Joel A. Middleton and Peter M. Aronow*

Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments

DOI 10.1515/spp-2013-0002

Abstract: Many estimators of the average treatment effect, including the difference-in-means, may be biased when clusters of units are allocated to treatment. This bias remains even when the number of units within each cluster grows asymptotically large. In this paper, we propose simple, unbiased, location-invariant, and covariate-adjusted estimators of the average treatment effect in experiments with random allocation of clusters, along with associated variance estimators. We then analyze a cluster-randomized field experiment on voter mobilization in the US, demonstrating that the proposed estimators have precision that is comparable, if not superior, to that of existing, biased estimators of the average treatment effect.

1 Introduction

In recent years, researchers have paid increased attention to the properties of treatment effect estimators for randomized experiments under the design-based model (see, e.g. Freedman 2008a,b). Under the design-based model (Neyman 1923, 1934; Sarndal 1978), potential outcomes are fixed and the only source of stochasticity lies in the random administration of a treatment to a finite population. Importantly, Freedman (2008a) demonstrated that, under a such a model, regression adjustment is generally biased (though consistent) and may reduce efficiency. Researchers have since derived methods that do not suffer from these problems (Lin 2013; Miratrix et al. 2013) and assessed the operating characteristics of common model-based estimators (Humphreys 2009; Samii and Aronow 2012) under the design-based paradigm. However, this

***Corresponding author: Peter M. Aronow,** Department of Political Science, Yale University, New Haven, CT, USA, e-mail: peter.aronow@yale.edu
Joel A. Middleton: Department of Political Science, University of California Berkeley, Berkeley, CA, USA

research has largely focused on experiments wherein treatment is randomized at the unit level.

Although extensively studied under the model-based paradigm (see, e.g. Donner and Klar 2000), comparatively little attention has been devoted to designs with random allocation of clusters under the design-based paradigm. The aforementioned estimators are not directly applicable to cluster-randomized designs. Even seemingly design-based estimators – such as the difference-in-means estimator – may suffer from bias even when all units have an equal probability of treatment assignment. Importantly, Middleton (2008) proves the bias of the difference-in-means estimator (and inconsistency under asymptotic scalings that entail a fixed number of clusters) for randomized experiments with unequal cluster sizes. Similarly, Imai et al. (2009) recognize the bias of the difference-in-means estimator and propose solutions that require altering the design of the experiment. The authors recommend pair matching on observables in order to reduce the amount of bias and variance that may result from a standard analysis of cluster-randomized experiments. The closest analogue to our proposed approach, however, may be found in Hansen and Bowers (2009), which proposes similar – though not necessarily unbiased – design-based estimators for cluster-randomized experiments.[1]

Bias is not the only statistical property that researchers are interested in. In choosing an estimator, researchers often consider efficiency (typically mean square error) to be of paramount importance. However, as we show below, the bias of estimators such as the difference-in-means estimator may not diminish with increasing study size under common designs. Estimators that are asymptotically biased are guaranteed to be relatively inefficient for a sufficiently large sample size, and we provide an empirical example where bias is critical in undermining the relative efficiency of common estimators. In sum, bias cannot always be ignored even when efficiency is a primary concern.

In this paper, we propose a simple and unbiased design-based estimator for the average treatment effect (ATE) for cluster-randomized experiments.[2] Drawing from classical sampling theory, we then propose a natural extension to improve efficiency and confer the property of location invariance: the Des Raj

[1] Hansen and Bowers (2008) also derives design-based balance tests for cluster-randomized experiments.

[2] As in Hansen and Bowers (2008), we consider estimation of the effect of assignment to treatment, which we refer to this simply as the ATE throughout. This quantity is also termed the intention to treat effect. Our approach circumvents the issue of compliance, but our estimators might be divided by suitable compliance rate estimates to estimate average treatment on treated effects, though this may introduce bias from ratio estimation (Hartley and Ross 1954).

(1965) difference estimator, which remains unbiased even in small samples. We also derive two different variance estimators. We then examine a field experiment designed to assess the effect of voter mobilization in a US presidential election, and use randomization inference to assess the bias and precision of a number of estimators under two different null hypotheses. Whereas many common treatment effect estimators, including the difference-in-means, ordinary least squares regression and random effects regression fail to unbiasedly recover the ATE, the proposed estimators are unbiased and are comparable (if not superior) in terms of efficiency.

2 Potential Outcomes

The foundation of our design-based approach is the model of potential outcomes introduced by Neyman (1923) and popularized by Rubin (1974). Define treatment indicator $D_i \in \{0, 1\}$ for units $i \in 1, 2, \ldots, N$ such that $D_i=1$ when unit i receives the treatment and $D_i=0$ otherwise. Assuming that the stable unit treatment value assumption (Rubin 1978, 2005) holds, let Y_{1i} be the potential outcome if unit i is exposed to the treatment, and let Y_{0i} be the potential outcome if unit i is not exposed to the treatment. The observed experimental outcome Y_i may be expressed as a function of the potential outcomes and the assigned treatment: $Y_i=D_iY_{1i}+(1-D_i)Y_{0i}$. The causal effect of the treatment on unit i, τ_i, is defined as the difference between the two potential outcomes for unit i: $\tau_i \equiv Y_{1i}-Y_{0i}$. And, by definition the ATE, denoted Δ, is the average value of τ_i for all units i. Under this model, the only random component of the experiment is the allocation of units to treatment and control groups.

Since $\tau_i \equiv Y_{1i}-Y_{0i}$, the ATE is equivalently

$$\Delta=\frac{\sum_{i=1}^{N}(Y_{1i}-Y_{0i})}{N}=\frac{1}{N}\left[\sum_{i=1}^{N}Y_{1i}-\sum_{i=1}^{N}Y_{0i}\right]=\frac{1}{N}[Y_1^T-Y_0^T],$$

where Y_1^T is the sum of potential outcomes if in the treatment condition and Y_0^T is the sum of potential outcomes if in the control condition. An estimator of Δ can be constructed using estimators of Y_0^T and Y_1^T:

$$\hat{\Delta}=\frac{1}{N}\left[\widehat{Y_1^T}-\widehat{Y_0^T}\right], \tag{1}$$

where $\widehat{Y_1^T}$ is the estimated sum of potential outcomes under treatment and $\widehat{Y_0^T}$ is the estimated sum of potential outcomes under control.

Formally, the bias of an estimator is the difference between the expected value of the estimator (over all randomizations) and the parameter of interest. If the estimators $\widehat{Y_0^T}$ and $\widehat{Y_1^T}$ are unbiased, the corresponding estimator of Δ is also unbiased since

$$\mathrm{E}[\hat{\Delta}]=\frac{1}{N}\left[\mathrm{E}\left[\widehat{Y_1^T}\right]-\mathrm{E}\left[\widehat{Y_0^T}\right]\right]=\frac{1}{N}[Y_1^T-Y_0^T]=\Delta.$$

3 Properties of the Difference-In-Means Estimator

In this section, we examine the properties of the difference-in-means estimator. We start with an examination of the difference-in-means for three reasons. First, the difference-in-means is one of the most commonly used esitmators of ATE in randomized experiments. Second, insights about the difference-in-means will help us understand the bias of other estimators. Third, the derivations will help us to identify conditions under which common estimators are not consistent and, hence, asymptotically inefficient.

Before discussing random allocation of clusters, we begin with a short derivation of the unbiasedness of the difference-in-means estimator under random allocation of individual units.[3] We then articulate the source of the bias for the difference-in-means estimator when applied to a cluster randomized experiment. Finally, we examine the asymptotic properties of the estimator.

3.1 Unbiased Estimation of Treatment Effects Under Random Allocation of Units

Define N and n_t as integers such that $0<n_t<N$. Random allocation of treatment implies that n_t, a fixed number, units are randomly assigned to treatment ($D_i=1$) and the remaining $n_c=N-n_t$ are in control ($D_i=0$). Define I_0 as the set of all i such that $D_i=0$ and I_1 as the set of all i such that $D_i=1$.

To derive an unbiased estimator of the ATE under random allocation, we can first posit estimators of Y_0^T and Y_1^T. Define an estimator of Y_0^T,

$$\widehat{Y_{0,S}^T}=\frac{N}{n_c}\sum_{i\in I_0}Y_{0i}=\frac{N}{n_c}\sum_{i\in I_0}Y_i \tag{2}$$

[3] Throughout, we use the term *random allocation* to refer to the assignment of a fixed number of units (or clusters) to treatment and a fixed number to control, following the terminology of Lachin (1988).

and, similarly, define an estimator of Y_1^T,

$$\widehat{Y_{1,S}^T}=\frac{N}{n_t}\sum_{i\in I_1}Y_{1i}=\frac{N}{n_t}\sum_{i\in I_1}Y_i. \tag{3}$$

It is easy to show that the estimators in equations 2 and 3 are unbiased under the random allocation rule:

$$\mathrm{E}\left[\widehat{Y_{0,S}^T}\right]=\mathrm{E}\left[\frac{N}{n_c}\sum_{i\in I_0}Y_i\right]=N\cdot\overline{Y_0}=Y_0^T, \tag{4}$$

where $\overline{Y_0}$ is the mean value of Y_{0i} over all *i* units (and is not an observable quantity). A proof for the unbiasedness of $\widehat{Y_{1,S}^T}$ directly follows the form of equation 4.

From equation 1, it follows that we may construct an unbiased estimator of Δ:

$$\widehat{\Delta_S}=\frac{1}{N}\left[\widehat{Y_{1,S}^T}-\widehat{Y_{0,S}^T}\right]=\frac{\sum_{i\in I_1}Y_i}{n_t}-\frac{\sum_{i\in I_0}Y_i}{n_c}, \tag{5}$$

where $\sum_{i\in I_1}Y_i/n_t$ is the mean value of Y_i for all units assigned to treatment and $\sum_{i\in I_0}Y_i/n_c$ is the mean value of Y_i for all units assigned to control. $\widehat{\Delta_S}$ is known as the difference-in-means estimator.

3.2 Properties of The Difference-In-Means Estimator Under Random Allocation of Clusters

Under random allocation of clusters, the difference-in-means estimator is no longer generally unbiased, despite all individuals having the same probability of entering into each treatment condition. The unit of randomization is no longer the individual: instead, clusters (or groups of individuals) are assigned to treatment. While random allocation of units may yield more efficient designs in principle, a number of settings may dictate clustered designs in practice. Some examples include when unit randomization is infeasible, when outcome measures are only available at the level of the cluster, or when unit interference (e.g. treatment synergies or spillover effects) is an important aspect of treatment.

In settings where unit randomization is infeasible or undesirable, the researcher rarely has control over the cluster size (e.g. household, village). As a consequence, bias can arise in estimation. We begin this section by deriving the bias associated with the difference-in-means estimator. As our derivation will show, the bias arises whenever outcomes are related to cluster size.

Formally, suppose each cluster $j=1, 2, \ldots, M$ is assigned to either treatment or control. Define m_t and M as (fixed) integers such that $0<m_t<M$. Now m_t clusters are randomly assigned to treatment ($D_j=1$) and the remaining $m_c=M-m_t$ clusters are assigned to control ($D_j=0$). Define J_0 as the set of all j such that $D_j=0$ and J_1 as the set of all j such that $D_j=1$. Let Y_{0ij} be the response of the i^{th} individual in the j^{th} cluster if the cluster is assigned to control and let Y_{1ij} be the response of the i^{th} individual in the j^{th} cluster if the cluster is assigned to treatment. Let n_j be the number of individuals in the j^{th} cluster. Note that all individuals have the same probability m_t/M of entering treatment.

The estimators in equations 2 and 3 can be rewritten as $\widehat{Y^T_{1,S}}=N\sum_{j\in J_1}\sum_{i=1}^{n_j}Y_{ij}/\sum_{j\in J_1}n_j$ and $\widehat{Y^T_{0,S}}=N\sum_{j\in J_0}\sum_{i=1}^{n_j}Y_{ij}/\sum_{j\in J_0}n_j$. The difference-in-means estimator in equation 5 can therefore be rewritten

$$\widehat{\Delta_S}=\frac{1}{N}\left[\widehat{Y^T_{1,S}}-\widehat{Y^T_{0,S}}\right]=\frac{\sum_{j\in J_1}\sum_{i=1}^{n_j}Y_{ij}}{\sum_{j\in J_1}n_j}-\frac{\sum_{j\in J_0}\sum_{i=1}^{n_j}Y_{ij}}{\sum_{j\in J_0}n_j}. \tag{6}$$

The double summations in the numerators make explicit that summation takes place across individuals in different clusters. In the denominators, the summations operate over clusters. While the estimator remains unchanged from equation 5, expressing it this way reveals a fundamental problem with its application.

