \mathbf{a}_*]$, where $\mathbf{a}_*$ is the separable optimizer for tests of $[A = a]$ against $[A \leq a]$; the next column refines these to obtain upper bounds for the maximum of p-values for $[\mathbf{A} = \mathbf{a}]$ against $[\mathbf{A} \leq \mathbf{a}]$ when $\mathbf{a} \in \{0 \leq \mathbf{a} \leq \mathcal{A}, \mathbf{a}'\mathbf{1} = a\}$. Of the upper and lower limits of 2/3 and 95% CIs for each of the three intervention effects, refinement increases only the upper limit of the 95% CI for mail (from 1059 to 1068).

this anti-conservatism can be replaced with a milder conservatism, slightly overestimating tests statistics' variance in the interests of simplifying joint optimization. Doing so with the Vote 98 study prompts very small revisions to confidence intervals.

6 Discussion

6.1 Methodology

Imai's re-analysis involved a propensity score matching of complying treatment group members to controls. In their response, Gerber and Green (2005a, pp.307–308) suggest that Imai's use of propensity scores caused his to conclusion to err. If his method is mistaken — more on this presently — then we believe a more likely mistake is his premise that contacted and noncontacted voters' potential outcomes are the same, conditional upon covariates. It makes his a form of "as-treated" analysis, which is known to perform unreliably in medical contexts (Lee et al. 1991). Using Gerber and Green's originally released replication archive, which did not include household identifiers, Bowers and Hansen (2005) propensity matched the New Haven study subjects who were and were not assigned to the various GOTV treatments, treating assignment as an instrument for receipt of treatment in the subsequent analysis. Results of their analysis closely resembled those obtained here; see also Hill et al. (2000) for discussion of the use of propensity scores in experiments.

Several facets of our methodology merit discussion. The tests and p-values that we have presented are not exact, even those that pertain to simple hypotheses, as they depend on an asymptotic Normal approximation. However, they more closely approximate exact p-values than do many large-sample p-values, since they are based on exact, not estimated, variances; see Hansen (2006a) for confirmation of this in simulation experiments. By the Berry-Esseen principle for simple random samples (Höglund 1978), the error of the Normal approximations on which these simple tests are based converges uniformly to zero as the sample size increases. It follows that the error of our tests of composites $[A = a]$ vanishes in large samples also.

Our analysis considered only hypotheses compatible with nonnegativity, $y \geq y_c$. A pronounced failure of nonnegativity, for instance nonpositivity ($y \leq y_c$) coupled with a sizable negative effect attributable to treatment, would lead to rejection of each nonnegative hypothesis, making the confidence interval empty. Had that result been obtained, we could have then tested hypotheses assuming nonpositivity, culminating in a confidence interval for a negative effect. Observe that this procedure does not risk swelling type I errors because of multiple comparisons, since the nonnegative and nonpositive hypotheses are mutually exclusive.

Section 4's covariate adjustment strategy is not necessarily the most efficient possible. Another form of covariate adjustment would appraise null hypotheses $[y_c = \tilde{y}_c]$ by fitting (2) to pre-experimental data and to the hypothesized outcomes $\tilde{y}_c$, and would then decide the hypothesis using a randomization test of similarity between control and treatment groups' residuals. Gail et al. (1996) test the strict null $[y_c = y_{obs}]$ in this way; Rosenbaum (2002b) and Raab and Butcher (2005) discuss adjustments of this type for studies with continuous outcomes, where the model of a constant treatment effect may be appropriate. We experimented with adapting this approach to estimation of attributable effects, finding it to be delicate, at times widening confidence intervals; using different methodology, Nixon and Thompson (2003) also find that covariance adjustment need not increase power in group randomized trials. Hansen (2006b) gives prognosis scoring strategies that make use of control-group outcomes, but these carry the possibility, at least in principle, of introducing bias, making them less suitable for our aim of resolving the methodological dispute over the Vote 98 study. More research is needed to determine which methods most reliably and effectively leverage covariates for increased precision in research designs of this type.

The attraction of § 4's use of prognosis scores is that it is equally applicable to study analysis and design. We used it for analysis, estimating a prognosis score *ex post facto* and then poststratifying the sample; but since our fit used only data from prior elections, precisely the same score could have been estimated *ex ante* and used in a block-randomized design. Such a design could have enjoyed a greater efficiency benefit.

