Improving Average Treatment Effect Estimates in Small-Scale Randomized Controlled Trials

Isaac M. Opper
RAND Corporation

Researchers often include covariates when they analyze the results of randomized controlled trials (RCTs), valuing the increased precision of the estimates over the potential of inducing small-sample bias when doing so. In this paper, we develop a sufficient condition which ensures that the inclusion of covariates does not cause small-sample bias in the effect estimates. Using this result as a building block, we develop a novel approach that uses machine learning techniques to reduce the variance of the average treatment effect estimates while guaranteeing that the effect estimates remain unbiased. The framework also highlights how researchers can use data from outside the study sample to improve the precision of the treatment effect estimate by using the auxiliary data to better model the relationship between the covariates and the outcomes. We conclude with a simulation, which highlights the value of using the proposed approach.

VERSION: January 2021

Improving Average Treatment Effect Estimates in Small-Scale Randomized Controlled Trials[*]

Isaac M. Opper[†]

January 5, 2021

Abstract

Researchers often include covariates when they analyze the results of randomized controlled trials (RCTs), valuing the increased precision of the estimates over the potential of inducing small-sample bias when doing so. In this paper, we develop a sufficient condition which ensures that the inclusion of covariates does not cause small-sample bias in the effect estimates. Using this result as a building block, we develop a novel approach that uses machine learning techniques to reduce the variance of the average treatment effect estimates while guaranteeing that the effect estimates remain unbiased. The framework also highlights how researchers can use data from outside the study sample to improve the precision of the treatment effect estimate by using the auxiliary data to better model the relationship between the covariates and the outcomes. We conclude with a simulation, which highlights the value of using the proposed approach.

[*] I thank John Engberg and William R. Johnston for early discussions on the question of how to use machine learning techniques to improve the effect estimates. Richardo Sanchez and Inez Khan took the lead in developing the R package that implements the method and both Troy Smith and Lisa Abraham provided valuable feedback on an early draft. Funding for this research was provided by gifts from RAND supporters and income from operations.

[†] RAND Corporation. Email: iopper@rand.org.

I Introduction

Randomized controlled trials (RCTs), in which treatment is randomly assigned to study participants, are a key tool for researchers and are often regarded as the gold standard of causal inference. One main advantage of RCTs is that simply comparing the average outcome of the treated observations to the average outcome of control observations provides an intuitive and unbiased estimate of the average treatment effect of the intervention. Unbiasedness is not the only criterion for an estimator, however, and there is a robust literature on alternative evaluation approaches that provide more precise estimates.

One of the most studied approaches involves controlling for covariates in the analysis of RCTs. The current literatures suggests a bias-variance tradeoff: a number of papers have illustrated that adjusting for a fixed number of covariates can reduce the asymptotic variance of the estimates, while other research illustrates that including covariates can lead to small-sample bias in the estimator, e.g. (Athey and Imbens, 2017; Freedman, 2008a; Lin, 2013; Berk et al., 2013; Freedman, 2008b; Zhang et al., 2008; Pitkin et al., 2013; Wager et al., 2016; Bloniarz et al., 2016; Lei and Ding, 2019). Given the small sample size of many RCTs, researchers are often stuck in an uncomfortable position: the inclusion of covariates is tempting due to the otherwise imprecise estimates, but the potential for inducing bias by doing so is non-trivial given the small sample size.

This debate on whether to include covariates in the analysis of RCTs is amplified when one considers using machine learning techniques such as lasso or random forest regressions to control for the covariates instead of ordinary least squares (OLS). The potential value of including controls is larger for these ML techniques than in a traditional OLS case, but the concern about small-sample bias is less well understood and potentially more of a concern due to the increased flexibility. A burgeoning literature has begun to study the performance of estimators that use machine learning approaches to evaluate RCTs, but all have made strong assumptions on the data generating process and/or taken an asymptotic perspective in which the size of the RCT grows, e.g. (Wager et al., 2016; Bloniarz et al., 2016; Lei and Ding, 2019; Wu and Gagnon-Bartsch, 2018).

In this paper, we use a design-based or randomization model in which the only uncertainty in the estimate is generated by the randomness of the treatment assignment.

We then focus on a set of estimators that retain the intuition behind the standard estimation approach by comparing the average residualized outcome, i.e. the outcome minus some function of the covariates, of the treated group to the average residualized outcome of the control group. We then derive a sufficient condition which, when satisfied, guarantees that using covariates to residualize the outcome will not add bias even in finite samples and discuss how this sufficient condition can be satisfied under most common treatment assignment mechanisms and estimation approaches.

We then turn our attention to the variance of the proposed estimator. A downside of the design-based model is that general statements about optimal residualization are difficult, as the optimal approach will depend on the characteristics of the particular (fixed) sample. Instead of an optimality condition, we show that as a general rule the uncertainty in the treatment effect decreases as one can better predict a weighted average of the two potential outcomes of the study sample.

Together, the above results suggest that the residualization can be done with the perspective that any approach can be judged purely based on how well one is able to (out-of-sample) predict the treated and control outcomes. We use this insight to propose a novel way to include covariates in the analysis of an RCT that starts by predicting the outcomes without distinguishing between treated and control observations, before separately modeling the relationship between the covariates and the resulting treatment and control residuals.

We also use the results to motivate the inclusion of information from observations outside the study, such as settings that did not participate in the RCT. Even if there is non-random selection of observations into the study, these observations may allow one to better residualize the outcome of observations inside the study and therefore improve the precision of the RCT. Our theoretical result suggests a way one can do so that does not risk adding bias to the estimates, turning the question of whether to include these observations in the analysis into an easily testable empirical one. This is in contrast to existing approaches, which generally add bias to the effect estimate as the cost of increasing its precision (Angrist et al., 2016; Kaizar, 2015).

Finally, we illustrate the potential value of the proposed approach by turning to an empirical application in which we randomly select schools from the New York City administrative data, randomly select some of them to be considered treated schools, and then add a known heterogeneous treatment effect to the treated school's true outcome. We show that the proposed approach reduces the variance of the estimates by anywhere from 55% to 70%, depending on the sample size and method used. This is equivalent to more than doubling the sample size of the RCT. Thus, the proposed method is a cost free way to improve the precision of treatment effect estimates, leading to more cost-effective RCTs and therefore more rapid scientific advancement.

II Conceptual Framework and Notation

We will use Rubin's Causal Model as the conceptual framework, in which we postulate the existence of one potential outcome for individual i if she is not treated and an alternative potential outcome for individual i if she is treated. We will use μ_i to denote individual i's outcome if she is not treated and τ_i is the causal effect of the treatment, i.e. the difference between her outcome if she is treated and her outcome if she is not treated. Using T_i, which equals one if she is treated and zero if not, to denote her treatment status, we can therefore write her observed outcome as $Y_i = \mu_i + \tau_i T_i$. We will also assume that there are a vector of covariates X_i that are observed by the researcher and are not impacted by the treatment.

We will take the perspective that the only randomness is in the treatment assignment, which implies that we can treat μ_i and τ_i as fixed parameters and T_i as the only random variable. We will let each individual potentially have different probabilities of being assigned to the treatment condition and only assume that $\mathbb{E}[T_i] = p_i \in (0, 1)$. We will not make any additional assumptions on how this assignment process works, allowing for correlations in the assignments as would be the case when a fixed number of units are treated or assignment is done via stratified random sampling, rerandomization, or when units within a cluster are all assigned the same treatment status.