The trouble with using the estimator in equation 6 is that the quantities $n_t=\sum_{j\in J_1}n_j$ and $n_c=\sum_{j\in J_0}n_j$ are no longer fixed numbers as they were in equation 5, but are now random variables. The total number of individuals in treatment and control now depends on the size of the particular clusters assigned to the experimental groups. To understand why this dependence is problematic, we need only examine equation 4: the term N/n_c may be moved to the outside of the expectation operator because it is a fixed constant. When n_c is a random variable, calculating the expectation is more involved. In general, for a ratio of two random variables u, v, (u/v),

$$\mathrm{E}\left[\frac{u}{v}\right]=\frac{1}{\mathrm{E}[v]}\left[\mathrm{E}[u]-\mathrm{Cov}\left(\frac{u}{v},\, v\right)\right] \tag{7}$$

if $v>0$ (Hartley and Ross 1954). Because the difference-in-means estimator is the difference between two ratios of random variables we can use the result in equation 7 to derive the bias of the difference-in-means estimator in equation 6. Following Middleton (2008),

$$\mathrm{E}[\widehat{\Delta_S}]=\frac{1}{N}[Y_1^T-Y_0^T]-\frac{M}{N}\left[\frac{1}{m_t}\mathrm{Cov}\left(\sum_{j\in J_1}\sum_{i=1}^{n_j}Y_{1ij}/\sum_{j\in J_1}n_j,\ \sum_{j\in J_1}n_j\right)\right.$$
$$\left.-\frac{1}{m_c}\mathrm{Cov}\left(\sum_{j\in J_0}\sum_{i=1}^{n_j}Y_{0ij}/\sum_{j\in J_0}n_j,\ \sum_{j\in J_0}n_j\right)\right].$$

It follows that the bias, $\mathrm{E}[\widehat{\Delta_S}]-\Delta=$

$$-\frac{M}{N}\left[\frac{1}{m_t}\mathrm{Cov}\left(\sum_{j\in J_1}\sum_{i=1}^{n_j}Y_{1ij}/\sum_{j\in J_1}n_j,\ \sum_{j\in J_1}n_j\right)-\frac{1}{m_c}\mathrm{Cov}\left(\sum_{j\in J_0}\sum_{i=1}^{n_j}Y_{0ij}/\sum_{j\in J_0}n_j,\ \sum_{j\in J_0}n_j\right)\right]. \qquad (8)$$

Inspection of this term reveals that, if the size of the cluster is correlated with the potential outcomes in the cluster, the difference-in-means estimator is biased. Moreover, the presence of the terms $1/m_t$ and $1/m_c$ shows that the magnitude (and even the direction) of the bias can depend on the relative number of clusters allocated to treatment and control.

In some special cases, there will be no bias, such as when the cluster size does not vary or when there is no covariance between cluster size and outcomes. Nonetheless, in applied research we might expect cluster size to be related to outcomes. For example, precinct size may be related to the characteristics of the precinct, such as partisan composition and voting rates. In Section 6 we show an example where cluster size is significantly related to treatment effect. Such an association between cluster size and treatment effect has been referred to as nonignorable cluster size (e.g. Hoffman et al. 2001).

3.3 Asymptotic Properties of the Difference-In-Means Estimator With Random Allocation of Clusters

In this section, we demonstrate two important facts about the difference-in-means estimator. First, in a proof adapted from Middleton (2008), we will show that the difference-in-means estimator is consistent as the number of clusters, M, grows. Second, we demonstrate that the difference-in-means estimator is not necessarily consistent as N grows.

Consistency of a statistic under a finite population is defined given a sequence of h finite populations H where $N_h<N_{h+1}$, $n_{th}<n_{th+1}$ and $n_{ch}<n_{ch+1}$ for $h=1, 2, 3, \ldots$. The estimator $\widehat{\Delta_S}$ is said to be a consistent estimator of Δ if $\widehat{\Delta_S}\xrightarrow{p}\Delta$ (converges in probability) as $h\rightarrow\infty$.

To show that the difference-in-means estimator is consistent with large M, we follow Brewer (1979) in assuming that as $h\rightarrow\infty$, the finite population H increases as follows: (1) the original population of M clusters is exactly copied (h–1) times; (2) from each of the h copies, m_t clusters are allocated to treatment (such that $0<m_t<M$) and the remaining $m_c=M-m_t$ are allocated to control; (3) the h subsets are collected in a single population of hM clusters, with hm_t clusters in treatment and $hm_c=hM-hm_t$ in control; and (4) $\widehat{\Delta_S}$ is defined as the difference-in-means estimator as in equation 5, only now summation takes place across all hm_c and hm_t clusters. Figure 1, Panel A illustrates this sort of asymptotic growth.

A less restrictive set of assumptions is possible, but this setup is convenient because H is easy to visualize and moment assumptions are built-in. We express the estimator as, $\widehat{\Delta_S}=\sum_{j\in J_1}\sum_{i=1}^{n_j}Y_{ij}/\sum_{j\in J_1}n_j-\sum_{j\in J_0}\sum_{i=1}^{n_j}Y_{ij}/\sum_{j\in J_0}n_j$, where in this case J_1 is defined as the set of hm_t treatment clusters and J_0 is defined as the set of hm_c control clusters. As $h\rightarrow\infty$, by the weak law of large numbers, $\frac{1}{h}\sum_{j\in J_1}\sum_{i=1}^{n_j}Y_{ij}\xrightarrow{p}Y_1^T\cdot\frac{hm_t}{hM}$, $\frac{1}{h}\sum_{j\in J_0}\sum_{i=1}^{n_j}Y_{ij}\xrightarrow{p}Y_0^T\cdot\frac{hm_c}{hM}$, $\frac{1}{h}\sum_{j\in J_1}n_j\xrightarrow{p}N\cdot\frac{hm_t}{hM}$ and $\frac{1}{h}\sum_{j\in J_0}n_j\xrightarrow{p}N\cdot\frac{hm_c}{hM}$. By Slutsky's theorem,

$$\widehat{\Delta_S}\xrightarrow{p}\frac{Y_1^T\cdot\frac{hm_t}{hM}}{N\cdot\frac{hm_t}{hM}}-\frac{Y_0^T\cdot\frac{hm_c}{hM}}{N\cdot\frac{hm_c}{hM}}=\frac{Y_1^T-Y_0^T}{N}. \tag{9}$$

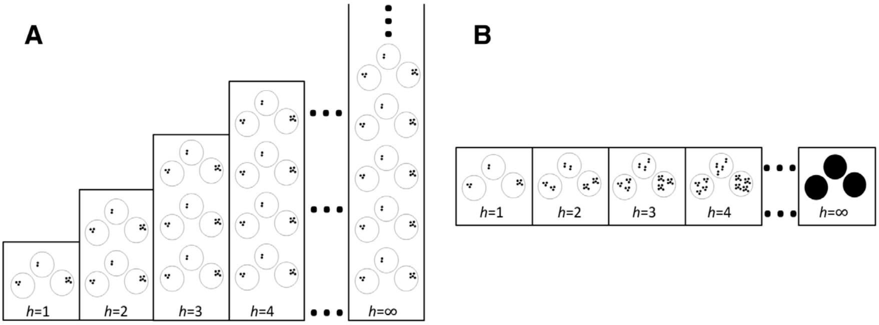


Figure 1: Two versions of Brewer's simple notion of asymptotic growth. The population is simply copied h–1 times. In Panel A, copies of the clusters are made and the number of clusters grows. In Panel B, the number of clusters is fixed and the individuals within are copied. An estimator is consistent under asymptotic growth if it converges to the parameter as $h\rightarrow\infty$.

This proves that the difference-in-means estimator is consistent as the number of clusters grows.

In the case where the size (rather than the number) of the clusters grows as $h\rightarrow\infty$, the finite population H increases as follows: (1) the original population of M units is exactly copied (h–1) times, but this time the h copies of a cluster are considered part of one supercluster; (2) m_t of the clusters are allocated to treatment (such that $0<m_t<M$) and the remaining $m_c=M-m_t$ are allocated to control; and (3) $\widehat{\Delta_S}$ is defined as the difference-in-means estimator as in equation 9, but now the inner summation takes place across all hn_j units in each cluster. Figure 1, Panel B illustrates this sort of asymptotic growth.

To show that the difference-in-means estimator is not necessarily consistent simply with large N, we express the estimator as,

$$\widehat{\Delta_S}=\frac{\sum_{j\in J_1}\sum_{i=1}^{hn_j}Y_{ij}}{\sum_{j\in J_1}hn_j}-\frac{\sum_{j\in J_0}\sum_{i=1}^{hn_j}Y_{ij}}{\sum_{j\in J_0}hn_j}=\frac{\sum_{j\in J_1}\sum_{i=1}^{n_j}Y_{ij}}{\sum_{j\in J_1}n_j}-\frac{\sum_{j\in J_0}\sum_{i=1}^{n_j}Y_{ij}}{\sum_{j\in J_0}n_j}. \tag{10}$$

As $h\rightarrow\infty$, the estimate remains unchanged with large N if the number of clusters is fixed. This proves that the bias articulated in equation 8 is unmitigated for increasingly large clusters.

3.4 Discussion

The results of this section highlight the fact that, for some designs, bias may not be mitigated with increased units. For example, imagine a study of the effect of state-level policy on public opinion. Increasing the number of surveys conducted does nothing to decrease bias in that case since the number of states is fixed.

More troubling, the above results also suggest that the bias of an estimator that averages together a number of biased sub-estimates will not diminish with increasing number of sub-estimates. Consider a block randomized design where clusters (e.g. houses, clinics, precincts) are randomized; if a fixed effects regression is used to "control" for groups, then adding more units by increasing the number of blocks (strata) does not diminish the bias. This is because the fixed effects estimator is simply a weighted average of group-level difference-in-means estimates estimates (cf. Angrist and Pischke 2009, Chapter 5).[4]

[4] However, as the formulas suggest, a way to mitigate such bias would be to block units based on cluster size as suggested by Imai et al. (2009).

4 Unbiased Estimation of Treatment Effects Under Random Allocation of Clusters

By understanding bias as a problem fundamental to ratio estimation, we can circumvent the bias with an alternative design-based estimator. Notationally, it helps to clarify the task if we consider cluster *totals* – i.e. the sum of the responses of the individuals in each cluster. Define $Y_{0j}^T=\sum_{i=1}^{n_j}Y_{0ij}$ as the sum of responses of the individuals in the j^{th} cluster if assigned to control and $Y_{1j}^T=\sum_{i=1}^{n_j}Y_{1ij}$ as the sum of responses of the individuals in the j^{th} cluster if assigned to treatment. For each individual, only one of the two possible responses, Y_{0ij} or Y_{1ij}, may be observed and, since individuals are assigned to treatment conditions in clusters, for any given cluster, only one of the possible totals Y_{0j}^T or Y_{1j}^T, may be observed. The observed cluster total for cluster j, Y_j^T, may be expressed as: $Y_j^T=D_jY_{1j}^T+(1-D_j)Y_{0j}^T$.

