Award:

R305D150040 - PRESIDENT AND FELLOWS OF HARVARD COLLEGE

Paper Title:

Insights on variance estimation for blocked matched pairs designs.

Authors:

Nicole E. Pashley and Luke Miratrix

Publication Date:

Jul 3, 2019 (arXiv)

Insights on Variance Estimation for Blocked and Matched Pairs Designs[*]

Nicole E. Pashley
Department of Statistics, Harvard University

Luke W. Miratrix
Graduate School of Education, Harvard University

July 2, 2019

Abstract

In the causal inference literature, evaluating blocking from a potential outcomes perspective has two main branches of work. The first focuses on larger blocks, with multiple treatment and control units in each block. The second focuses on matched pairs, with a single treatment and control unit in each block. These literatures not only provide different estimators for the standard errors of the estimated average impact, but they are also built on different sets of assumptions. Additionally, neither literature handles cases with blocks of varying size that contain singleton treatment or control units, a case which can occur with different forms of matching or post-stratification. Differences in the two literatures have also created some confusion regarding the benefits of blocking in general. In this paper, we first reconcile the literatures by carefully examining the performance of different estimators of treatment effect and of associated variance estimators under several different frameworks. We then use these insights to derive novel variance estimators for experiments containing blocks of different sizes. We also assess in which situations blocking is not guaranteed to reduce precision.

Keywords: Causal inference; Potential outcomes; Precision; Finite sample inference; Randomization inference; Neymanian Inference

[*] *Email:* `npashley@g.harvard.edu`. The authors would like to thank Guillaume Basse, Avi Feller, Colin Fogarty, Michael Higgins, Luke Keele, and Lo-Hua Yuan for their comments and edits. We would also like to thank members of Luke Miratrix's and Donald B. Rubin's research labs for their useful feedback on the project and Peter Schochet and Kosuke Imai for insightful discussion of this material. Finally, we thank anonymous reviewers. The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D150040. This material is also based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences or the U.S. Department of Education.

1 Introduction

Beginning with Neyman and Fisher, there is a long literature of analyzing randomized experiments by focusing on the assignment mechanism rather than some generative model of the data. One major family of experimental designs in this literature is blocked randomized experiments, where units are grouped to hopefully create homogenous collections, and then treatment assignment is randomized within each group (see Fisher, 1926). Ideally, this process gives a higher precision estimate of the overall average treatment effect, as compared to a completely randomized design.

We follow the potential outcome causal literature, (as in Imbens and Rubin, 2015; Rosenbaum, 2010), as opposed to the experimental design literature (as in Cochran and Cox, 1950; Wu and Hamada, 2000). Much of the prior work on randomized experiments within the potential outcomes framework has focused on two forms of blocking: blocking where there are several treated and control units in each block and blocking where there is exactly one treated and one control unit in each block (matched pairs). See, for example Imai et al. (2008) or Imbens (2011) for treatments of large blocks and Abadie and Imbens (2008) or Imai (2008) for treatments of matched pairs. The literature has not, however, treated the cases where researchers have generated groups of varying size but where there is still only one treated and/or one control in some of the blocks, which we call the "hybrid design." For instance, even more recent textbooks such as *Field Experiments: Design, Analysis and Interpretation* (Gerber and Green, 2012) and *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Imbens and Rubin, 2015) do not propose a clear answer for Neyman-style variance estimation in this case. While obtaining a point estimate for the overall average treatment effect is straightforward in this context, assessing the uncertainty of such an estimate is not. We believe this gap likely exists because much of the current work uses different frameworks tailored to the specific special cases of large blocks or matched pairs. We therefore build on prior work to solve this more general hybrid design, providing tools for conducting a Neyman-style analysis. Currently one would instead have to turn to Fisher-style permutation tests, which typically rely on constant treatment effect assumptions, or regression-based approaches, which can be biased and often require making assumptions as to the residual error structure.

This gap is important as hybrid experiments with blocks of different sizes, and different numbers of treated and control units within the blocks, can easily arise in many modern social science experiments. For example, multisite trials in education could easily have several sites (e.g., districts) with only a few schools in each site. These designs are also common in many matching methods used in observational studies. For instance, using Coarsened Exact Matching (CEM) (Iacus et al., 2012) on an observational study may lead to many variable-sized blocks, some of which have singleton treatment or control units. "Full matching," which identifies collections of units that are similar on some baseline covariates (see Hansen (2004), Rosenbaum (1991)), creates variable-sized blocks with either one treated or one control unit. Our approach allows for a Neyman-style analysis that takes full advantage of the matching itself. See Section 7 for more on these applications.

Due to the focus on the hybrid case, and consequent focus on choice of sampling framework and block types, we also resolve some apparent contradictions found in the literature on the classical blocked and matched pairs designs. In particular, we derive expressions of blocking vs. complete randomization for multiple frameworks and show which frameworks give guarantees on the benefits of blocking. We correctly separate out sampling variance from assignment variance, giving more precise statements of the benefits of blocking than given in prior literature. We also clarify how these comparisons depend on the block types and sampling mechanism, not the block sizes. We provide novel methods for estimating variance in the hybrid case, and analyze the performance of uncertainty estimation in multiple frameworks for all cases.

Recent work by Fogarty (2018) has addressed some of these issues. In particular, Fogarty presents a method for estimating variance with small blocks of variable size, not just matched pairs. He also makes explicit the issue of differing results under different population and sampling frameworks by comparing multiple settings. However, he does not directly address the issue of creating a cohesive hybrid estimator for experiments with large and small blocks. In our paper, we tackle this issue and also discuss the comparison of blocking and complete randomization. Additionally, we do not focus on the use of covariates to model treatment heterogeneity. As far as we know, this is the first work that combines the matched pairs and blocked designs to derive more general results for the hybrid design. The approach to causal inference used in this work has strong connections to the survey sampling literature, as treated in, e.g., Särndal et al. (2003) or Cochran (1977). This is particularly true when considering the infinite population frameworks, where the sampling of units (or clusters of units) from a larger population plays a role in the precision of the estimators. For connections to the core finite sample context see Lin (2013).

We start with the finite sample framework because it acts as a building block for infinite population frameworks, as we can obtain expectations and variances in the infinite population by first conditioning on the sample. In Section 2 we set out our notation and discuss complete randomization and blocking. We then propose methods for estimating uncertainty in the case of large blocks, small blocks, and the hybrid of the two, and derive their bias under the finite sample framework in Section 3. We examine several infinite population frameworks, evaluating the performance of the variance estimators in each in Section 4. In Section 5 we systematically compare blocking to complete randomization under the full range of frameworks and discuss how the findings differ. Finally, in Section 6, we illustrate estimator performance with some finite sample simulation studies (although the findings are broadly applicable to the other sampling regimes) and, in Section 7, illustrate estimation in two data examples. Throughout, we provide various formulae for the performance, or relative performance, of the various estimators in terms of bias or variance for the different frameworks considered. For clarity in presentation, we have moved the derivations of these formulae to the Supplementary Material.

2 Overall setup and notation

We use the Neyman-Rubin model of potential outcomes (Rubin, 1974; Splawa-Neyman et al., 1923/1990). We assume the Stable Unit Treatment Value Assumption of no differential forms of treatment and no interference between units (Rubin, 1980). We will discuss both completely randomized and block randomized experiments. Let there be n units in the sample. In a completely randomized experiment, the entire collection of the units in the sample is divided into a treatment group and a control group by taking a simple random sample of pn units as the treatment group and leaving the remainder as control. In a blocked randomized experiment, we have K blocks, formed based on some pretreatment covariate(s), in the sample, with n_k units in block k. Each block k is then treated as a mini-experiment, with a fixed number of $p_k n_k$ units being assigned to treatment and the rest to control, independently of the other blocks.

The sample average treatment effect (SATE) is the typical estimand in so-called finite sample inference, which takes our sample as fixed, leaving the assignment mechanism as the only source of randomness. Under blocking, the SATE within block k, for $k = 1, ..., K$, is

$$\tau_{k,\mathcal{S}} = \frac{1}{n_k} \sum_{i:b_i=k} \big(Y_i(t) - Y_i(c)\big),$$

where $Y_i(t)$ and $Y_i(c)$ are the potential outcomes for unit i under treatment and control, respectively, and where b_i indicates the block that unit i belongs to. The overall SATE (see Imbens and Rubin (2015), p. 86) is then

$$\tau_{\mathcal{S}} = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i(t) - Y_i(c)\big).$$

In this work, we consider two estimators for the SATE (and later the population average treatment effect), one typically used for complete randomization and one for blocked randomization. Define the variable Z_i as $Z_i = t$ if unit i is assigned treatment and $Z_i = c$ if unit i is assigned control, for $i = 1, ..., n$. Let $\mathbb{I}_{Z_i=t}$ be the indicator that unit i received treatment, n_t be the total number of treated units, and n_c be the total number of control units. So, $n_t = \sum_{i=1}^{n} \mathbb{I}_{Z_i=t}$, $n_c = n - n_t$. Similarly, let $n_{t,k}$, $n_{c,k}$ indicate these values within block k. Define $Y_i^{obs} = Y_i(Z_i)$ as the outcome we observe for unit i given a specific treatment Z_i. The blocked randomization estimator is then a weighted average of simple difference estimators for each block

$$\widehat{\tau}_{(BK)} = \sum_{k=1}^{K} \frac{n_k}{n} \widehat{\tau}_k,$$

with the

$$\widehat{\tau}_k = \frac{1}{n_{t,k}} \sum_{i:b_i=k} \mathbb{I}_{Z_i=t} Y_i(t) - \frac{1}{n_{c,k}} \sum_{i:b_i=k} (1 - \mathbb{I}_{Z_i=t}) Y_i(c),$$

$k = 1, ..., K$, being simple difference estimators within each block. The complete randomization estimator, $\widehat{\tau}_{(CR)}$, is

$$\widehat{\tau}_{(CR)} = \frac{1}{n_t} \sum_{i=1}^{n} \mathbb{I}_{Z_i=t} Y_i(t) - \frac{1}{n_c} \sum_{i=1}^{n} (1 - \mathbb{I}_{Z_i=t}) Y_i(c).$$

We will often take the expectation over the randomization of units to treatment for a fixed sample. In particular, $\mathbb{E}\left[\widehat{M}|\mathcal{S}, \mathbf{P}\right]$ is the expected value of some estimator $\widehat{M}$ for a given, fixed, finite sample $\mathcal{S}$ and for some assignment mechanism $\mathbf{P}$, which may be complete randomization or blocked randomization. To reduce clutter, we drop the $\mathbf{P}$ and simply write $\mathbb{E}\left[\widehat{M}|\mathcal{S}\right]$ if the estimator makes the assignment mechanism clear.

In general, our estimators are unbiased, with

$$\mathbb{E}\left[\widehat{\tau}_{(CR)}|\mathcal{S}\right] = \mathbb{E}\left[\widehat{\tau}_{(BK)}|\mathcal{S}\right] = \tau_{\mathcal{S}}.$$

It is assessing variance, and the precision gains of blocking, that is more tricky. This assessment is the goal of the paper, but first we need to introduce a few more useful concepts.

An important aspect of blocking is how the blocks are formed. Explicit articulation of block formation will be useful when we discuss asymptotic properties of our estimators and will also be used to differentiate the various population frameworks in Section 4. We identify three primary ways that blocks are formed:

(a) Fixed blocks: Occurs when the total number of blocks and the covariate distribution of blocks is fixed before looking at the sample covariates. E.g., blocking that occurs on a single categorical covariate.

(b) Flexible blocks: Occurs when the covariate distribution and total number of blocks may not be known before looking at the sample's covariates. E.g. if there are many covariates or continuous covariates and matching or discretizing is used to form blocks.

(c) Structural blocks: Occurs when units have some natural grouping such that the blocks are self-contained. The members of each block are fixed and if a block is represented in the sample, typically all members of that block are in the sample. E.g., twins or classrooms.

3 Variance estimation

The prior section outlined the different experiments and treatment effect estimators in the finite sample. We next discuss how to estimate the estimators' variances, which are integral to obtaining standard errors and confidence intervals. We discuss estimators built from a Neyman-Rubin randomization perspective. See Supplementary Material B for a discussion of alternative variance estimators (such as from linear models) that make additional assumptions on the data structure. We first investigate bias of variance estimators under a finite sample framework and extend to other frameworks in Section 4.

We start by giving the true variance in the finite sample for each of the designs. To do so, we need some additional notation. The mean of the potential outcomes for the units in the sample under treatment z is

$$\bar{Y}(z) = \frac{1}{n}\sum_{i=1}^{n} Y_i(z).$$

The sample variance of potential outcomes under treatment z is

$$S^2(z) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(z) - \bar{Y}(z))^2.$$

The sample variance of the individual level treatment effects is

$$S^2(tc) = \frac{1}{n-1}\sum_{i=1}^{n}\Big(Y_i(t) - Y_i(c) - \tau_{\mathcal{S}}\Big)^2.$$

$\bar{Y}_k(z)$, $S^2_k(z)$, and $S^2_k(tc)$ are defined analogously over the units in block k.

For the finite sample, the variances of $\widehat{\tau}_{(CR)}$ and $\widehat{\tau}_{(BK)}$ are well known (see Imbens and Rubin (2015)). For complete randomization,

$$\text{var}(\widehat{\tau}_{(CR)}|\mathcal{S}) = \frac{S^2(t)}{n_t} + \frac{S^2(c)}{n_c} - \frac{S^2(tc)}{n}. \tag{1}$$

Extending this to blocked randomization (see Imbens (2011)), the overall variance is

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) = \sum_{k=1}^{K}\frac{n_k^2}{n^2}\text{var}(\widehat{\tau}_k|\mathcal{S}) = \sum_{k=1}^{K}\frac{n_k^2}{n^2}\Big(\frac{S^2_k(t)}{n_{t,k}} + \frac{S^2_k(c)}{n_{c,k}} - \frac{S^2_k(tc)}{n_k}\Big). \tag{2}$$

For complete randomization we use the Neyman-style variance estimator of

$$\widehat{\sigma}^2_{(CR)} = \widehat{\text{var}}(\widehat{\tau}_{(CR)}) = \frac{s^2(c)}{n_c} + \frac{s^2(t)}{n_t}$$

with $s^2(z)$ the sample variance of the units which received treatment z.

For blocked experiments, the type of variance estimator one would use depends on the sizes of blocks one has. In cases where we have at least two treated and two control units in each block we can directly extend the completely randomized estimator strategy by using it within each block and weighting. Defining $s_k^2(z)$ analogously to $s^2(z)$ within block k, the typical variance estimator is

$$\widehat{\sigma}^2_{(BK)} = \widehat{\text{var}}(\widehat{\tau}_{(BK)}) = \sum_{k=1}^{K} \frac{n_k^2}{n^2} \left(\frac{s_k^2(c)}{n_{c,k}} + \frac{s_k^2(t)}{n_{t,k}} \right). \tag{3}$$

See Imbens (2011) for a more in depth discussion of blocking of this form. This is the "big block" style of blocking, and the "big block" estimator.

For the "small blocks" case, where our blocks have only one treated unit or one control unit, we need to use an alternative approach as we cannot estimate the variance for a treatment arm with a single unit. This approach is presented below. To give some background, the analytical problems that arise when estimating the variance in matched pairs experiments, especially when working in the finite sample framework, have been lamented by many statisticians (see, e.g., Imbens, 2011). The issues arise from the fact that there is no way to estimate the within pair variance with only one unit assigned to treatment and one unit assigned to control in each pair. Previous work has found conservative estimators, however, which we build on. For instance, Imai (2008) showed that the standard matched pairs estimator was biased in the finite sample setting and put bounds on the true variance. The RCT-Yes R package and documentation (Schochet, 2016) also provides a conservative variance estimator for the matched pairs design (as well as estimators for blocked designs); this is discussed more in Supplementary Material B.3.

For a hybrid experiment with both big and small blocks, we combine results to create an overall variance estimator.

3.1 Small block experiments with equal size blocks

When we have small blocks of the same size, we can directly use the usual variance estimator in the matched pairs literature (e.g., Imai, 2008) as a variance estimator for $\widehat{\tau}_{(BK)}$, no matter what the block sizes are, as also noted by Fogarty (2018). This gives a variance estimator of

$$\widehat{\sigma}^2_{(SMALL/s)} = \frac{1}{K(K-1)} \sum_{k=1}^{K} (\widehat{\tau}_k - \widehat{\tau}_{(BK)})^2. \tag{4}$$

This estimator directly estimates the variance of the overall block treatment effect estimator, rather than estimating the variance for each individual block and then weighting. We will see that, depending on the framework used, this estimator can give positively biased estimates if the true τ_k tends to differ across blocks.

3.2 Small block experiments with varying size blocks

For experiments with small blocks of varying sizes we offer two variance estimators. The first directly extends the standard matched pairs estimator by grouping the blocks by size into J groups and using Equation 4 for each group. We then weight and combine to get an overall variance estimator.

Stratified Small Block Variance Estimator:

$$\widehat{\sigma}^2_{(SMALL/m)} = \frac{1}{\left(\sum_{j=1}^{J} m_j K_j\right)^2} \sum_{j=1}^{J} (m_j K_j)^2 \widehat{\sigma}^2_{(SMALL),j}, \tag{5}$$

where K_j is the number of blocks of size m_j and

$$\widehat{\sigma}^2_{(SMALL),j} = \frac{1}{K_j(K_j-1)} \sum_{k:n_k=m_j} (\widehat{\tau}_k - \widehat{\tau}_{(SMALL),j})^2 \tag{6}$$

with $\widehat{\tau}_{(SMALL),j} = \sum_{k:n_k=m_j} \widehat{\tau}_k / K_j$. That is, grouping by the same size allows for using the equal size block estimator above. While straightforward, this is not ideal because it requires at least two blocks of each size in the overall experiment to estimate each $\widehat{\sigma}^2_{(SMALL),j}$. See Supplementary Material D.1 for more details on this approach.

The second approach allows the variance of all of the small blocks to be estimated at the same time, without requiring multiple blocks of the same size.

Unified Small Block Variance Estimator:

$$\widehat{\sigma}^2_{(SMALL/p)} = \sum_{k=1}^{K} \frac{n_k^2}{(n-2n_k)(n + \sum_{i=1}^{K} \frac{n_i^2}{n-2n_i})} (\widehat{\tau}_k - \widehat{\tau}_{(BK)})^2. \tag{7}$$

For $\widehat{\sigma}^2_{(SMALL/p)}$ to be defined and guaranteed conservative, no one block can make up half or more of the units. We derived this estimator using the basic form of the matched pairs variance estimator as a weighted sum of the squared differences between the estimated average block treatment effects and the estimated overall average treatment effect. The weights then come from a simple optimization (see Supplementary Material E), and partially account for the different blocks having different levels of precision when estimating the variance of the block-level impacts. This estimator has similar finite sample properties to the standard estimator for blocks of the same size (Equation (4)). In particular, it is also conservative and unbiased when the block average treatment effects are all the same.

3.3 Hybrid experiments

When doing variance estimation in a hybrid blocked design, we can split the blocks up into small blocks and big blocks. Grouping the big and small blocks together allows us to write the causal effect estimand as a combination of two estimands for our two different types of block sizes. Let there be n_{sb} total units in small blocks in the sample. Then

$$\tau_{\mathcal{S}} = \frac{n - n_{sb}}{n}\tau_{(BIG),\mathcal{S}} + \frac{n_{sb}}{n}\tau_{(SMALL),\mathcal{S}}$$

where

$$\tau_{(BIG),\mathcal{S}} = \frac{1}{n - n_{sb}} \sum_{k: n_{t,k} \geq 2, n_{c,k} \geq 2} n_k \tau_k \quad \text{and} \quad \tau_{(SMALL),\mathcal{S}} = \frac{1}{n_{sb}} \sum_{k: n_{t,k} = 1 \text{ or } n_{c,k} = 1} n_k \tau_k.$$

The estimator for the overall treatment effect can also be written as

$$\widehat{\tau}_{(BK)} = \frac{n - n_{sb}}{n}\widehat{\tau}_{(BIG)} + \frac{n_{sb}}{n}\widehat{\tau}_{(SMALL)}.$$

For finite sample inference, we can similarly break down the variance, and estimator of the variance, of $\widehat{\tau}_{(BK)}$ because the block estimators are independent due to the block randomized treatment assignment.

Hybrid Variance Estimator:

$$\widehat{\text{var}}\left(\widehat{\tau}_{(BK)}\right) = \frac{(n - n_{sb})^2}{n^2}\widehat{\text{var}}\left(\widehat{\tau}_{(BIG)}\right) + \frac{n_{sb}^2}{n^2}\widehat{\text{var}}\left(\widehat{\tau}_{(SMALL)}\right).$$

Thus, when we have small blocks, we can estimate the variance for those small blocks separately and use the usual blocking estimator on the larger blocks. Alternatively, one could use $\widehat{\sigma}^2_{(SMALL/m)}$ or $\widehat{\sigma}^2_{(SMALL/p)}$ for all blocks, but we do not recommend this.