Finally, we will restrict our attention to estimators that consist of differencing the mean residuals in the treatment group and the mean residuals in the control group, potentially with weights to adjust for the fact that different observations may have different probabilities of being assigned to the treatment. More precisely, we will focus on estimators of the form:

$$\hat{\tau} = \frac{1}{N} \sum_{\forall i} \frac{Y_i - \hat{g}_i(X_i)}{p_i} T_i - \frac{Y_i - \hat{g}_i(X_i)}{1 - p_i} (1 - T_i) \tag{1}$$

for some function $\hat{g}_i(X_i)$ that can be estimated from the observed data (Aronow and

Middleton, 2013). Note that in cases where each observation has the same probability of treatment, this is equivalent to:

$$\hat{\tau} = \left(\frac{1}{N_T} \sum_{\forall i \text{ s.t. } T_i=1} Y_i - \hat{g}_i(X_i) \right) - \left(\frac{1}{N_C} \sum_{\forall i \text{ s.t. } T_i=0} Y_i - \hat{g}_i(X_i) \right)$$

where N_T is the number of treated observations and N_C is the number of control observations.[1] Since we take this form as given, the focus of this paper is solely on how best to estimate $\hat{g}_i(X_i)$.

III Bias and Variance of Estimator

Bias of Estimator

We first turn our attention to the bias of the estimator. Given the definition of the estimator in Equation (1), we have that

$$\mathbb{E}[\hat{\tau}] = \frac{1}{N} \sum_{\forall i} \mathbb{E}\left[\frac{Y_i - \hat{g}_i(X_i)}{p_i} T_i - \frac{Y_i - \hat{g}_i(X_i)}{1 - p_i} (1 - T_i) \right] \tag{2}$$

Replacing Y_i with $\tau_i T_i + \mu_i$ and noting that $T_i \cdot T_i = T_i$ and that $T_i \cdot (1 - T_i) = 0$, we can write this expression as:

$$\begin{aligned} \mathbb{E}[\hat{\tau}] = \frac{1}{N} \sum_{\forall i} \mathbb{E}\Big[\frac{\tau_i T_i}{p_i} \Big] + \frac{1}{N} \sum_{\forall i} \mathbb{E}\Big[\frac{\mu_i T_i}{p_i} - \frac{\mu_i (1 - T_i)}{1 - p_i} \Big] \\ - \frac{1}{N} \sum_{\forall i} \mathbb{E}\Big[\frac{\hat{g}_i(X_i) T_i}{p_i} - \frac{\hat{g}_i(X_i)(1 - T_i)}{1 - p_i} \Big] \end{aligned}$$

Because the only randomness is in the treatment assignment, the terms τ_i, μ_i and p_i can be viewed as fixed with respect to the expectation. Therefore $\mathbb{E}\left[\frac{\mu_i T_i}{p_i}\right] = \frac{\mu_i \mathbb{E}[T_i]}{p_i} = \frac{\mu_i \mathbb{E}[1 - T_i]}{1 - p_i} = \mathbb{E}\left[\frac{\mu_i (1 - T_i)}{1 - p_i}\right]$, which ensures that $\frac{1}{N} \sum_{\forall i} \mathbb{E}\left[\frac{\mu_i T_i}{p_i} - \frac{\mu_i (1 - T_i)}{1 - p_i}\right] = 0$.

[1] Technically, N_T is the expected number of treated observations, or $N_T = pN$, where p is the probability of being assigned to the treatment, and N_C is the expected number of control observations, or $(1 - p)N$.

Similarly, $\mathbb{E}\Big[\frac{\tau_i T_i}{p_i}\Big] = \tau_i$ which means the formula simplifies to:

$$\mathbb{E}[\hat{\tau}] = \frac{1}{N}\sum_{\forall i}\tau_i - \frac{1}{N}\sum_{\forall i}\mathbb{E}\Big[\frac{\hat{g}_i(X_i)T_i}{p_i} - \frac{\hat{g}_i(X_i)(1-T_i)}{1-p_i}\Big] \tag{3}$$

which means the bias of the estimator defined in Equation (1) is given by:

$$\mathbb{B}\Big(\hat{\tau}\Big) = -\frac{1}{N}\sum_{\forall i}\mathbb{E}\Big[\hat{g}_i(X_i)\cdot\Big(\frac{T_i}{p_i} - \frac{(1-T_i)}{1-p_i}\Big)\Big] \tag{4}$$

Note that unlike μ_i, the value of $\hat{g}_i(X_i)$ is not fixed. If, for example, one assumes that $\hat{g}_i(X_i) = \beta' X_i$ and estimates β using an OLS regression of Y_i on X_i, the estimated values of β will depend in part on the value of Y_i, which in turn depends on whether individual i is assigned to the treatment group or control group. This dependence means that, unlike the case of μ_i, the value of $\hat{g}_i(X_i)$ is different when i is treated than when i is control, which means that $\mathbb{E}\big[\hat{g}_i(X_i)\cdot\frac{T_i}{p_i}\big]$ will not necessarily be equal to $\mathbb{E}\big[\hat{g}_i(X_i)\cdot\frac{1-T_i}{1-p_i}\big]$. However, this also implies the following sufficient condition:

Theorem 1. *Suppose that for every i the value of* $\mathbb{E}\big[\hat{g}_i(X_i)|T_i = 1\big] = \mathbb{E}\big[\hat{g}_i(X_i)|T_i = 0\big]$. *Then* $\mathbb{B}\Big(\hat{\tau}\Big) = 0$.

This sufficient condition highlights the fact that any bias that controlling for covariates induces in the estimates comes from the fact that $\hat{g}_i(X)$ is estimated and the potential that the estimation approach causes dependencies between the estimate of $\hat{g}_i(X)$ and i's treatment assignment. This suggests that if one exercises care in the estimation of $\hat{g}_i(X)$, it might be possible to ensure that controlling for covariates does not lead to any bias in the treatment effect estimate. A natural way to do this is to partition observations into K subgroups in a way that knowing all the treatment assignments of all individuals outside subgroup K' does not provide any information on the treatment assignment of individuals within subgroup K', and then estimate $\hat{g}_i(X)$ using observations that are in a different subgroup than i. To formalize this, we note the following theorem:

Theorem 2. *Let $\mathscr{A}$ be any algorithm that inputs outcomes and covariates and outputs a bounded function which itself inputs covariates and outputs a real number. Then define an estimator that:*

1. *Partitions the N observations into K subgroups such that $\mathbb{E}[T_i|T^{(-K)}] = p_i$ for all $i \in K$, where $T^{(-K)}$ be the vector of treatment assignments for all observations not in subgroup K;*

2. *Calculates $\hat{g}_i(X)$ using the algorithm $\mathscr{A}$ on observations that are in a different subgroup than i;*

3. *Estimates the treatment effect using Equation (1).*

Then that estimator generates an unbiased estimate of the average treatment effect.

A natural follow-up question is whether it is possible to partition N observations into K subgroups such that $\mathbb{E}[T_i|T^{(-K)}] = p_i$ for all $i \in K$. Luckily, this partitioning is possible under many of the common randomization approaches. Four examples of partitions that satisfy this criteria are described below:

- Suppose that the treatment assignment is independently determined for each observation. Then the criteria holds for any choice of K subgroups.

- Suppose that treatment assignment is done by drawing a fixed number of observations, which become the treated observations, from the set of all observations in the study. Furthermore, suppose that the number of treated units is such that is possible to partition the N observations into K subgroups with N_K observations, each of which has $p \cdot N_K$ treated observations and $(1 - p) \cdot N_K$ control observations. Then the criteria holds when there are exactly $p \cdot N_K$ treated observations and $(1 - p) \cdot N_K$ control observations in each subgroup.[2]

- Suppose that treatment is determined by stratified sampling, in which the full set of observations are divided into a number of strata and the treatment assignment is done independently across the strata. Then the criteria holds when using the strata as the K subgroups.