Using this new notation, the ATE may be expressed as

$$\Delta=\frac{\sum_{j=1}^{M}\sum_{i=1}^{n_j}(Y_{1ij}-Y_{0ij})}{\sum_{j=1}^{M}\sum_{i=1}^{n_j}1}=\frac{\sum_{j=1}^{M}Y_{1j}^T-\sum_{j=1}^{M}Y_{0j}^T}{\sum_{j=1}^{M}n_j}=\frac{1}{N}[Y_1^T-Y_0^T].$$

We can again construct an unbiased estimator for Δ with unbiased estimators of Y_0^T and Y_1^T. Following the logic of equation 4,

$$\widehat{Y_{0,HT}^T}=\frac{M}{m_c}\sum_{j\in J_0}Y_{0j}^T=\frac{M}{m_c}\sum_{j\in J_0}Y_j^T. \tag{11}$$

One can think of this estimator as estimating the average of the cluster totals (among control clusters) and then multiplying by the number of clusters M to get the estimated total for all units in the study. Likewise,

$$\widehat{Y_{1,HT}^T}=\frac{M}{m_t}\sum_{j\in J_1}Y_{1j}^T=\frac{M}{m_t}\sum_{j\in J_1}Y_j^T. \tag{12}$$

Following the same steps as equation 4, it can be shown that $\widehat{Y_{0,HT}^T}$ and $\widehat{Y_{1,HT}^T}$ are unbiased estimators of Y_0^T and Y_1^T, respectively. The terms M/m_t and M/m_c are fixed; when taking the expectations of equations 11 and 12, they can be moved outside the expectation operator. Note that the random variables at the root of the ratio estimation problem above, n_t and n_c, do not appear in either estimator. From these two unbiased estimators, we may therefore construct an estimator of the ATE:

$$\widehat{\Delta_{HT}} = \frac{1}{N}\left[\widehat{Y^T_{1,HT}} - \widehat{Y^T_{0,HT}}\right] = \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} Y^T_j - \frac{1}{m_c}\sum_{j\in J_0} Y^T_j\right]. \tag{13}$$

We refer to this estimator as the Horvitz-Thompson (HT) estimator because it is a special case of the well-known estimator from sampling theory (Horvitz and Thompson 1952; Chaudhuri and Stenger 2005).

The HT estimator can be criticized on two grounds. First, as Imai et al. (2009) suggest, this estimator is not location invariant. We offer a proof of the non-invariance of the HT estimator in Section 4.1. Second, the HT estimator can be highly imprecise; cluster sums tend to vary a great deal because there are more individuals in some clusters than in others. In large clusters, totals may tend to be large and in small clusters, totals may tend to be smaller. In Section 5.1, we will develop an estimator that addresses both these limitations.

4.1 Non-Invariance of the Horvitz-Thompson estimator

To show that the estimator in equation 13 is not invariant to location shifts, let Y^*_{1ij} be a linear transformation of the treatment outcome for the i^{th} person in the j^{th} cluster such that $Y^*_{1ij} \equiv b_0 + b_1 \cdot Y_{1ij}$ and likewise, the control outcomes, $Y^*_{0ij} \equiv b_0 + b_1 \cdot Y_{0ij}$. Invariance to this transformation would imply that, when analyzing the transformed data, we achieve the relationship between the old estimate and new estimate such that

$$\widehat{\Delta^*_{HT}} = b_1 \cdot \widehat{\Delta_{HT}}, \tag{14}$$

i.e. the ATE estimated from linearly transformed outcomes will be equal to the ATE estimated from non-transformed outcomes multiplied by the scaling factor b_1. In Appendix A, we demonstrate that the HT estimator is not location-invariant because the estimate based on the transformed data will be

$$\widehat{\Delta^*_{HT}} = b_0 \cdot \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} n_j - \frac{1}{m_c}\sum_{j\in J_0} n_j\right] + b_1 \cdot \widehat{\Delta_{HT}}. \tag{15}$$

Unless b_0=0, the term on the left does not generally reduce to zero but instead varies across treatment assignments, so equation 15 is not generally equivalent to equation 14 for a given randomization. Note that, while a multiplicative scale change (e.g. transforming feet to inches) need not be a concern, a linear transformation that includes a location shift (e.g. reversing a binary indicator variable or transforming Fahrenheit to Celsius) will lead to a violation of invariance. For any given randomization, linearly transforming the data such that the intercept changes can yield different estimates.

4.2 Deriving Estimators of the Variance of the Horvitz-Thompson Estimator Under Random Allocation of Clusters

In our derivation of variances, we follow the general formulations of Freedman et al. (1998), which follow from a long tradition dating from Neyman (1923). The variance of the estimator in equation 13 is

$$\mathrm{V}(\hat{\Delta})=\frac{1}{N^2}\left[\mathrm{V}\left(\widehat{Y_0^T}\right)+\mathrm{V}\left(\widehat{Y_1^T}\right)-2\mathrm{Cov}\left(\widehat{Y_0^T},\ \widehat{Y_1^T}\right)\right]. \tag{16}$$

This expression is the true, not estimated, variance. To construct an unbiased estimator of this variance, we must have unbiased estimators of each of the quantities in equation 16. While unbiased estimators may be constructed for $\mathrm{V}\left(\widehat{Y_0^T}\right)$ and $\mathrm{V}\left(\widehat{Y_1^T}\right)$, there does not generally exist an unbiased estimator for $\mathrm{Cov}\left(\widehat{Y_0^T},\ \widehat{Y_1^T}\right)$ because the joint distribution of potential outcomes is not observable.

We may, however, derive a generally conservative estimator of the variance. First, we derive the components of the true variance from equation 16. From the principles of finite population sampling,

$$\mathrm{V}\left(\widehat{Y_{0,HT}^T}\right)=\frac{M^2}{m_c}\left(\frac{M-m_c}{M-1}\right)\sigma^2(Y_{0j}^T),$$

$$\mathrm{V}\left(\widehat{Y_{1,HT}^T}\right)=\frac{M^2}{m_t}\left(\frac{M-m_t}{M-1}\right)\sigma^2(Y_{1j}^T),$$

and

$$\mathrm{Cov}\left(\widehat{Y_{0,HT}^T},\ \widehat{Y_{1,HT}^T}\right)=-\frac{M^2}{M-1}\sigma(Y_{0j}^T,\ Y_{1j}^T),$$

where, given features v_j and w_j for $j\in 1,\ \ldots,\ M$, finite population variance $\sigma^2(v_j)=\frac{1}{M}\sum_{j=1}^{M}\left(v_j-\frac{1}{M}\sum_{j=1}^{M}v_j\right)^2$ and finite population covariance $\sigma(v_j,\ w_j)=\frac{1}{M}\sum_{j=1}^{M}\left(v_j-\frac{1}{M}\sum_{j=1}^{M}v_j\right)\left(w_j-\frac{1}{M}\sum_{j=1}^{M}w_j\right)$.

From equation 16,

$$\mathrm{V}(\widehat{\Delta_{HT}})=\frac{1}{N^2}\left[\frac{M^2}{m_c}\left(\frac{M-m_c}{M-1}\right)\sigma^2(Y_{0j}^T)+\frac{M^2}{m_t}\left(\frac{M-m_t}{M-1}\right)\sigma^2(Y_{1j}^T)+\frac{2M^2}{M-1}\sigma(Y_{0j}^T,\ Y_{1j}^T)\right]$$
$$=\frac{M^2}{N^2}\left[\frac{M}{M-1}\left[\frac{\sigma^2(Y_{0j}^T)}{m_c}+\frac{\sigma^2(Y_{1j}^T)}{m_t}\right]+\frac{1}{M-1}[2\sigma(Y_{0j}^T,\ Y_{1j}^T)-\sigma^2(Y_{0j}^T)-\sigma^2(Y_{1j}^T)]\right]. \quad (17)$$

Since $2\sigma(Y_{0j}^T,\ Y_{1j}^T)-\sigma^2(Y_{0j}^T)-\sigma^2(Y_{1j}^T)\le 0$, it follows that

$$\mathrm{V}(\widehat{\Delta_{HT}})\le \mathrm{V}_{\mathrm{apx}}\left(\widehat{\Delta_{HT}}\right)=\frac{M^3}{N^2(M-1)}\left[\frac{\sigma^2(Y_{0j}^T)}{m_c}+\frac{\sigma^2(Y_{1j}^T)}{m_t}\right].$$

Substituting unbiased estimators of $\sigma^2(Y_{0j}^T)$ and $\sigma^2(Y_{1j}^T)$ (Cochran 1977, theorem 2.4), we may derive an unbiased estimator of the quantity $\mathrm{V}_{\mathrm{apx}}(\widehat{\Delta_{HT}})$,

$$\hat{\mathrm{V}}(\widehat{\Delta_{HT}})=\frac{M^2}{N^2}\left[\frac{\sum_{j\in J_0}\left(Y_j^T-\overline{Y_{cj}^T}\right)^2}{m_c(m_c-1)}+\frac{\sum_{j\in J_1}\left(Y_j^T-\overline{Y_{tj}^T}\right)^2}{m_t(m_t-1)}\right],$$

where $\overline{Y_{mj}^T}=\sum_{j\in J_0}Y_j^T/m_c$, the mean value of Y_j^T over all $j\in J_0$ and $\overline{Y_{tj}^T}=\sum_{j\in J_1}Y_j^T/m_t$, the mean value of Y_j^T over all $j\in J_1$.[5]

The bias of the variance estimator is always nonnegative, thus ensuring the variance estimator is conservative. However, while $\hat{\mathrm{V}}(\widehat{\Delta_{HT}})$ is conservative, it may also be imprecise. In addition, when m_c or m_t is 1 the estimate is undefined. In general, it is impossible to consistently estimate $\mathrm{V}(\widehat{\Delta_{HT}})$ in experiments performed on finite populations (Aronow et al. 2014) and thus it may be the case that no single variance estimator is generally adequate. This issue is compounded when N is small and asymptotic approximations may be poor.

We propose an alternative estimator of the variance by assuming sharp null hypothesis and either analytically or computationally calculating the variance of the estimator. One common sharp null hypothesis is that of the sharp null hypothesis of no treatment effect: $H_0:\tau_i=0,\ \forall i$. H_0 implies that the treatment has no effect whatsoever on the outcome, i.e. that both potential outcomes are

[5] When M is large, researchers may encounter numerical problems computing M^2 and, later, M^4. This problem may be obviated by replacing M^2/N^2 with $\left(\frac{1}{M}\sum_{j=1}^{M}n_j\right)^{-2}$, the reciprocal of the square of the average number of units per cluster.

identical: $Y_0i=Y_{1i}=Y_i$. When the sharp null hypothesis of no effect holds, we know two important facts: $\sigma^2(Y_{0j})=\sigma^2(Y_{1j})=\sigma^2(Y_j)$ and $\sigma(Y_{0j}, Y_{1j})=\sigma^2(Y_j)$. By substituting σ^2 into the last line of equation 17, we may calculate the true variance under this null hypothesis,

$$\begin{aligned} \mathrm{V}^N(\widehat{\Delta_{HT}}) &= \frac{M^2}{N^2}\left[\frac{M}{M-1}\left[\frac{\sigma^2(Y_j^T)}{m_c}+\frac{\sigma^2(Y_j^T)}{m_t}\right]+\frac{1}{M-1}[2\sigma^2(Y_j^T)-\sigma^2(Y_j^T)-\sigma^2(Y_j^T)]\right] \\ &= \frac{M^4\sigma^2(Y_j^T)}{N^2(M-1)m_c m_t}. \end{aligned}$$

Note that if the sharp null hypothesis of no effect holds, $\mathrm{V}^N(\widehat{\Delta_{HT}})$ is the *true* variance, which can be calculated from the data exactly or by way of resampling. When the sharp null hypothesis of no effect does not necessarily hold, $\mathrm{V}^N(\widehat{\Delta_{HT}})$ may be construed as an estimator of $\mathrm{V}(\widehat{\Delta_{HT}})$. We therefore refer to a variance estimator constructed by assuming the sharp null hypothesis of no effect as $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$.

The primary benefit of using $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$ is that it tends to be more stable than $\hat{\mathrm{V}}(\widehat{\Delta_{HT}})$, particularly when either n_c or n_t is small, because it combines the variance of the treatment and control groups. In cases where $\hat{\mathrm{V}}(\widehat{\Delta_{HT}})$ is imprecise, $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$ may be preferable. Highly imprecise standard errors may be downwardly biased even when the associated variance estimator is conservative. The square root is a concave function so, by Jensen's inequality, $\mathrm{E}[\hat{\mathrm{V}}(\widehat{\Delta_{HT}})^{0.5}]\leq(\mathrm{E}[\hat{\mathrm{V}}(\widehat{\Delta_{HT}})])^{0.5}$. Since the estimates from $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$ will tend to remain stable across randomizations, its use may therefore avoid the bias resulting from Jensen's inequality. However, when effect sizes are large, $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$ will tend to overestimate the true sampling variability.