3.4 Finite sample bias of the variance estimators

In the finite setting all of the above estimators are conservative, and are only unbiased in specific circumstances. First, $\widehat{\sigma}^2_{(CR)}$ is known (Imbens and Rubin, 2015, p. 92; Splawa-

Neyman et al., 1923/1990) to have bias

$$\mathbb{E}\left[\widehat{\sigma}^2_{(CR)}|\mathcal{S}\right] - \text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) = \frac{S^2(tc)}{n}.$$

If all of the blocks have at least two treated and two control units, we can extend this result to $\widehat{\sigma}^2_{(BK)}$, which has bias

$$\mathbb{E}\left[\widehat{\sigma}^2_{(BK)}|\mathcal{S}\right] - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) = \sum_{k=1}^{K} \frac{n_k}{n^2} S_k^2(tc).$$

For the small blocks of varying sizes, we have two corollaries. See Supplementary Material D.2.1 and E for proofs. The first is

Corollary 3.4.1. *The bias of $\widehat{\sigma}^2_{(SMALL/m)}$ under the finite framework is*

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/m)}|\mathcal{S}\right] - var\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right) = \sum_{j=1}^{J} \frac{K_j m_j^2}{n_{sb}^2(K_j - 1)} \sum_{k:n_k=m_j} \left(\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S},j}\right)^2.$$

The above extends prior results for $\widehat{\sigma}^2_{(SMALL/s)}$ for matched pairs (see Imai (2008), Imbens and Rubin (2015), p. 227, or, for a more general case, Fogarty (2018)). $\widehat{\sigma}^2_{(SMALL/m)}$ is conservative and unbiased when the average treatment effect is the same for all blocks of the same size (similar to the unbiased result from Imai (2008) for $\widehat{\sigma}^2_{(SMALL/s)}$).

For the second estimator we have

Corollary 3.4.2. *The bias of $\widehat{\sigma}^2_{(SMALL/p)}$ under the finite framework is*

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/p)}|\mathcal{S}\right] &- var\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right)\\ &= \sum_{k=1}^{K} \frac{n_k^2}{(n_{sb} - 2n_k)(n_{sb} + \sum_{i=1}^{K} \frac{n_i^2}{n_{sb}-2n_i})} (\tau_{k,\mathcal{S}} - \tau_{(SMALL),\mathcal{S}})^2.\end{aligned}$$

If the average treatment effect is the same across all small blocks then this estimator is unbiased, and if there is heterogeneity, it is conservative.

Remark. Both small block estimators are conservative, which raises the question of whether one is superior. The constant in front of each term of the bias of both estimators is of order n_k^2/n^2. Then we expect the bias of $\widehat{\sigma}^2_{(SMALL/m)}$ to be less than the bias of $\widehat{\sigma}^2_{(SMALL/p)}$ when the treatment effects of blocks of similar sizes are similar because the variance of impacts within blocks of a given size will be smaller than across all of the blocks. However, $\widehat{\sigma}^2_{(SMALL/m)}$ has the drawback that it can only be used when we have at least two blocks of each small size.

The improved potential performance of the first estimator when there is homogeneity within block sizes does suggest that we could group blocks in some other way if we had prior knowledge of which blocks were most similar. The first estimator relies on the blocks being equal size so the weights factor out of the sum to give the expression for the cross-block estimate of variation. But we could first subdivide our blocks based on some similarity measure and apply the second estimator to each group, combining the parts with the hybrid weighting approach. This could make the estimator less conservative while maintaining its validity.

It is also worth considering the extent of conservatism of the estimators. For the case where all blocks are the same size, when we have blocks with m control units and 1 treated unit, as m increases the variance of the treatment effect estimator will decrease, as we are getting a more precise estimate for the control units. However, the form of the bias of $\widehat{\sigma}^2_{(SMALL/s)}$ remains the same. Therefore, with large m the bias of $\widehat{\sigma}^2_{(SMALL/s)}$ due to treatment heterogeneity becomes larger relative to the true variance. This intuition extends to the variable size case as well. In these cases alternative variance estimation strategies, such as discussed in Supplementary Material B, may become more appealing.

It is important to note that the type of blocks will impact whether the bias of these estimators go to zero as sample size increases. For instance, one might argue for the use of $\widehat{\sigma}^2_{(SMALL/p)}$ instead of $\widehat{\sigma}^2_{(BK)}$ even if we have big blocks, because the condition for unbiasedness for $\widehat{\sigma}^2_{(SMALL/p)}$ (that all blocks have the same average treatment effect) could be considered less stringent than for $\widehat{\sigma}^2_{(BK)}$ (that there is zero treatment variation within each block). However, with fixed blocks, the number of units within each block increases as sample size increases and the bias of $\widehat{\sigma}^2_{(BK)}$ will go to zero, the standard result, but the bias of $\widehat{\sigma}^2_{(SMALL/p)}$ will not unless all of the blocks have the same average treatment effect. In this case, as the blocks grow to be big, we would use $\widehat{\sigma}^2_{(BK)}$.

In the hybrid setting the overall bias will be a weighted sum of the biases for the big and small block components. Therefore, because the overall weighting depends on the block sizes, having a poor estimator for the small blocks may not have a large effect on the overall bias if small blocks make up only a small proportion of the sample.

There is no way to unbiasedly estimate variance within small blocks without additional structure or covariates. If we think that the treatment effects of different strata are not too far apart, then we suggest using one of the previous estimators. We at least know that the bias incurred is positive. However, if we have reason to believe that the treatment effects of different strata will be very far apart, a plug-in estimator, as discussed in Supplementary Material B, may be more appropriate.

4 Infinite Population Frameworks

Up to this point we have examined blocking in a finite sample framework, conditioning on the units in the experiment in question. In the literature, however, blocking has often been examined under a variety of infinite population frameworks. In particular, the matched pairs literature uses a framework where the blocks themselves are sampled from an infinite population of blocks, whereas the big block literature typically assumes stratified random sampling from a finite number of infinite size strata. Using different population frameworks will give different answers to important questions of what the true variance of the treatment effect estimate is and what the bias of our variance estimators are. In this section, we first discuss the literature related to variance estimation for infinite populations, identifying the apparent tensions that exist. We then systematically discuss different frameworks, deriving the true variance of the treatment effect estimators under each of them. We also evaluate the bias of the variance estimators introduced in Section 3. We focus on infinite superpopulations; finite superpopulations substantially larger than the sample would give similar results. We explore work pertaining to the use of linear models, such as Cochran (1953) and Lin (2013), in Supplementary Material B.1. An important note is that in some cases these sampling schemes are chosen for convenience and that the generalizability of the experiment to the population will depend upon the assumptions made in them being true. The sampling model may also be considered to serve as a conservative approach to finite sample inference (see Ding et al., 2017).

Related work

For matched pairs experiments, Imai (2008) showed that with a superpopulation of an infinite number of structural blocks, specifically matched pairs, from which pairs are randomly sampled, the standard matched pairs variance estimator (Equation (4)), is unbiased for the population average treatment effect (PATE). On the other hand, Imbens (2011) showed that the standard matched pairs variance estimator is biased in the setting where we have fixed blocks and units are drawn using stratified random sampling (see Section 4.3 for more on this setting). This is a clear example of how the population framework being used matters. We therefore advise practitioners to carefully consider what population and sampling structure they are assuming and to not simply assume a framework for convenience.

The general blocked design has been previously discussed in various forms. Imbens (2011) discussed blocking in the context of a superpopulation with a fixed number of strata from which units are sampled using a stratified sampling method. He formed unbiased estimators for the variance in this context, assuming that the blocks each have at least two units assigned to treatment and control. These results are similar to finite sample results discussed in Section 3 and will be discussed more in Section 4.3. Imai et al. (2008) analyzed estimation error and variance with the blocked design. Scosyrev (2014) also analyzed the blocked experiment in the finite sample and under two sampling frameworks, recognizing that the different settings resulted in different outcomes. Sävje (2015) analyzed flexible "threshold" blocking and made critical points about the importance of block structure and sampling design when analyzing blocked experiments, which we will echo and expand on.

4.1 Infinite populations in general

Inference for the population average treatment effect (PATE) typically takes the sample as a random sample from some larger population, as opposed to inference for the SATE discussed earlier which held the sample of potential outcomes as fixed. This makes estimation an implicit two-step process, estimating the treatment effect for the sample and extrapolating this estimate to the population. Frequently, in fact, the estimators themselves are the same as for finite sample inference even though the estimands are different.

Define the PATE as

$$\tau = \mathbb{E}[Y_i(t) - Y_i(c)|\mathcal{F}],$$

where $\mathcal{F}$ both indicates the block type and sampling framework. This is the same as the direct average of the unit-level treatment effects for all of the units in the population, as is commonly used (see Imbens and Rubin, 2015, p. 99), as long as our sampling mechanism is not biased. Here we will only consider frameworks where the sampling scheme provides a sample that, on average, has the same average treatment effect as the population but note that bias from the sampling mechanism can be fixed using weighting if the sampling mechanism is known (see Miratrix et al., 2018).

Under blocking, the PATE within block k is

$$\tau_k = \mathbb{E}[Y_i(t) - Y_i(c)|b_i = k, \mathcal{F}],$$

where, again, b_i indicates the block that unit i belongs to. It is possible that k indexes a (countably) infinite set of blocks in the case of some infinite population models.

Overall, using the law of total expectation and variance decompositions, we can generally obtain the properties of our estimators with respect to population estimands by first obtaining expressions for a finite sample and then averaging these expressions across the sampling distributions. In other words, we heavily exploit $\mathbb{E}\left[\widehat{M}|\mathcal{F}\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{M}|\mathcal{S}\right]|\mathcal{F}\right]$, where $\mathcal{S}$ is a sample obtained from $\mathcal{F}$, our population and sampling framework. Under any unbiased framework $\mathcal{F}$, we have the typical result (e.g. see Imbens, 2011)

$$\mathbb{E}\left[\widehat{\tau}_{(CR)}|\mathcal{F}\right] = \mathbb{E}\left[\widehat{\tau}_{(BK)}|\mathcal{F}\right] = \mathbb{E}\left[\tau_{\mathcal{S}}|\mathcal{F}\right] = \tau.$$

There are several different frameworks that one might assume. These can generally be characterized by two primary features: the block types, which also dictates the population strata structure, and the sampling scheme. Note that the term strata is used for the population here analogously to blocks in the sample. We may obtain a sample using simple random sampling and then form blocks based on covariates post-sampling and pre-randomization, i.e. flexible blocks. Or we may have fixed blocks (e.g. blood types) and use stratified sampling where we sample units from each population strata. Finally, we may have structural blocks and conceptualize a population of an infinite number of these blocks (e.g. schools in an "infinite" population of schools) from which we randomly select a fixed number of blocks. As we show next, the bias of the variance estimators can differ depending on the framework assumed. We refer to frameworks using their sampling method as a shorthand, leaving the block type and population structure implicit.

4.2 Simple random sampling, flexible blocks

In this framework, denoted SRS, units are sampled at random, without regard to block membership, from the population. This gives the classic result for complete randomization (see Imbens and Rubin, 2015, p. 101) of

$$\text{var}(\widehat{\tau}_{(CR)}|SRS) = \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t}$$

where $\sigma^2(z)$ is the population variance of the potential outcomes under treatment z. The classic variance estimator $\widehat{\sigma}^2_{(CR)}$ is unbiased in this setting.

For blocking with SRS, we focus on the use of flexible blocks, e.g. blocking using clustering on a continuous covariate or based on observed covariates in the sample obtained, as we believe they are most useful in this setting. Structural blocks do not make sense in this framework (e.g. one would always sample pairs of twins not individuals who are twins if we wish to run a twin study) and fixed blocks give rise to difficulties when the sample does not have units from all population strata. Note however that this sampling framework was examined for blocked experiments with fixed blocks in Scosyrev (2014).

For a blocked experiment, the variance in this framework, using the basic variance decomposition, is

$$\text{var}(\widehat{\tau}_{(BK)}|SRS) = \mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] + \text{var}\left(\tau_{\mathcal{S}}|SRS\right).$$

The expectation is across the sampling and blocking process.

Theorem 4.2.1. *The estimator*

$$\widehat{\sigma}^2_{SRS} = \sum_{k=1}^{K} \frac{n_k(n_k-1)}{n(n-1)}\left(\frac{s_k^2(c)}{n_{c,k}} + \frac{s_k^2(t)}{n_{t,k}}\right) + \sum_{k=1}^{K} \frac{n_k}{n(n-1)}\left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2. \tag{8}$$

is an unbiased estimator for var($\widehat{\tau}_{(BK)}|SRS$).

See Supplementary Materials F for a derivation. The first term in the estimator looks similar to our usual big block estimator and captures part of the first term in our variance decomposition. The second term in the estimator looks similar to our proposed small block estimator and accounts for the rest of the variation. This is very similar to the estimator found in Scosyrev (2014), however we make adjustments to achieve unbiasedness of the estimator whereas Scosyrev (2014) focuses on consistency. Scosyrev (2014) also works with fixed blocks where the number of blocks is assumed known before sampling and weights are used to match the sample to the population proportions, as opposed to flexible blocks which allow random numbers of blocks that are created post-sampling.

Remark. If we use $\widehat{\sigma}^2_{(BK)}$ our bias will be

$$\mathbb{E}\left[\widehat{\sigma}^2_{(BK)}|SRS\right] - \text{var}(\widehat{\tau}_{(BK)}|SRS) = \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}\frac{S^2_k(tc)}{n} - \frac{S^2(tc)}{n}\Big|SRS\right].$$

This can be seen easily from derivations in Supplementary Materials F. This implies that $\widehat{\sigma}^2_{(BK)}$ could be anti-conservative in this setting if units put within the same block are identical in terms of treatment effects (which would imply $S^2_k(tc) = 0$) but there is greater variation across the whole population (which would imply $S^2(tc) > 0$). Achieving identical units with flexible blocks would be hard to do, but we may consider that the limiting case.

Similarly, if we use either of the small block variance estimators, the bias will be the difference between the expected finite sample bias for those estimators (which for both our small block estimators depends on treatment effect heterogeneity between blocks) and $\mathbb{E}\left[\frac{S^2(tc)}{n}\Big|SRS\right]$, which corresponds to treatment effect heterogeneity across the whole population. Therefore whether these estimators are conservative or not depends upon the structure of the population and how the blocks are formed.

4.3 Stratified sampling, fixed blocks

In the "stratified sampling" framework, denoted $\mathcal{F}_1$, there are K fixed strata of infinite size in the population. Then n_k units are randomly sampled from strata k (i.e., stratified random sampling is used). Here we have fixed blocks. We assume that n_k is fixed and that n_k/n is the population proportion of units in stratum k, for simplicity. Otherwise, a weighting scheme, as mentioned in Section 4.1, would be needed to create an unbiased estimator of the direct average of treatment effects in the population. This is the framework used in Imbens (2011) and Miratrix et al. (2013), who show the following result under equal proportions treated within each block, which simplifies the weights.

As in the finite sample, overall variance is a weighted sum of within block variances:

$$\text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) = \sum_{k=1}^{K} \frac{n_k^2}{n^2}\text{var}\Big(\hat{\tau}_k|\mathcal{F}_1\Big) = \sum_{k=1}^{K} \frac{n_k^2}{n^2}\Big(\frac{\sigma_k^2(c)}{n_{c,k}} + \frac{\sigma_k^2(t)}{n_{t,k}}\Big). \tag{9}$$

where $\sigma_k^2(z)$ is the population variance of the potential outcomes under treatment z in strata k.

As noted in Imbens (2011), the variance estimator of big blocks, $\widehat{\sigma}^2_{(BK)}$, is unbiased in this framework. The estimators for the variance of the small blocks, however, can have bias. We have two results pertaining to this.

First, as with the finite sample, we can extend results for $\widehat{\sigma}^2_{(SMALL/s)}$ (see Imbens, 2011) to $\widehat{\sigma}^2_{(SMALL/m)}$.

Corollary 4.3.1. *The bias of* $\widehat{\sigma}^2_{(SMALL/m)}$ *under the stratified sampling framework is*

$$\mathbb{E}\Big[\widehat{\sigma}^2_{(SMALL/m)}|\mathcal{F}_1\Big] - var(\widehat{\tau}_{(SMALL)}|\mathcal{F}_1) = \sum_{j=1}^{J} \frac{K_j m_j^2}{n_{sb}^2(K_j-1)} \sum_{k:n_k=m_j} \Big(\tau_k - \tau_{(SMALL)}\Big)^2.$$

See Supplementary Material D.2.1 for the derivation. As with finite sample inference, this shows that $\widehat{\sigma}^2_{(SMALL/m)}$ is a conservative estimator unless the average treatment effect is the same across all small blocks of the same size.

Second, for our new variance estimator we have the following result:

Corollary 4.3.2. *The bias of* $\widehat{\sigma}^2_{(SMALL/p)}$ *under the stratified sampling framework is*

$$\begin{aligned}&\mathbb{E}\Big[\widehat{\sigma}^2_{(SMALL/p)}|\mathcal{F}_1\Big] - var(\widehat{\tau}_{(SMALL)}|\mathcal{F}_1)\\ &= \sum_{k=1}^{K} \frac{n_k^2}{(n_{sb}-2n_k)(n_{sb} + \sum_{i=1}^{K}\frac{n_i^2}{n_{sb}-2n_i})}\Big(\tau_k - \tau_{(SMALL)}\Big)^2.\end{aligned}$$

This shows that $\widehat{\sigma}^2_{(SMALL/p)}$ is also a conservative estimator (given no block makes up more than half the sample) and it is unbiased when the average treatment effect is the same across all small blocks. See Supplementary Material E for a derivation.

4.4 Random sampling of strata, structural blocks

In the "random sampling of strata" framework, denoted $\mathcal{F}_2$, there are an infinite number of strata of finite size, i.e. an infinite number of structural blocks. K strata are then randomly chosen to be in the sample and randomization is done within each of the sample blocks. This setting, with equal block sizes, is often used in the matched pairs literature, such as in Imai (2008).

Under this framework, the variance of $\widehat{\tau}_{(BK)}$, is

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_2\right) = \mathbb{E}\left[\sum_{k:B_k=1}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)|\mathcal{F}_2\right] + \text{var}\left(\tau_{\mathcal{S}}|\mathcal{F}_2\right), \qquad (10)$$

where B_k is the indicator that stratum k is included in the sample, with $B_k = 1$ indicating sample membership and $B_k = 0$ otherwise.

In this framework, which blocks are included in the sample is random. Therefore, the variance estimator needs to capture not only the within strata variance but also the variance due to which strata are chosen to be in the sample.

When blocks are of the same size, we can simplify the expression with $\frac{n_k^2}{n^2} = \frac{1}{K^2}$, which is no longer random. If we have all blocks of the same size, then we can rewrite $\widehat{\sigma}^2_{(SMALL/s)}$ (Equation 4) using sample inclusion indicators as

$$\widehat{\sigma}^2_{(SMALL/s)} = \frac{1}{K(K-1)}\sum_k B_k(\widehat{\tau}_k - \widehat{\tau}_{(BK)})^2,$$

and this is an unbiased estimator for $\text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_2)$. This is simply the variance of the estimated block effect in the sample. Imai (2008) showed that this estimator is unbiased in this setting with an infinite population of matched pairs. See Supplementary Material D.2.2 for the proof of this result extended to other small block types.

Considering the case of strata of varying sizes makes things more complicated because the variable sizes induce a random number of units in each stratum and in total. In particular, under this framework there is a chance that there is only a single block of a given size, making the first variance estimator infeasible. To simplify, we could condition on the number of strata drawn of each possible strata size, assuming that there are multiple strata of the each size in the sample.

Corollary 4.4.1. *In the conditioned case, assuming it is defined, $\widehat{\sigma}^2_{(SMALL/m)}$ is an unbiased estimator for var$(\widehat{\tau}_{(BK)}|\mathcal{F}_2)$.*

This result can be seen directly from the results in Supplementary Material D.2.2.

Alternatively, if we are willing to assume that block size is independent of treatment effect, then we have the following more general result:

Theorem 4.4.1 (Unbiasedness of $\widehat{\sigma}^2_{(SMALL/p)}$ given independence)**.** *In the case the random sampling of strata setting where block sizes are independent of block average treatment effects, $\widehat{\sigma}^2_{(SMALL/p)}$ is an unbiased estimator for var$(\widehat{\tau}_{(BK)}|\mathcal{F}_2)$.*

The proof is in Supplementary Materials E.2. Alternatively, the RCT-YES R package (Schochet, 2016) estimator discussed in Supplementary Material B.3 could be used.