6.2 Revisiting the Vote 98 controversy

Using model-based methods, Gerber and Green (2000) produced 95% confidence intervals of 5 to 14 percent increases in turnout among contacted voters, for direct face-to-face appeals, but 8 to 0 percent *decreases* in turnout among voters contacted by telephone; a revision accounting for clustering (Gerber and Green 2005a) slightly widened these confidence intervals and shifted them toward zero. Imai's (2005) analysis gave similar results for face-to-face contact but found telephone entreatments to increase turnout, with a 95% CI of 1 to 13 percentage points. The finding of a benefit for phone GOTV has a precedent in earlier literature on canvassing by telephone (Eldersveld 1956; Miller et al. 1981), although experiments conducted since the mid-1990s, when survey organizations began to see steep declines in telephone response rates (Curtin et al. 2005), have been unable to replicate it (Gerber and Green 2005b).

Imai's adjustment assumes that conditional on available covariates, compliers' likelihood of voting equals non-compliers'. Acknowledging that this premise is open to debate, he maintains his adjustment to be "certainly more appropriate than the method of instrumental variables used by Gerber and Green" (2005, p.295). This claim has been favorably received among at least some political professionals. Grenzke and Watts (2004), for example, write approvingly that Imai's article "corrects errors in Green and Gerber's 1998 New Haven study and finds that even with weak, nonpartisan scripts, phones increased turnout," alluding to unspecified "post-election analysis" in support of telephone GOTV.

Our randomization analysis speaks more directly to the question of which of Gerber and Green's (2005a) and Imai's (2005) corrections of the earlier analysis by Gerber and Green (2000) is more correct, estimating the same quantities without making either of the later papers' assumptions. Imai's claim is mistaken; results of the randomization-based analysis are in accord with Gerber and Green's, and certainly contradict those of the propensity analysis without the instrumental variable. Its 95% CI for the in-person benefit contains and slightly widens Gerber and Green's; more important, it allows no more than 5% of voters contacted by telephone to have turned out because of the call. With this, results of recent telephone canvassing experiments are definitive. The era of anonymous telephone mobilization has passed.

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

Barndorff-Nielsen, O. E. and Cox, D. R. (1994), *Inference and asymptotics*, Chapman & Hall Ltd.

Bates, D. and Maechler, M. (2006), *Matrix: A Matrix package for R*, r package version 0.995-16.

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.

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.

Copas, J. B. (1973), “Randomization models for the matched and unmatched 2×2 tables,” *Biometrika*, 60, 467–476.

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

Curtin, R., Presser, S., and Singer, E. (2005), “Changes in Telephone Survey Nonresponse over the Past Quarter Century,” *Public Opinion Quarterly*, 69, 87–98.

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

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.

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

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996), “On Design Considerations and Randomization-based Inference for Community Intervention Trials,” *Statistics in Medicine*, 15, 1069–1092.

Gastwirth, J., Krieger, A., and Rosenbaum, P. R. (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.

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

— (2005b), "Do Phone Calls Increase Voter Turnout?: An Update," *The Annals of the American Academy of Political and Social Science*, 601, 142–154.

Greevy, R., Silber, J. H., Cnaan, A., and Rosenbaum, P. R. (2004), "Randomization Inference with Imperfect Compliance in the ACE-Inhibitor After Anthracycline Randomized Trial," *Journal of the American Statistical Association*, 99, 7–15.

Grenzke, J. and Watts, M. (2004), "Hold the Phones; Taking Issue with a get-out-the-vote strategy," *Campaigns and Elections*.

Hamilton, M. A. (1979), "Choosing the Parameter for a 2 × 2 Table or a 2 × 2 × 2 Table Analysis," *American Journal of Epidemiology*, 109, 362–375.

Hansen, B. B. (2006a), "Appraising Covariate Balance after Assignment to Treatment by Groups," Tech. Rep. 436, University of Michigan, Statistics Department.

— (2006b), "Bias Reduction in Observational Studies via Prognosis Scores," Tech. Rep. 441, University of Michigan, Statistics Department.

Hill, J., Rubin, D. B., and Thomas, N. (2000), "The Design of the New York School Choice Scholarship Program Evaluation," in *Research Design: Donald Campbell's Legacy*, ed. Bickman, L., Sage Publications.