Recent theoretical results suggest that $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$ may be adequate as a conservative approximation. In general, $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$ will be conservative relative to the true variance if effects are constant (at the cluster scale) or if the number of clusters is balanced, in a result that directly follows from theorem 3 of Ding (2014) and Samii and Aronow (2012) (by way of the relationship between pooled and combined variance). These results indicate $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$ will have a higher value than that of the true variance if treatment effects are in fact constant at the cluster scale. For these reasons, choosing the sharp null of no effect as an approximation will generally be conservative among the class of hypotheses such that effects are constant at the cluster scale.[6]

[6] Researchers may seek to calculate separate variance estimators for each of a grid of hypothesized, constant treatment effects, and use these to form a confidence interval by way of inverting hypothesis tests. We thank an anonymous reviewer for this suggestion.

Computational approximations of exact bias and variance terms may be computed for any estimator under any given sharp null hypothesis. Another noteworthy benefit of using $\hat{\mathrm{V}}^N(\widehat{\Delta_{HT}})$ is that it can be computed under designs where $\hat{\mathrm{V}}(\widehat{\Delta_{HT}})$ cannot be computed, such as pair randomized designs.

4.3 Block Randomized Designs

In this section we consider how to generalize the HT estimator to block randomized designs. In a block randomized design, clusters are first classified in to one of B blocks, often on the basis of homogeneity of the clusters. In the b^{th} block, a fixed number of clusters, m_{tb}, are assigned to treatment and the rest, m_{cb}, to control.

As each block represents an independent randomized experiment, the HT estimator and variance estimators may be applied to each block separately. For the b^{th} block the estimator of the ATE can be written $\widehat{\Delta^b_{HT}}$. An unbiased estimate of the ATE for all the units in the study can be written as a weighted average of the block-level estimates,

$$\widehat{\Delta^B_{HT}}=\sum_{b=1}^{B}\frac{N_b}{N}\widehat{\Delta^b_{HT}}, \tag{18}$$

where N_b is the number of units in the b^{th} block. From first principles, the variance of the estimator is

$$\mathrm{V}\left(\widehat{\Delta^B_{HT}}\right)=\sum_{b=1}^{B}\frac{N_b^2}{N^2}\mathrm{V}\left(\widehat{\Delta^b_{HT}}\right), \tag{19}$$

and conservative variance estimation can be achieved by "plugging in" conservative estimators of the variance for each of the $\mathrm{V}\left(\widehat{\Delta^b_{HT}}\right)$. Alternatively, Monte Carlo simulations may be used as an approximation.

5 Difference Estimators

In this section, we propose a simple extension of the HT estimator to improve the efficiency of the estimator as well as confer the important property of location invariance.

5.1 Des Raj Difference Estimator for Cluster Size

A major source of variability with the HT estimator is the variation in the number of individuals in each cluster. Clusters with large n_j will tend to have larger values of Y_j^T – that is, in many applications, as clusters get larger, the sum of the outcomes for that cluster will also tend to get larger. We use the Des Raj (1965) difference estimator to reduce this variability. To derive the Des Raj difference estimator in this context, we first derive our estimates of the study population totals, Y_{0j}^T and Y_{0j}^T by "differencing" off some of the variability:

$$\widehat{Y_{0,R1}^T} = \frac{M}{m_c}\sum_{j\in J_0}(Y_j^T - k(n_j - N/M)), \tag{20}$$

where constant k is a *prior* estimate of the regression coefficient from a regression of Y_j^T on n_j, and $(n_j - N/M)$ is the difference between the size of cluster j and the average cluster size.[7] k is also roughly equivalent to an estimate of the average value of Y_{ij} for all units and does not have a causal interpretation. In Section 5.3, we derive an exact expression for the optimal value of k, which depends on both potential outcomes and the specifics of the experimental design. Similarly,

$$\widehat{Y_{1,R1}^T} = \frac{M}{m_t}\sum_{j\in J_1}(Y_j^T - k(n_j - N/M)). \tag{21}$$

To develop an intuition about this method, note that it is equivalent to defining a new "differenced" variable U_j^T, where $U_j^T = Y_j^T - k(n_j - N/M)$ and conducting the analysis based on U_j^T instead of Y_j^T. So long as k is fixed before analysis, this strategy does not lead to bias because

$$\mathrm{E}k[n_j - N/M] = k\mathrm{E}[n_j - N/M] = k\cdot 0 = 0. \tag{22}$$

It follows that the HT and Des Raj estimators have the same expected value. Since $\widehat{Y_{0,R1}^T}$ and $\widehat{Y_{1,R1}^T}$ are unbiased, it follows that the Des Raj estimator,

$$\widehat{\Delta_{R1}} = \frac{1}{N}\left[\widehat{Y_{1,R1}^T} - \widehat{Y_{0,R1}^T}\right],$$

is also unbiased.[8]

[7] A similar estimator is proposed by Hansen and Bowers (2009), differing primarily in that it contains a random denominator.

[8] Note that estimating k from the same data set can lead to bias, as we demonstrate in Appendix B, raising the question of where to obtain a suitable value. In Section 6, we suggest using data from other blocks in experiments with blocking. Another option would be to find an auxilliary data source from which a trustworthy value of k can be estimated. In survey

Deriving a conservative estimator of the variance of the Des Raj estimator follows directly from Section 4.2:

$$\hat{\mathrm{V}}(\widehat{\Delta_{R1}})=\frac{M^2}{N^2}\left[\frac{\sum_{j\in J_0}\left(U_j^T-\overline{U_{cj}^T}\right)^2}{m_c(m_c-1)}+\frac{\sum_{j\in J_1}\left(U_j^T-\overline{U_{tj}^T}\right)^2}{m_t(m_t-1)}\right], \tag{23}$$

where $\overline{U_{cj}^T}=\sum_{j\in J_0}U_j^T/m_c$, the mean value of U_j^T in the control condition and $\overline{U_{tj}^T}=\sum_{j\in J_1}U_j^T/m_t$, the mean value of U_j^T in the treatment condition. Similarly, from Section 4.2, we may easily construct a variance estimator for $\widehat{\Delta_{R1}}$ by assuming the sharp null hypothesis of no treatment effect:

$$\hat{\mathrm{V}}^N(\widehat{\Delta_{R1}})=\frac{M^4\sigma^2(U_j^T)}{N^2(M-1)m_c m_t}.$$

As an alternative, Monte Carlo simulations can be used to compute this quantity.

5.2 Invariance of the Des Raj Difference Estimator

One benefit of the Des Raj estimator is that it has invariance to location transformation, regardless of the accuracy of the researcher's choice of *k*. In this section, we prove the invariance of the Des Raj estimator. When Y_{0ij} and Y_{1ij} are linearly transformed, *k* will also change: the same transformation must be applied to *k* as to Y_{0ij}^T and Y_{1ij}^T. Since *k* is on the same scale as the outcome variable, when the outcome variable is transformed, *k* will also be transformed:

$$k^*=(b_0+b_1\cdot k). \tag{24}$$

Using this new k^*, we may again define new differenced treatment outcomes,

$$\begin{aligned}U_{1j}^{T*}&=Y_{1j}^{T*}-k^*\cdot(n_j-N/M)\\&=\sum_{i=1}^{n_j}(b_0+b_1\cdot Y_{1ij})-(b_0+b_1\cdot k)\cdot(n_j-N/M)\\&=n_j\cdot b_0+b_1\cdot Y_{1j}^T-(b_0+b_1\cdot k)\cdot(n_j-N/M)\\&=b_0\cdot N/M+b_1\cdot U_{1j}^T.\end{aligned}$$

sampling, researchers sometimes accept the bias of estimating *k* with regression (Sarndal 1978), but the focus of the current paper is on unbiased estimation so regression estimation is outside our scope. We recommend that either the value of or procedure for choosing *k* be specified in a preanalysis planning document, so as to reduce the uncertainty associated with researcher discretion.

And, likewise, we may define new differenced control outcomes, $U^{T*}_{0j} = b_0 \cdot N/M + b_1 \cdot U^T_{0j}$. The estimate based on these transformed variables will be

$$\begin{aligned}
\widehat{\Delta^*_{R1}} &= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} U^{T*}_j - \frac{1}{m_c}\sum_{j\in J_0} U^{T*}_j\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}(b_0 \cdot N/M + b_1 \cdot U^T_{1j}) - \frac{1}{m_c}\sum_{j\in J_0}(b_0 \cdot N/M + b_1 \cdot U^T_{1j})\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t} b_1 \sum_{j\in J_1} U^T_{1j} - \frac{1}{m_c} b_1 \sum_{j\in J_0} U^T_{0j}\right] \\
&= b_1 \cdot \widehat{\Delta_{R1}}.
\end{aligned} \tag{25}$$

The Des Raj estimator is therefore invariant to linear transformation because any linear transformation to the outcome will necessarily be reflected in k.

Note that the HT estimator may be considered a special case of the Des Raj estimator when $k=0$. However, unlike the HT estimator, the explicit assumption that $k=0$ ensures that when the scale of the outcome changes, the scale of k also changes. The non-invariance of the HT estimator may therefore be thought of as a failure to recognize the implicit assumption that $k=0$ and to transform to k^* when the scale of the outcome changes.

5.3 Optimal Selection of *k*

To derive the optimal value of k, we begin by noting that the variance of U^T_{0j} is

$$\begin{aligned}
\sigma^2(U^T_{0j}) &= \frac{\sum_j \left(U^T_{0j} - \overline{U^T_{0j}}\right)^2}{M} \\
&= \frac{\sum_j \left(Y^T_{0j} - k(n_j - N/M) - \overline{Y^T_{0j}}\right)^2}{M} \\
&= \sigma^2(Y^T_{0j}) + k^2\sigma^2(n_j) - 2k\sigma(n_j,\, Y^T_{0j}),
\end{aligned}$$

where $\overline{U^T_{0j}}$ is the mean value of U^T_{0j} over all j clusters. k_{optim_c}, the value of k that minimizes $\sigma^2(U^T_{0j})$, can be found using simple optimization. Since the second derivative with respect to k, $2\sigma^2(n_j)$, must be positive, we may set the first derivative equal to zero and solve for k, so that

$$k_{optim_c} = \frac{\sigma(n_j, Y_{0j}^T)}{\sigma^2(n_j)}. \tag{26}$$

Equation 26 should look familiar to the reader: the best fitting *k* is the ordinary least squares coefficient.

Likewise, the optimal value of *k* for the potential outcomes under treatment is $k_{optim_t} = \sigma(n_j, Y_{1j}^T) / \sigma^2(n_j)$. Given that k_{optim_t} does not generally equal k_{optim_c}, a researcher could justifiably identify different values of *k* for treatment and control groups. In practice, however, this would require a good deal of prior knowledge (including knowledge about treatment effects); for this reason, a single value of *k* will typically be preferable. In Appendix C, we derive a single optimal value of k, $k_{optim*} = m_t k_{optim_c} / M + m_c k_{optim_t} / M$.

Unlike a structural parameter, the value of k_{optim*} will depend on the number of clusters assigned to treatment and to control. Perhaps counterintuitively, when there are fewer clusters in the control condition, k_{optim*} is more heavily weighted toward k_{optim_c}, the value of *k* that minimizes $\sigma^2(U_{0j}^T)$ (and vice versa). A simple intuition for this weighting is that the condition with fewer clusters will contribute more to the overall variance of the estimator. Consequently, the greatest increase in precision comes from adjustments made to units in that condition.

The chosen value of *k* will reduce the variability of the Des Raj estimator, $\widehat{\Delta_{R1}}$, relative to the HT estimator when, for $k_{optim*} > 0$, $0 < k < 2k_{optim*}$ and, for $k_{optim*} < 0$, $0 > k > 2k_{optim*}$. In other words, the Des Raj estimator will have better precision than the HT estimator unless the researcher picks a *k* with the wrong sign or chooses a *k* that is more than twice the magnitude of k_{optim*}. Because k_{optim*} will tend to be close to the average outcome for all individuals, the researcher will usually have prior knowledge about the mean individual-level outcome.[9]

Under the sharp null hypothesis of no treatment effect, $k_{optim*} = k_{optim_c} = k_{optim_t} = \sigma(n_j, Y_j^T) / \sigma^2(n_j)$, and thus the optimal *k* would be the ordinary least squares coefficient from regressing Y_j^T on n_j. *Prima facie*, the intuitive next step would be to try to estimate *k* from the data, utilizing ordinary least squares on the observed data (perhaps controlling for D_j). However, regression estimates of *k* can lead to bias in the estimation of treatment effects. In Appendix B, we demonstrate that the bias from estimating *k* from within-sample data is

$$\mathrm{E}\left[\frac{\widehat{Y_{1,R1}^T} - \widehat{Y_{0,R1}^T}}{N}\right] - \Delta = \frac{M}{N}(\mathrm{Cov}(\hat{k}, \overline{n_{cj}}) - \mathrm{Cov}(\hat{k}, \overline{n_{tj}})),$$

[9] Note that our fundamental uncertainty about the optimal value of *k* does not itself contribute to the uncertainty of our estimate since *k* is treated as a fixed constant, e.g. in equation 23.

where $\hat{k}$ is an estimator of k, $\overline{n_{tj}}$ is the mean value of n_j for clusters in the treatment condition in a given randomization and $\overline{n_{cj}}$ is the mean value of n_j for units in the control condition in a given randomization.