Remark. We may also consider an infinite number of strata of infinite size, as is commonly used in multisite randomized trials. This is the setting considered in Schochet (2016). The sampling scheme then has two steps: first sample the strata, then sample units from the strata. To discuss variance, we need to add a bit of notation. Let $\tau^*_{\mathcal{S}}$ denote the expectation of the treatment effect estimator given the blocks in the sample. That is, we fix which strata are in the sample and take the expectation over the sampling of units from the infinite size strata. So conditioning on which strata are in the sample we are in a stratified sampling set up. Let this framework be denoted by $\mathcal{F}_3$. Then the variance of $\widehat{\tau}_{(BK)}$ is

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_3\right) = \mathbb{E}\left[\sum_{k:B_k=1} \frac{n_k^2}{n^2}\left(\frac{\sigma_k^2(c)}{n_{c,k}} + \frac{\sigma_k^2(t)}{n_{t,k}}\right)|\mathcal{F}_3\right] + \text{var}\left(\tau^*_{\mathcal{S}}|\mathcal{F}_3\right).$$

It is straightforward to extend the results of Corollary 4.4.1 and Theorem 4.4.1 to this case.

4.5 Discussion

While the variance formulas that we presented above share a similar structure with each other and the finite sample forms, there are important differences. In the finite sample framework (Equation (2)), there is a term regarding treatment effect variation that reduces the variance due to the correlation of potential outcomes. This term is retained in the random sampling of strata framework of Section 4.4 but not in the stratified sampling framework of Section 4.3. This difference in the true variance implies that different variance estimators may be more appropriate in different settings and that comparisons of blocking to complete randomization under these different assumptions will also diverge. In fact, this difference explains much of apparent discrepancy between the matched pairs literature and the blocking literature.

Relatedly, different variance estimators can have different amounts of bias depending on the framework being used. The small blocks estimators ($\widehat{\sigma}^2_{(SMALL/m)}$ and $\widehat{\sigma}^2_{(SMALL/p)}$) in the finite sample and the stratified sampling framework are unbiased if the average treatment effect is the same across all of the small blocks (or all of the small blocks of the same size for $\widehat{\sigma}^2_{(SMALL/m)}$) and otherwise are more conservative as the variance of the average treatment effects across blocks increases. For the infinite number of strata framework, under some assumptions all of our small block variance estimators are unbiased. We have no small block estimator that is guaranteed to be unbiased or conservative for the simple random sampling (flexible block) framework.

The big blocks estimator ($\widehat{\sigma}^2_{(BK)}$) in the finite sample is unbiased if the treatment effect is additive within each block and otherwise depends on the treatment effect heterogeneity within each block. In the stratified sampling framework, however, $\widehat{\sigma}^2_{(BK)}$ will be unbiased.

Overall, only the framework of Section 4.4, with the additional assumption given there, has unbiased variance estimators for a mixture of big and small blocks. This means that, without additional assumptions allowing for plug-in approaches, the hybrid estimators will always be conservative in the other settings discussed.

5 Comparing blocking to complete randomization

Much confusion in the literature surrounding the benefits of blocking can be attributed to researchers performing isolated investigations of estimators' properties under specific sampling frameworks. Although researchers typically focus on finite vs. infinite population as a distinction when assessing if blocking may be beneficial or harmful, the block types and sampling framework being used also matter. This comparison has been made from many perspectives (e.g. Snedecor and Cochran, 1989) but we will discuss the prior work that has been done within the causal inference potential outcomes framework. For instance, Imai (2008) compared the true variance for the matched pairs design to the variance of the estimator for the completely randomized design under two sampling schemes, recognizing the important role that the sampling scheme plays. From this he concluded that "the relative efficiency of the matched-pair design depends on whether matching induces positive or negative correlations regarding potential outcomes within each pair" (Imai, 2008, p. 4865), a comment similar to one made on p. 101 of Snedecor and Cochran (1989). In contrast, Imbens (2011), assuming a stratified sampling superpopulation model, claimed that "In experiments with randomization at the unit-level, stratification is superior to complete randomization, and pairing is superior to stratification in terms of precision of estimation" (p. 1). Imai et al. (2008) similarly concluded that the variance under the blocked design was lower than under complete randomization for a superpopulation set-up. Although these conclusions are correct for the context the authors were working in, the use of different frameworks for blocking and matched pairs can make the results seem inconsistent.

For our primary discussion, we assume that $p_k = p$ for all k, i.e. the proportion of treated units is constant across blocks and is the same as the proportion of treated units under complete randomization. We compare the variances of a blocked design vs. complete randomization under several population frameworks to clarify the similarities and differences under these different sampling methods. All derivations are in Supplementary Material C. The Supplementary Material also contains further results on comparing blocking to complete randomization; Section C.5 discusses extensions to unequal proportions treated and Section G discusses the performance of the complete randomization variance estimator when treatment assignment is done according to blocked randomization, i.e., the performance of the variance estimator that ignores blocking. Supplementary Material I.1 contains a related discussion about when the *variances* of our variance estimators are higher for blocking than for complete randomization.

5.1 Finite framework

For the finite setting, we find a result similar to those presented in other papers, such as Imai et al. (2008) and Miratrix et al. (2013). The difference in variance of the treatment effect estimator between the completely randomized design and the blocked design, in the finite sample, is

$$\begin{aligned} &\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \qquad (11)\\ &= \frac{1}{n-1}\left[\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right) - \sum_{k=1}^{K}\frac{n_k}{n}\frac{n-n_k}{n}\text{var}(\widehat{\tau}_k|\mathcal{S})\right]\end{aligned}$$

where

$$\text{Var}_k\left(X_k\right) \equiv \sum_{k=1}^{K}\frac{n_k}{n}\left(X_k - \sum_{j=1}^{K}\frac{n_j}{n}X_j\right)^2. \qquad (12)$$

Whether this quantity is positive or negative depends on whether some form of between block variation is larger than a form of within block variation. Most prior work state that although the difference in the brackets can be negative, as the sample size grows this difference will go to a non-negative quantity. However, this statement depends on the type of blocks we have. In particular, if we have structural blocks such that as n grows, the number of blocks K also grows, the difference in the brackets of Equation (11) will not necessarily go to zero or become positive as $n \to \infty$. This has ties to the random sampling of strata framework (see Section 4.4 and Section 5.4).

5.2 Simple random sampling, flexible blocks

The difference between the variance under the completely randomized design and the blocked design in the simple random sampling framework with flexible blocks is

$$\begin{aligned}&\text{var}\left(\widehat{\tau}_{(CR)}|SRS\right) - \text{var}\left(\widehat{\tau}_{(BK)}|SRS\right)\\ =&\mathbb{E}\left[\frac{1}{n-1}\left[\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right) - \sum_{k=1}^{K}\frac{n_k}{n}\frac{n-n_k}{n}\text{var}(\widehat{\tau}_k|\mathcal{S})\right]|SRS\right].\end{aligned}$$

If we have a fixed set of covariates and a fixed algorithm used to do blocking, then even if the covariates used to form blocks are independent from outcomes, blocking will not increase variance and in fact this difference will be zero, as shown in Supplementary Materials C.2. Hence, in the "worst case" of blocking on something irrelevant, blocking will still not increase variance on average. However, if the algorithm or the set of covariates used to block is allowed to change from sample to sample, we can end up in a worse but less realistic "worst case." In that setting, one could hypothetically choose the worst possible blocking for each sample, such that sample variance for blocking is higher than the complete randomization for every sample. Then blocking would do worse than complete randomization on average over the sampling.

5.3 Stratified sampling, fixed blocks

Theorem 5.3.1 (Variance comparison under stratified sampling)**.** *The difference in variance between complete randomization and blocked randomization under the stratified sampling* *framework is*

$$
\begin{aligned}
&var\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) - var\left(\widehat{\tau}_{(BK)}|\mathcal{F}_1\right)\\
&= \frac{1}{n-1}\, Var_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right)\\
&\geq 0
\end{aligned}
$$

where $\mu_k(z)$ is the population mean of potential outcomes under treatment z in strata k.

The above expression is very similar to the positive term in Equation (11) for the finite sample framework. Now, however, we no longer have the negative term. Hence, it is not possible for blocking to be harmful in the stratified sampling framework with equal proportion treated in all blocks.

Interestingly, when comparing blocking to completely randomized experiments in an infinite population setting, researchers will typically evaluate the completely randomized design under the simpler sampling mechanism of simple random sampling and analyze the blocked design under the stratified sampling framework (e.g., see Imbens, 2011). Then the difference between the two estimators is an agglomeration of differences in the characteristics of the samples under the two different sampling regimes as well as the difference in doing a blocked experiment versus a completely randomized experiment. In general, the difference between $\text{var}(\widehat{\tau}_{(CR)}|SRS) - \text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right)$ may be positive or negative, showing that the traditional estimates of the benefits of blocking can either be under or overstated in this context. For further discussion and mathematical formulation, see Supplementary Material C.4.

5.4 Random sampling of strata, structural blocks

The difference between the variance under the completely randomized design and the blocked design in the random sampling of strata framework is

$$
\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_2\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_2\right)\\
=&\mathbb{E}\left[\frac{1}{n-1}\left[\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right) - \sum_{k=1}^{K}\frac{n_k}{n}\frac{n-n_k}{n}\text{var}(\widehat{\tau}_k|\mathcal{S})\right]|\mathcal{F}_2\right].
\end{aligned}
$$

As in the finite framework, because the strata themselves are finite, it is possible that blocking could result in higher variance. In particular, if the block means do not vary, blocking will not be beneficial. Although this formulation looks essentially identical to that obtained under simple random sampling, there is an important difference here. In the simple random sampling framework we assumed sampling of individual units and the use of flexible blocks. Here the blocks themselves are sampled with block membership fixed, so the expectation can be thought of as over all blocks in the population. In this framework it is possible to have systematically poor blocks. For instance, if we use elementary school classrooms as blocks, we may find that schools break up students into classes such that the classrooms all look similar to each other, in that they have similar proportions of high and low achieving students, but by the same token have higher within classroom variability.

Remark. Again, we may instead consider an infinite number of blocks of infinite size. Then the difference between the variance under the completely randomized design and the blocked design under this framework is

$$\begin{aligned}&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_3\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_3\right)\\&= \mathbb{E}\left[\frac{1}{n-1}\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right)|\mathcal{F}_3\right]\\&\geq 0.\end{aligned}$$

Hence, making the population from which units in the blocks are sampled infinite guarantees that blocking is always beneficial or at least not harmful in terms of variance compared to complete randomization.

5.5 Discussion

We see that in the stratified sampling framework we are guaranteed that blocking will not increase variance. Simple random sampling with flexible blocks, by contrast, does not have such guarantees although it will not increase variance compared to complete randomization in most reasonable cases. In the finite sample or random sampling of strata frameworks, the benefits (or harm) of blocking depend on the difference of within and between block variation. However, this result for the random sampling of strata framework changes if we instead have infinite size strata. Finite sample numerical studies in Supplementary Material H.2 show an example where even in the worst case for blocking, when all blocks have the same distribution of potential outcomes, the increase in variance is not too great. Here, as treatment heterogeneity between the blocks increases, blocking quickly provides large gains.

The types of blocks used can have a large impact. Structural blocks can be a systematically poor choice for blocking such that even when we have an infinite population of such blocks, blocking will increase variance compared to complete randomization. With fixed blocks and stratified sampling, however, this is impossible. Increased variance is also unlikely when we have simple random sampling and flexible blocks. That being said, such poor structural blocks would be an extreme case and subject matter knowledge should help avoid such circumstances.

These findings, however, assume equal treatment proportions across blocks. With unequal proportions we do not have such guarantees (see Supplementary Material C.5). In our simulations (see Supplementary Material I.1) we in fact see degradation in the benefits of blocking on weakly predictive covariates when the proportion treated is only approximately equal, rather than precisely. This comes from the additional variation induced by different units being weighted differently, and is akin to the increase in variance from weighted survey sampling. This suggests that in many realistic scenarios, one might not want to block on covariates that are only weakly predictive.

6 Simulations

We compare different estimators of the variance for hybrid blocked experiments where there are a few big blocks and many small blocks. Here, 50% of our units are in small blocks with only one treated unit and the remainder are in big blocks with at least two treated units. In none of our blocks were there many treated units due to only having approximately 20% of the units treated. The 20% was approximate in order to create varying size small blocks to see the different performance of the hybrid estimators. The sizes of the fifteen blocks ranged from 3 to 20.

The simulations presented here are for the finite framework. However, the results are largely applicable to the other settings. For instance, the biases for the small blocks variance estimators have the same form for the finite sample and the stratified sampling frameworks.

We considered our two hybrid estimators, which correspond to estimating the variance of the small blocks two different ways. We also considered two regression estimators: the HC1 sandwich estimate (Hinkley, 1977) from a linear model with fixed effects and no interaction between treatment indicator and blocking factor and the standard variance estimate from a weighted regression, weighting each unit by the inverse probability of being assigned to its given treatment status in its block, multiplied by the overall proportion of units in its treatment group. Note that the HC1 estimator is the "robust" estimator used in Stata (StataCorp, 2017) for estimating standard deviations. We finally considered a plug in estimator that uses a weighted average of the big block variance estimates for the treated for each missing small block variance estimate for the treated.

In our simulations, we varied how well blocking separated units based on the potential outcomes under control and on the treatment effect. The average potential outcome under control and the treatment effect for each block were both negatively correlated with block size, so that smaller blocks had larger control potential outcomes and larger treatment effects. The correlation of potential outcomes within blocks was also varied between $\rho = 0, 0.5,$ and 1. See Supplementary Material H.1 for more on the data generating process.

We compared all of the variance estimators to the actual variance of the corresponding blocking treatment estimator in Figure 1 by looking at the percent relative bias ($[\text{mean}(\hat{\sigma}^2_*) - \text{var}(\widehat{\tau}_{(BK)}|\mathcal{S})]/\text{var}(\widehat{\tau}_{(BK)}|\mathcal{S})$).[1] The variation due to changing the between block difference in the mean of control potential outcomes was found to be minimal so these are not separated out on the plots. We see that the plug-in (Plug-in) estimator generally performs well. This is unsurprising considering that in these simulations this plug-in method is well specified. The two hybrid estimators, the one using $\widehat{\sigma}^2_{(SMALL/m)}$ (Hybrid$_\text{m}$) for the small blocks and the one using $\widehat{\sigma}^2_{(SMALL/p)}$ (Hybrid$_\text{p}$) for the small blocks, outperform the linear model estimators, especially as the treatment effect variation across the blocks increases. We see that Hybrid$_\text{m}$ also has lower bias than Hybrid$_\text{p}$ as treatment heterogeneity increases. This is because the value of treatment effects are correlated with block size and $\widehat{\sigma}^2_{(SMALL/m)}$ groups variance estimation by block size. Weighted regression performance was generally similar to that of Hybrid$_\text{p}$. It was slightly anti-conservative for samples with low treatment heterogeneity when $\rho = 1$.

[1] We compare all estimators to the variance of $\widehat{\tau}_{(BK)}$ to put everything on the same scale, even though the sandwich estimate for a linear model is estimating variance for the linear model estimator.

For discussion of the variance of the variance estimators, see Supplementary Material I.2. The variance estimators' variances were found to be comparable, with weighted regression generally having the lowest variance.

When comparing the performance of estimators, there is an important note about the linear model estimator: the sandwich estimate for a linear model is associated with a different treatment effect estimator than the others. In particular, a linear model with fixed effects is estimating a precision weighted estimate of the treatment effect across the blocks. It is well known that as treatment heterogeneity increases, this estimator can become increasingly biased. This is not an issue for the weighted regression which, similar to adding interactions between treatment and block dummy variables, will recover $\widehat{\tau}_{(BK)}$.

Another important note is that the plug-in estimator depends on a model, while the hybrid estimators do not. In these simulations, the model is correctly specified because all blocks have the same variance. However, if the big blocks have systematically smaller variance than the small blocks, the plug-in variance estimator could be anti-conservative.

7 Data Example

One area where analysts are often faced with many small blocks of varying sizes is found in the matching literature. In particular, full matching (see Hansen (2004), Rosenbaum (1991)) finds sets of similar units, with either one treated and several control or vice versa, that could be considered as-if randomized. After matching, a researcher could then analyze these data using permutation tests and associated sensitivity checks (see, e.g. Rosenbaum (2010)), but in this context generating confidence intervals or standard errors using permutation inference would typically rely on a constant treatment effect assumption across the blocks. One might alternatively wish for a Neyman-style randomization analysis such as would be typically done for large block experiments to obtain inference for the average effect in the presence of treatment variation. The average treatment effect estimate is easy to obtain; it is the uncertainty estimation that causes the trouble. Our small block variance estimators fills this gap. To illustrate, we analyze a data set from the National Health and Nutrition Examination Survey (NHANES) 2013-2014 given in the `CrossScreening` package (Rosenbaum and Zhao, 2017) in `R` statistical software (R Core Team, 2016). This data set was also used by Zhao

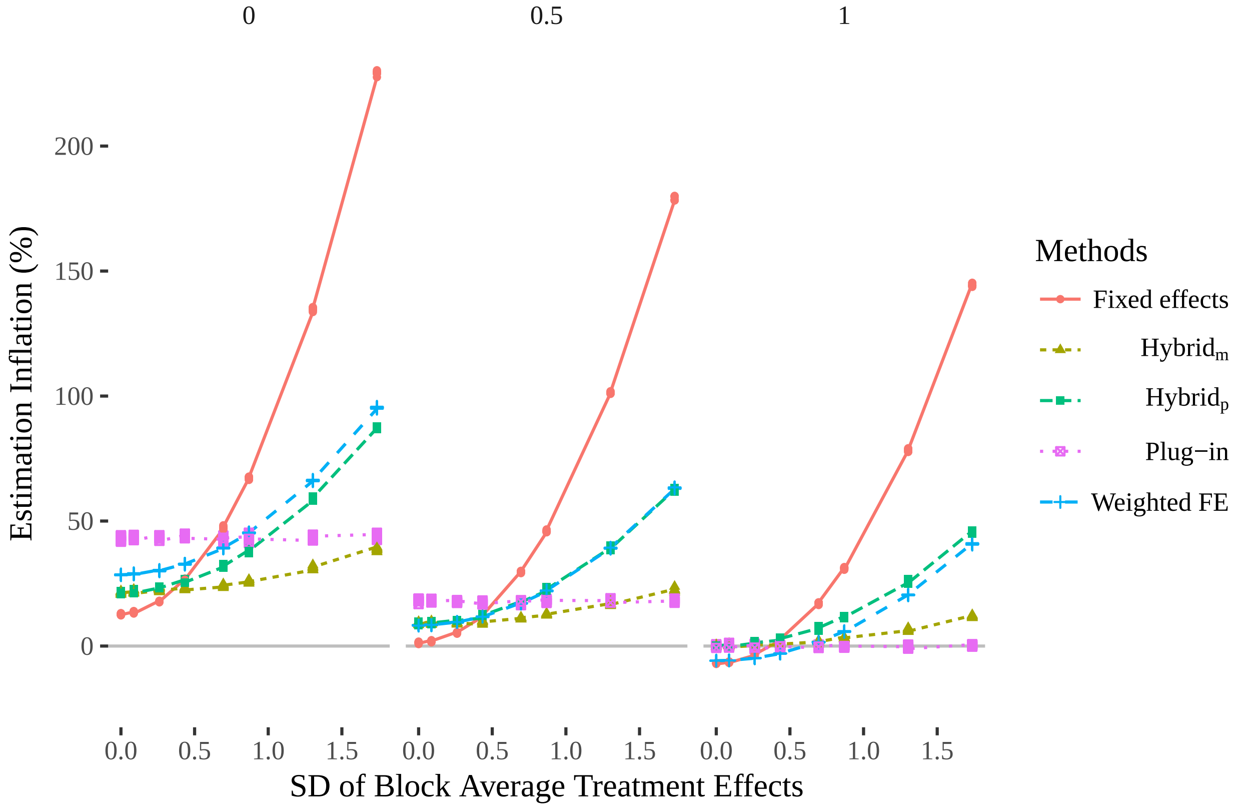

Figure 1: Simulations to assess variance estimators' relative bias as a function of treatment variation across blocks. Each column represents a different value of ρ, with values denoted at the top of the graph. The x-axis shows the standard deviation of block treatment effects. Dots indicate specific finite samples. FE stands for fixed effects.

et al. (2018) to analyze the effect of high fish consumption (defined as 12 or more servings of fish or shellfish in the previous month) versus low fish consumption (defined as 0 or 1 servings of fish or shellfish in the previous month) on a number of biomarkers.