- Suppose that the researchers observe some a set of observations that are not in the study. Then splitting the sample into two subgroups, one which consists of observations in the study and one of observations outside of the study, satisfies the criteria regardless of the treatment assignment mechanism.

[2] See proof in appendix. The idea is also easily extended to cluster randomized trial in which a fixed number of clusters are randomly chosen and all individuals within the chosen cluster receive the treatment and those in other clusters are considered controls. In this case, the clusters should be partitioned in subgroups and each individual within a cluster is in the same subgroup.

As can be seen in the examples above, while it is often possible to satisfy the condition it does require some care in how researchers partition the sample. For example, image a simple RCT in which half the 100 total observations are randomly chosen and assigned to the treatment. In this case, the commonly used leave-one-out approach of completely partitioning the sample into 100 subgroups each with one observation would not satisfy the assumption, as knowing the treatment assignment of the other 99 observations in the study provides enough information to determine the treatment assignment of the 100^{th} observation. However, partitioning the sample into 50 subgroups of two observations each, one of which is treated and one is control, would satisfy the assumptions of Theorem 2.

Variance of Estimator

The previous discussion suggests that under a number of common treatment assignments it is possible to control for covariates in a RCT framework without adding bias to the estimates. Perhaps surprisingly, it showed that if one is careful about determining the K folds, there are virtually no restrictions on how to control for covariates. While the previous discussion suggests that it is possible to control for covariates without biasing the estimate, however, the flexibility means the result above provides little guidance on how one should do so. In this section, we calculate the variance of the estimator, in order to guide the optimal choice of $\hat{g}_i(X_i)$.

To do so, it will help to denote ϵ_i as the amount that $\hat{g}_i(X_i)$ differs from $\mu_i + (1 - p_i)\tau_i$, i.e. $\epsilon_i = \hat{g}_i(X_i) - \Big(\mu_i + (1 - p_i)\tau_i\Big)$. We can substitute this expression and the fact that $Y_i = \mu_i + \tau_i T_i$ into Equation (1) to get that the variance of the estimate corresponds to:[3]

$$\mathbb{V}\Big(\hat{\tau}\Big) = \mathbb{V}\Bigg(\frac{1}{N}\sum_{\forall i}\Bigg[\frac{\epsilon_i}{p_i}T_i - \frac{\epsilon_i}{1 - p_i}(1 - T_i)\Bigg]\Bigg) \tag{5}$$

Equation (5) highlights that as a general rule the more closely we can set $\hat{g}_i(X_i)$ to be equal to $\mu_i + (1 - p_i)\tau_i$, the more precisely we can measure the average effect of the treatment. For example, if treatment assignment is done independently for each observation and the probability for each observation to be assigned to the treatment

[3] See appendix for full derivation.

was p, this expression reduces an sum of the squared-residuals times a constant, or:

$$\mathbb{V}\Big(\hat{\tau}\Big) = \frac{1}{p(1-p)}\frac{1}{N^2}\sum_{\forall i}\epsilon_i^2 \tag{6}$$

More broadly, even with some correlations in the treatment assignment, it is likely minimizing the sum of squared-residuals also comes close to minimizes the variance of the average treatment effect estimate, i.e. a function chosen as:

$$\hat{g}_i(X_i) = \arg\min_{g_i(X_i)}\Bigg(\Big(\mu_i + (1-p_i)\tau_i\Big) - g_i(X_i)\Bigg)^2 \tag{7}$$

is likely near optimal. Note that though the motivation and derivation is different, this result gives the same formula as used in the doubly-robust estimation of average treatment effects in observational studies, e.g. Equation 13.37 in (Tsiatis, 2006). There is an important distinction that in RCTs the propensity scores are known rather than estimated. This means that the treatment effect estimate is unbiased even if $\hat{g}_i(X_i) \neq \mu_i + (1-p_i)\tau_i$, subject to the constraints discussed in Section III. As discussed more below, this gives more freedom in how one can estimate $\hat{g}_i(X_i)$.

IV Proposed Method

Guided by the result of Theorem 2, our proposed method consists of four steps:

1. Partition the Sample;
2. Estimate Relationship Between Covariates and Outcomes;
3. Estimate the Treatment Effect;
4. Conduct Inference on the Estimated Effect.

Steps one and three are straightforward. Step one is to partition the N observations into K subgroups in a way that satisfies the criteria specified in Theorem 2; we discussed in Section III how this is possible in most randomization schemes. Similarly, step three is to estimate the treatment effect, which can be done by using the results of step two in Equation (1). Steps two and four, in contrast, are more nuanced; we discuss them in more detail below.

Estimating the Relationship Between the Covariates and Outcomes

As we discuss above, we aim to choose $\hat{g}_i(X_i)$ to be as close as possible to $\mu_i+(1-p_i)\tau_i$, in a way that ensures whatever approach is used satisfies the constraints discussed in Section III to ensure that the resulting treatment effect estimate is unbiased. A challenge, of course, is that we do not observe both μ_i and τ_i for each individual. It will therefore be helpful to relate $\mu_i + (1-p_i)\tau_i$ to the two potential outcomes for individual i. We can do so by re-writing $\mu_i + (1-p_i)\tau_i$ as $p_i\mu_i + (1-p_i)(\mu_i+\tau_i)$ since we observe μ_i for control individuals and $\mu_i+\tau_i$ for treated individuals. A natural approach is therefore to set $\hat{g}_i(X_i) = p_i\hat{g}_{i,0}(X_i) + (1-p_i)\hat{g}_{i,1}(X_i)$, where:

$$\hat{g}_{i,0}(X_i) = \arg\min_{g_0\in G} \sum_{\forall j s.t. K_j\neq K_i and. T_j=0} \Big(Y_j - g_0(X_j)\Big)^2 + \lambda J(g_0) \tag{8}$$

$$\hat{g}_{i,1}(X_i) = \arg\min_{g_1\in G} \sum_{\forall j s.t. K_j\neq K_i and. T_j=1} \Big(Y_j - g_1(X_j)\Big)^2 + \lambda J(g_1) \tag{9}$$

and K_i denotes the subgroup that i is partitioned into, G is the set of potential functions, and $\lambda J(g)$ is a penalization term meant to smooth g_0 and g_1.

There are, however, reasons to believe it is possible to improve on this approach. Most notably, even when there is substantial treatment effect heterogeneity, it is likely dwarfed by the amount of heterogeneity across individuals, i.e. variation across individuals in τ_i is substantially lower than variance across individuals in μ_i. In this case, the estimates of $\hat{g}_{i,0}(X_i)$ and $\hat{g}_{i,1}(X_i)$ should be similar and incorporating this prior into the estimates is likely to improve the predictions.

We therefore propose a boosting-like approach, in which one starts by pooling all the data and estimating a function $\hat{h}_i(X_i)$ that predicts all outcomes well. It then estimates two separate functions that separately predict the residuals in the treatment and control. More formally, the first step is to choose $\hat{h}(\tilde{X}_i)$ such that

$$\hat{h}_i(X_i) = \arg\min_{h\in G} \sum_{\forall j s.t. K_j\neq K_i} \Big(Y_j - h(X_i)\Big)^2 + \lambda J(h) \tag{10}$$

We then use this function to calculate the residuals, i.e. $r_i = Y_i - \hat{h}_i(X_i)$. We next estimate a function $\hat{h}_{i,0}(X_i)$ that predicts the residuals well in the control group and another function $\hat{h}_{i,1}(X_i)$ that predicts the residuals well in the treatment group, i.e.

$$\hat{h}_{i,0}(X_i) = \arg\min_{h_0 \in G} \sum_{\forall j s.t. K_j \neq K_i and T_j = 0} \Big(r_j - h_0(X_j)\Big)^2 + \lambda J(h_0) \tag{11}$$

$$\hat{h}_{i,1}(X_i) = \arg\min_{h_1 \in G} \sum_{\forall j s.t. K_j \neq K_i and T_j = 1} \Big(r_j - h_1(X_j)\Big)^2 + \lambda J(h_1) \tag{12}$$

We then set $\hat{g}_{i,0}(X_i) = \hat{h}_i(X_i) + \hat{h}_{i,0}(X_i)$ and $\hat{g}_{i,1}(X_i) = \hat{h}_i(X_i) + \hat{h}_{i,1}(X_i)$.