Knowing the optimal value of k under the sharp null hypothesis of no treatment effect is nevertheless informative as we seek to construct principled prior estimates for k. By using the ordinary least squares estimator on auxiliary data with similar potential outcomes, we can approximate k_{optim*} with out-of-sample data.

As we will demonstrate in our empirical example, such auxiliary data can come from the *other* blocks in an block randomized experiment. If one was concerned that estimating the values of k from other blocks of an experiment would lead to additional stochasticity in the values of U_{0j}^T and U_{1j}^T, Monte Carlo simulations (whereby the values of k are recomputed for each simulation) may be used to compute the sharp null variance estimate.

5.4 Des Raj Difference Estimator for Cluster Size and Covariates

The Des Raj estimator may also be extended to include other covariates which may further reduce the sampling variability of the estimator. Assume the researcher has access to A covariates for each individual i in cluster j, denoted by X_{aij}^T, $a\in 1, 2, \ldots, A$. Define the cluster total of the covariate, $X_{aj}^T=\sum_{i=1}^{n_j} X_{aij}$, and define the sum of the X_{aij} across all individuals in all clusters, $X_a^T=\sum_{j=1}^{M}\sum_{i=1}^{n_j} X_{aij}$. It is simple to adapt the Des Raj estimator to incorporate these additional covariates. Define constants k' and k_a $(\forall a)$ as prior estimates of the coefficients associated with a regression of Y_j on cluster size and cluster-level covariates, respectively. Again, k' and k_a do not have causal interpretations. It follows that we may define

$$\widehat{Y_{0,R2}^T}=\frac{M}{m_c}\sum_{j\in J_0}\left(Y_j^T-\underbrace{k'(n_j-N/M)}_{\text{adjusting for size}}-\underbrace{\sum_{a=1}^{A}k_a(X_{aj}^T-X_a^T/M)}_{\text{adjusting for other covariates}}\right)$$

and

$$\widehat{Y_{1,R2}^T}=\frac{M}{m_t}\sum_{j\in J_1}\left(Y_j^T-k'(n_j-N/M)-\sum_{a=1}^{A}k_a(X_{aj}^T-X_a^T/M)\right).$$

By the logic of equation 22, $\widehat{Y_{0,R2}^T}$ and $\widehat{Y_{1,R2}^T}$ are unbiased estimators of Y_0^T and Y_1^T, respectively. It follows that we may again construct an unbiased estimator of Δ,

$$\widehat{\Delta_{R2}} = \frac{1}{N}\left[\widehat{Y_{1,R2}^{T}} - \widehat{Y_{0,R2}^{T}}\right].$$

Following the same steps as in equation 25, it can be shown that as long as k' undergoes the same linear transformation as the original data and k_a $(\forall a \in A)$ undergoes the same multiplicative scale shift, the Des Raj estimator with covariates will also be invariant. It will also be more efficient than the preceding estimators if the researcher's estimates for k' and k_a are reasonable; constructing variance estimators for $\widehat{\Delta_{R2}}$ is simple and follows directly from Section 5.1.

Note that the efficiency characteristics of this Des Raj estimator may be derived as in Section 5.3, where the same intuitions about efficiency hold. In principle, a researcher should choose covariates that together do the best job of predicting values of the potential outcomes to achieve the values of $Y_{0,R2}^{T}$ and $Y_{1,R2}^{T}$ with the lowest variability across randomizations. In practice, a researcher might apply a variable selection method such as penalized regression techniques using an auxiliary data set to identify suitable covariates and values of k.

6 Application

In this application we reanalyze the data from Green and Vavreck (2008) who used a cluster randomized design to examine the effectiveness of television ads on voter turnout among 18- and 19-year-old voters in the 2004 presidential election. The study randomized television cable districts to either a treatment group, in which advertisements encouraging young people to vote were shown, or to the control group. The original experiment included a total of 23,869 voters in 85 television cable districts in blocks (strata) of size 2 or 3. Because we wanted to use prior turnout in the cable district as a covariate in our analysis, we limited the analysis to the 80 cable districts for which this information was available from the authors. This yielded 40 blocks of two cable districts each (one in treatment, the other in control) and a total of 22,733 individual voters.

The outcome measure of interest, Y_{ij}, is whether or not the individual i in cluster (cable district) j voted in the 2004 American presidential election (coded 1 if the individual voted, 0 if the individual did not vote). Because 18 and 19 year olds are new registrants they have no prior voter history, so individual voter history could not be used for covariates. However, we use turnout rate in the cable district in the 2000 election as a covariate as well as age. While the covariates are somewhat less than ideal because they are unlikely to be particularly predictive, they provide us with an opportunity to examine how the Raj difference estimator performs when covariates are not particularly informative. In such a situation we might expect k_{optim} to be near zero and values of k chosen may actually reduce the efficiency of the Raj difference estimator since it is less likely to be the case that $2k_{optim}>k>0$ when $k_{optim}=0$.

6.1 Randomization inference

Randomization inference (RI) will allow us to assess the bias and variance of any given estimator. In addition, RI allows the researcher to perform completely nonparametric significance testing (see, e.g. Rosenbaum 2002). We refer to the estimate produced by a given estimator as the test statistic. RI assumes that a given sharp null hypothesis holds and evaluates the test statistic for every possible random assignment of units to treatment and control. By recalculating the test statistic for each possible treatment assignment, the reference distribution of the test statistic is constructed. Fisher's exact test is a well-known form of RI for significance testing, but the method is much more general.

Because the total possible permutations increase rapidly with population size, RI may be computationally infeasible. We may use Monte Carlo simulations to approximate RI by repeatedly assigning units to treatment and control groups randomly and estimating the test statistic that would be observed for each repetition. The distribution of the test statistic across randomizations forms the reference distribution of the statistic. As the number of repetitions gets large, the distribution of the test statistic based on repeated randomizations converges to that of full RI. This method can achieve results arbitrarily close to RI by increasing the number of repetitions.

We use randomization inference to examine the behavior of our estimators and compare them with the behavior of three commonly used estimators. We conduct randomization inference for two scenarios (5000 iterations). The first scenario examines the behavior of the estimators under the sharp null hypothesis of absolutely no treatment effect. The second scenario examines the behavior of estimators under heterogeneous treatment effects.

6.2 Imputing Missing Potential Outcomes

Computing the test statistics under repeated randomizations requires that we can observe both potential outcomes for each unit. Since in reality we only observe the response of unit i under one of the treatments, we must impute the value of the missing potential outcome before conducting RI. We conduct RI using two different methods of imputation.

The first method assumes the sharp null hypothesis of no treatment effect. This effectively imputes the missing potential outcome with the observed potential outcome.

In the second method we simulate heterogeneous treatment effects, first modeling the data using logistic regression in order to impute missing potential outcomes. This method looks to the data as a guide to creating realistic potential outcomes that have a similar structure to the original data. We used the logistic regression model,

$$P(Y_{ij}=1)=1-\left(1+\exp\left(\alpha+\tau t_j+\beta n_j+\phi n_j t_j+\sum_{f=1}^{F-1}\gamma_f\Gamma_f\right)\right)^{-1}$$

where t_j is a treatment indicator for cluster j, n_j is the cluster size, F is the number of blocks (in this case, 40), Γ_f is an indicator variable indicating whether cluster j is in block f. The terms α, τ, β, ϕ and γ_f are coefficients estimated from the data using maximum likelihood methods. Note the coefficient ϕ is responsible for the heterogeneous treatment effects. We estimate τ as 0.4, β as 1.4 and ϕ as −0.9.

We used this model to impute missing potential outcomes for each individual. To do so, the latent probability of response (voting) was first computed for each unit when treated, p_{ti}, and when not treated, p_{ci}, using the estimated model. Each missing Y_{ci} and Y_{ti} was imputed using a random draw from a Bernoulli random variable with probability estimated from the logistic regression model. The imputation process was conducted for each iteration of the RI. Marginalizing over the imputation process, the ATE is in expectation $\frac{\sum_i(p_{ti}-p_{ci})}{N}=0.007$, or 0.7 percentage points.

6.3 Treatment effect estimates

In this section, we define the estimators that will be compared. We will consider four regression-based estimators as well as the three design-based estimators proposed in this paper. We begin by detailing each of these estimators.

The first estimator is the regression without covariates, also known as the difference-in-means. The model can be written:

$$Y_{ij}=\beta_0+\beta_1 D_j+e_{ij}, \tag{27}$$

where β_0 is a constant, D_j is an indicator for treatment, β_1 is an estimate of Δ, and e_{ij} is an error term. Our estimate of β_1 follows from fitting the model with ordinary least squares. The next estimator under consideration is the fixed-effects regression estimator, $\widehat{\Delta_{FE}}$, which includes fixed effects for each of the blocks (strata):

$$Y_{ij} = \beta_0 + \beta_1 D_j + \sum_{f=1}^{F-1} \gamma_f \Gamma_f + e_{ij}, \tag{28}$$

where β_0, β_1, D_j and e_{ij} are as above and Γ_f represents the dummy variable for the f^{th} block, and the model is again fitted with ordinary least squares. We then consider the fixed-effects regression estimator that also adjusts for the covariates: average turnout in 2000 and age. The model can be written:

$$Y_{ij} = \beta_0 + \beta_1 D_j + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \sum_{f=1}^{F-1} \gamma_f \Gamma_f + e_{ij},$$

where β_2 is the coefficient on precinct-level voter turnout in 2000, β_3 is the coefficient on age, and the model is fitted with ordinary least squares.

As Freedman (2008a) notes, even without clustering, the models above that include covariates may be biased due to regression adjustment if treatment assignment is imbalanced (i.e. $n_t \neq n_c$) or there exists treatment effect heterogeneity (i.e. $\exists i, j$ s.t. $\tau_i \neq \tau_j$). For both fixed effects models, Huber-White "robust" cluster standard errors are estimated. While often sufficient for inference, these standard errors may be unreliable in finite samples (Freedman 2006; Angrist and Pischke 2009) and they may also fail to address larger issues of model misspecification (King and Roberts 2014).

Our next estimator adds a random effect for cluster to the above specification. Random effects estimation was the recommended analytical technique in Green and Vavreck (2008). However, as we will show, this estimator is not guaranteed to be unbiased. We use the following specification:

$$Y_{ij} = \beta_0 + \beta_1 D_j + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \sum_{f=1}^{F-1} \gamma_f \Gamma_f + e_j + e_{ij},$$

where e_j is a normally distributed cluster-level disturbance (and e_{ij} is also distributed normally). This model is estimated using the `lmer()` function in the lme4 (Bates and Maechler 2010) package in R (R Development Core Team 2010) using the default settings. Standard errors are empirical Bayes estimates also produced by the `lmer()` function.

And, finally, we present treatment effect estimates for the HT estimator, the Des Raj difference estimator (with n_j) and the Des Raj difference estimator (with n_j and covariates). For all three the standard error estimates are the square root of our estimated "sharp null" variances $\hat{\mathrm{V}}^N(\hat{\Delta})$ are used as opposed to $\hat{\mathrm{V}}(\hat{\Delta})$ as the latter is not identified in the pair-randomized design.

In this application we use the alternate blocks of the experiment to derive the values of k, k' and k_a from the data. For a given block, the values are estimated by dropping that block from the data and regressing the outcome on the covariates using data from the remaining 39 blocks.