Although Zhao et al. (2018) analyzed numerous outcomes, we focus on a measure of mercury (LBXGM), converted to the $\log_2$ scale, as a simple illustration of our methods. We use full matching to obtain a set of all small blocks of varying size. As in Zhao et al. (2018), we matched on smoking, age, gender, race, income, and education. We used Bayesian logistic regression through the `brglm` package (Kosmidis, 2017) and `optmatch` (Hansen and Klopfer, 2006) in `R` (R Core Team, 2016). This resulted in 197 blocks with only one treated or one control unit in each. Sizes of blocks ranged from 2 to 47.

There were some block sizes that were unique, so the hybrid estimator with $\widehat{\sigma}^2_{(SMALL/m)}$

	NHANES		Lalonde	
Estimator	Estimate	$\widehat{SE}$	Estimate	$\widehat{SE}$
Hybrid blocking with $\widehat{\sigma}^2_{(SMALL/m)}$	2.453	N/A	\$560	\$570
Hybrid blocking with $\widehat{\sigma}^2_{(SMALL/p)}$	2.453	0.202	\$560	\$606
Weighted regression	2.453	0.109	\$560	\$560
Fixed effects regression (HC1)	2.746	0.134	\$425	\$601

Table 1: Results of NHANES (full matching), and Lalonde (CEM) for different estimation strategies.

could not be used. The blocking treatment effect estimate ($\widehat{\tau}_{(BK)}$) was 2.453 but using a fixed effects model with no interaction the treatment effect estimate was 2.746. Looking at Table 1, we see that our hybrid estimator using $\widehat{\sigma}^2_{(SMALL/p)}$ gave a much larger variance estimate (relative to the scale of the precision estimates) than the two linear model based variance estimators.

A second method for analyzing observational datasets where our variance estimators could be useful is coarsened exact matching (CEM). CEM coarsens covariates used to match and then exactly matches to these coarsened variables (Iacus et al., 2012). We follow the example from the vignette of the `cem` package (Iacus et al., 2016) in `R` using the most automated version of CEM on the classic LaLonde data set (LaLonde, 1986), available in the `cem` package. This data set consists of individuals who received or did not receive a job training program with the outcome of interest as earnings in 1978. We use the unmodified version of the LaLonde dataset, but otherwise follow the automated process for CEM laid out in the vignette to create blocks (we do not follow the analysis). This resulted in the creation of 69 blocks, some small and some big, with some ungrouped units being dropped. The blocking treatment effect estimate ($\widehat{\tau}_{(BK)}$) was \$560 but using a fixed effects model with no interaction the treatment effect estimate was \$425. From Table 1, the precision estimates from all methods were similar though, again, the hybrid estimator using $\widehat{\sigma}^2_{(SMALL/p)}$ was the largest and likely the most conservative.

8 Discussion

Blocking can be viewed under a wide variety of population frameworks ranging from a fixed, finite-sample model to one where we envision the units as being sampled from a larger population in pre-set groups. Because some findings regarding blocking change depending on what framework is used, the current literature can seem confusing and contradictory. Furthermore, because different types of blocking tend to use different frameworks, there is a lack of clarity on how one should proceed when faced with a randomized trial containing blocks of all different sizes.

We have worked to clarify these subtleties and to fill this gap. We identified and compared the true variance of a blocking-based estimator under multiple settings, and created corresponding estimators of the impact estimator's variance. We also provide simple, model-free variance estimators for two types of experiments that have not received much attention: blocked experiments with variable-sized blocks containing singleton treatment or control units, and hybrid blocked experiments with large and small blocks combined. These contexts are quite common, frequently appearing in, for example, the matching literature. We analyzed the performance of both our new variance estimators and the classic variance estimators under different frameworks, identifying when they are unbiased or conservative. This investigation again illustrates how different sampling frameworks and block types can impact assessments of an estimator's performance.

We also carefully compared complete randomization to blocking, identifying that prior literature has often collapsed the sampling step and randomization step. Overall, we showed that blocking often, but not always, improves precision, and that guarantees about blocking depend on the framework adopted. Blocking will not reduce precision, compared to a complete randomization, when working in the stratified sampling framework with equal proportion of units treated across blocks, no matter how small the blocks are or how poorly they separate the units. Similarly, we found that in the simple random sampling setting, given a fixed algorithm for creating flexible blocks, and when blocking on covariates that are independent of potential outcomes, the blocking estimator will have equal variance to the estimator under complete randomization. In the other two main frameworks considered, however, blocking is not guaranteed to reduce variance. Without prior knowledge about the distribution of potential outcomes, it is impossible to know before an experiment is conducted whether blocking will improve the estimates in these latter two settings. This being said, the above assumes the blocks have equal proportions of units treated; if the proportions treated differ, blocking is more costly regardless of framework.

We do not discuss in this paper two other issues which would be useful areas of further investigation: degrees of freedom and stability of variance estimators. The tradeoff between increased precision and reduced degrees of freedom by using blocked or matched pairs designs has been noted by others (e.g., Box et al., 2005, p. 93; Imbens, 2011; Snedecor and Cochran, 1989, p. 101) and is an important practical limitation to consider when using these designs. How variable our variance estimators is an important question in determining the best design and analysis to use. Although we touch on the stability of our variance estimators in the Supplementary Materials, a more in depth exploration is needed. One might think, given the above, to implement blocking to realize some gains, but then analyze as a completely randomized experiment to avoid these concerns. Unfortunately, this does not result in a guaranteed valid or conservative analysis: even when proportion treated across blocks is constant, such a move is not necessarily conservative, depending on the framework adopted (see Supplementary Materials G). Future work should investigate how the real costs of degrees of freedom loss and instability in variance estimation depend on the experimental design within these frameworks.

Future work includes extending these results to other population settings and sampling methods, in particular finding small block estimators for the setting of constructing blocks post-sampling and pre-randomization. Variance estimation is also a missing and needed piece in post-stratification research, as noted in Miratrix et al. (2013). Although conditional answers for post-stratification would correspond to the estimators presented in this work, the unconditional case remains open.

References

Abadie, A. and Imbens, G. W. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Économie et de Statistique*, 91/92:175–187.

Box, G. E., Hunter, J. S., and Hunter, W. G. (2005). Statistics for experimenters. In *Wiley Series in Probability and Statistics*. John Wiley & Sons, Hoboken, NJ.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS) (2013-2014). National Health and Nutrition Examination Survey Data. *Hyattsville, MD: U.S. Department of Health and Human Services, CDC.*

Cochran, W. G. (1953). Matching in analytical studies. *American Journal of Public Health and the Nations Health*, 43(6_Pt_1):684–691.

Cochran, W. G. (1977). *Sampling techniques*. Wiley Series in Probability and Mathematical Statistics-Applied. John Wiley & Sons, New York, 3d edition.

Cochran, W. G. and Cox, G. M. (1950). *Experimental Designs*. John Wiley & Sons, New York, NY.

Ding, P., Li, X., and Miratrix, L. W. (2017). Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2).

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of Ministry of Agriculture*, 33:503–513.

Fogarty, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *J. Roy. Statist. Soc. Ser. B*, 80(5):1035–1056.

Freedman, D. A. (2008a). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193.

Freedman, D. A. (2008b). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.*, 2(1):176–196.

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis and Interpretation*. Norton, New York.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.*, 99(467):609–618.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist*, 15(3):609–627.

Higgins, M. J., Sävje, F., and Sekhon, J. S. (2015). Blocking estimators and inference under the Neyman-Rubin model. *arXiv preprint arXiv:1510.01103*.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19(3):285–292.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24.

Iacus, S. M., King, G., and Porro, G. (2016). *cem: Coarsened Exact Matching.* R package version 1.1.17.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Stat. Med.*, 27(24):4857–4873.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *J. Roy. Statist. Soc. Ser. A*, 171(2):481–502.

Imbens, G. W. (2011). Experimental design for unit and cluster randomid trials. *Conf. International Initiative for Impact Evaluation, Cuernavaca.*

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, New York.

Kosmidis, I. (2017). *brglm: Bias Reduction in Binary-Response Generalized Linear Models.* R package version 0.6.1.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.*, 7(1):295–318.

Lohr, S. L. (2009). *Sampling: Design and Analysis.* Cengage Learning, Boston, 2nd edition.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth weighting? How to think about and use weights in survey experiments. *Political Analysis*, 26(3):275–291.

Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. Roy. Statist. Soc. Ser. B*, 75(2):369–396.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B*, 53(3):597–610.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer Series in Statistics. Springer, New York.

Rosenbaum, P. R. and Zhao, Q. (2017). *CrossScreening: Cross-Screening in Observational Studies that Test Many Hypotheses*. R package version 0.1.1.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.*, 75(371):591–593.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer, New York.

Sävje, F. (2015). The performance and efficiency of threshold blocking. *arXiv preprint arXiv:1506.02824*.

Schochet, P. Z. (2016). Statistical theory for the RCT-YES software: Design-based causal inference for RCTs, Second Edition. Technical Report (NCEE 2015-4011), Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Scosyrev, E. (2014). Causal inference in block-randomized experiments: Analysis based on Neyman's stochastic causal model. Unpublished.

Snedecor, G. and Cochran, W. (1989). *Statistical Methods*. Iowa State University Press, 8th edition.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.*, 5(4):465–472.

StataCorp (2017). Stata Statistical Software: Release 15. *StataCorp. College Station, TX: StataCorp*.

Wu, C. F. J. and Hamada, M. S. (2000). *Experiments : Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, New York, NY.

Zhao, Q., Small, D. S., and Rosenbaum, P. R. (2018). Cross-screening in observational studies that test many hypotheses. *J. Amer. Statist. Assoc.*, 113(523):1070–1084.

Supplementary Material

for

“Insights on Variance Estimation for Blocked and Matched Pairs Designs”

Nicole E. Pashley and Luke W. Miratrix

These Supplementary Material primarily contain detailed derivations of the results in the main paper, as well as some additional results and discussion. We first give some detail on the notational elements in the main paper. We then proceed with a non-technical discussion of alternate estimators for variance estimation. We next give derivations for the results in the main paper. We finally provide additional results and further details on simulations. More detailed proofs for some of these sections are available upon request.

A Details on notation and definitions

We here make the elements used in the presented results in the main paper more explicit. First, the mean of the potential outcomes for the units in the sample under treatment z is defined as

$$\bar{Y}(z) = \frac{1}{n}\sum_{i=1}^{n} Y_i(z).$$

The sample variance of potential outcomes under treatment z is

$$S^2(z) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(z) - \bar{Y}(z))^2.$$

The sample variance of the individual level treatment effects is

$$S^2(tc) = \frac{1}{n-1}\sum_{i=1}^{n}\Big(Y_i(t) - Y_i(c) - \tau_{\mathcal{S}}\Big)^2.$$

$\bar{Y}_k(z)$, $S^2_k(z)$, and $S^2_k(tc)$ are defined analogously over the units in block k. The blocked estimator is as in the main paper. The complete randomization simple difference estimator is then, using $Y_i^{obs} = Y_i(Z_i)$,

$$\widehat{\tau}_{(CR)} = \frac{1}{n_t}\sum_{i=1}^{n}\mathbb{I}_{Z_i=t}Y_i(t) - \frac{1}{n_c}\sum_{i=1}^{n}(1-\mathbb{I}_{Z_i=t})Y_i(c).$$

When looking at observed data, we have the variance of the observed outcomes in treatment arm z as

$$s^2(z) = \frac{1}{n_z - 1} \sum_{i:Z_i = z} \left(Y_i(z) - \bar{Y}^{obs}(z) \right)^2 . \tag{13}$$

These are the components of the big block variance estimates and Neyman's classic estimator for uncertainty.

B Alternative strategies for variance estimation

In the main paper we examined strategies for variance estimation that put no structure on how the individual blocks may differ from each other. At root, the focus is on estimating the residual variance of units around their block means, and aggregating appropriately.

This section discusses alternatives along with when they might be more or less appropriate. The first two subsections describe estimators that require model assumptions. Therefore these estimators may perform well in certain circumstances, but involve assumptions that we do not make in our analysis. The third subsection describes an estimator used for matched pairs that is proposed in the RCT-YES software documentation. The RCT-YES estimator assumes a specific population model that we do not consider in our paper.

B.1 Linear regression

Perhaps the most common method of estimation for randomized trials is to simply fit a linear model to the data with a treatment indicator and a dummy variable for each block. If there is no interaction between the treatment and block dummies, this approach will produce a precision-weighted estimate of the treatment effect, with an overall implicit estimand of a weighted average of the average impacts within each block, weighted by the estimated block precision under a homoscedasticity assumption. If there is impact variation correlated with this precision, then this precision-weighted estimand could be different than the overall ATE, resulting in a biased estimator. Furthermore, if blocks have different proportions of treated units and different sizes, this weighting might not correspond to any easily interpretable quantity. As pointed out by Freedman, this regression model is also not justified by randomization, which results in complications with the corresponding standard errors (Freedman, 2008a,b). Unfortunately, however, this approach is likely the most common in the field.

The above issues are, in part, repairable. Lin (2013) shows that the estimator from a linear model including interactions between the treatment indicator and block dummies is unbiased. In fact, this estimator is equivalent to the blocked estimator presented in this paper. The question is then how to estimate the standard errors from within the ordinary least squares framework. Lin advocates a Huber-White sandwich estimator for the general covariate case, but these have problems when the blocks have single treated or single control units. In particular, several variants of these estimators, such as HC2 and HC3, will not even be defined due to the characteristics of the corresponding design matrix. The HC0 estimator can still be heavily biased if there is systematic heteroscedasticity across the blocks. Gerber and Green (2012) (p. 116-117) advocate a weighted estimator, but this can also fail in the presence of blocks with singleton treated or control units.

B.2 Pooling variance estimates

As we have seen, the most straightforward way to get an overall variance is to obtain variance estimates for all of the block specific estimates, and then combine them in a weighted average. For small blocks, obtaining these block specific estimates is difficult because we do not have enough units to estimate variances of the treatment arms. In this case we have two general options: When treatment effects are considered to be homogenous, we can use the variance of the estimates across the blocks knowing that the variability is dictated purely by block mean variability and not treatment effect variability. An alternative strategy is to use the variance estimates in other blocks to estimate the variances in the tractable blocks. We discuss this next.

One such estimator is to, given a means of assessing how similar blocks are in terms of variance, simply use the variance of the closest big block for each small block. This typically requires some assumptions that, based on covariate values of the blocks, the variances of the potential outcomes are the same or similar. Similarly, Abadie and Imbens created an estimator of the variance for matched pairs that involves pairing the closest matched pairs and creating a pooled variance estimator for the two blocks together (Abadie and Imbens, 2008). They found that their estimator was asymptotically unbiased given certain conditions, such as the closeness of pairs increasing as the sample size grew. Although the asymptotic results derived in their paper are not necessarily appropriate here, this could be a reasonable plug-in estimator under the assumptions that (i) the covariate(s) that create the strata are related to the potential outcomes and variance and that (ii) the small strata are more similar to each other than the larger strata.

Covariates could also be exploited using linear regression to create variance estimates; see, for example, Fogarty (2018). Or, if we believe that the variance of the estimator in each block is related to the block size, we could fit a linear regression for the big blocks, of variance versus their size, and then extrapolate to the small blocks. Alternatively, if nothing is known and there are very few small blocks, an average (or the largest) of the big block variance estimates might be used. This type of plug-in is used in simulations in Section 6.

Any of these plug-in estimators could be used, instead of plugging in for the overall variance, to simply fill in the missing component for small blocks. For instance, if all of the small blocks are such that they have multiple controls but only one treated unit, we can calculate $s_k^2(c)$ as usual but approximate $s_k^2(t)$ based on one of the previously mentioned methods.

There are many other plug-ins that might be appropriate, based on what assumptions the researcher is able to make. The choice of plug-in estimator should be chosen prior to running the experiment and should be based on the researchers assumptions and knowledge at that time. Trying several plug-in estimators and using the smallest will create bias.

B.3 The RCT-YES estimator

One might also consider an estimator suggested in the RCT-YES manuscript (Schochet, 2016, p. 83). The form of this estimator, using block sizes as weights, is

$$\widehat{\sigma}^2_{RCT} = \frac{1}{K\,(K-1)\,(\frac{n}{K})^2} \sum_{k=1}^{K} \left(n_k \widehat{\tau}_k - \frac{n}{K} \widehat{\tau}_{(BK)} \right)^2 .$$

As discussed and proven in the RCT-YES documentation, this estimator is consistent under an infinite population of an infinite number of strata of infinite size, where we sample strata and then units within strata. This is the random sampling of strata setting in the remark of Section 4.4. This estimator differs from our estimators $\widehat{\sigma}^2_{(SMALL/m)}$ and $\widehat{\sigma}^2_{(SMALL/p)}$ by putting the weights inside the square. Unfortunately, moving the weighting inside the squares can cause large bias in the finite setting and the stratified sampling framework. In fact, in the simulations comparing variance estimator performance in the finite sample, presented in Section 6, the RCT-YES bias and variance was high enough that it was not comparable to the other estimators presented. This estimator is targeting a superpopulation quantity, thus the standard errors are larger in part to capture the additional variation of the strata being a random sample.

We discussed the performance of the original RCT-YES estimator with Dr. Schochet (personal correspondence, April 2018), and he proposed an alternate estimator. Again using the block sizes as weights, this estimator has the form

$$\widehat{\sigma}^2_{RCT,2} = \frac{1}{K\left(K-1\right)\left(\frac{n}{K}\right)^2}\sum_{k=1}^{K} n_k^2\left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2$$

and is rooted in survey sampling methods (Cochran, 1977). This estimator is more stable because the weights are not inside the parentheses. This estimator is still motivated by a superpopulation sampling framework, and takes the variability of the blocks into account. Finite sample performance using this estimator on all of the blocks does not perform well, unless all blocks have the same τ_k, which aligns with what we expect from explorations of our small block estimators. If used in combination as a hybrid estimator, its performance is very similar to that of the hybrid using $\widehat{\sigma}^2_{(SMALL/p)}$.

C Derivations of blocking versus complete randomization differences

Unless otherwise noted, the following section assumes that $p_k = p$ for all $k = 1, \ldots, K$.

C.1 Finite sample, Equation (11)

We start by deriving the result of Equation (11),

$$
\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right)\\
&\quad=\sum_{k=1}^{K}\frac{1}{n(n-1)}\Bigg[n_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c)+\sqrt{\frac{1-p}{p}}\bar{Y}_k(t)-\left(\sqrt{\frac{p}{1-p}}\bar{Y}(c)+\sqrt{\frac{1-p}{p}}\bar{Y}(t)\right)\right)^2\\
&\quad-\frac{n-n_k}{n(n_k-1)}\sum_{i:b_i=k}\left(\sqrt{\frac{p}{1-p}}Y_i(c)+\sqrt{\frac{1-p}{p}}Y_i(t)-\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c)+\sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right)\right)^2\Bigg].
\end{aligned}
$$

We have from usual results (Imbens and Rubin, 2015, p. 89), in addition to independence of treatment assignment within blocks and the assumption that $p_k = p$ for $k = 1, ..., K$, that

$$
\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) = \frac{S^2(c)}{n_c} + \frac{S^2(t)}{n_t} - \frac{S^2(tc)}{n} = \frac{S^2(c)}{(1-p)n} + \frac{S^2(t)}{pn} - \frac{S^2(tc)}{n}
$$

and

$$
\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) = \sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right) = \sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{(1-p)n_k} + \frac{S_k^2(t)}{pn_k} - \frac{S_k^2(tc)}{n_k}\right).
$$

To take the difference, we need to write the complete randomization variance in terms of the block components. First we look at the expansion of $S^2(z)$:

$$
\begin{aligned}
S^2(z) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i(z)-\bar{Y}(z)\right)^2\\
&= \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i:b_i=k}\left(Y_i(z)-\bar{Y}_k(z)+\bar{Y}_k(z)-\bar{Y}(z)\right)^2\\
&= \frac{1}{n-1}\sum_{k=1}^{K}\left[(n_k-1)S_k^2(z)+n_k\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2\right]\\
&= \sum_{k=1}^{K}\frac{n_k-1}{n-1}S_k^2(z)+\sum_{k=1}^{K}\frac{n_k}{n-1}\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2.
\end{aligned}
$$

Now do the same for $S^2(tc)$:

$$
S^2(tc) = \sum_{k=1}^{K}\frac{n_k-1}{n-1}S_k^2(tc)+\sum_{k=1}^{K}\frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}}\right)^2.
$$

$$
\begin{aligned}
&\operatorname{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \operatorname{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\
&= \frac{S^2(c)}{(1-p)n} + \frac{S^2(t)}{pn} - \frac{S^2(tc)}{n} - \left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{(1-p)n_k} + \frac{S_k^2(t)}{pn_k} - \frac{S_k^2(tc)}{n_k}\right)\right] \\
&= \frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(c) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n} + \frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(t) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn} \\
&- \frac{\sum_{k=1}^{K} \frac{n_k-1}{n-1} S_k^2(tc) + \sum_{k=1}^{K} \frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2}{n} \\
&- \left[\sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{(1-p)n_k} + \frac{S_k^2(t)}{pn_k} - \frac{S_k^2(tc)}{n_k}\right)\right] \\
&= \underbrace{\frac{\sum_{k=1}^{K} n_k\left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n(n-1)} + \frac{\sum_{k=1}^{K} n_k\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn(n-1)} - \frac{\sum_{k=1}^{K} n_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2}{n(n-1)}}_{\mathbf{A}} \\
&\underbrace{- \sum_{k=1}^{K}\Big[\left(\frac{n_k}{(1-p)n^2} - \frac{n_k-1}{(1-p)n(n-1)}\right) S_k^2(c) + \left(\frac{n_k}{pn^2} - \frac{n_k-1}{pn(n-1)}\right) S_k^2(t) - \left(\frac{n_k}{n^2} - \frac{n_k-1}{n(n-1)}\right) S_k^2(tc)\Big]}_{\mathbf{B}}
\end{aligned}
$$

We have now split our calculation into the between block variation and within block variation pieces.