Note that the method is so far agnostic regarding the specific technique used to estimate $\hat{h}_i$, $\hat{h}_{i,0}$, and $\hat{h}_{i,1}$. Rather than focusing on whether to estimate these functions using lasso or random forest, which will depend on the form of μ_i and τ_i, we instead focus on what variables the algorithms should aim to predict. Nearly any supervised learning algorithm can then be used to estimate these functions (Hastie et al., 2009; Tibshirani, 1996; Zou and Hastie, 2005; Breiman, 2001; Hill, 2011). In the simulations below, we estimate the functions once using lasso and once using random forest and then average the two estimates to generate our estimate of $\hat{g}$. Finally, it is worth noting that in an unpenalized linear model, this approach is equivalent to to the first approach defined in Equations (8) - (9). In our simulations, however, we find that the boosting approach significantly improves the out-of-sample predictions and lowers the variance of the treatment effect estimates.

Employing an Auxiliary Sample

It is often the case that even in small-scale RCTs researchers have access to a much larger auxiliary sample consisting of observations that did not participate in the study, such as administrative data on the full population of interest. From Theorem 2, we see that using these observations to estimate $\hat{g}$ will not bias the treatment effect estimate, as long as they are not included in the final estimate given by Equation (1). As we discuss in the Appendix and show in the simulations, even if there is some non-random selection into the study use of the auxiliary sample can further reduce the variance.

Conducting Inference on the Estimated Effect

Understanding the precision of the effect estimates presents its own set of challenges. One approach is to calculate an exact p-value of the sharp null hypothesis, i.e. the hypothesis that the treatment effect is zero for everyone or $\tau_i = 0$ for all i, using randomization inference. Under this null, none of the individuals' outcomes depend on whether they were assigned to the treatment or control, so it is possible to determine the distribution of estimated effects under the null by simulating the randomization process multiple times and each time repeating the process above to generate an estimated "effect" given the specified treatment assignment. The p-value then corresponds to the proportion of these estimate effects that are larger (in absolute value) than the estimated effect using the true treatment assignment. Versions of this approach can similarly be used to generate confidence intervals or to test more complex null hypotheses (Ding et al., 2014; Caughey et al., 2017). While randomization inference gives exact p-values, however, it can be computationally intensive; we discuss approximation approaches in the Appendix.

V Empirical Simulations

Simulation Specification

To explore how well the method works, we conduct a simulation using school-level New York City administrative data for our simulation. We also conducted additional simulations, which are discussed in the Appendix. For our hypothetical experiment, we randomly select a subset of the schools to participate in the hypothetical study aimed at improving students' math scores. To capture the potential of non-random selection into the study, we only employ schools that applied for the Attendance Improvement and Dropout Prevention (AIDP) program as potential treated schools. We then randomly assign one-quarter of the participating schools to the treatment; the other three-quarters were considered the control schools.

We then simulate a treatment effect that captures the idea that hypothetical treatment has a bigger impact on schools that: a) have higher poverty; b) struggle more initially with student attendance. To do so, we define:

$$\tau_i = \alpha + \frac{C}{1 + atrate_i * (1 - pov_i)} \tag{13}$$

$atrate_i$ is the baseline attendance rate at the school, pov_i is the baseline poverty rate at the school. The two constants α and C chosen to make the average treatment effect equal to 0.15 and the standard deviation of the treatment effect equal to 0.05. Note that the functional form we use is different than the functional form we use to estimate treatment effect heterogeneity, so we are if anything biasing against our proposed approach being useful.

We assume that μ_i, school i's outcome if they are untreated, is equal to the outcome we observe in the data. Again, this means that we are not biasing the simulation toward our advantage by assuming a functional form for μ_i.

For covariates, we use baseline measures (i.e. measures for the two years prior to the hypothetical intervention) of the: racial/ethnic composition of the school, gender composition of the school, fraction of the school in poverty, diagnosed with a disability, or in temporary housing; average math and English test scores, attendance rate and fraction of students who are chronically absent.

Estimators Used

After each simulated random assignment, we estimate the effect using three approaches described below:

1. No Covariates: As a benchmark, we estimate the effect using the standard approach of comparing treatment to control outcomes, without adjusting for covariates.

2. Proposed Approach: We use the algorithm defined in Section IV using the schools that participated in the study.

3. Employing an Auxiliary Sample: We also estimate the effect using the algorithm defined in Section IV using administrative data from all NYC schools, regardless of whether they participated in the study. Note that while we use administrative data from all NYC schools to estimate $\hat{g}(X_i)$, we only use observations in the study to estimate the treatment effect via Equation (1). Despite using data outside the study to estimate $\hat{g}(X_i)$, the resulting ATE is therefore still unbiased. We discuss more details on how this is implemented in the Appendix.

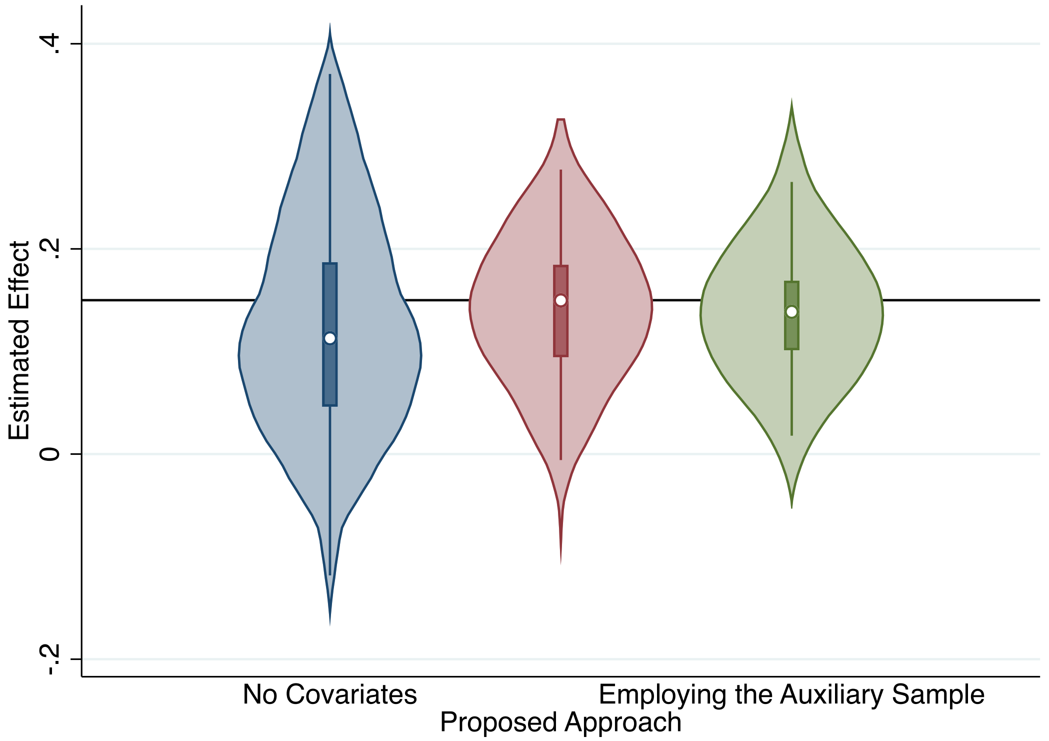
For approaches two and three, we estimate the relationship between the covariates and the outcome using the method described twice, once using lasso and once using random forest regressions. We then average the two results and use these as the ultimate values for $\hat{g}_{i,0}(X_i)$ and $\hat{g}_{i,1}(X_i)$.