To estimate k for the Des Raj estimator with only n_j, we use the following model:

$$Y_j^T = \alpha + kn_j + e_j,$$

where α is a constant, n_j is cluster size, and e_j is a random disturbance. This estimation procedure yields a principled estimate for k. To estimate k' and k_a for the Des Raj estimator with both n_j and covariates, we use the following model:

$$Y_j^T = \alpha' + k'n_j + k_1 X_{1j}^T + k_2 X_{2j}^T + e_j,$$

where α' is a constant, X_{1j}^T is the total turnout in cluster j in the 2000 election, X_{2j}^T is the sum of ages in cluster j.

Note that in the sharp null scenario the estimated values of k, k', k_1 and k_2 are the same for all randomizations for a given block. For the heterogeneous treatment effect scenario, however, these values can vary across randomizations as the observed values of Y_j^T change depending on whether cluster j is in treatment or not. As mentioned above, this sort of variability in these values contributes to the variability of the Raj difference estimator. In our application, the variance estimators remain conservative nonetheless. In practice, if the contribution of k to the uncertainty is a concern, Monte Carlo simulations could be used to estimate the variance.

6.4 Randomization Inference With the Sharp Null Hypothesis of No Treatment Effect

Figure 2 displays the results for the point estimators assuming the sharp null hypothesis of no treatment effect. Solid vertical lines indicate the mean of the sampling distributions.

Results show that all estimators are unbiased under the sharp null. The HT estimator is the least precise estimator by far. The rest perform very competitively with the random effects regression and Raj's difference estimator being the most precise.

Figure 3 displays the results for the standard error estimators under the sharp null hypothesis. In the case of the regression with no covariates (difference-in-means) the standard errors are biased upwards due to the failure

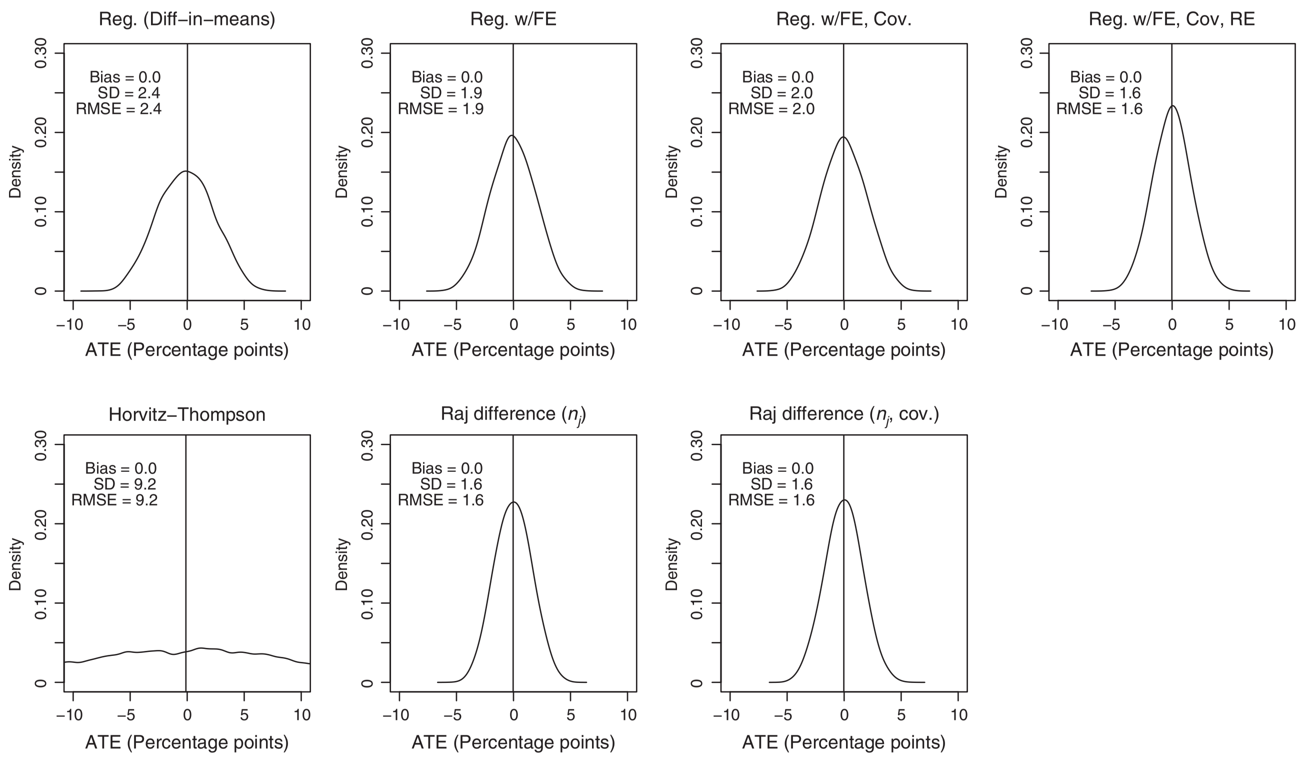


Figure 2: Sampling distributions associated with the ATE estimators under the sharp null hypothesis of no treatment effect detailed in Section 6. Five thousand randomizations were used to estimate the desity of the sampling distributions. Density plots were generated using the `density()` function in R (R Development Core Team 2010), with the default settings and a bandwidth of 0.5 percentage points. Each estimator is detailed in Section 6.3. The vertical line indicates the expected value, and therefore bias, of the estimator. Bias and SE estimates in the upper-right of each plot are computed from each empirical distribution.

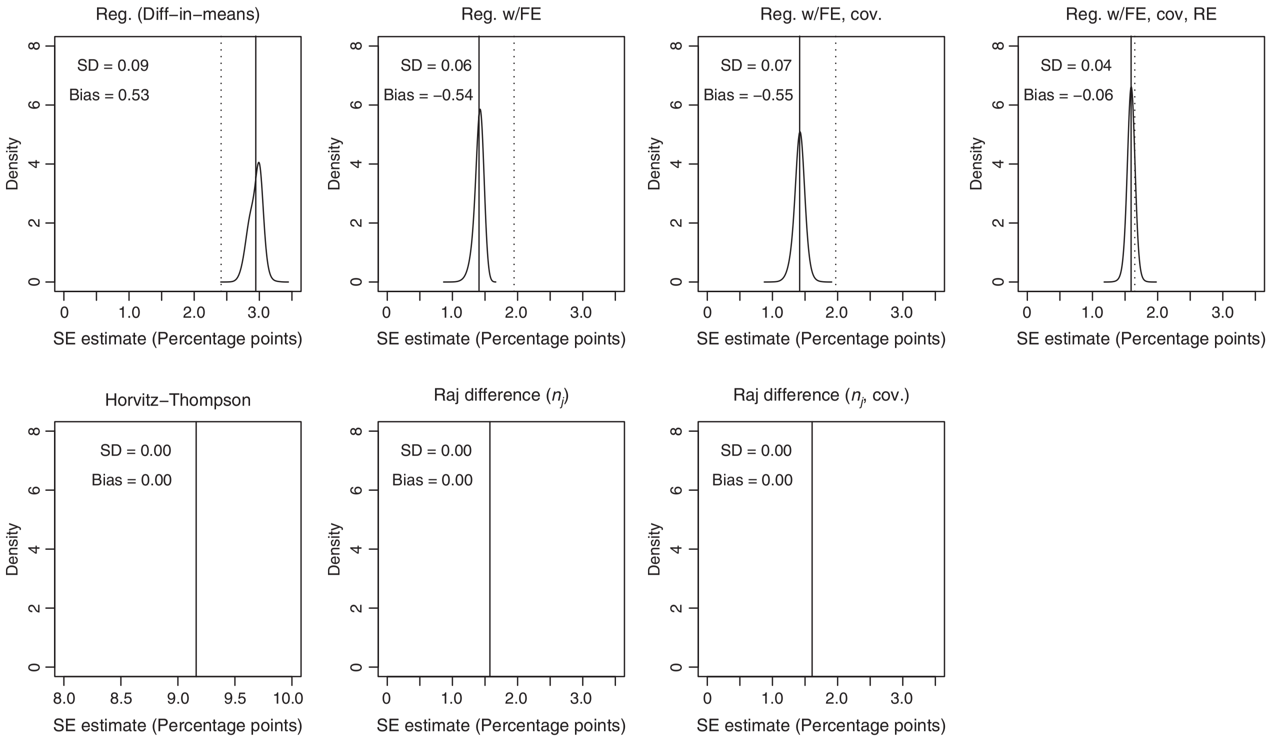


Figure 3: Sampling distributions associated with the SE estimators under the sharp null hypothesis of no treatment effect detailed in Section 6. Five thousand randomizations were used to estimate the sampling distributions. Density plots were generated using the `density()` function in R (R Development Core Team 2010), with the default settings and a bandwidth of 0.05 percentage points. Each SE estimator is detailed in Section 6.3. The solid vertical line indicates the expected value of the SE estimator. The dotted vertical line indicates the true standard error of the estimator. Bias and SD estimates are computed from each empirical distribution. Distributions for the SEs under the sharp null hypothesis of no treatment effect were too narrow to display.

of this model to account for blocking.[10] However, for the regression models that include fixed effects the standard errors are badly biased downward as we might expect given that sandwich type estimators tend to be unreliable in finite samples. The standard errors associated with the random effects model perform reasonably well, being only slightly biased downwards. Meanwhile, under the sharp null, the standard errors associated with the HT estimator and the Raj Difference estimators are exact, being both unbiased and having no sampling variability.

6.5 Randomization Inference with Treatment Effect Heterogeneity

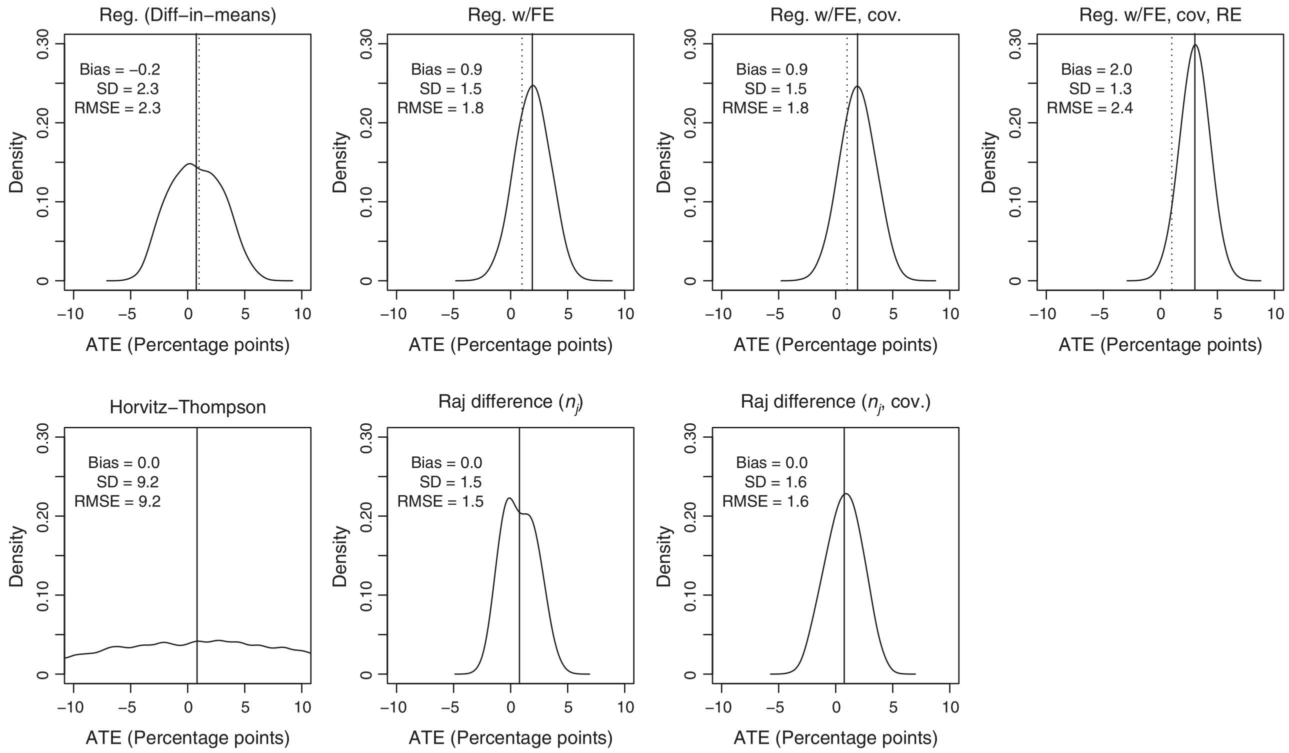
Figure 4 displays results under treatment effect heterogeneity. Solid vertical lines indicate the mean of the sampling distributions. Dotted vertical lines indicate the true treatment effect (0.7 percentage points).