Hodges, J. L. and Lehmann, E. L. (1963), "Estimates of location based on rank tests (Ref: V42 p1450-1451)," *The Annals of Mathematical Statistics*, 34, 598–611.

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.

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.

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*, 2, 445–460.

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

Nixon, R. and Thompson, S. (2003), "Baseline adjustments for binary data in repeated cross-sectional cluster randomized trials," *Statistics in Medicine*, 22, 2673–2692.

Raab, G. M. and Butcher, I. (2005), "Randomization inference for balanced cluster-randomized trials," *Clinical Trials*, 2, 130–140.

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.

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

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

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.

Stoker, L. and Jennings, M. (1995), "Life-Cycle Transitions and Political Participation: The Case of Marriage," *The American Political Science Review*, 89, 421–433.

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.

Appendix

Proof of Proposition 2.1

Consider first that among compatible hypotheses $[t_c = \tilde{t}_c]$ such that $\sum_{i\in I} t_i - \tilde{t}_{ci} = \sum_{k=1}^{N} t_k - \tilde{t}_{ck} = a$, only the $\sum_k \tilde{t}_{ck}^2$ term on the right of

$$\mathrm{Var}(\sum_{j\in\mathcal{I}} \tilde{t}_{cj}) = n\left(1 - \frac{n-1}{N-1}\right)\left(\frac{\sum_{k=1}^{N}\tilde{t}_{ck}^2}{N} - \left(\frac{\sum_{k=1}^{N}\tilde{t}_{ck}}{N}\right)^2\right)$$

may vary. Second, the existence of a γ_0 as in the proposition is logically equivalent to the statement that there is no pair $l, k \leq N$ such that $0 < \tilde{t}_l, \tilde{t}_k < t_l$. Third, if such a pair existed, then $\sum_k \tilde{t}_{ck}^2$ would not be at its largest subject to $\sum_{k=1}^{N} t_k - \tilde{t}_{ck} = a$ and compatibility constraints: ensure that $\tilde{t}_k \leq \tilde{t}_l$, by switching k and l if need be; then construct $\tilde{t}^*_{cm}$ such that $\tilde{t}^*_{cl} = \min(\tilde{t}_{cl} + \tilde{t}_{ck}, t_l)$, $\tilde{t}^*_{ck} = \max(0, t_l - \tilde{t}_{cl} - \tilde{t}_{ck})$, and $\tilde{t}^*_{cm} = \tilde{t}_{cm}$, for $m \neq l, k$. Then $\tilde{t}^*_{cl} > \max(\tilde{t}_{ck}, \tilde{t}_{cl})$, so that $(\tilde{t}^*_{ck})^2 + (\tilde{t}^*_{cl})^2 > (\tilde{t}_{ck})^2 + (\tilde{t}_{cl})^2$, and consequently $\sum_k (\tilde{t}^*_{ck})^2 > \sum_k \tilde{t}_{ck}^2$. This establishes (i); the remainder of the proposition follows.□

Proof of Proposition 3.1

First, a lemma. Write δ_i for the unit S_κ-vector with 0's in all but the ith position.

Lemma 6.1 $\tilde{h}_\kappa(\mathbf{a}) = \tilde{\mathrm{V}}\mathrm{ar}(t_\kappa; \mathbf{a})$ *has directional derivatives*

$$\begin{aligned}
\partial\tilde{h}_\kappa(\mathbf{a}; \delta_i) &= C_{\kappa i}\left[2\left(\bar{t}_{\kappa i} - \frac{a_{\kappa i}}{N_{\kappa i}}\right) - \gamma(i, \kappa, a)^+\right], \\
-\partial\tilde{h}_\kappa(\mathbf{a}; -\delta_j) &= C_{\kappa j}\left[2\left(\bar{t}_{\kappa j} - \frac{a_{\kappa j}}{N_{\kappa j}}\right) - \gamma(j, \kappa, a)^-\right],
\end{aligned}$$

where $C_{\kappa s} = (n_{\kappa s}/N_{\kappa s})[1 - (n_{\kappa s} - 1)/(N_{\kappa s} - 1)]$, $\bar{t}_{\kappa s} = \sum_{k\in U_{\kappa s}} t^{(s)}_{\kappa k}/N_{\kappa s}$, $\gamma(s, \kappa, a)^+ = \max\{\gamma : \sum[t_{\kappa k} : k \in C, t_{\kappa k} < \gamma] \leq a\}$ *and* $\gamma(s, \kappa, a)^- = \max\{\gamma : \sum[t_k : k \in C, t_{\kappa k} < \gamma] <$