Results

Figure 1 is a violin plot showing the distribution of estimated effects over 100 simulations using each of the three estimation approaches defined above, when the overall study includes forty schools; the white dots indicate the median estimate, the box indicates the 25^{th} and 75^{th} percentile-range, and the outer shape indicates the kernel density estimate. As can be seen, the proposed approach significantly reduces the variance of the estimated effects. More concretely, the variance of the proposed approach is approximately 45% the variance of a simple difference-in-means estimator. Employing the auxiliary sample to better estimate the relationship between the outcome and covariates further reduces the variance; the variance of this approach is about one-third as large as the simple difference-in-means estimator. Stated differently, the proposed approach when using the auxiliary sample provides an effect estimate in a study with 40 schools with the same level of precision as a study with 120 schools that used the traditional approach to estimate the effect.

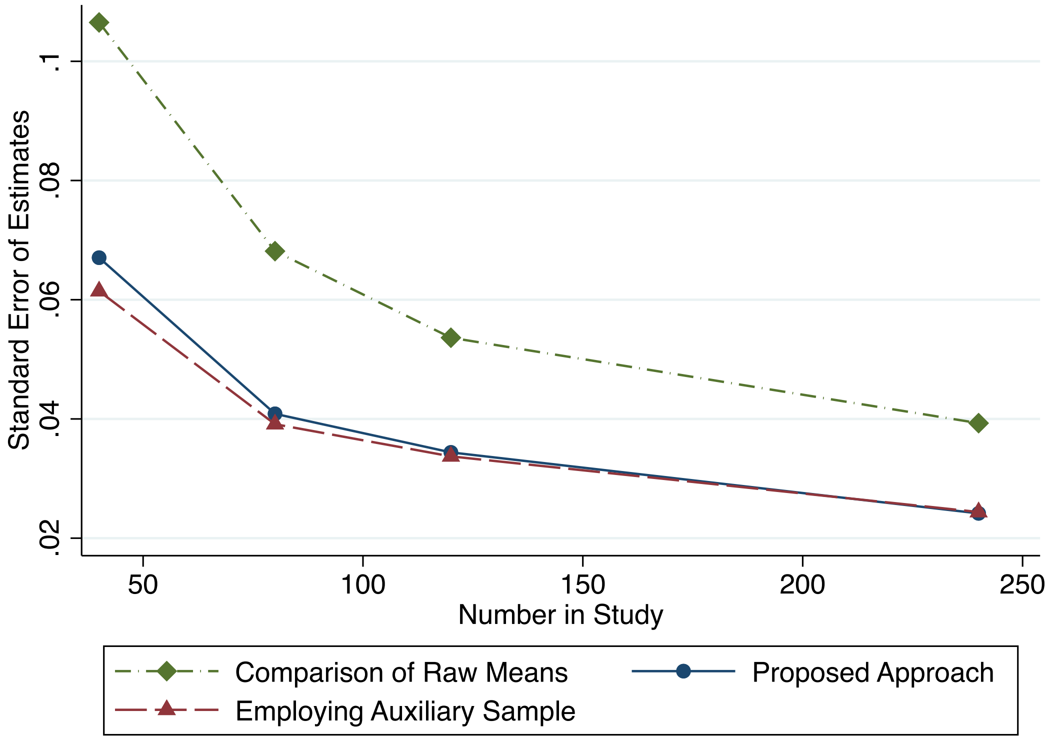
Figure 2 shows that the benefits of using the proposed approach extends to cases where the study size is larger. As the sample size increases, the variance of the simple difference-in-means decreases; the variance of the proposed approaches also decreases, however, at a roughly equivalent rate. Table 1 in the Appendix shows that the ratio of the variation of the proposed approach to the variance of the simple difference-in-means approach actually decreases as the sample size gets larger, due to the fact that larger sample sizes improve the estimation of how the outcome is related to the available covariates. The value of reducing the variance may be larger for smaller sample sizes, when the results of the simple difference-in-means estimates are more likely to be inconclusive. In addition, the value of employing the auxiliary sample decreases as the sample size gets larger; in large studies, the relationship between the covariates and outcomes can be well-estimated without relying on the auxiliary sample.

Figure 1: Estimated Effect Distributions When N = 40



Note: This graph shows the results of 100 simulations in a context when 10 schools out of 40 schools are treated. For each simulation, we estimated the effect using three approaches described in Section V. The results of the simulations are shown in violin plots, in which the median estimate is shown as a white dot, the box indicates the 25^{th} and 75^{th} percentile-range, and the area shows the kernel density of the estimates.

Figure 2: Effect Standard Error for Different Study Sizes



Note: This graph shows the results of 100 simulations for a range of sample sizes; regardless of the sample size, one-quarter of the schools were randomly chosen to be treated. For each simulation, we estimated the effect using three approaches described in Section V.

VI Conclusion

In this paper, we study how machine learning methods can be used to improve the precision of small-scale RCTs. To start, we show that under common randomization schemes it is possible to control for covariates in a way that does not bias the resulting effect estimate, even in finite samples. We use this result, along with our examination of the variance of the estimate, to propose a new way to include covariates in the analysis of an RCT. The degree to which employing these methods will increase the precision of the estimates will depend on the context, and in particular how well the available covariates can explain variation in the outcome of interest. In our simulation, the new approach reduces the variance of the treatment effect estimate by over 50%, relative to a traditional comparison of the treatment and control means, without adding bias.

We also discuss how the result that the treatment effect estimates can be unbiased regardless of how one controls for the covariates hints at a way researchers can use data on observations outside of the RCT to help improve the precision of the RCT itself; these observations can be used to help model the relationship between the outcome and covariates. When we implement this in our simulation, we find that this approach reduces the variance of the estimates by an additional 20% in the smallest experiment we simulate, relative to the approach that uses machine learning but not the auxiliary sample.

These results highlight an important point: while most of the discussion surrounding machine-learning methods tends to focus on "big data" contexts, many of the key concepts such as sample-splitting and coefficient penalization are particularly relevant when the sample size is small. This paper illustrates one way to apply these insights to the analysis of RCTs, leading to more precise and still unbiased estimates of the treatment effect.

References

Angrist, Joshua, Peter Hull, Parag A. Pathak, and Christopher R. Walters, "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, 2016.

Aronow, Peter M. and Joel A. Middleton, "A Class of Unbiased Estimators of the Average Treatment Effect in Randomzied Experiments," *Journal of Causal Inference*, June 2013, *1* (1), 135–154.

Athey, S. and G.W. Imbens, "Chapter 3 - The Econometrics of Randomized Experiments," in Abhijit Vinayak Banerjee and Esther Duflo, eds., *Handbook of Field Experiments*, Vol. 1 of *Handbook of Economic Field Experiments*, North-Holland, 2017, pp. 73 – 140.

Berk, Richard, Emil Pitkin, Lawrence Brown, Andreas Buja, Edward George, and Linda Zhao, "Covariance Adjustments for the Analysis of Randomized Field Experiments," *Evaluation Review*, 2013, *37*, 170–196.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S. Sekhon, and Bin Yu, "Lasso adjustments of treatment effect estimates in randomized experiments," *Proceedings of the National Academy of Sciences*, 2016, *113* (27), 7383–7390.