The results demonstrate that the regressions tend to be biased to varying degrees. Interestingly, the regression without covariates (difference-in-means) is only slightly biased downward. That the difference-in-means is not terribly biased can be understood as a result of the sample size (80 clusters) being sufficiently large (recall the consistency proof in Section 3.3).

When the regressions include fixed-effects, however, the bias actually increases. This can be understood in light of the fact that fixed-effects regression estimates yield variance-weighted averages of the block-level estimates (Angrist and Pischke 2009). In other words, the fixed-effects estimator is equivalent to taking the difference-in-means for each block and then taking a weighted average of them. Since the block-level estimates are each biased, the overall average is similarly biased. As discussed in Section 3.4 above, this is a particularly troubling property of the fixed-effects estimator because it will also be inconsistent for increasing numbers of blocks. In other words, adding more blocks to the experiment will not necessarily diminish the overall bias.

Again, the HT estimator is unbiased but has very poor precision. And while the random effects estimator has the lowest standard deviation, Des Raj's difference estimators are the most precise in terms of RMSE.

[10] Although we consider the bias of the standard error estimator, in practice, bias is not an ideal loss function for evaluating standard error estimators. However, given the size of the sample and the typical rate of convergence for variance estimators, we expect that bias serves as an approximation for asymptotic bias, which is of greater interest for constructing confidence intervals.



Reg. (Diff–in–means)

Reg. w/FE

Reg. w/FE, cov.

Reg. w/FE, cov, RE

Horvitz–Thompson

Raj difference (n_j)

Raj difference (n_j, cov.)

Figure 4: ATE estimator sampling distributions associated with heterogeneous treatment effects detailed in Section 6. Five thousand randomizations were used to estimate the desity of the sampling distributions. Density plots were generated using the `density()` function in R (R Development Core Team 2010), with the default settings and a bandwidth of 0.5 percentage points. Each estimator is detailed in Section 6.3. The vertical line indicates the expected value, and therefore bias, of the estimator. Bias and SE estimates in the upper-right of each plot are computed from each empirical distribution.

Note also that the addition of the covariates (age and turnout rate in 2000) actually increases the variability in Raj's difference estimator. This is because the covariates are not particularly predictive of the outcome and so the estimated values of k's tend to miss their mark by a large extent.

Finally, Figure 5 displays the performance of the standard error estimators in the case of heterogeneous treatment effects. Results again show that the "robust cluster" standard errors can perform very badly, being substantially downwardly biased in the case of the regressions with fixed effects. The standard error estimator associated with the random effects regression performs well, being only slightly upwardly biased. The standard error estimators for the HT and Raj difference estimators are conservative, being biased only slightly upwards.

7 Conclusion

The unbiased estimation of the ATE in cluster-randomized experiments has been elusive. In unpacking the source of the bias in the difference-in-means estimator, this paper has also identified some common design-estimator combinations where the bias of estimators will not diminish with sample size such as pair-randomized designs combined with regression estimators with fixed effects for block. This paper has returned to the first principles of randomization and sampling theory, showing that the fundamental statistical properties of randomization can be applied to modern causal inferential problems. Not only does the Des Raj estimator provide the basis for an unbiased and location-invariant estimator for the analysis of cluster-randomized experiments, compared to the HT estimator it also achieves improved precision through covariate adjustment.

There are a number of theoretical implications of this return to the first principles of randomization. First, machinery based solely on sampling-theoretic ideas can be sufficient for precise and unbiased estimation of causal parameters. Second, researchers need not feel that achieving precise and unbiased causal estimates requires an up-to-date knowledge of complex statistical models: we may easily derive estimators with good statistical properties using only fundamental concepts. Third, utilizing such estimators serves to remind us of the importance of this distinction between observational studies and randomized experiments. The importance of the logic of the experiment, with its reliance on randomization, may be lost when researchers rely on model-based estimators that may or may not reflect the experimental design.

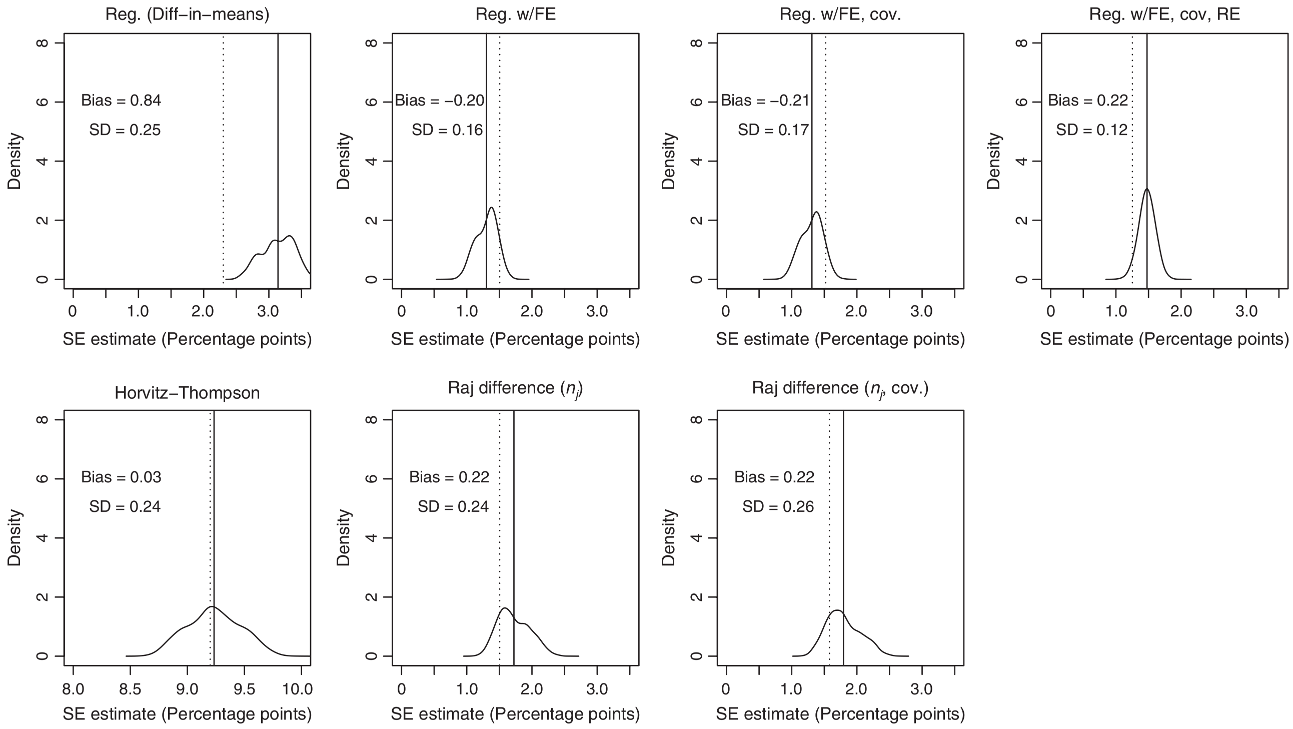


Figure 5: SE estimator sampling distributions associated with the heterogeneous treatment effect detailed in Section 6. Five thousand randomizations were used to estimate the sampling distributions. Density plots were generated using the `density()` function in R (R Development Core Team 2010), with the default settings and a bandwidth of 0.05 percentage points. Each SE estimator is detailed in Section 6.3. The solid vertical line indicates the expected value of the SE estimator. The dotted vertical line indicates the true standard error of the estimator. Bias and SD estimates are computed from each empirical distribution. Distributions for the SEs under the sharp null hypothesis of no treatment effect were too narrow to display.

Acknowledgments: The authors acknowledge support from the Yale University Faculty of Arts and Sciences High Performance Computing facility and staff. The authors would also like to thank Allison Carnegie, Adam Dynes, Don Green, Jennifer Hill, Mary McGrath, David Nickerson, Cyrus Samii and two helpful reviewers for helpful comments. The authors thank Kyle Peyton for research assistance and manuscript preparation. Any errata are the sole responsibility of the authors.

Appendix

A Proof of non-invariance of the Horvitz-Thompson estimator

To prove that the HT estimator is not invariant to location shifts, we need only replace Y_j^T with its linear transformation:

$$\begin{aligned}
\widehat{\Delta^*_{HT}} &= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} Y_j^{T*} - \frac{1}{m_c}\sum_{j\in J_0} Y_j^{T*}\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\sum_{i=1}^{n_j} Y_{ij}^{*} - \frac{1}{m_c}\sum_{j\in J_0}\sum_{i=1}^{n_j} Y_{ij}^{*}\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\sum_{i=1}^{n_j} (b_0 + b_1\cdot Y_{ij}) - \frac{1}{m_c}\sum_{j\in J_0}\sum_{i=1}^{n_j} (b_0 + b_1\cdot Y_{ij})\right] \\
&= \frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\left(n_j\cdot b_0 + \sum_{i=1}^{n_j} b_1\cdot Y_{ij}\right) - \frac{1}{m_c}\sum_{j\in J_0}\left(n_j\cdot b_0 + \sum_{i=1}^{n_j} b_1\cdot Y_{ij}\right)\right] \\
&= b_0\cdot\frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} n_j - \frac{1}{m_c}\sum_{j\in J_0} n_j\right] + b_1\cdot\frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1}\sum_{i=1}^{n_j} Y_{ij} - \frac{1}{m_c}\sum_{j\in J_0}\sum_{i=1}^{n_j} Y_{ij}\right] \\
&= b_0\cdot\frac{M}{N}\left[\frac{1}{m_t}\sum_{j\in J_1} n_j - \frac{1}{m_c}\sum_{j\in J_0} n_j\right] + b_1\cdot\hat{\Delta}_{HT}.
\end{aligned}$$

B Bias from estimating *k* from within-sample data

Consider the situation where one wishes to improve upon the HT estimator by adjusting for cluster size; in other words, one wishes to estimate *k* in equations 20 and 21 from the data to approximate the optimal value of k with an estimator $\hat{k}$. In this scenario, the expected value of equation 20 yields

$$\begin{aligned}
\mathrm{E}\left[\widehat{Y^T_{1,R1}}\right] &= \mathrm{E}\left[\frac{M}{m_t}\sum_{j\in J_1}(Y_j^T-\hat{k}(n_j-N/M))\right] \\
&= \frac{M}{m_t}\left(\mathrm{E}\left[\sum_{j\in J_1}Y_j^T\right]-\mathrm{E}\left[\sum_{j\in J_1}\hat{k}n_j\right]+\mathrm{E}\left[\sum_{j\in J_1}\hat{k}N/M\right]\right) \\
&= \frac{M}{m_t}\left(\mathrm{E}\left[m_t\overline{Y^T_{1j}}\right]-\mathrm{E}[\hat{k}m_t\overline{n_{tj}}]+\mathrm{E}[\hat{k}m_tN/M]\right) \\
&= Y_1^T-M(\mathrm{E}[\hat{k}\overline{n_{tj}}]-\mathrm{E}[\hat{k}]\mathrm{E}[\overline{n_{tj}}]) \\
&= Y_1^T-M\mathrm{Cov}(\hat{k},\overline{n_{tj}}),
\end{aligned} \tag{29}$$

where $\overline{n_{tj}}$ is the mean value of n_j for clusters in the treatment condition in a given randomization. In the third line of equation 29, $\hat{k}$ moves outside the summation operator because it is a constant for a given randomization. Likewise,

$$\mathrm{E}\left[\widehat{Y^T_{0,R1}}\right]=Y_0^T-M\mathrm{Cov}(\hat{k},\overline{n_{cj}}), \tag{30}$$

where $\overline{n_{cj}}$ is the mean value of n_j for units in the control condition in a given randomization. So the expected value of the estimator will be

$$\mathrm{E}\left[\frac{\widehat{Y^T_{1,R1}}-\widehat{Y^T_{0,R1}}}{N}\right]=\Delta+\frac{M}{N}(\mathrm{Cov}(\hat{k},\overline{n_{cj}})-\mathrm{Cov}(\hat{k},\overline{n_{cj}})). \tag{31}$$

The term on the right of equation 31 represents the bias. A special case with no bias is when the sharp null hypothesis of no treatment effect holds and treatment and control groups have equal numbers of clusters. We refer the reader to Williams (1961), Freedman (2008a) and Freedman (2008b) for additional reading on the particular bias associated with the regression adjustment of random samples and experimental data.