A is the between block variation:

We need to expand $\sum_{k=1}^{K} n_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2$:

$$
\begin{aligned}
\sum_{k=1}^{K} n_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 &= \sum_{k=1}^{K} n_k\left(\bar{Y}_k(t) - \bar{Y}_k(c) - (\bar{Y}(t) - \bar{Y}(c))\right)^2 \\
&= \sum_{k=1}^{K} n_k\left[\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2 + \left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2 - 2\left(\bar{Y}_k(t) - \bar{Y}(t)\right)\left(\bar{Y}(c) - \bar{Y}(c)\right)\right]
\end{aligned}
$$

So then

$$
\begin{aligned}
\mathbf{A} =& \frac{\sum_{k=1}^K n_k \left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n(n-1)} + \frac{\sum_{k=1}^K n_k \left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn(n-1)} \\
&- \frac{\sum_{k=1}^K n_k \left[\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2 + \left(\bar{Y}(c) - \bar{Y}(c)\right)^2 - 2\left(\bar{Y}_k(t) - \bar{Y}(t)\right)\left(\bar{Y}(c) - \bar{Y}(c)\right)\right]}{n(n-1)} \\
=& \frac{\sum_{k=1}^K pn_k \left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n(n-1)} + \frac{\sum_{k=1}^K (1-p)n_k \left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn(n-1)} + 2\frac{\left(\bar{Y}_k(t) - \bar{Y}(t)\right)\left(\bar{Y}(c) - \bar{Y}(c)\right)}{n(n-1)} \\
=& \frac{1}{n-1}\sum_{k=1}^K \frac{n_k}{n}\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}(c) + \sqrt{\frac{1-p}{p}}\bar{Y}(t)\right)\right)^2 \\
=& \frac{1}{(n-1)}\mathrm{Var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right).
\end{aligned}
$$

B is the within block variation:

$$
\begin{aligned}
\mathbf{B} =& \sum_{k=1}^K \Big[\left(\frac{n-n_k}{(1-p)n^2(n-1)}\right)S_k^2(c) + \left(\frac{n-n_k}{pn^2(n-1)}\right)S_k^2(t) - \left(\frac{n-n_k}{n^2(n-1)}\right)S_k^2(tc)\Big] \\
=& \frac{1}{n^2(n-1)}\sum_{k=1}^K (n-n_k)n_k\Big[\frac{S_k^2(c)}{(1-p)n_k} + \frac{S_k^2(t)}{pn_k} - \frac{S_k^2(tc)}{n_k}\Big] \\
=& \frac{1}{n^2(n-1)}\sum_{k=1}^K (n-n_k)n_k \mathrm{var}(\widehat{\tau}_k|\mathcal{S})
\end{aligned}
$$

Then we can write the difference as

$$
\begin{aligned}
&\mathrm{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \mathrm{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) \\
&\quad= \frac{1}{n-1}\left[\mathrm{var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right) - \sum_{k=1}^K \frac{(n-n_k)n_k}{n^2}\mathrm{var}(\widehat{\tau}_k|\mathcal{S})\right] \\
&\quad= \sum_{k=1}^K \frac{n_k}{n(n-1)}\Big[\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}(c) + \sqrt{\frac{1-p}{p}}\bar{Y}(t)\right)\right)^2 \\
&\qquad - \frac{n-n_k}{n}\mathrm{var}(\widehat{\tau}_k|\mathcal{S})\Big].
\end{aligned}
$$

To write this another way, note that

$$
\begin{aligned}
S_k^2(tc) =& \frac{1}{n_k-1}\sum_{i:b_i=k}\left(\tau_i - \tau_{fs,k}\right)^2 \\
=& \frac{1}{n_k-1}\sum_{i:b_i=k}\Big[\left(Y_i(t) - \bar{Y}_k(t)\right)^2 + \left(Y_i(c) - \bar{Y}_k(c)\right)^2 - 2\left(Y_i(t) - \bar{Y}_k(t)\right)\left(Y_i(c) - \bar{Y}_k(c)\right)\Big].
\end{aligned}
$$

So then

$$
\begin{aligned}
&\frac{S_k^2(c)}{1-p} + \frac{S_k^2(t)}{p} - S_k^2(tc)\\
&= \frac{\frac{1}{n_k-1}\sum_{i:b_i=k}\left(Y_i(c)-\bar{Y}_k(c)\right)^2}{1-p} + \frac{\frac{1}{n_k-1}\sum_{i:b_i=k}\left(Y_i(t)-\bar{Y}_k(t)\right)^2}{p}\\
&- \frac{1}{n_k-1}\sum_{i:b_i=k}\left[\left(Y_i(t)-\bar{Y}_k(t)\right)^2 + \left(Y_i(c)-\bar{Y}_k(c)\right)^2 - 2\left(Y_i(t)-\bar{Y}_k(t)\right)\left(Y_i(c)-\bar{Y}_k(c)\right)\right]\\
&= \sum_{i:b_i=k}\frac{1}{n_k-1}\left[\frac{p\left(Y_i(c)-\bar{Y}_k(c)\right)^2}{1-p} + \frac{(1-p)\left(Y_i(t)-\bar{Y}_k(t)\right)^2}{p} + 2\left(Y_i(t)-\bar{Y}_k(t)\right)\left(Y_i(c)-\bar{Y}_k(c)\right)\right]\\
&= \sum_{i:b_i=k}\frac{1}{n_k-1}\left(\sqrt{\frac{p}{1-p}}\left(Y_i(c)-\bar{Y}_k(c)\right) + \sqrt{\frac{1-p}{p}}\left(Y_i(t)-\bar{Y}_k(t)\right)\right)^2\\
&= \sum_{i:b_i=k}\frac{1}{n_k-1}\left(\sqrt{\frac{p}{1-p}}Y_i(c) + \sqrt{\frac{1-p}{p}}Y_i(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right)\right)^2.
\end{aligned}
$$

So we get

$$
\begin{aligned}
&\operatorname{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \operatorname{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right)\\
&\qquad= \sum_{k=1}^{K}\frac{1}{n(n-1)}\Bigg[n_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}(c) + \sqrt{\frac{1-p}{p}}\bar{Y}(t)\right)\right)^2\\
&\qquad\quad - \frac{n-n_k}{n(n_k-1)}\sum_{i:b_i=k}\left(\sqrt{\frac{p}{1-p}}Y_i(c) + \sqrt{\frac{1-p}{p}}Y_i(t) - \left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right)\right)^2\Bigg].
\end{aligned}
$$

This allows us to see the similarity in the two terms.

C.2 Simple random sampling

The simple random sampling framework has two steps: First, obtain a random sample of units for the experiment. Then, we form blocks. We assume there is a procedure for every sample that clusters on the observed covariate matrix $\boldsymbol{X}$ to form blocks for the given sample. The number and size of blocks can be sample dependent. Now think about the the "worst" case, that this set of covariates, $\boldsymbol{X}$, are actually independent of the potential outcomes.

Then we have, assuming equal proportions treated in each block,

$$
\begin{aligned}
&\text{var}\left(\widehat{\tau}_{(CR)}|SRS\right) - \text{var}\left(\widehat{\tau}_{(BK)}|SRS\right)\\
&\quad= \mathbb{E}\left[\text{var}(\widehat{\tau}_{(CR)}|\mathcal{S})|SRS\right] + \text{var}\left(\mathbb{E}[\widehat{\tau}_{(CR)}|\mathcal{S}]|SRS\right) - \mathbb{E}\left[\text{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\text{SRS}\right] - \text{var}\left(\mathbb{E}[\widehat{\tau}_{(BK)}|\mathcal{S}]|SRS\right)\\
&\quad= \mathbb{E}\left[\frac{S^2(c)}{n_c} + \frac{S^2(t)}{n_t} - \frac{S^2(tc)}{n}\Big|SRS\right] - \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right]\\
&\quad= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n} - \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}\left(\frac{S_k^2(c)}{n_c} + \frac{S_k^2(t)}{n_t} - \frac{S_k^2(tc)}{n}\right)\Big|SRS\right]\\
&\quad= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n} - \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}\left(\frac{S_k^2(c)}{n_c} + \frac{S_k^2(t)}{n_t} - \frac{S_k^2(tc)}{n}\right)\Big|SRS, \boldsymbol{X}\right]\Big|SRS\right]\\
&\quad= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n} - \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}\left(\frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n}\right)\Big|SRS\right]\\
&\quad= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n} - \left(\frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t} - \frac{\sigma^2(tc)}{n}\right)\\
&\quad= 0.
\end{aligned}
$$

Note that conditioning on $\boldsymbol{X}$ fixes the number of units in each block and number of blocks, but potential outcomes are independent of this value of the covariates so we can push the expectation through.

C.3 Proof of Theorem 5.3.1: Variance comparison under stratified sampling

Proof. First use decomposition of variance and then simplify using results from the derivation in Appendix C.1. Let $\mu(z)$ and $\sigma^2(z)$ be the population mean and variance, respectively, of the potential outcomes of all units under treatment z. Let $\mu_k(z)$ and $\sigma_k^2(z)$ be the population mean and variance of the potential outcomes of units in stratum k under treatment z.

$$
\begin{aligned}
&\operatorname{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) - \operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) \\
=&\mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(CR)}|\mathcal{S})|\mathcal{F}_1\right] + \operatorname{var}\left(\mathbb{E}[\widehat{\tau}_{(CR)}|\mathcal{S}]|\mathcal{F}_1\right) - \mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\mathcal{F}_1\right] + \operatorname{var}\left(\mathbb{E}[\widehat{\tau}_{(BK)}|\mathcal{S}]|\mathcal{F}_1\right) \\
=&\mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(CR)}|\mathcal{S})|\mathcal{F}_1\right] - \mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\mathcal{F}_1\right] \\
=&\mathbb{E}\Big[\underbrace{\frac{\sum_{k=1}^K n_k\left(\bar{Y}_k(c)-\bar{Y}(c)\right)^2}{(1-p)(n-1)n} + \frac{\sum_{k=1}^K n_k\left(\bar{Y}_k(t)-\bar{Y}(t)\right)^2}{p(n-1)n}}_{\mathbf{A}} - \underbrace{\frac{\sum_{k=1}^K n_k(\tau_{\mathcal{S}}-\tau_{k,\mathcal{S}})^2}{n(n-1)}}_{\mathbf{B}}|\mathcal{F}_1\Big] \\
&-\sum_{k=1}^K \frac{n-n_k}{n^2(n-1)}\left(\frac{\sigma_k^2(c)}{(1-p)} + \frac{\sigma_k^2(t)}{p} - \sigma_k^2(tc)\right) \qquad (14)
\end{aligned}
$$

We start with expectation of the numerators in term **A**:

$$
\begin{aligned}
\mathbb{E}\left[\left(\bar{Y}_k(z)-\bar{Y}(z)\right)|\mathcal{F}_1\right]^2 &= \mathbb{E}\left[\bar{Y}_k(z)^2 - 2\bar{Y}_k(z)\bar{Y}(z) + \bar{Y}(z)^2|\mathcal{F}_1\right] \\
&= \underbrace{\mathbb{E}\left[\bar{Y}_k(z)^2|\mathcal{F}_1\right]}_{\mathbf{A.1}} - 2\mathbb{E}\left[\bar{Y}_k(z)\bar{Y}(z)|\mathcal{F}_1\right] + \underbrace{\mathbb{E}\left[\bar{Y}(z)^2|\mathcal{F}_1\right]}_{\mathbf{A.2}}.
\end{aligned}
$$

A.1:

$$
\mathbb{E}\left[\bar{Y}_k(z)^2|\mathcal{F}_1\right] = \operatorname{var}\left(\bar{Y}_k(z)|\mathcal{F}_1\right) + \mathbb{E}\left[\bar{Y}_k(z)|\mathcal{F}_1\right]^2 = \frac{\sigma_k^2(z)}{n_k} + \mu_k(z)^2
$$

A.2:

$$
\mathbb{E}\left[\bar{Y}(z)^2|\mathcal{F}_1\right] = \operatorname{var}\left(\bar{Y}(z)|\mathcal{F}_1\right) + \mathbb{E}\left[\bar{Y}(z)|\mathcal{F}_1\right]^2 = \sum_{k=1}^K \frac{n_k}{n^2}\sigma_k^2(z) + \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(z)\right)^2
$$

Putting **A.1** and **A.2** together and combining like terms:

$$
\begin{aligned}
\sum_{k=1}^k \frac{n_k}{n}\mathbb{E}\left[\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2|\mathcal{F}_1\right] &= \sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\left[\bar{Y}_k(z)^2|\mathcal{F}_1\right] - \mathbb{E}\left[\bar{Y}(z)^2|\mathcal{F}_1\right] \\
&= \sum_{k=1}^K \frac{n-n_k}{n^2}\sigma_k^2(z) + \sum_{k=1}^K \frac{n_k}{n}\mu_k(z)^2 - \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(z)\right)^2.
\end{aligned}
$$

The expectation of **B** follows in a very similar manner, except now we have treated and control units in the calculation.

B becomes

$$
\begin{aligned}
&\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\left[\left(\tau_{\mathcal{S}}-\tau_{k,\mathcal{S}}\right)^2|\mathcal{F}_1\right] \\
=&\sum_{k=1}^K \frac{n-n_k}{n^2}\sigma_k^2(tc) + \sum_{k=1}^K \frac{n_k}{n}\left(\mu_k(t)^2-\mu_k(c)\right)^2 - \left[\left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(t)\right) - \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(c)\right)\right]^2.
\end{aligned}
$$

Putting **A** and **B** into Equation 14 and simplifying:

$$
\begin{aligned}
&\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) - \text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) \\
&= \frac{\sum_{k=1}^K \frac{n_k}{n}\mu_k(c)^2 - \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(c)\right)^2}{(1-p)(n-1)} + \frac{\sum_{k=1}^K \frac{n_k}{n}\mu_k(t)^2 - \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(t)\right)^2}{p(n-1)} \\
&- \frac{\sum_{k=1}^K \frac{n_k}{n}\left(\mu_k(t) - \mu_k(c)\right)^2 - \left(\sum_{k=1}^K \frac{n_k}{n}\left(\mu_k(t) - \mu_k(c)\right)\right)^2}{n-1} \\
&= \frac{p}{1-p}\frac{\sum_{k=1}^K \frac{n_k}{n}\mu_k(c)^2 - \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(c)\right)^2}{n-1} + \frac{1-p}{p}\frac{\sum_{k=1}^K \frac{n_k}{n}\mu_k(t)^2 - \left(\sum_{k=1}^K \frac{n_k}{n}\mu_k(t)\right)^2}{n-1} \\
&+ 2\frac{\sum_{k=1}^K \frac{n_k}{n}\mu_k(t)\mu_k(c) - \left(\sum_{k=1}^K \frac{n_k}{n}\mu(t,k)\right)\left(\sum_{k=1}^K \frac{n_k}{n}\mu(c,k)\right)}{n-1} \\
&= \frac{1}{n-1}\left[\frac{p}{1-p}\text{Var}_k\left(\mu_k(c)\right) + \frac{1-p}{p}\text{Var}_k\left(\mu_k(t)\right) + 2Cov_k\left(\mu_k(c), \mu_k(t)\right)\right] \\
&= \frac{1}{n-1}\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\mu_k(t) + \sqrt{\frac{1-p}{p}}\mu_k(c)\right) \geq 0.
\end{aligned}
$$

The variance in the last line is the variance over the blocks, as defined in Equation 12. Therefore we have that $\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) - \text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) \geq 0$ so we are always doing better with blocking in this setting. □

C.4 Stratified sampling vs SRS comparisons

Corollary C.4.1. *The difference between $var(\widehat{\tau}_{(CR)}|SRS) - var\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right)$ may be positive or negative.*

Proof. Compare the previous result to when we assume SRS for complete randomization and $\mathcal{F}_1$ for blocked randomization.

$$
\begin{aligned}
&\text{var}(\widehat{\tau}_{(CR)}|SRS) - \text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1) \\
&\quad = \frac{1}{n_c}\left(\sum_{k=1}^K \frac{n_k}{n}\left(\mu_k(c) - \mu(c)\right)^2\right) + \frac{1}{n_t}\left(\sum_{k=1}^K \frac{n_k}{n}\left(\mu_k(t) - \mu(t)\right)^2\right) \\
&\quad \geq 0
\end{aligned}
$$

A form of this result can be found in Imbens (2011).

We can then take the difference, $\text{var}(\widehat{\tau}_{(CR)}|SRS) - \text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right)$, between the results under the two frameworks to see whether using different sampling frameworks over or under estimates the benefits of blocking. First consider the form of the variance under the two different models.

$$\begin{aligned}\text{var}(\widehat{\tau}_{(CR)}|SRS) &= \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(t)}{n_t}\\ &= \frac{\sum_{k=1}^{K} \frac{n_k}{n}\sigma_k^2(c) + \sum \frac{n_k}{n}\left(\mu_k(c) - \mu(c)\right)^2}{n_c} + \frac{\sum_{k=1}^{K} \frac{n_k}{n}\sigma_k^2(t) + \sum \frac{n_k}{n}\left(\mu_k(t) - \mu(t)\right)^2}{n_t}\end{aligned}$$

$$\begin{aligned}\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) &= \mathbb{E}\left[\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right)|\mathcal{F}_1\right] + \text{var}\left(\mathbb{E}\left[\widehat{\tau}_{(CR)}|\mathcal{S}\right]|\mathcal{F}_1\right)\\ &= \underbrace{\mathbb{E}\left[\frac{S^2(c)}{n_c} + \frac{S^2(t)}{n_t} - \frac{S^2(tc)}{n}|\mathcal{F}_1\right]}_{\mathbf{A}} + \underbrace{\text{var}\left(\tau_{\mathcal{S}}|\mathcal{F}_1\right)}_{\mathbf{B}}\end{aligned}$$

For **A** we can use similar techniques from previous proofs to break the pieces up by block:

$$\begin{aligned}\mathbb{E}\left[\frac{S^2(tc)}{n}|\mathcal{F}_1\right] &= \sum_{k=1}^{K}\left[\frac{n_k}{n^2}\sigma_k^2(tc) + \frac{n_k}{n(n-1)}\left(\tau_k - \tau\right)^2\right]\\ \mathbb{E}\left[\frac{S^2(z)}{n}|\mathcal{F}_1\right] &= \frac{1}{n_z}\sum_{k=1}^{K}\left[\frac{n_k}{n}\sigma_k^2(z) + \frac{n_k}{n-1}\left(\mu_k(z) - \mu(z)\right)^2\right]\end{aligned}$$

For **B** we can split up by block and then use classical results from sampling theory on variation of sample means under SRS (Lohr, 2009, Chapter 2).

$$\text{var}\left(\tau_{\mathcal{S}}|\mathcal{F}_1\right) = \sum_{k=1}^{K}\frac{n_k^2}{n^2}\frac{\sigma_k^2(tc)}{n_k}$$

Putting **A** and **B** together and collecting similar terms, we have

$$\begin{aligned}\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_1) = &\sum_{k=1}^{K}\left[\frac{n_k}{n}\frac{\sigma_k^2(c)}{n_c} + \frac{n_k}{n}\frac{\sigma_k^2(t)}{n_t}\right] + \sum_{k=1}^{K}\left[\frac{n_k}{(n-1)n_c}\left(\mu_k(c) - \mu(c)\right)^2 + \frac{n_k}{(n-1)n_t}\left(\mu_k(t) - \mu(t)\right)^2\right]\\ &- \sum_{k=1}^{K}\frac{n_k}{n(n-1)}\left(\tau_k - \tau\right)^2.\end{aligned}$$

So then we can do the subtraction:

$$\begin{aligned}
&\mathrm{var}(\widehat{\tau}_{(CR)}|SRS) - \mathrm{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right)\\
&= \frac{1}{n(n-1)}\sum_{k=1}^{K} n_k\left[(\tau_k-\tau)^2 - \frac{(\mu_k(c)-\mu(c))^2}{n_c} - \frac{(\mu_k(t)-\mu(t))^2}{n_t}\right]\\
&= \sum_{k=1}^{K}\frac{n_k}{n(n-1)}\Bigg[\frac{(n_c-1)\left(\mu_k(c)-\mu(c)\right)^2}{n_c} + \frac{(n_t-1)\left(\mu_k(t)-\mu(t)\right)^2}{n_t}\\
&\qquad - 2\left(\mu_k(c)-\mu(c)\right)\left(\mu_k(t)-\mu(t)\right)\Bigg].
\end{aligned}$$

We see that this difference could be positive or negative. In particular, if the blocks all have the same average treatment effect but different control and treatment means, then this expression will be negative, indicating that comparing the variance of blocking under $\mathcal{F}_1$ to complete randomization under SRS is an underestimate of the benefits of blocking. On the other hand, if n_c and n_t are large compared to the variation of block average control and treatment outcomes, then we would expect the negative terms in the first expression to be small, resulting in the difference being positive. This would mean that there is an overestimate of the benefits of blocking when comparing the variance of blocking under $\mathcal{F}_1$ to complete randomization under SRS.