Breiman, Leo, "Random Forests," *Machine Learning*, 2001, *45* (5-32).

Caughey, Devin, Allan Dafoe, and Luke Miratrix, "Beyond the Sharp Null: Randomization Inference, Bounded Null Hypothesis, and Confidence Intervals for Maximum Effects," *arXiv:1709.07339*, 2017.

Ding, Peng, Avi Feller, and Luke Miratrix, "Randomization Inference for Treatment Effect Variation," *arXiv:1412.5000*, 2014.

Eskreis-Winkler, Lauren, Katherine L. Milkman, Dena M. Gromet, and Angela L. Duckworth, "A large-scale field experiment shows giving advice improves academic outcomes for the advisor," *Proceedings of the National Academy of Sciences*, 2019, *116* (30), 14808–14810.

Freedman, David A., "On regression adjustments in experiments with several treatments," *Annals of Applied Statistics*, 2008, *2* (1), 176–196.

_ , "On regression adjustments to experimental data.," *Advanced in Applied Mathematics*, 2008, *40*, 180–193.

Hastie, Trevor, Robert J. Tibshirani, and Jersome Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Science and Business Media, 2009.

Hill, Jennifer L., "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 2011, *20* (1), 217–240.

Kaizar, Eloise E., "Incorporating Both Randomized and Observational Data into a Single Analysis," *Annual Review of Statistics and Its Application*, 2015, *2*, 49–72.

Lei, Lihua and Peng Ding, "Regression adjustment in completely randomized experiments with a diverging number of covariates," *Working Paper*, 2019.

Lin, Winston, "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *Annals of Applied Statistics*, 2013, *7* (1), 295–318.

Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence R. Katz, Ronald Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu, "Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults," *Science*, 2012, *337* (6101), 1505–1510.

_ , _ , _ , _ , _ , _ , and _ , *Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults From All Five Sites of the Moving to Opportunity Experiment, 2008-2010 [Public Use Data]* Inter-university Consortium for Political and Social Research [distributor] 2013-03-14.

Pitkin, Emil, Richard Berk, Larry Brown, Andreas Buja, Ed George, Kai Zhang, and Linda Zhao, "Improved Precision in Estimating Average Treatment Effects," *arXiv:1311.0291*, 2013.

Tibshirani, Robert, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, *58* (1), 267–288.

Tsiatis, Anastasios A., *Semiparametric Theory and Missing Data* Springer Series in Statistics, Springer, 2006.

Wager, Stefan, Wenfei Du, Jonathan Taylor, and Robert J. Tibshirani, "High-dimensional regression adjustments in randomized experiments," *Proceedings of the National Academy of Sciences*, 2016, *113* (45), 12673–12678.

Wu, Edward and Johann A. Gagnon-Bartsch, "The LOOP Estimator: Adjusting for Covariates in Randomized Experiments," *Evaluation Review*, 2018, *42* (4), 458–488.

Zhang, Min, Anastasios A. Tsiatis, and Marie Davidian, "Improving efficiency of inferences in randomized clinical trials using auxiliary covariates," *Biometrics*, 2008, *64* (3), 707–715.

Zou, Hui and Trevor Hastie, "Regularization and variable selection via the Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005, *67*, 201–320.

A Appendix

Variance Derivation

To derive the variance, we substitute the $\epsilon_i = \hat{g}_i(X_i) - (\mu_i + (1-p_i)\tau_i)$ and the fact that $Y_i = \mu_i + \tau_i T_i$ into Equation (1) to get that the estimator is equal to:

$$\begin{aligned}\hat{\tau} &= \frac{1}{N}\sum_{\forall i}\frac{\mu_i + \tau_i T_i - \big(\mu_i + (1-p_i)\tau_i + \epsilon_i\big)}{p_i}T_i \\ &-\frac{1}{N}\sum_{\forall i}\frac{\mu_i + \tau_i T_i - \big(\mu_i + (1-p_i)\tau_i + \epsilon_i\big)}{1-p_i}(1-T_i)\end{aligned}$$

Some algebra reduces this expression to:

$$\hat{\tau} = \frac{1}{N}\sum_{\forall i}\tau_i - \frac{1}{N}\sum_{\forall i}\left[\frac{\epsilon_i}{p_i}T_i - \frac{\epsilon_i}{1-p_i}(1-T_i)\right] \tag{14}$$

Thus, the variance of the estimate corresponds to:

$$\mathbb{V}\Big(\hat{\tau}\Big) = \mathbb{V}\left(\frac{1}{N}\sum_{\forall i}\left[\frac{\epsilon_i}{p_i}T_i - \frac{\epsilon_i}{1-p_i}(1-T_i)\right]\right) \tag{15}$$

Proof of Theorem 1

Proof. From Equation (4), we get that:

$$\mathbb{B}\Big(\hat{\tau}\Big) = -\frac{1}{N}\sum_{\forall i}\mathbb{E}\Big[\hat{g}_i(X_i)\cdot\Big(\frac{T_i}{p_i} - \frac{(1-T_i)}{1-p_i}\Big)\Big] \tag{16}$$

Using the law of iterated expectations, we get that:

$$\mathbb{B}\Big(\hat{\tau}\Big) = -\frac{1}{N}\sum_{\forall i}\mathbb{E}\Big[\mathbb{E}\big[\hat{g}_i(X_i)|T_i\big]\cdot\Big(\frac{T_i}{p_i} - \frac{(1-T_i)}{1-p_i}\Big)\Big] \tag{17}$$

□

Under the assumption that $\mathbb{E}\big[\hat{g}_i(X_i)|T_i = 1\big] = \mathbb{E}\big[\hat{g}_i(X_i)|T_i = 0\big]$, we then get that

$\mathbb{E}\big[\hat{g}_i(X_i)|T_i\big] = \mathbb{E}\big[\hat{g}_i(X_i)\big]$ and can write that:

$$\mathbb{B}\Big(\hat{\tau}\Big) = -\frac{1}{N}\sum_{\forall i}\mathbb{E}\big[\hat{g}_i(X_i)\big]\cdot\mathbb{E}\Big[\Big(\frac{T_i}{p_i} - \frac{(1-T_i)}{1-p_i}\Big)\Big] \tag{18}$$

which equals zero, since $\mathbb{E}\Big[\Big(\frac{T_i}{p_i} - \frac{(1-T_i)}{1-p_i}\Big)\Big] = 0$ for all i.

Proof of Theorem 2

Proof. From above, if we can show that:

$$\mathbb{E}\left[\hat{g}_i(X)\cdot\left(\frac{T_i}{p_i} - \frac{1-T_i}{1-p_i}\right)\right] = 0 \tag{19}$$

for all i, then it follows that the estimator defined in Equation (1) is unbiased. To prove this, we will denote $T^{(-k)}$ to be the vector of treatment assignments for all observations not in subgroup K.