C Derivation of the optimal value of *k*

To identify a single optimal value of k, k_{optim*}, we refer to the first line of equation 17,

$$v\mathrm{V}(\widehat{\Delta_{R1}})=c\sigma^2(U_{j0}^T)+t\sigma^2(U_{j1}^T)+2\sigma(U_{j0}^T,U_{j1}^T) \tag{32}$$

where $v=\dfrac{(M-1)N^2}{M^2}$, $c=\dfrac{M-m_c}{m_c}$, and $t=\dfrac{M-m_t}{m_t}$. Now note that the terms $\sigma^2(U_{j0}^T)$, $\sigma^2(U_{j0}^T)$, and $\sigma(U_{j0}^T, U_{j1}^T)$ in equation 32 can be written as follows:

$$\sigma^2(U_{j1}^T)=\sigma^2(Y_{j1}^T)+k^2\sigma^2(n_j)-2k\sigma(Y_{j1}^T, n_j), \tag{33}$$

$$\sigma^2(U_{j0}^T)=\sigma^2(Y_{j0}^T)+k^2\sigma^2(n_j)-2k\sigma(Y_{j0}^T, n_j), \tag{34}$$

and, defining $\delta_j=(n_j-N/M)$,

$$\begin{aligned}\sigma(U_{j0}^T, U_{j1}^T)&=\mathrm{E}\,[U_{j0}^T U_{j1}^T]-\overline{U_0^T U_1^T}\\ &=\mathrm{E}\,[Y_{j0}^T-k\delta_j](Y_{j1}^T-k\delta_j)-\overline{Y_0^T Y_1^T}\\ &=\mathrm{E}\,[Y_{j0}^T Y_{j1}^T-Y_{j0}^T k\delta_j-Y_{j1}^T k\delta_j+k^2\delta_j^2]-\overline{Y_0^T Y_1^T}\\ &=\mathrm{E}[Y_{j0}^T Y_{j1}^T]-\overline{Y_0^T Y_1^T}-\mathrm{E}[Y_{j0}^T k\delta_j]-\mathrm{E}[Y_{j1}^T k\delta_j]+\mathrm{E}[k^2\delta_j^2]\\ &=\sigma(Y_{j0}^T, Y_{j1}^T)-k[\sigma(Y_{j0}^T, n_j)+\mathrm{E}[Y_{j0}^T]\mathrm{E}[\delta_j]]\\ &\quad-k[\sigma(Y_{j1}^T, n_j)+\mathrm{E}[Y_{j1}^T]\mathrm{E}[\delta_j]]+k^2\sigma^2(n_j)\\ &=\sigma(Y_{j0}^T, Y_{j1}^T)-k[\sigma(Y_{j0}^T, n_j)+\mathrm{E}[Y_{j0}^T]\cdot 0]\\ &\quad-k[\sigma(Y_{j1}^T, n_j)+\mathrm{E}[Y_{j1}^T]\cdot 0]+k^2\sigma^2(n_j)\\ &=\sigma(Y_{j0}^T, Y_{j1}^T)-k\sigma(Y_{j0}^T, n_j)-k\sigma(Y_{j1}^T, n_j)+k^2\sigma^2(n_j),\end{aligned} \tag{35}$$

respectively. Substituting equations 33, 34, and 35 into equation 32,

$$\begin{aligned}v\mathrm{V}(\widehat{\Delta_{R1}})=&c[\sigma^2(Y_{j0}^T)+k^2\sigma^2(n_j)-2k\sigma(Y_{j0}^T, n_j)]+t[\sigma^2(Y_{j1}^T)+k^2\sigma^2(n_j)-2k\sigma(Y_{j1}^T, n_j)]\\ &+2[\sigma(Y_{j0}^T,Y_{j1}^T)-k\sigma(Y_{j0}^T,n_j)-k\sigma(Y_{j1}^T,n_j)+k^2\sigma^2(n_j)].\end{aligned}$$

Setting the first derivative with respect to k equal to zero,

$$\begin{aligned}0=&c[2k_{optim*}\sigma^2(n_j)-2\sigma(Y_{j0}^T, n_j)]+t[2k_{optim*}\sigma^2(n_j)-2\sigma(Y_{j1}^T, n_j)]\\ &+2[-\sigma(Y_{j0}^T, n_j)-\sigma(Y_{j1}^T, n_j)+2k_{optim*}\sigma^2(n_j)],\end{aligned}$$

$$\begin{aligned}ck_{optim*}\sigma^2(n_j)+tk_{optim*}\sigma^2(n_j)+2k_{optim*}\sigma^2(n_j)=&c\sigma(Y_{j0}^T, n_j)+t\sigma(Y_{j1}^T,n_j)\\ &+\sigma(Y_{j0}^T, n_j)+\sigma(Y_{j1}^T,n_j)\end{aligned}$$

$$\left(\frac{M-m_c}{m_c}+\frac{M-m_t}{m_t}+\frac{m_c}{m_c}+\frac{m_t}{m_t}\right)k_{optim*}\sigma^2(n_j)=\left(\frac{M-m_c}{m_c}+\frac{m_c}{m_c}\right)\cdot\sigma(Y_{j0}^T,\,n_j)+\left(\frac{M-m_t}{m_t}+\frac{m_t}{m_t}\right)\cdot\sigma(Y_{j1}^T,\,n_j)$$

$$\left(\frac{M}{m_c}+\frac{M}{m_t}\right)k_{optim*}\sigma^2(n_j)=\left(\frac{M}{m_c}\right)\sigma(Y_{j0}^T,\,n_j)+\left(\frac{M}{m_t}\right)\sigma(Y_{j1}^T,\,n_j)$$

$$k_{optim*}=\left(\frac{1}{m_c}+\frac{1}{m_t}\right)^{-1}\left[\left(\frac{1}{m_c}\right)\frac{\sigma(Y_{j0}^T,\,n_j)}{\sigma^2(n_j)}+\left(\frac{1}{m_t}\right)\frac{\sigma(Y_{j1}^T,\,n_j)}{\sigma^2(n_j)}\right]$$

$$k_{optim*}=\left(\frac{1}{m_c}+\frac{1}{m_t}\right)^{-1}\left[\left(\frac{1}{m_c}\right)k_{optim_c}+\left(\frac{1}{m_t}\right)k_{optim_t}\right]$$

$$k_{optim*}=\frac{m_t}{M}k_{optim_c}+\frac{m_c}{M}k_{optim_t}.$$

The Des Raj estimator will be more efficient than the HT estimator when

$$(c+t+2)k^2\sigma^2(n_j)<2k[(c+1)\sigma(Y_{j0}^T,\,n_j)+(t+1)\sigma(Y_{j1}^T,\,n_j)]$$

$$(c+t+2)k^2<2k\left[(c+1)\frac{\sigma(Y_{j0}^T,\,n_j)}{\sigma^2(n_j)}+(t+1)\frac{\sigma(Y_{j1}^T,\,n_j)}{\sigma^2(n_j)}\right]$$

$$(c+t+2)k^2<2k[(c+1)k_{optim_c}+(t+1)k_{optim_t}]$$

$$k^2<2k\left[\frac{m_t}{M}k_{optim_c}+\frac{m_c}{M}k_{optim_t}\right]$$

$$k^2<2k\cdot k_{optim*}.$$

References

Angrist, J. D. and J. Pischke (2009) *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Aronow, P. M., D. P. Green and D. K. K. Lee (2014) "Sharp Bounds on the Variance in Randomized Experiments," Annals of Statistics, 42(3):850–871.

Bates, D. and M. Maechler (2010) lme4: Linear mixed-effects models using S4 classes. R package, version 0.999375-37.

Brewer, K. R. W. (1979) “A Class of Robust Sampling Designs for Large-Scale Surveys,” Journal of the American Statistical Association, 74:911–915.

Chaudhuri, A. and H. Stenger (2005) *Survey Sampling*. Boca Raton: Chapman and Hall.

Cochran, W. G. (1977) *Sampling Techniques*, 3rd ed. New York: John Wiley.

Des Raj. (1965) “On A Method of Using Multi-Auxiliary Information in Sample Surveys,” Journal of The American Statistical Association, 60:270–277.

Ding, P. (2014) “A Paradox from Randomization-Based Causal Inference,” arXiv preprint arXiv:1402.0142.

Donner, A. and N. Klar (2000) *Design and Analysis of Cluster Randomization Trials in Health Research.* New York: Oxford Univ. Press.

Freedman, D. A. (2006) “On the So-Called ‘Huber Sandwich Estimator’ and ‘Robust’ Standard Errors,” American Statistician, 60:299–302.

Freedman, D. A. (2008a) “On Regression Adjustments to Experimental Data,” Advances in Applied Mathematics, 40:180–193.

Freedman, D. A. (2008b) “On Regression Adjustments in Experiments with Several Treatments.” Annals of Applied Statistics, 2:176–196.

Freedman, D. A., R. Pisani and R. A. Purves (1998) *Statistics*, 3rd ed. New York: W. W. Norton, Inc.

Green, D. P. and L. Vavreck (2008) “Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches,” Political Analysis, 16:138–152.

Hansen, B. and J. Bowers (2008) “Covariate Balance in Simple, Stratified and Clustered Comparative Studies,” Statistical Science, 23:219–236.

Hansen, B. and J. Bowers (2009) “Attributing Effects to a Cluster-Randomized Get-Out-the-Vote Campaign,” Journal of the American Statistical Association, 104:873–885.

Hartley, H. O. and A. Ross (1954) “Unbiased Ratio Estimators,” Nature, 174:270.

Hoffman, E. B., P. K. Sen and C. R. Weinberg (2001) “Within-Cluster Resampling,” Biometrika, 88: 1121–1134.

Horvitz, D. G. and D. J. Thompson (1952) “A Generalization of Sampling Without Replacement From a Finite Universe,” Journal of the American Statistical Association, 47:663–684.

Humphreys, M. (2009) *Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogeneous Assignment Probabilities*. Working paper. Available at: http://www.columbia.edu/~mh2245/papers1/monotonicity4.pdf.

Imai, K., G. King and C. Nall (2009) “The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation,” Statistical Science, 24:29–53.

King, G. and M. Roberts (2014) “How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It,” Political Analysis, 1–12.

Lachin, J. M. (1988) “Properties of Simple Randomization in Clinical Trials,” Controlled Clinical Trials, 9(4):312–326.

Lin, W. (2013) “Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman’s Critique,” Annals of Applied Statistics, 7(1):295–318.

Middleton, J. A. (2008) “Bias of the Regression Estimator for Experiments Using Clustered Random Assignment,” Statistics and Probability Letters, 78:2654–2659.

Miratrix, L., J. Sekhon and B. Yu (2013) “Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments,” Journal of the Royal Statistical Society. Series B (Methodological), 75(2):369–396.

Neyman, J. (1923) “On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9,” Statistical Science, 5:465–480. (Translated in 1990).

Neyman, J. (1934) “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection,” Journal of the Royal Statistical Society, 97(4):558–625.

R Development Core Team. (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. Version 2.12.0.

Rosenbaum, P. R. (2002) *Observational Studies*, 2nd ed. New York: Springer.

Rubin, D. (1974) “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” Journal of Educational Psychology, 66:688–701.

Rubin, D. B. (1978) “Bayesian Inference for Causal Effects: The Role of Randomization,” The Annals of Statistics, 6:34–58.

Rubin, D. B. (2005) “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions,” Journal of the American Statistical Association, 100:322–331.

Samii, C. and P. M. Aronow (2012) “On Equivalencies Between Design-Based and Regression-Based Variance Estimators for Randomized Experiments,” Statistics and Probability Letters, 82:365–370.

Sarndal, C.-E. (1978) “Design-Based and Model-Based Inference in Survey Sampling,” Scandinavian Journal of Statistics, 5(1):27–52.

Williams, W. H. (1961) “Generating Unbiased Ratio and Regression Estimators,” Biometrics, 17:267–274.