If we have n_c and n_t large so $\frac{n_c-1}{n_c} \approx 1 \approx \frac{n_t-1}{n_t}$ then we have

$$\mathrm{var}\left(\widehat{\tau}_{(CR)}|SRS\right) - \mathrm{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) = \frac{1}{n(n-1)}\sum_{k=1}^{K} n_k\left(\tau_k-\tau\right)^2.$$

□

C.5 Unequal treatment proportions

Our comparison of complete randomization to blocking in the main paper only applies to the small slice of possible experiments in which the treatment proportion is equal across all blocks. In practice, however, the proportion treated, p_k, may be unequal across blocks, and in this case the above results are not guaranteed to hold. In particular, with blocks of variable size, it can be difficult to have the same proportion treated within each block due to the discrete nature of units.

With varying p_k, the units within each block are weighted differently than they would be in a complete randomization when calculating a treatment effect estimate. That is, in a complete randomization, the treated units are all weighted proportional to $1/p$ but here the treated units in each block get weighted instead by $1/p_k$, meaning units with low probability of treatment will "count more" towards the overall treatment mean and their variability will have greater relevance for the overall variance of the estimator. Sävje (2015) also noted the effect of variable proportions treated on variance and Higgins et al. (2015) explored estimators for blocked designs with possibly unequal treatment proportions, but also multiple treatments. The costs here are similar to the costs of variable selection probabilities in survey sampling (see Särndal et al., 2003).

When different blocks have different proportions of units treated, it is possible to systematically have blocks and treatment groups with more variance to also have more weight, which could cause blocking to be harmful even in the stratified sampling setting of Section 5.3.

Theorem C.5.1 (Variance comparison with unequal treatment proportions)**.**

$$\begin{aligned}
&\mathit{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) - \mathit{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_1\right)\\
&= \frac{1}{n-1}\,\mathit{Var}_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right) + \sum_{k=1}^{K}\frac{(p-p_k)n_k}{n^2}\left[\frac{\sigma_k^2(c)}{(1-p_k)(1-p)} - \frac{\sigma_k^2(t)}{p_k p}\right].
\end{aligned}$$

Proof. Let the proportion of units treated in a complete randomization be p and in block randomization be p_k for block k. Again, we need to break the complete randomization variance into block components. For the finite sample, the complete randomization variance, from Appendix C.1 is

$$\begin{aligned}
\text{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) =& \frac{\sum_{k=1}^{K}\frac{n_k-1}{n-1}S_k^2(c) + \sum_{k=1}^{K}\frac{n_k}{n-1}\left(\bar{Y}_k(c) - \bar{Y}(c)\right)^2}{(1-p)n}\\
&+ \frac{\sum_{k=1}^{K}\frac{n_k-1}{n-1}S_k^2(t) + \sum_{k=1}^{K}\frac{n_k}{n-1}\left(\bar{Y}_k(t) - \bar{Y}(t)\right)^2}{pn}\\
&- \frac{\sum_{k=1}^{K}\frac{n_k-1}{n-1}S_k^2(tc) + \sum_{k=1}^{K}\frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2}{n}.
\end{aligned}$$

For block randomization, the variance is

$$\begin{aligned}
\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) &= \sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right)\\
&= \sum_{k=1}^{K}\frac{n_k}{n^2}\left(\frac{S_k^2(c)}{1-p_k} + \frac{S_k^2(t)}{p_k} - S_k^2(tc)\right).
\end{aligned}$$

Then the difference is

$$
\begin{aligned}
&\operatorname{var}\left(\widehat{\tau}_{(CR)}|\mathcal{S}\right) - \operatorname{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right)\\
&= \sum_{k=1}^{K}\left[\frac{(1-p_k)(n_k-1)n-(1-p)n_k(n-1)}{(1-p_k)(1-p)n^2(n-1)}S_k^2(c) + \frac{p_k(n_k-1)n-pn_k(n-1)}{p_kpn^2(n-1)}S_k^2(t)\right]\\
&+\sum_{k=1}^{K}\left[\frac{n-n_k}{(n-1)n^2}S_k^2(tc)\right] + \frac{\sum_{k=1}^{K} n_k\left[\frac{\left(\bar{Y}_k(c)-\bar{Y}(c)\right)^2}{1-p}+\frac{\left(\bar{Y}_k(t)-\bar{Y}(t)\right)^2}{p}-\left(\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}}\right)^2\right]}{(n-1)n}\\
&= \frac{1}{n^2(n-1)}\sum_{k=1}^{K}\Bigg(\frac{(1-p_k)(n_k-1)n-(1-p)n_k(n-1)}{(1-p_k)(1-p)}S_k^2(c) + \frac{p_k(n_k-1)n-pn_k(n-1)}{p_kp}S_k^2(t)\\
&\qquad + (n-n_k)S_k^2(tc)\Bigg) + \frac{\sum_{k=1}^{K} n_k\left[\frac{\left(\bar{Y}_k(c)-\bar{Y}(c)\right)^2}{1-p}+\frac{\left(\bar{Y}_k(t)-\bar{Y}(t)\right)^2}{p}-\left(\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}}\right)^2\right]}{(n-1)n}.
\end{aligned}
$$

For the stratified random sampling ($\mathcal{F}_1$), we use results from the proof in Appendix C.3 and the above derivation, and combine like terms to get

$$
\begin{aligned}
&\operatorname{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) - \operatorname{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_1\right)\\
&= \frac{1}{n-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\sqrt{\frac{1-p}{p}}\mu_k(c)+\sqrt{\frac{p}{1-p}}\mu_k(t)-\left[\sqrt{\frac{1-p}{p}}\mu(c)+\sqrt{\frac{p}{1-p}}\mu(t)\right]\right)^2\\
&+\sum_{k=1}^{K}\left[\frac{(p-p_k)n_k}{(1-p_k)(1-p)n^2}\sigma_k^2(c)+\frac{(p_k-p)n_k}{p_kpn^2}\sigma_k^2(t)\right].
\end{aligned}
$$

More details on the derivation can be given upon request.

This can also be written as

$$
\begin{aligned}
&\operatorname{var}\left(\widehat{\tau}_{(CR)}|\mathcal{F}_1\right) - \operatorname{var}\left(\widehat{\tau}_{(BK)}|\mathcal{F}_1\right)\\
&= \frac{1}{n-1}\operatorname{Var}_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c)+\sqrt{\frac{1-p}{p}}\mu_k(t)\right)+\sum_{k=1}^{K}\frac{(p-p_k)n_k}{n^2}\left[\frac{\sigma_k^2(c)}{(1-p_k)(1-p)}-\frac{\sigma_k^2(t)}{p_kp}\right].
\end{aligned}
\tag{15}
$$

The first term in Equation 15 is equal to the result in Proposition 5.3.1 and the second term, capturing the additional variability due to varying proportions, is zero when $p_k = p$ for all k. In general, the second term of Equation 15 can be positive or negative.

In particular, if by some bad luck, $pp_k\sigma_k^2(c) > (1-p)(1-p_k)\sigma_k^2(t)$ for all blocks where $p_k > p$ and $pp_k\sigma_k^2(c) < (1-p)(1-p_k)\sigma_k^2(t)$ for all blocks where $p_k < p$, this term will be negative. If, in this case, the population mean potential outcomes for all blocks are approximately equal, the entire expression will be negative. The two terms are of the same order with respect to sample size n, making comparison easier. If $p_k \approx p$ then the second term should not be too large. In small blocks, where one unit more in treatment or control can dramatically change the proportion of treated units, we would expect unequal proportions to have more of an impact. □

D Creation and bias of $\widehat{\sigma}^2_{(SMALL/m)}$

D.1 Creation of $\widehat{\sigma}^2_{(SMALL/m)}$, Equation (5)

Here we give an expanded explanation of $\widehat{\sigma}^2_{(SMALL/m)}$, Equation (5) from Section 3.2. To formally state this method, we first express our estimands and estimators in terms of weighted averages of estimates within collections of same-size blocks. In the following, let there be J unique block sizes in the sample. Let m_j be the jth block size and let K_j be the number of blocks in the population of size m_j. So then $n = \sum_{j=1}^J m_jK_j$. In particular, the sample average treatment effect for all units in blocks of size m_j is

$$\tau_{(SMALL),\mathcal{S},j} = \frac{1}{K_j}\sum_{k:n_k=m_j}\tau_{k,\mathcal{S}}.$$

Let $N_j = m_jK_j$ be the total number of units in the small blocks of size m_j. Then the overall sample average treatment effect in terms of these $\tau_{(SMALL),\mathcal{S},j}$ is

$$\tau_{(SMALL),\mathcal{S}} = \frac{1}{\sum_{i=1}^J N_j}\sum_{j=1}^J N_j\tau_{(SMALL),\mathcal{S},j}. \qquad (16)$$

$\tau_{(SMALL),\mathcal{S}}$ is the same as $\tau_{\mathcal{S}}$ as before; we add the subscript "small" here to clarify the notation when we discuss hybrid experiments. Note that these definitions are analogous for the infinite population, which are indicated by removing the $\mathcal{S}$ subscript.

The treatment effect estimators can be written in analogous form to the above. We have unbiased estimators $\widehat{\tau}_{(SMALL),j} = \frac{1}{K_j}\sum_{k:n_k=m_j}\widehat{\tau}_k$ for the average treatment effects in blocks of size m_j. Simply plug them into Equation 16 to obtain an overall treatment effect estimator.

As discussed in Section 3.2, within each piece j, use a variance estimator with the same form as Equation 4:

$$\widehat{\sigma}^2_{(SMALL),j} = \frac{1}{K_j(K_j-1)} \sum_{k:n_k=m_j} (\widehat{\tau}_k - \widehat{\tau}_{(SMALL),j})^2.$$

Then combine to create an overall variance estimator:

$$\widehat{\sigma}^2_{(SMALL/m)} = \frac{1}{\left(\sum_{j=1}^J N_j\right)^2} \sum_{j=1}^J N_j^2 \widehat{\sigma}^2_{(SMALL),j}.$$

D.2 Bias of $\widehat{\sigma}^2_{(SMALL/m)}$

D.2.1 Proof of Corollary 3.4.1 and Corollary 4.3.1

Proof. For this section assume that we are in the finite sample framework. The results for the stratified sampling from an infinite population framework follow directly by changing the expectations and notation.

First we will focus on $\widehat{\sigma}^2_{(SMALL),j}$ which is the variance estimator for $\widehat{\tau}_{(SMALL),j}$. Note that

$$\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right) = \text{var}\left(\frac{1}{K_j}\sum_{k:n_k=m_j}\widehat{\tau}_k\middle|\mathcal{S}\right) = \frac{1}{K_j^2}\sum_{k:n_k=m_j}\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right).$$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{S}\right]\\
&\quad= \mathbb{E}\left[\frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\widehat{\tau}_k-\widehat{\tau}_{(SMALL),j}\right)^2\Big|\mathcal{S}\right]\\
&\quad= \frac{1}{K_j(K_j-1)}\mathbb{E}\left[\sum_{k:n_k=m_j}\left(\widehat{\tau}_k^2-2\widehat{\tau}_k\widehat{\tau}_{(SMALL),j}+\widehat{\tau}^2_{(SMALL),j}\right)\Big|\mathcal{S}\right]\\
&\quad= \frac{1}{K_j(K_j-1)}\left(\sum_{k:n_k=m_j}\left[\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right)+\tau^2_{k,\mathcal{S}}\right]-K_j\left[\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)+\tau^2_{(SMALL),\mathcal{S},j}\right]\right)\\
&\quad= \frac{1}{K_j(K_j-1)}\left(\sum_{k:n_k=m_j}\left[\frac{K_j-1}{K_j}\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right)+\tau^2_{k,\mathcal{S}}\right]-K_j\tau_{(SMALL),\mathcal{S},j}\right)\\
&\quad= \frac{1}{K_j^2}\sum_{k:n_k=m_j}\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right)+\frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}}-\tau_{(SMALL),\mathcal{S},j}\right)^2\\
&\quad= \text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)+\frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}}-\tau_{(SMALL),\mathcal{S},j}\right)^2.
\end{aligned}
$$

So the bias is

$$
\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{S}\right]-\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)=\frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}}-\tau_{(SMALL),\mathcal{S},j}\right)^2.
$$

Now we move our attention to $\widehat{\sigma}^2_{(SMALL/m)}$ which is a variance estimator for $\widehat{\tau}_{(SMALL)}$. We have

$$
\begin{aligned}
\text{var}\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right)&=\text{var}\left(\frac{1}{\sum_{i=1}^J m_iK_i}\sum_{j=1}^J m_jK_j\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)\\
&=\frac{1}{\left(\sum_{i=1}^J m_iK_i\right)^2}\sum_{j=1}^J\left(m_jK_j\right)^2\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right).
\end{aligned}
$$

So then

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/m)}|\mathcal{S}\right]\\
&\quad=\frac{1}{\left(\sum_{i=1}^J m_iK_i\right)^2}\sum_{j=1}^J\left(m_jK_j\right)^2\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{S}\right]\\
&\quad=\text{var}\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right)+\sum_{j=1}^J\frac{m_j^2K_j}{\left(\sum_{i=1}^J m_iK_i\right)^2(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}}-\tau_{(SMALL),\mathcal{S},j}\right)^2.
\end{aligned}
$$

So the bias is

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL/m)}|\mathcal{S}\right]-\text{var}\left(\widehat{\tau}_{(SMALL)}|\mathcal{S}\right)=\sum_{j=1}^{J}\frac{m_j^2K_j}{\left(\sum_{i=1}^{J}m_iK_i\right)^2(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}}-\tau_{(SMALL),\mathcal{S},j}\right)^2.$$

□

D.2.2 Proof of Corollary 4.4.1

Proof. Assume that we are in the random sampling of strata framework of Section 4.4. We are focusing in on just the set of strata of size m_j. We can either consider that there only exist strata of this size or we can imagine a sampling mechanism that draws these strata independently from strata of other sizes (e.g. stratified sampling by strata size), which is the same as conditioning on the number of strata of each size in the sample. Also,

$$\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{F}_2\right)=\mathbb{E}\left[\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)|\mathcal{F}_2\right]+\text{var}\left(\mathbb{E}\left[\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right]|\mathcal{F}_2\right).$$

From Appendix D.2.1, we can see that

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{S}\right]\quad=\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)+\frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}}-\tau_{(SMALL),\mathcal{S},j}\right)^2.$$

From standard results from sampling theory (see (Lohr, 2009, Chapter 2)) we have

$$\mathbb{E}\left[\frac{1}{K_j(K_j-1)}\sum_{k:n_k=m_j}\left(\tau_{k,\mathcal{S}}-\tau_{(SMALL),\mathcal{S},j}\right)^2|\mathcal{F}_2\right]=\frac{\sigma_\tau^2}{K_j}=\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{F}_2\right).$$

Hence, we end up with

$$\mathbb{E}\left[\widehat{\sigma}^2_{(SMALL),j}|\mathcal{F}_2\right]=\mathbb{E}\left[\text{var}\left(\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right)|\mathcal{F}_2\right]+\text{var}\left(\mathbb{E}\left[\widehat{\tau}_{(SMALL),j}|\mathcal{S}\right]|\mathcal{F}_2\right).$$

Thus, this is an unbiased variance estimator in this setting.

If we have varying size but condition on the number of strata of each size that we sample, then we use the fact that stratified sampling causes the estimators for each strata size to be independent.

The proof of this result for an infinite number of infinite size strata is direct by replacing the conditioning on $\mathcal{S}$ by conditioning on $\boldsymbol{B}$ and using results from the stratified sampling framework. □

E Creation and bias of $\widehat{\sigma}^2_{(SMALL/p)}$

E.1 Proof of Corollary 3.4.2 and Corollary 4.3.2

Proof. For this section assume that we are in the finite sample framework. The results for the stratified sampling framework follow directly by changing what we are taking the expectation with respect to and exchanging notation. We explain how $\widehat{\sigma}^2_{(SMALL/p)}$ was derived which also provides the bias of the estimator in these two frameworks.

To begin we consider, for an experiment with all small blocks, a variance estimator of the form

$$X \equiv \sum_{k=1}^{K} a_k \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2$$

for some collection of a_k. We then wish to find non-negative a_k's that would make this estimator as close to unbiased as possible. In particular, we aim to create an estimator with similar bias to $\widehat{\sigma}^2_{(SMALL/m)}$ but that allows for blocks of varying size. That is, we are looking to create a similarly conservative estimator that is unbiased when the average treatment effect is constant across blocks. This also means that we are creating the minimally biased conservative estimator of this form, without further assumptions.

The expected value of an estimator of this form is

$$\begin{aligned}
&\mathbb{E}\left[X|\mathcal{S}\right]\\
=&\mathbb{E}\left[\sum_{k=1}^{K} a_k \left(\widehat{\tau}_k - \tau_{k,\mathcal{S}} + \tau_{k,\mathcal{S}} - \tau_{\mathcal{S}} + \tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)^2 |\mathcal{S}\right]\\
=&\mathbb{E}\Big[\sum_{k=1}^{K} a_k \Big(\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)^2 + \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 + \left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)^2\\
&+ 2\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right) + 2\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)\left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right) + 2\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)\left(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)}\right)\Big)|\mathcal{S}\Big]\\
=&\sum_{k=1}^{K} a_k \left(\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \mathbb{E}\left[\left(\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}}\right)^2|\mathcal{S}\right] + \underbrace{\mathbb{E}\left[\left(\tau_{\mathcal{S}}-\widehat{\tau}_{(BK)}\right)^2|\mathcal{S}\right]}_{\mathbf{A}} + 2\underbrace{\mathbb{E}\left[\left(\widehat{\tau}_k-\tau_{k,\mathcal{S}}\right)\left(\tau_{\mathcal{S}}-\widehat{\tau}_{(BK)}\right)|\mathcal{S}\right]}_{\mathbf{B}}\right)
\end{aligned}$$

For **A**:

$$\mathbb{E}\left[(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)})^2|\mathcal{S}\right] = \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right) = \sum_{k=1}^{K} \frac{n_k^2}{n^2} \text{var}\left(\widehat{\tau}_k|\mathcal{S}\right)$$

For **B**:

$$\begin{aligned}
\mathbb{E}\left[(\widehat{\tau}_k - \tau_{k,\mathcal{S}})(\tau_{\mathcal{S}} - \widehat{\tau}_{(BK)})|\mathcal{S}\right] &= \mathbb{E}\left[(\widehat{\tau}_k - \tau_{k,\mathcal{S}})\sum_{j=1}^{K}\frac{n_j}{n}\left(\tau_{j,\mathcal{S}} - \widehat{\tau}_j\right)\Big|\mathcal{S}\right]\\
&= \mathbb{E}\left[-\frac{n_k}{n}\left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)^2 + \left(\widehat{\tau}_k - \tau_{k,\mathcal{S}}\right)\sum_{j\neq k}\frac{n_j}{n}\left(\tau_{j,\mathcal{S}} - \widehat{\tau}_{j,\mathcal{S}}\right)\Big|\mathcal{S}\right]\\
&= -\frac{n_k}{n}\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right)
\end{aligned}$$

Due to the assignment mechanism, $\widehat{\tau}_j$ will be independent of $\widehat{\tau}_k$ so the cross terms are all zero in the above equation.