We then note that

$$\mathbb{E}\left[\hat{g}_i(X)\cdot\left(\frac{T_i}{p_i} - \frac{1-T_i}{1-p_i}\right)\middle|T^{(-k)}\right] = \hat{g}_i(X)\cdot\mathbb{E}\left[\left(\frac{T_i}{p_i} - \frac{1-T_i}{1-p_i}\right)\middle|T^{(-k)}\right] \tag{20}$$

since conditioning on treatment assignments for those outside of k is equivalent to conditioning on outcomes for those outside of k, at which point $\hat{g}_i(X)$ is fixed and can be pulled out of the expectation.[4]

The assumption regarding the choice of subgroup K's then ensures that $\mathbb{E}\left[\frac{T_i}{p_i}\middle|T^{(-k)}\right] = 1 = \mathbb{E}\left[\frac{1-T_i}{1-p_i}\middle|T^{(-k)}\right]$ and so $\mathbb{E}\left[\left(\frac{T_i}{p_i} - \frac{1-T_i}{1-p_i}\right)\middle|T^{(-k)}\right] = 0$. Since this is true for generic i, it is true for all observations and therefore the sum of the expectations over all observations is also zero. □

Lemma 1. *Suppose that treatment assignment is done by drawing a fixed number of observations, which become the treated observations, from the set of all observations in the study. Then if the number of treated units is such that is possible to partition*

[4] This assumes that $\hat{g}_i(X)$ is a deterministic function; however, this assumption can be relaxed as long as the uncertainty in $\hat{g}_i(X)$ conditional on the outcomes is uncorrelated with the treatment assignment. This would be the case in estimators that incorporate some uncertainty into the estimates such as random forests.

the N observations into K subgroups with N_K observations, each of with has $p \cdot N_K$ treated observations and $(1-p) \cdot N_K$ control observations, it is possible to partition the sample such that $\mathbb{E}[T_i | T^{(-K)}] = p_i$ for all $i \in K$.

Proof. There are exactly $\binom{N_K}{p \cdot N_K}$ possible treatment assignments with assignments for all individuals not in K being $T^{(-K)}$ and, given the way the treatment assignment is done, we know that each one is equally likely. Of these, there are $\binom{N_K - 1}{p \cdot N_K - 1}$ in which individual i is chosen to be treated and $\binom{N_K - 1}{p \cdot N_K}$ in which individual i is not chosen to be treated. Thus, we have that:

$$\mathbb{E}[T_i | T^{(-K)}] = \frac{\binom{N_K - 1}{p \cdot N_K - 1}}{\binom{N_K}{p \cdot N_K}} = \frac{p \cdot N_K}{N_K} = p$$

□

Inference Approximations

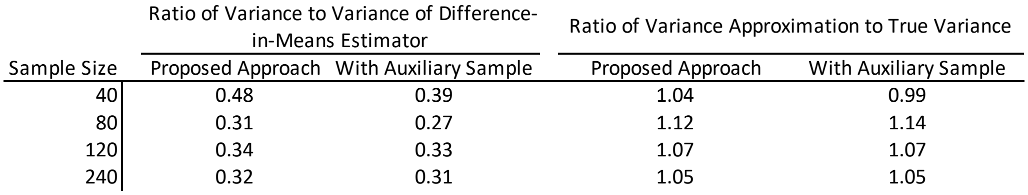
As a shortcut, we might be able to use Equation (5) as a guide. The challenge is that we cannot calculate $\epsilon_i \equiv \hat{g}_i(X_i) - \big(p_i Y_i(0) + (1-p_i) Y_i(1)\big)$, since we only observe one of $Y_i(0)$ or $Y_i(1)$. We can, however, either calculate $\epsilon_{i,0} \equiv \hat{g}_{i,0}(X_i) - Y_i(0)$ or $\epsilon_{i,1} \equiv \hat{g}_{i,1}(X_i) - Y_i(1)$, depending on whether i is treated or not. A feasible approach is therefore to simply use $\epsilon_{i,0}$ or $\epsilon_{i,1}$ in place of ϵ_i, roughly assuming that the ability to predict two potential outcomes for individual i is similar. This also implicitly assumes that the randomization does not have a large impact on $\hat{g}_{i,1}(X_i)$ or $\hat{g}_{i,0}(X_i)$, since they are considered fixed in the approach. Further assuming that the treatment assignment is sufficiently independent, we could then approximate the variance of the estimate using the following equation:

$$\hat{V} = \frac{1}{N}\left(\frac{1}{N}\sum \frac{\hat{\epsilon}_i^2}{p_i(1-p_i)}\right) \tag{21}$$

where $\hat{\epsilon}_i = \epsilon_{i,1}$ if treated and $\epsilon_{i,0}$ if control. Despite the many assumptions implicit in this approximation, in the simulation discussed below it does a good job reflecting the true variance of the estimated effect.

To see how well this approach works, Table 1 compares the standard error estimates given by Equation 21 to the true standard error of the estimates as calculated over the 100 simulations. Despite the numerous assumptions implicit in Equation 21, in this simulation the standard error estimates tend to be quite close to the true standard errors regardless of the sample size. The estimates tend to be slightly conservative, overestimating the standard error but never by more than 15%.

Table 1: Variance Reduction and Approximation for Different Study Sizes



	Ratio of Variance to Variance of Difference-in-Means Estimator		Ratio of Variance Approximation to True Variance	
Sample Size	Proposed Approach	With Auxiliary Sample	Proposed Approach	With Auxiliary Sample
40	0.48	0.39	1.04	0.99
80	0.31	0.27	1.12	1.14
120	0.34	0.33	1.07	1.07
240	0.32	0.31	1.05	1.05

Note: The left two columns of the table shows how the proposed method, with and without an auxiliary sample, impacts the variance of the estimated effects as the sample size grows from 40 observations to 240 observations. For each sample size, we conduct 100 simulations in which one-quarter of the sample is randomly chosen to be treated schools. The variance is calculated as the variance of the estimated effect over the 100 simulations and the number reported is the ratio of the variance when using the proposed method to the variance when simply comparing the treatment and control averages. The right two columns compare this approximated variance, using the formula discussed in Section III to the variance of the estimates.

Employing an Auxiliary Sample

The theoretical analysis suggests that using observations outside the study to model the relationship between the covariates and outcomes may help improve the precision of the estimates. We now provide more details on how this can be done in practice, using the simulation described in Section V as the example. To do so, we define $\tilde{X}_i$ as the covariate vector X_i plus an dummy variable S_i the indicates whether observation i was part of RCT study or in the auxiliary sample and, in the case of a linear model, any interactions between the sample indicator and the covariates. We furthermore create an additional fold and assign all of the data in the auxiliary sample into the $K+1^{th}$ fold. Since none of the auxiliary sample receives the treatment, we finally set $T_i = 0$ for all of the auxiliary sample.[5]

[5] While we believe it is usually the case that none of the observations in the auxiliary sample will have received the treatment, the method allows for some or all of the auxiliary sample to have

With this set-up, we estimate $\hat{g}(\tilde{X}_i)$ using Equations (10 -12) using both the data in the RCT study and the data in the auxiliary sample. We then use only the observations that were part of the RCT study to estimate the treatment effect using Equation (1).

By including both both S_i and the interactions between S_i and X_i, we allow for the possibility of non-random selection into the study. In fact, if $\hat{g}(\tilde{X}_i)$ is estimated as a linear function of $\tilde{X}_i$ and no penalization term is included, the auxiliary sample is essentially used only to estimate selection into the study. Thus, the auxiliary sample would have no impact on the value of $\hat{g}(\tilde{X}_i)$ for the observations in the RCT study and therefore no impact on the estimated treated effect. However, when a penalization term is included or a when a non-linear function is estimated, the model does use the auxiliary sample to estimate the relationship between the covariates and the outcome, unless that relationship is sufficiently different in the auxiliary sample than in the RCT sample to warrant including the interaction term in the estimated model.

Additional Simulations

We conduct two additional simulations, using data from two relatively large RCTs. In the first study, Eskreis-Winkler et. al. (Eskreis-Winkler et al., 2019), about half of 2,000 high school students were randomly chosen to give advice to middle school students. They then showed that giving advice increased students' grades is the quarter after they gave advice. The second study, Ludwig et. al. (Ludwig et al., 2012), is an analysis of the Moving to Opportunity (MTO). Here, roughly two-thirds of 4,000 households were given a voucher to move from a poor neighborhood to a rich neighborhood; the Ludwig et. al. (Ludwig et al., 2012) paper we use shows that moving increased individuals' subjective well-being 10-15 years after being given the voucher.