$a\}$. *Also* $h_\kappa(\mathbf{a}) = \mathrm{Var}(t_\kappa;\mathbf{a})$ *has directional derivatives*

$$\begin{aligned} \partial h_\kappa(\mathbf{a};\delta_i) &= 2C_{\kappa i}\left\{\left(\bar{t}_{\kappa i} - \frac{a_{\kappa i}}{N_{\kappa i}}\right) - \gamma(i,\kappa,a)^+\left[1 - \left(\frac{r(i,\kappa,a)^+}{\gamma(i,\kappa,a)^+} - \left\lfloor\frac{r(i,\kappa,a)^+}{\gamma(i,\kappa,a)^+}\right\rfloor\right)\right]\right\}, \\ -\partial h_\kappa(\mathbf{a};-\delta_j) &= 2C_{\kappa j}\left\{\left(\bar{t}_{\kappa j} - \frac{a_{\kappa j}}{N_{\kappa j}}\right) - \gamma(j,\kappa,a)^-\left[1 - \left(\frac{r(i,\kappa,a)^-}{\gamma(i,\kappa,a)^-} - \left\lfloor\frac{r(i,\kappa,a)^-}{\gamma(i,\kappa,a)^-}\right\rfloor\right)\right]\right\}, \end{aligned}$$

where $r(s,\kappa,a)^+ = a - \sum[t_k : k \in C, t_k < \gamma(s,\kappa,a)^+]$ *and* $r(s,\kappa,a)^- = a - \sum[t_k : k \in C, t_k < \gamma(s,\kappa,a)^-]$. *In particular,* $|\partial h_\kappa(\mathbf{a};\delta_i)|$, $|\partial h_\kappa(\mathbf{a};-\delta_j)|$, $|\partial \tilde{h}_\kappa(\mathbf{a};\delta_i)|$, *and* $|\partial \tilde{h}_\kappa(\mathbf{a};-\delta_j)|$ *are uniformly bounded, as are* $|\partial h_\kappa(\mathbf{a};\delta_i - \delta_j)|$ *and* $|\partial \tilde{h}_\kappa(\mathbf{a};\delta_i - \delta_j)|$.

Lemma 6.1 follows from the fact that $\tilde{s}^2(t_\kappa^{(s)};\cdot)$ and $s^2(t_\kappa^{(s)};\cdot)$ are piecewise differentiable with

$$\begin{aligned} \frac{\partial^+}{\partial a}\tilde{s}^2(t_\kappa^{(s)};a) &= -\frac{\gamma(s,\kappa,a)^+}{N_{\kappa s}-1} + \overbrace{\frac{2}{N_{\kappa s}-1}\left(\frac{\sum_{k\in U_s} t_{\kappa k}^{(s)}}{N_{\kappa s}} - \frac{a}{N_{\kappa s}}\right)}^{*}, \\ \frac{\partial^+}{\partial a}s^2(t_\kappa^{(s)};a) &= -\frac{2\gamma(s,\kappa,a)^+}{N_{\kappa s}-1}\left[1 - \left(\frac{r(i,\kappa,a)^+}{\gamma(i,\kappa,a)^+} - \left\lfloor\frac{r(i,\kappa,a)^+}{\gamma(i,\kappa,a)^+}\right\rfloor\right)\right] + *, \\ \frac{\partial^-}{\partial a}\tilde{s}^2(t_\kappa^{(s)};a) &= -\frac{\gamma(s,\kappa,a)^-}{N_{\kappa s}-1} + *, \;\; \text{and} \\ \frac{\partial^-}{\partial a}s^2(t_\kappa^{(s)};a) &= -\frac{2\gamma(s,\kappa,a)^-}{N_{\kappa s}-1}\left[1 - \left(\frac{r(i,\kappa,a)^-}{\gamma(i,\kappa,a)^-} - \left\lfloor\frac{r(i,\kappa,a)^-}{\gamma(i,\kappa,a)^-}\right\rfloor\right)\right] + *, \end{aligned}$$

which follow directly from the definition of $\tilde{s}$.