Putting **A** and **B** together, we get

$$\begin{aligned}
\mathbb{E}\left[X|\mathcal{S}\right] &= \sum_{k=1}^{K} a_k\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K} a_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 + \sum_{k=1}^{K} a_k\sum_{j=1}^{K}\frac{n_j^2}{n^2}\text{var}\left(\widehat{\tau}_j|\mathcal{S}\right)\\
&\quad - 2\sum_{k=1}^{K} a_k\frac{n_k}{n}\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right)\\
&= \sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K} a_j\right)\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K} a_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2
\end{aligned}$$

We now select a_k to make the above as close to the true variance as possible. The second term will be small if the $\tau_{k,\mathcal{S}}$ do not vary much. But this is unknown and thus we cannot select universal a_k to control it. We would like

$$a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K} a_j = \frac{n_k^2}{n^2}$$

so that the first term is the true variance. Then the second term, the bias, would be similar to that of the standard matched pairs variance estimator.

If we solve the a_k as above we will obtain a conservative estimator that is unbiased when we have equal average treatment effect for all blocks, for the stratified sampling or finite framework. To show this, consider the bias:

$$\begin{aligned}
&\mathbb{E}\left[X|\mathcal{S}\right] - \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}\right)\\
&\qquad = \sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K} a_j - \frac{n_k^2}{n^2}\right)\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K} a_k\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2. \qquad (17)
\end{aligned}$$

We know $\text{var}\left(\widehat{\tau}_k|\mathcal{S}\right) \geq 0$ for all k. We also have $\sum_{k=1}^{K} a_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \geq 0$ so at a minimum it is 0. This implies that to always be conservative, the first term in the above expression must always be at least 0. Hence, to minimize the bias but remain conservative, we set $a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{j=1}^{K} a_j - \frac{n_k^2}{n^2} = 0$. Note that $\tau_{k,\mathcal{S}}$ and $\tau_{\mathcal{S}}$ are unknown and so we cannot optimize with respect to them.

Denote $C = \sum_{k=1}^{K} a_k$. Then we want to solve

$$\begin{aligned} a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}C &= \frac{n_k^2}{n^2} \\ a_k(1 - 2\frac{n_k}{n}) &= \frac{n_k^2}{n^2}(1 - C) \\ a_k &= \frac{n_k^2}{n}\frac{1 - C}{n - 2n_k} \end{aligned}$$

But then

$$\begin{aligned} C = \sum_{k=1}^{K} a_k &= \sum_{k=1}^{K} \frac{n_k^2(1 - C)}{n(n - 2n_k)} \\ \left(1 + \frac{1}{n}\sum_{k=1}^{K}\frac{n_k^2}{n - 2n_k}\right) C &= \frac{1}{n}\sum_{k=1}^{K}\frac{n_k^2}{n - 2n_k} \end{aligned}$$

$$C = \frac{1}{n}\sum_{k=1}^{K}\frac{n_k^2}{n - 2n_k}\left(\frac{1}{1 + \frac{1}{n}\sum_{j=1}^{K}\frac{n_j^2}{n - 2n_j}}\right) = \frac{\sum_{k=1}^{K}\frac{n_k^2}{n - 2n_k}}{n + \sum_{j=1}^{K}\frac{n_j^2}{n - 2n_j}}$$

Then we have

$$a_k = \frac{n_k^2}{n}\left(\frac{1 - C}{n - 2n_k}\right) = \frac{n_k^2}{(n - 2n_k)(n + \sum_{j=1}^{K}\frac{n_j^2}{n - 2n_j})}.$$

So then

$$\sum_{k=1}^{K}\frac{n_k^2}{(n - 2n_k)\left(n + \sum_{j=1}^{K}\frac{n_j^2}{n - 2n_j}\right)}\left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2$$

has bias

$$\sum_{k=1}^{K}\frac{n_k^2}{(n - 2n_k)\left(n + \sum_{j=1}^{K}\frac{n_j^2}{n - 2n_j}\right)}\left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2.$$

Bigger strata get weighted more heavily.

As a check, in the case where the n_k are all the same, so that $n = Kn_k$, the above boils down to $a_k = \frac{1}{K(K-1)}$ and $C = \frac{1}{K-1}$, giving us the classic matched pairs variance estimator. □

E.2 Proof of Theorem 4.4.1: Unbiasedness of $\widehat{\sigma}^2_{(SMALL/p)}$ given independence

Proof. We start from the Equation 17. As in Section E, let

$$X \equiv \sum_{k=1}^{K} a_k \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2 .$$

Then

$$\mathbb{E}\left[X|\mathcal{S}\right] = \sum_{k=1}^{K} \left(a_k - 2a_k \frac{n_k}{n} + \frac{n_k^2}{n^2} \sum_{j=1}^{K} a_j \right) \operatorname{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K} a_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 .$$

Previously we were concerned with getting the first term correct. But in the Random Sampling of Strata setting, the second term is trickier. This is especially the case if we have large blocks and thus can estimate the first term. So first we focus on the second term. Keeping the variance decomposition in mind, we ultimately want the expectation of this second term to look like var$(\tau_{\mathcal{S}})$.

As a reminder, the variance decomposition is

$$\begin{aligned} \operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_2) &= \mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\mathcal{F}_2\right] + \operatorname{var}\left(\mathbb{E}[\widehat{\tau}_{(BK)}|\mathcal{S}]|\mathcal{F}_2\right) \\ &= \mathbb{E}\left[\operatorname{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|\mathcal{F}_2\right] + \operatorname{var}\left(\tau_{\mathcal{S}}|\mathcal{F}_2\right). \end{aligned}$$

We assume, for simplicity, that the block sizes are independent from the treatment effects. This implies that $\mathbb{E}[\tau_{k,\mathcal{S}}|\mathcal{F}_2] = \tau$.

The expected value of the second term in this setting is

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K} a_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \Big| \mathcal{F}_2\right] \\
=&\mathbb{E}\left[\sum_{k=1}^{K} a_k \tau_{k,\mathcal{S}}^2 - 2\tau_{\mathcal{S}} \sum_{k=1}^{K} a_k \tau_{k,\mathcal{S}} + \tau_{\mathcal{S}}^2 \sum_{k=1}^{K} a_k \Big| \mathcal{F}_2\right] \\
=&\mathbb{E}\Bigg[\sum_{k=1}^{K} a_k \tau_{k,\mathcal{S}}^2 - 2\left(\sum_{k=1}^{K} a_k \frac{n_k}{n} \tau_{k,\mathcal{S}}^2 + \sum_{k=1}^{K}\sum_{j\neq k} a_k \frac{n_j}{n} \tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\right) \\
&+ \left(\sum_{k=1}^{K} \frac{n_k^2}{n^2}\tau_{k,\mathcal{S}}^2 + \sum_{k=1}^{K}\sum_{j\neq k} \frac{n_k n_j}{n^2}\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\right) \sum_{k=1}^{K} a_k \Big| \mathcal{F}_2\Bigg] \\
=&\mathbb{E}\left[\sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{i=1}^{K} a_i\right)\tau_{k,\mathcal{S}}^2 - \sum_{k=1}^{K}\sum_{j\neq k}\left(2a_k\frac{n_j}{n} - \frac{n_k n_j}{n^2}\sum_{i=1}^{K} a_i\right)\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}} \Big| \mathcal{F}_2\right] \\
=&\mathbb{E}\left[\sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{i=1}^{K} a_i\right)\tau_{k,\mathcal{S}}^2 \Big| \mathcal{F}_2\right] - \mathbb{E}\left[\sum_{k=1}^{K}\sum_{j\neq k}\left(2a_k\frac{n_j}{n} - \frac{n_k n_j}{n^2}\sum_{i=1}^{K} a_i\right) \Big| \mathcal{F}_2\right]\tau^2.
\end{aligned}$$

Now consider the true variance, which we are trying to estimate.

$$\begin{aligned}
\mathrm{var}(\tau_{\mathcal{S}}|\mathcal{F}_2) &= \mathrm{var}\left(\sum_{k=1}^{K}\frac{n_k}{n}\tau_{k,\mathcal{S}}\Big|\mathcal{F}_2\right) \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{K}\frac{n_k}{n}\tau_{k,\mathcal{S}} - \tau\right)^2 \Big|\mathcal{F}_2\right] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{K}\frac{n_k}{n}\tau_{k,\mathcal{S}}\right)^2 \Big|\mathcal{F}_2\right] - \tau^2 \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\tau_{k,\mathcal{S}}^2\Big|\mathcal{F}_2\right] + \mathbb{E}\left[\sum_{k=1}^{K}\sum_{j\neq k}\frac{n_k n_j}{n^2}\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\Big|\mathcal{F}_2\right] - \tau^2 \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\tau_{k,\mathcal{S}}^2\Big|\mathcal{F}_2\right] - \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\Big|\mathcal{F}_2\right]\tau^2
\end{aligned}$$

We have the last equality because

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K}\sum_{j\neq k}\frac{n_k n_j}{n^2}\tau_{k,\mathcal{S}}\tau_{j,\mathcal{S}}\Big|\mathcal{F}_2\right] &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{j\neq k}\frac{n_k n_j}{n}\Big|\mathcal{F}_2\right]\tau^2 \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}(1-\frac{n_k}{n})\Big|\mathcal{F}_2\right]\tau^2 \\
&= \mathbb{E}\left[1-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\Big|\mathcal{F}_2\right]\tau^2.
\end{aligned}$$

Matching this up with the expectation of our estimator, we want

$$\frac{n_k^2}{n^2} = a_k - 2a_k\frac{n_k}{n} + \frac{n_k^2}{n^2}\sum_{i=1}^{K} a_i$$

and

$$\sum_{k=1}^{K}\frac{n_k^2}{n^2} = \sum_{k=1}^{K}\sum_{j\neq k}\left(2a_k\frac{n_j}{n} - \frac{n_k n_j}{n^2}\sum_{i=1}^{K} a_i\right).$$

The first equation we solved for before in Section E. So we will get

$$a_k = \frac{n_k^2}{(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}.$$

Let's see if this weight works for the second term.

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{j\neq k}\left(2a_k\frac{n_j}{n}-\frac{n_kn_j}{n^2}\sum_{i=1}^{K}a_i\right)\\
&=\sum_{k=1}^{K}2a_k(1-\frac{n_k}{n})-(1-\sum_{k=1}^{K}\frac{n_k^2}{n^2})\sum_{i=1}^{K}a_i\\
&=\sum_{k=1}^{K}\frac{2n_k^2(1-\frac{n_k}{n})}{(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}-(1-\sum_{k=1}^{K}\frac{n_k^2}{n^2})\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2(1-2\frac{n_k}{n})}{(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}+\sum_{k=1}^{K}\frac{n_k^2}{n^2}\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2(n-2n_k)}{n(n-2n_k)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}+\sum_{k=1}^{K}\frac{n_k^2}{n^2}\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2}{n(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}+\sum_{k=1}^{K}\frac{n_k^2}{n^2}\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2(n+\sum_{i=1}^{K}\frac{n_i^2}{(n-2n_i)})}{n^2(n+\sum_{j=1}^{K}\frac{n_j^2}{n-2n_j})}\\
&=\sum_{k=1}^{K}\frac{n_k^2}{n^2}
\end{aligned}
$$

Which is exactly what we wanted. So this weight works. And because it is the same as the weight for the finite sample (where we wanted to get the first term in the variance decomposition correct), this also takes care of the first term in the variance decomposition.

The proof of this result for an infinite number of infinite size strata is direct by replacing the conditioning on $\mathcal{S}$ by conditioning on $\boldsymbol{B}$ and using results from the stratified sampling framework. □

F Creation of $\widehat{\sigma}^2_{SRS}$, Equation (8)

First, we start with the basic variance decomposition to examine what we are trying to estimate.

$$
\begin{aligned}
\mathrm{var}(\widehat{\tau}_{(BK)}|SRS) &= \mathbb{E}\left[\mathrm{var}(\widehat{\tau}_{(BK)}|\mathcal{S})|SRS\right] + \mathrm{var}\left(\mathbb{E}[\widehat{\tau}_{(BK)}|\mathcal{S}]|SRS\right) \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}}+\frac{S_k^2(t)}{n_{t,k}}-\frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] + \mathrm{var}\left(\tau_{\mathcal{S}}|SRS\right) \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}}+\frac{S_k^2(t)}{n_{t,k}}-\frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] + \frac{\sigma^2(tc)}{n} \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}}+\frac{S_k^2(t)}{n_{t,k}}-\frac{S_k^2(tc)}{n_k}\right)\Big|SRS\right] + \mathbb{E}\left[\frac{S^2(tc)}{n}\Big|SRS\right] \\
&= \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}}+\frac{S_k^2(t)}{n_{t,k}}\right)\Big|SRS\right]}_{\mathbf{A}} + \underbrace{\mathbb{E}\left[\frac{S^2(tc)}{n}-\sum_{k=1}^{K}\frac{n_k}{n}\frac{S_k^2(tc)}{n}\Big|SRS\right]}_{\mathbf{B}}
\end{aligned}
\tag{18}
$$

Let's examine $S^2(tc)$ so we can simplify term **B**.

$$
S^2(tc) = \sum_{k=1}^{K}\frac{n_k-1}{n-1}S_k^2(tc) + \sum_{k=1}^{K}\frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}}\right)^2
$$

So term **B** can simplify as follows:

$$
\begin{aligned}
S^2(tc) - \sum_{k=1}^{K}\frac{n_k}{n}S_k^2(tc) &= \sum_{k=1}^{K}\frac{n_k-1}{n-1}S_k^2(tc) + \sum_{k=1}^{K}\frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}}\right)^2 - \sum_{k=1}^{K}\frac{n_k}{n}S_k^2(tc) \\
&= \sum_{k=1}^{K}\frac{n_k}{n-1}\left(\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}}\right)^2 - \sum_{k=1}^{K}\frac{n-n_k}{n(n-1)}S_k^2(tc).
\end{aligned}
$$

Now recall from Supplementary Materials E that

$$
\mathbb{E}\left[\sum_{k=1}^{K}a_k\left(\widehat{\tau}_k-\widehat{\tau}_{(BK)}\right)^2|\mathcal{S}\right] = \sum_{k=1}^{K}\left(a_k - 2a_k\frac{n_k}{n}+\frac{n_k^2}{n^2}\sum_{j=1}^{K}a_j\right)\mathrm{var}\left(\widehat{\tau}_k|\mathcal{S}\right) + \sum_{k=1}^{K}a_k\left(\tau_{k,\mathcal{S}}-\tau_{\mathcal{S}}\right)^2.
$$

Letting $a_k = n_k$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K} n_k \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2 \middle| \mathcal{S}\right] \\
&= \sum_{k=1}^{K} \left(n_k - \frac{n_k^2}{n}\right) \text{var}\left(\widehat{\tau}_k | \mathcal{S}\right) + \sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \\
&= \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n} \text{var}\left(\widehat{\tau}_k | \mathcal{S}\right) + \sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \\
&= \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}} - \frac{S_k^2(tc)}{n_k}\right) + \sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 \\
&= \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) + \sum_{k=1}^{K} n_k \left(\tau_{k,\mathcal{S}} - \tau_{\mathcal{S}}\right)^2 - \sum_{k=1}^{K} \frac{(n-n_k)}{n} S_k^2(tc) \\
&= \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) + (n-1)\left(S^2(tc) - \sum_{k=1}^{K} \frac{n_k}{n} S_k^2(tc)\right).
\end{aligned}$$

This means that

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k}{n(n-1)} \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2 \middle| \mathcal{S}\right] = \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n^2(n-1)} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) + \frac{S^2(tc)}{n} - \sum_{k=1}^{K} \frac{n_k}{n} \frac{S_k^2(tc)}{n}.$$

So we have a way to estimate term **B** of Equation 18, which means we just need to add in a correction to get term **A**.

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{n_k^2}{n^2} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) - \sum_{k=1}^{K} \frac{n_k(n-n_k)}{n^2(n-1)} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) \\
&= \sum_{k=1}^{K} \frac{n_k(n_k-1)}{n(n-1)} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right)
\end{aligned}$$

Putting it all together, we have

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{n_k(n_k-1)}{n(n-1)} \left(\frac{S_k^2(c)}{n_{c,k}} + \frac{S_k^2(t)}{n_{t,k}}\right) + \sum_{k=1}^{K} \frac{n_k}{n(n-1)} \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2 \middle| SRS\right] = \text{var}(\widehat{\tau}_{(BK)} | SRS).$$

So

$$\widehat{\sigma}^2_{SRS} = \sum_{k=1}^{K} \frac{n_k(n_k-1)}{n(n-1)} \left(\frac{s_k^2(c)}{n_{c,k}} + \frac{s_k^2(t)}{n_{t,k}}\right) + \sum_{k=1}^{K} \frac{n_k}{n(n-1)} \left(\widehat{\tau}_k - \widehat{\tau}_{(BK)}\right)^2$$

is an unbiased variance estimator.

G Consequences of ignoring blocking

In this section we explore what happens when a blocked randomization was implemented but then the experiment was analyzed as if complete randomization was used. There is a misconception that implementing a blocked design and then ignoring the blocking when calculating the variance estimator will result in an estimator that is conservative for the variance of the $\widehat{\tau}_{(BK)}$.

Theorem G.0.1 (Using completely randomized variance estimator for blocked experiment)**.** *In the finite sample setting, analyzing a blocked experiment as if it were completely randomized could give anti-conservative estimators for variance.*

That is, it is possible to have $\mathbb{E}\left[\widehat{\sigma}^2_{(CR)}|\mathcal{S}, \mathbf{P}_{blk}\right] \leq \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}, \mathbf{P}_{blk}\right)$, where $\mathbf{P}_{blk}$ is a blocked randomization assignment mechanism. See Supplementary Material G.1 for a derivation that proves this result (assuming $p_k = p$ for all k and with a positive correlation of potential outcomes).

However, in the stratified sampling framework, ignoring blocking when a blocked design was run will always result in a conservative estimator for the variance of $\widehat{\tau}_{(BK)}$.

Corollary G.0.1 (Using completely randomized variance estimator for blocked experiment)**.** *Analyzing a blocked experiment as if it were completely randomized will not give anti-conservative estimators for variance if we are analyzing for a superpopulation with fixed blocks and stratified random sampling.*

See Supplementary Material G.2 for more on this result (assuming $p_k = p$ for all k).

This discussion is related to the misconception that the complete randomization variance estimator is always more stable than the blocking variance estimator, which is discussed in Supplementary Material I.1.

If we do not have $p_k = p$ for all k, on the other hand, then it is possible to have $\mathbb{E}\left[\widehat{\tau}_{(CR)}|\mathcal{S}, \mathbf{P}_{blk}\right] \neq \tau_{\mathcal{S}}$. This means $\widehat{\tau}_{(CR)}$ could be biased, even under a constant treatment effect assumption, because some units will be weighted more heavily than others in both the treatment and control arms. In this case, variance comparison is less relevant.

G.1 Proof of Theorem G.0.1

Proof. We perform a block randomization but then use the variance estimator from a complete randomization, $\frac{s^2(c)}{n_c}+\frac{s^2(t)}{n_t}$. We will condition on $\mathbf{P}_{blk}$, the assignment mechanism being blocked randomization, throughout to make this clear. For the finite sample framework, the true variance would still be

$$\text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S},\mathbf{P}_{blk}\right)=\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}}+\frac{S_k^2(t)}{n_{t,k}}-\frac{S_k^2(tc)}{n_k}\right).$$

Again, we assume that $p_k=p$ for all $k=1,\dots,K$. Then

$$\bar{Y}^{obs}(z)=\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\bar{Y}_k^{obs}(z)=\sum_{k=1}^{K}\frac{n_k}{n}\bar{Y}_k^{obs}(z).$$

We have

$$\begin{aligned}
s^2(z)=&\frac{1}{n_z-1}\sum_{i:Z_i=z}\left(Y_i(z)-\bar{Y}^{obs}(z)\right)^2\\
=&\frac{1}{n_z-1}\sum_{k=1}^{K}\sum_{i:Z_i=z,b_i=k}\left(Y_i(z)-\bar{Y}_k(z)+\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)^2\\
=&\frac{1}{n_z-1}\sum_{k=1}^{K}\sum_{i:Z_i=z,b_i=k}\Big[\left(Y_i(z)-\bar{Y}_k(z)\right)^2+2\left(Y_i(z)-\bar{Y}_k(z)\right)\left(\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)\\
&+\left(\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)^2\Big]\\
=&\frac{1}{n_z-1}\Big[\underbrace{\sum_{k=1}^{K}\sum_{i:Z_i=z,b_i=k}\left(Y_i(z)-\bar{Y}_k(z)\right)^2}_{\mathbf{A}}+\underbrace{2\sum_{k=1}^{K}n_{z,k}\left(\bar{Y}_k^{obs}(z)-\bar{Y}_k(z)\right)\left(\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)}_{\mathbf{B}}\\
&+\underbrace{\sum_{k=1}^{K}n_{z,k}\left(\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)^2}_{\mathbf{C}}\Big].
\end{aligned}$$

Now expand and take the expectation of each term separately. We start with **A** and note that $\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right]$ is the same for all units because the proportion treated is assumed to be the same in all blocks.