For both studies, we conduct the simulation as follows. We first randomly sample a set of observations to be in the hypothetical study, stratifying by treatment status to ensure that the same fraction of individuals are treated in our sample as in the full study. We also stratify by school for the Eskreis-Winkler et. al. study and by city in the Ludwig et. al. study, to ensure that the distribution across sites is

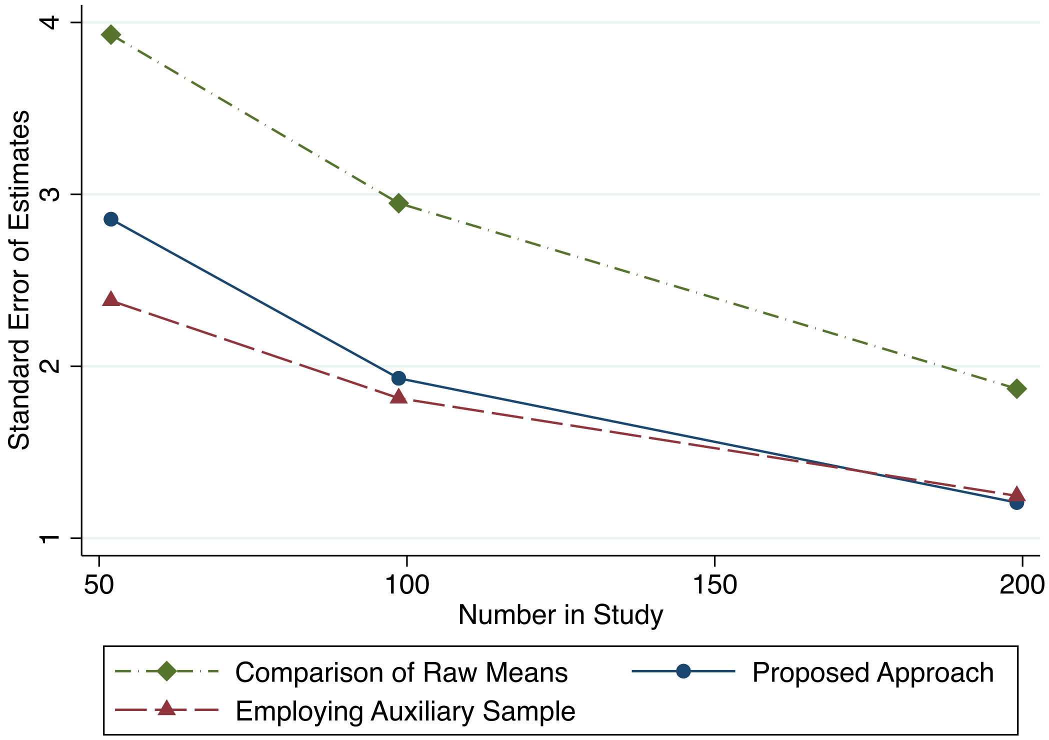
$T_i = 1$.

the same in our sample as in the overall study. We then assume that researchers observe all the observations randomly sampled to be in the hypothetical study as well as the rest of control observations not randomly sample; these additional control observations are what we consider to be the auxiliary sample. We then estimate the effect for hypothetical study by: a) taking the mean-difference between the treatment average and control average; b) implementing the approach developed in this paper using only observations in the hypothetical study; and c) implementing the approach developed in this paper using both observations in the hypothetical study and in the auxiliary sample. As discussed above, we only use observations in the auxiliary sample to estimate $\hat{g}(X_i)$ and in all cases only use observations in the hypothetical study to estimate the treatment effect using Equation (1). When conducting the residualization, we use as covariates all of the exogenous variables the researchers include in their publicly available data.[6]

We repeat this simulation 100 times for three different study sizes. A downside of these simulations is that we do not know the "true" average treatment effect; however, as we prove in Theorem 2 all three approaches we use are unbiased and so we can judge the estimators based on the standard deviation of the estimated treated effect. The results, illustrated in Figures 3 and 4, show a similar pattern as the simulation we discuss in Section V. The proposed approach significantly decreases standard deviation of the estimated effect relative to a simple mean-difference, and this reduction holds regardless of the sample size of the study. Furthermore, employing the auxiliary sample to improve residualization further reduces the standard errors, especially when the sample size is quite small.

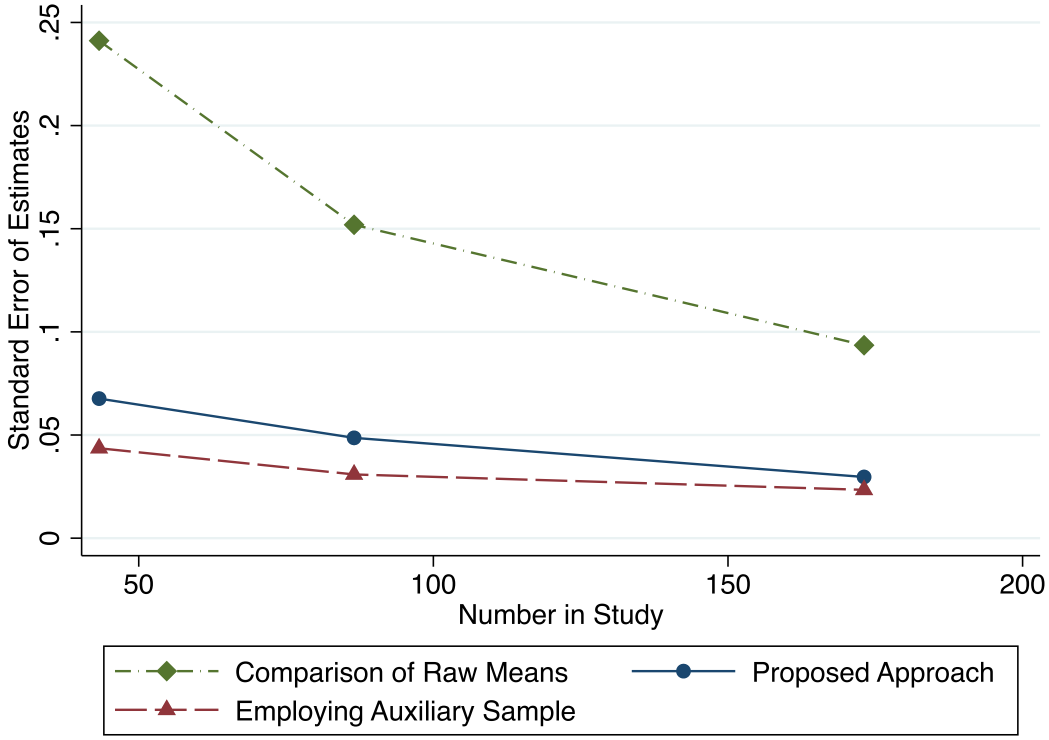
[6] The publicly available data for Ludwig et. al. is synthetic data, modified slightly to ensure confidentiality. Using these as covariates is thus likely a lower-bound on the benefit of using our method on the original individual-level data. See (Ludwig et al., 2013-03-14) for more information on how the synthetic data was created.

Figure 3: Eskreis-Winkler et. al. Simulation



Note: This graph shows the results of 100 simulations for a range of sample sizes; regardless of the sample size, one-quarter of the schools were randomly chosen to be treated. For each simulation, we estimated the effect using three approaches described in Section V.

Figure 4: Ludwig et. al. Simulation



Note: This graph shows the results of 100 simulations for a range of sample sizes; regardless of the sample size, one-quarter of the schools were randomly chosen to be treated. For each simulation, we estimated the effect using three approaches described in Section V.