Turning to Proposition 3.1, its 'if' part is immediate. We prove the other implication as applied to tests of $[A = a]$ against $[A > a]$; its demonstration for tests of $[A = a]$ against $[A < a]$ is analogous. For $a > 0$ let the S_κ-vector of positive integers $\mathbf{a}(a,\kappa)$ be a separable optimizer. For S_κ-vectors $\mathbf{a}$ write $z(\mathbf{a})$ for $d(t_\kappa, I_\kappa;\mathbf{a})/\tilde{\mathrm{V}}\mathrm{ar}(t_\kappa;\mathbf{a})^{1/2}$. We show that for sufficiently large κ, if $[\mathbf{A} = \mathbf{a}(a,\kappa)]$ is rejected then $\mathbf{a}(a,\kappa)$ attains the minimum of $f_{a,\kappa}(\cdot) = \tilde{g}_\alpha(t_\kappa, I_\kappa;\cdot)$ over $\{\mathbf{a} : 0 \le \mathbf{a} \le \mathcal{A}_\kappa, \mathbf{a}'\mathbf{1} = a\} =: \Theta$. Since $[\mathbf{A} = \mathbf{a}(a,\kappa)]$ is rejected, this minimum must then be positive, and $[\mathbf{A} = \mathbf{a}]$ is rejected for all $\mathbf{a} \in \Theta$.

From $\mathbf{a}(a,\kappa)$, any $\mathbf{a}^* \in \Theta$ can be reached by a path along line segments of the form $(\mathbf{a}, \mathbf{a} + \delta_s - \delta_t)$, where $\mathbf{a}(a,\kappa) + \delta_s - \delta_t \in \Theta$ — *i.e*, $\delta_s - \delta_t$ points inside of the box $\{\mathbf{a} : 0 \le \mathbf{a} \le \mathcal{A}\}$ from $\mathbf{a}(a,\kappa)$. Also, we may chose the path so that any steps in directions $\delta_s - \delta_t$ such that $n_{\kappa s}/N_{\kappa s} = n_{\kappa t}/N_{\kappa t}$ are taken first. We show that the net change of $f_{a,\kappa}(\cdot)$ along all of these first steps is nonnegative, after which we show that each subsequent step results in an increase in $f_{a,\kappa}(\cdot)$ (at least if κ is sufficiently large).

For any $(s_1, t_1), \ldots, (s_m, t_m)$ s.t. $n_{\kappa s_i}/N_{\kappa s_i} = n_{\kappa t_i}/N_{\kappa t_i}$, each i, one has

$$f_{a,\kappa}(\mathbf{a} + \delta_{s_i} - \delta_{t_i}) - f_{a,\kappa}(\mathbf{a}) = z_{1-\alpha}^2 \left[\mathrm{Var}(t_\kappa; \mathbf{a} + \delta_{s_i} - \delta_{t_i}) - \mathrm{Var}(t_\kappa; \mathbf{a})\right],$$

all $\mathbf{a}$ and i, so that

$$f_{a,\kappa}(\mathbf{a}(a;\kappa) + \sum_i (\delta_{s_i} - \delta_{t_i})) - f_{a,\kappa}(\mathbf{a}(a;\kappa)) = z_{1-\alpha}^2 \left[\mathrm{Var}(t_\kappa; \mathbf{a}(a;\kappa) + \sum_i (\delta_{s_i} - \delta_{t_i})) - \mathrm{Var}(t_\kappa; \mathbf{a}(a;\kappa))\right].$$

But the separability algorithm has been so chosen $\mathbf{a}(a;\kappa)$ as to maximize $\mathrm{Var}(t;\cdot)$ over a set of the form $\Theta \cap \{\mathbf{a} : \forall p, \sum\{a_s : n_{\kappa s}/N_{\kappa s} = p\} = \gamma_p\}$; so this difference must be positive.

Now consider the later steps $\delta_s - \delta_t$ of the path, for which $n_{\kappa s}/N_{\kappa s} \neq n_{\kappa t}/N_{\kappa t}$. By construction of $\mathbf{a}(a, \kappa)$, $\delta_s - \delta_t$ points outside the box from $\mathbf{a}(a, \kappa)$ if $n_{\kappa s}/N_{\kappa s} > n_{\kappa t}/N_{\kappa t}$, so we may assume $n_{\kappa s}/N_{\kappa s} < n_{\kappa t}/N_{\kappa t}$; by hypothesis, this difference is no smaller than $\delta > 0$. Also, coupled with the assumption that $\alpha < \frac{1}{2}$, rejection of $[\mathbf{A} = \mathbf{a}(a, \kappa)]$ entails $d(t_\kappa, I_\kappa; \mathbf{a}(a, \kappa)) > 0$. Since the separable optimizer is so constructed that $d(t_\kappa, I_\kappa; \mathbf{a}(a, \kappa)) = \min_{\mathbf{a}\in\Theta} d(t_\kappa, I_\kappa; \mathbf{a})$, this means $d(t_\kappa, I_\kappa; \cdot)$ is positive throughout Θ. So the sign of $\partial f_{a,\kappa}(\mathbf{a}; \mathbf{v})$ is the same as that of $\partial f_{a,\kappa}(\mathbf{a}; \mathbf{v})/d(t_\kappa, I_\kappa; \mathbf{a})$, for all $\mathbf{a}$ in the convex closure of Θ. We show that for sufficiently large κ, $\partial f_{a,\kappa}(\cdot; \delta_s - \delta_t)/d(t_\kappa, I_\kappa; \cdot)$ is positive on Θ.

Since $\mathbf{a} \mapsto d(t_\kappa, I_\kappa; \mathbf{a})^2$ has a total derivative, Lemma 6.1 entails that $f_{a,\kappa}$ has directional derivatives in all directions, at each a and κ for which $[A = a]$ is compatible. From the lemma and (4), one has

$$\frac{\partial f_{a,\kappa}(\mathbf{a}; \delta_s - \delta_t)}{2d(t_\kappa, I_\kappa; \mathbf{a})} - \left[\frac{n_{\kappa t}}{N_{\kappa t}} - \frac{n_{\kappa s}}{N_{\kappa s}}\right] = -\frac{z_{1-\alpha}^2}{2d(t_\kappa, I_\kappa; \mathbf{a})} \partial\tilde{h}_\kappa(\mathbf{a}; \delta_i - \delta_j) \qquad \text{(A-1)}$$

Since $[\mathbf{A} = \mathbf{a}(a, \kappa)]$ is rejected, $z_{1-\alpha} \leq z(\mathbf{a}(a, \kappa))$ and $z_{1-\alpha}/d(t_\kappa, I_\kappa; \mathbf{a}(a, \kappa)) \leq \mathrm{Var}(t_\kappa, \mathbf{a}(a, \kappa))^{-1/2}$. But $\mathrm{Var}(t_\kappa, \mathbf{a}(a, \kappa)) \geq N_\kappa \min(\psi_l(1 - \psi_l), \psi_u(1 - \psi_u))\tilde{\sigma}^2$, so by assumption on $\tilde{\sigma}$ and ψ_l, ψ_u, the left-hand side of (A-1) with $\mathbf{a} = \mathbf{a}(a, \kappa)$ must be $O(N_\kappa^{-1/2})$, uniformly in a for which $[A = a]$ is compatible. Recall that $n_{\kappa t}/N_{\kappa t} - n_{\kappa s}/N_{\kappa s} \geq \delta > 0$; thus if we choose κ_0 such that for $\kappa > \kappa_0$, the left-hand side of (A-1) is uniformly smaller than δ, then for $\kappa > \kappa_0$ the sign of $\partial f_{a,\kappa}(\mathbf{a}; \delta_s - \delta_t)/d(t_\kappa, I_\kappa; \mathbf{a})$ will be that of $n_{\kappa t}/N_{\kappa t} - n_{\kappa s}/N_{\kappa s}$, or $+1$.

This completes the proof. As the only properties of $\mathrm{Var}(t_\kappa; \mathbf{a})$ it has depended on were the uniform boundedness of its partial derivatives in $\mathbf{a}$ and its increasing as $O(N_\kappa)$ as $\kappa \uparrow \infty$, it applies equally well when $\tilde{\mathrm{V}}\mathrm{ar}(t_\kappa; \mathbf{a})$ is substituted for $\mathrm{Var}(t_\kappa; \mathbf{a})$

throughout, as in § 5.