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K}\sum_{i:Z_i=z,b_i=k}\left(Y_i(z)-\bar{Y}_k(z)\right)^2|\mathcal{S},\mathbf{P}_{blk}\right]&=\mathbb{E}\left[\sum_{k=1}^{K}\sum_{i:b_i=k}\mathbb{I}_{Z_i=z}\left(Y_i(z)-\bar{Y}_k(z)\right)^2|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=\sum_{k=1}^{K}\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n_k-1)S_k^2(z)
\end{aligned}$$

Now **B**,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K} n_{z,k}\left(\bar{Y}_k^{obs}(z)-\bar{Y}_k(z)\right)\left(\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=\mathbb{E}\left[\sum_{k=1}^{K} n_{z,k}\left(\bar{Y}_k^{obs}(z)\bar{Y}_k(z)-\bar{Y}_k(z)^2+\bar{Y}_k(z)\bar{Y}^{obs}(z)-\bar{Y}_k^{obs}(z)\bar{Y}^{obs}(z)\right)|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=\mathbb{E}\left[\sum_{k=1}^{K} n_{z,k}\left(\bar{Y}_k(z)\bar{Y}^{obs}(z)-\bar{Y}_k^{obs}(z)\bar{Y}^{obs}(z)\right)|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=n_z\left(\bar{Y}(z)^2-\mathbb{E}\left[\bar{Y}^{obs}(z)^2|\mathcal{S},\mathbf{P}_{blk}\right]\right)\\
&=n_z\left(-\mathrm{var}\left(\bar{Y}^{obs}(z)|\mathcal{S},\mathbf{P}_{blk}\right)\right)\\
&=-n_z\sum_{k=1}^{K}\frac{n_k^2}{n^2}\mathrm{var}\left(\bar{Y}_k^{obs}(z)|\mathcal{S},\mathbf{P}_{blk}\right)\\
&=-n_z\sum_{k=1}^{K}\frac{n_{z,k}^2}{n_z^2}\frac{n_k-n_{z,k}}{n_k}\frac{S_k^2(z)}{n_{z,k}}\\
&=-\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk},b_i=k\right]\right)S_k^2(z).
\end{aligned}$$

Now **C**,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K} n_{z,k}\left(\bar{Y}_k(z)-\bar{Y}^{obs}(z)\right)^2|\mathcal{S}\right]\\
&=\sum_{k=1}^{K} n_{z,k}\bar{Y}_k(z)^2-2n_z\bar{Y}(z)^2+n_z\mathbb{E}\left[\bar{Y}^{obs}(z)^2|\mathcal{S},\mathbf{P}_{blk}\right]\\
&=\sum_{k=1}^{K} n_{z,k}\bar{Y}_k(z)^2-2n_z\bar{Y}(z)^2+n_z\mathrm{var}\left(\bar{Y}^{obs}(z)|\mathcal{S},\mathbf{P}_{blk}\right)+n_z\mathbb{E}\left[\bar{Y}^{obs}(z)|\mathcal{S},\mathbf{P}_{blk}\right]^2\\
&=\sum_{k=1}^{K} n_{z,k}\bar{Y}_k(z)^2-n_z\bar{Y}(z)^2+\sum_{k=1}^{K}\frac{n_{z,k}(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right])}{n_z}S_k^2(z).
\end{aligned}$$

Putting it all back together,

$$\begin{aligned}
&\mathbb{E}\left[s^2(z)|\mathcal{S}, \mathbf{P}_{blk}\right] \\
=&\frac{1}{n_z-1}\Big[\sum_{k=1}^{K}\mathbb{E}\left[\mathbb{I}_{Z_i=z}\right](n_k-1)S_k^2(z) - 2\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z|\mathbf{P}_{blk}}\right]\right)S_k^2(z) + \sum_{k=1}^{K}n_{z,k}\bar{Y}_k(z)^2 - n_z\bar{Y}(z)^2 \\
&+\sum_{k=1}^{K}\frac{n_{z,k}(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right])}{n_z}S_k^2(z)\Big] \\
=&\frac{1}{n_z-1}\Big[\sum_{k=1}^{K}\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n_k-1)S_k^2(z) - \sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(1-\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right]\right)S_k^2(z) \\
&+\sum_{k=1}^{K}n_{z,k}\bar{Y}_k(z)^2 - n_z\bar{Y}(z)^2\Big] \\
=&\sum_{k=1}^{K}\left(\frac{n_k}{n} - \frac{\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n-n_k)}{n(n_z-1)}\right)S_k^2(z) + \frac{1}{n_z-1}\sum_{k=1}^{K}n_{z,k}\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2.
\end{aligned}$$

So then the bias is

$$\begin{aligned}
&\mathbb{E}\left[\frac{s^2(c)}{n_c}+\frac{s^2(t)}{n_t}|\mathcal{S},\mathbf{P}_{blk}\right]-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{n_{c,k}}+\frac{S_k^2(t)}{n_{t,k}}-\frac{S_k^2(tc)}{n_k}\right)\\
&=\sum_{k=1}^{K}\left(\frac{n_k}{(1-p)n^2}-\frac{(1-p)(n-n_k)}{(1-p)n^2(n_c-1)}\right)S_k^2(c)+\frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_{c,k}}{n_c}\left(\bar{Y}_k(c)-\bar{Y}(c)\right)^2\\
&+\sum_{k=1}^{K}\left(\frac{n_k}{pn^2}-\frac{p(n-n_k)}{pn^2(n_t-1)}\right)S_k^2(t)+\frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_{t,k}}{n_t}\left(\bar{Y}_k(t)-\bar{Y}(t)\right)^2\\
&-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{S_k^2(c)}{(1-p)n_k}+\frac{S_k^2(t)}{pn_k}-\frac{S_k^2(tc)}{n_k}\right)\\
&=\frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\bar{Y}_k(c)-\bar{Y}(c)\right)^2+\frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\bar{Y}_k(t)-\bar{Y}(t)\right)^2\\
&-\left(\sum_{k=1}^{K}\frac{n-n_k}{n^2(n_c-1)}S_k^2(c)+\sum_{k=1}^{K}\frac{n-n_k}{n^2(n_t-1)}S_k^2(t)-\sum_{k=1}^{K}\frac{n_k}{n^2}S_k^2(tc)\right)\\
&=\frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\bar{Y}_k(c)-\bar{Y}(c)\right)^2+\frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\bar{Y}_k(t)-\bar{Y}(t)\right)^2\\
&+\sum_{k=1}^{K}\frac{(1-p)n_k-1}{n((1-p)n-1)}S_k^2(c)+\sum_{k=1}^{K}\frac{pn_k-1}{n(pn-1)}S_k^2(t)\\
&-2\sum_{k=1}^{K}\frac{1}{n_k-1}\sum_{i:b_i=k}\left(Y_i(c)-\bar{Y}_k(c)\right)\left(Y_i(t)-\bar{Y}_k(t)\right).
\end{aligned}$$

A positive correlation of potential outcomes could cause this difference to be negative. □

G.2 Proof of Corollary G.0.1

Proof. We start from the result of Appendix G.1 and utilize work done in Appendix C.3 to simplify things.

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\frac{s^2(z)}{n_z}|\mathcal{S}, \mathbf{P}_{blk}\right]|\mathcal{F}_1\right]\\
=&\mathbb{E}\left[\sum_{k=1}^{K}\left(\frac{n_k}{nn_z} - \frac{\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n-n_k)}{n_z n(n_z-1)}\right)S_k^2(z) + \frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2|\mathcal{F}_1\right]\\
=&\sum_{k=1}^{K}\left(\frac{n_k}{nn_z} - \frac{\mathbb{E}\left[\mathbb{I}_{Z_i=z}|\mathbf{P}_{blk}\right](n-n_k)}{n_z n(n_z-1)}\right)\sigma_k^2(z) + \mathbb{E}\left[\frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n_{z,k}}{n_z}\left(\bar{Y}_k(z)-\bar{Y}(z)\right)^2|\mathcal{F}_1\right]\\
=&\sum_{k=1}^{K}\left(\frac{n_k}{nn_z} - \frac{n-n_k}{n^2(n_z-1)}\right)\sigma_k^2(z) + \frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n-n_k}{n^2}\sigma_k^2(z) + \frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(z)-\mu(z)\right)^2\\
=&\sum_{k=1}^{K}\frac{n_k}{nn_z}\sigma_k^2(z) + \frac{1}{n_z-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(z)-\mu(z)\right)^2
\end{aligned}
$$

Now to get the bias we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{s^2(c)}{n_c} + \frac{s^2(t)}{n_t}|\mathbf{P}_{blk}, \mathcal{F}_1\right] - \sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{\sigma_k^2(c)}{n_{c,k}} + \frac{\sigma_k^2(t)}{n_{t,k}}\right)\\
=&\sum_{k=1}^{K}\frac{n_k}{nn_c}\sigma_k^2(c) + \frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(c)-\mu(c)\right)^2 + \sum_{k=1}^{K}\frac{n_k}{nn_t}\sigma_k^2(t) + \frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(t)-\mu(t)\right)^2\\
&-\sum_{k=1}^{K}\frac{n_k^2}{n^2}\left(\frac{\sigma_k^2(c)}{n_{c,k}} + \frac{\sigma_k^2(t)}{n_{t,k}}\right)\\
=&\frac{1}{n_c-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(c)-\mu(c)\right)^2 + \frac{1}{n_t-1}\sum_{k=1}^{K}\frac{n_k}{n}\left(\mu_k(t)-\mu(t)\right)^2.
\end{aligned}
$$

□

H Details on the numerical studies

H.1 Data generating process for simulations

The data generating process used both for the simulations comparing the variance estimators and also those comparing blocking to complete randomization gives us a single finite data set. We then repeatedly randomize units to treatment, according to blocked randomization, multiple times to assess finite sample behavior. Note that we found that the simulation values agreed with biases calculated for our variance estimators via the bias formulas presented in the paper, in Section 3.4. Potential outcomes for the units in each block were drawn from a bivariate normal distribution, with the means and covariance matrix as follows (shown for a unit in block k):

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \sim MVN\left(\begin{pmatrix} \alpha_k \\ \alpha_k + \beta_k \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

The correlation of potential outcomes, ρ, was varied among 0, 0.5, and 1. We controlled how differentiated the blocks were, and how heterogeneous the treatment effects across blocks were, by varying α_k and β_k. We set α_k as $\alpha_k = \Phi^{-1}\left(1 - \frac{k}{K+1}\right) a$. Similarly, $\beta_k = 5 + \Phi^{-1}\left(1 - \frac{k}{k+1}\right) b$. The larger the a, the more the mean control potential outcomes for the blocks were spread apart. The larger the b, the more heterogeneous the treatment impacts. The parameters a and b were varied among the values (0,0.1,0.3,0.5,0.8,1,1.5,2). We keep the number and sizes of blocks fixed. The blocks were ordered by size with the smallest block as block one. As a consequence, the smaller blocks have both larger means under control and larger average treatment effects.

Simulations were run over assignment of units to treatments under a blocked design, which was done 5000 times for each combination of factors.

H.2 Blocking vs. complete randomization

Here we have numerical examples to explore the potential benefits and costs of blocking in the finite sample framework. These numerical studies look at the actual variances of the treatment effect estimators, not the costs and differences of estimating these variances. The data generating mechanism was the same as in the simulations of Section 6. However, we kept the correlation of potential outcomes at $\rho = 0.5$ in the interest of keeping the plots uncluttered. Other values of ρ gave similar results. We examined a series of scenarios ranging from a collection of blocks where there is little variation from block to block (causing blocking to be less beneficial) to scenarios where the blocks are well separated and blocking is critical for controlling variation. In our first numerical study we treat 20% of all of the blocks, which enforces specific block sizes of 5, 10, 15, and 20. 50% of the units were in small blocks. In the second numerical study we kept the sizes of the blocks the same as in the first numerical study but allowed the proportion treated to vary from block to block, from 0.1 to 0.4. We kept the overall proportion treated the same so that the completely randomized design (which randomized the same total number of units to treatment as the blocked design) was the same in both numerical studies. To do so, the big blocks all had less than 20% treated. The first numerical study corresponds to the mathematical argument presented in Section 5 and examines how much blocking can hurt in a variety of scenarios. The second numerical study is to illustrate what changes when proportions treated are not the same across blocks, as discussed in Supplementary Material C.5.

We see on the x-axis of Figure 2 an R^2-like measure of how predictive blocks are of outcome, calculated for each finite data set investigated. R^2 was varied by manipulating the spread of block means under control and the spread of block treatment effects. The y-axis is the ratio of variances of the different treatment effect estimators. Generally, as expected, we see large gains in blocking for moderate to large R^2, and a slight penalty to blocking when the R^2 is relatively small. In this example with varying proportions treated, the gains of blocking is muted. The difference between equal and unequal proportions is potentially augmented by the fact that, to hold the overall proportion treated constant, the four largest blocks, of sizes 20, 15 and 10, (which have a larger overall impact on variance) had roughly proportion 0.1 treated.

I The variance of the variance estimators

I.1 Blocking versus complete randomization

Discussions of blocking versus complete randomization often include a discussion on the performance of the variance estimators in terms of their own variance. There is a misconception that the variance of the blocking variance estimator will be larger than that of the complete randomization variance estimator. For instance, the standard big block variance estimator ($\widehat{\sigma}^2_{(BK)}$) under the stratified sampling framework was assessed in Imbens (2011). In particular, in this piece there is a discussion about how the true variance of the blocking estimator may be lower than under complete randomization but that the estimate of that variance may be more variable under blocking than under complete randomization. We consider that

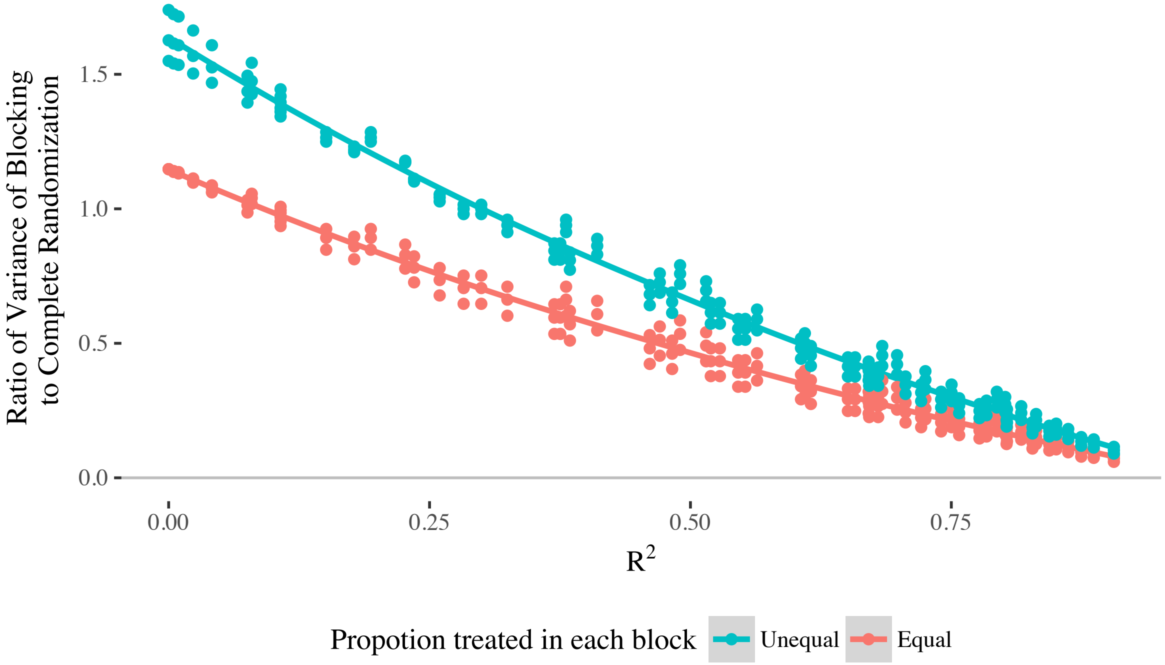

Figure 2: Numerical study to assess completely randomized versus blocked design, when $p_k = 0.2$ (equal proportions) or unequal proportions across blocks. The y axis is $\frac{\text{Var}(\widehat{\tau}_{(BK)}|\mathcal{S})}{\text{Var}(\widehat{\tau}_{(CR)}|\mathcal{S})}$.

discussion here. First, assume that $n_{z,k} \geq 2$ for all blocks k and all treatment assignments z. Then, as we have seen, $\widehat{\sigma}^2_{(BK)}$ and $\widehat{\sigma}^2_{(CR)}$ are unbiased estimators of $\text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_1)$ and $\text{var}(\widehat{\tau}_{(CR)}|SRS)$, respectively. This implies, based on results from Section C.4, that on average the variance estimator under the blocked design is less than that under the completely randomized design, as noted by Imbens (2011).

In Imbens (2011), p. 11, a further claim is made that

$$\text{var}(\widehat{\sigma}^2_{(BK)}) \geq \text{var}(\widehat{\sigma}^2_{(CR)})$$

It is not entirely clear whether the variance of $\widehat{\sigma}^2_{(CR)}$ is with respect to SRS or $\mathcal{F}_1$ but let us assume that it is with respect to $\mathcal{F}_1$ (our example extends easily to considering complete randomization with respect to SRS instead). This statement is only true if $\text{var}(s^2(c)|\mathcal{F}_1) \leq \text{var}(s^2_k(c)|\mathcal{F}_1)$ and $\text{var}(s^2(t)|\mathcal{F}_1) \leq \text{var}(s^2_k(t)|\mathcal{F}_1)$ for most $k = 1, ..., K$. Imbens (2011) gave an example in which this might be true. In his example, there is no variance in the potential outcomes under control ($\sigma^2_k(c) = 0$ for all $k = 1, ...K$) and the distribution of the potential outcomes under treatment is the same in all of the strata ($\sigma^2_k(t) = \sigma^2(t)$ for all $k = 1, .., K$).

Then, Imbens argues, because $s^2(t)$ is a less noisy estimator of $\sigma^2(t)$ than any of the $s_k^2(t)$, the variance of the variance estimator would be smaller under the completely randomized design. This notion of the variance estimator for complete randomization being less noisy as it is using more data may be true in many situations. However, this result does not hold in general. For instance, consider a population with four strata. Within each stratum, there is zero treatment effect and all units are identical. Between strata, however, the potential outcomes differ. For convenience, say in stratum one $Y_i(c) = Y_i(t) = 1$, in stratum two $Y_i(c) = Y_i(t) = 2$, in stratum three $Y_i(c) = Y_i(t) = 3$, and in stratum four $Y_i(c) = Y_i(t) = 4$. Now assume that four units are sampled from each of the strata. In a blocked design, our variance estimate would always be 0. But in a completely randomized design, the variance estimate would change based on which units were assigned to treatment and control. Thus, the blocking variance estimator would have 0 variance whereas the completely randomized variance estimator would have non-zero variance.

So, when blocking is "good", we expect the true variance of our treatment effect estimator to be lower under blocking than complete randomization and the variance of our variance estimator to also be lower. However, when the blocking is "bad" we would expect blocking to not be beneficial in terms of variance of our estimator and in this case the variance of our variance estimator could also be higher than under complete randomization.

I.2 Variance simulations

The previous discussion relates how variances of variance estimators differ under blocking and complete randomization. This raises the question of how the blocking variance estimators compare amongst themselves in terms of variance. To assess this, we examine the variance of our variance estimators from our simulation study in Section 6 with data generating process given in Section H.1.

Results are on Figure 3. We see that, in terms of variance, the estimators are generally comparable, with the plug-in approach (where an average of the big block treatment variances is used as a plug-in) giving the highest variance. We expect more instability from estimators that utilize only information from the estimated average treatment effects, not from the variation of the individual units. We see that the weighted regression estimator has the lowest variability.

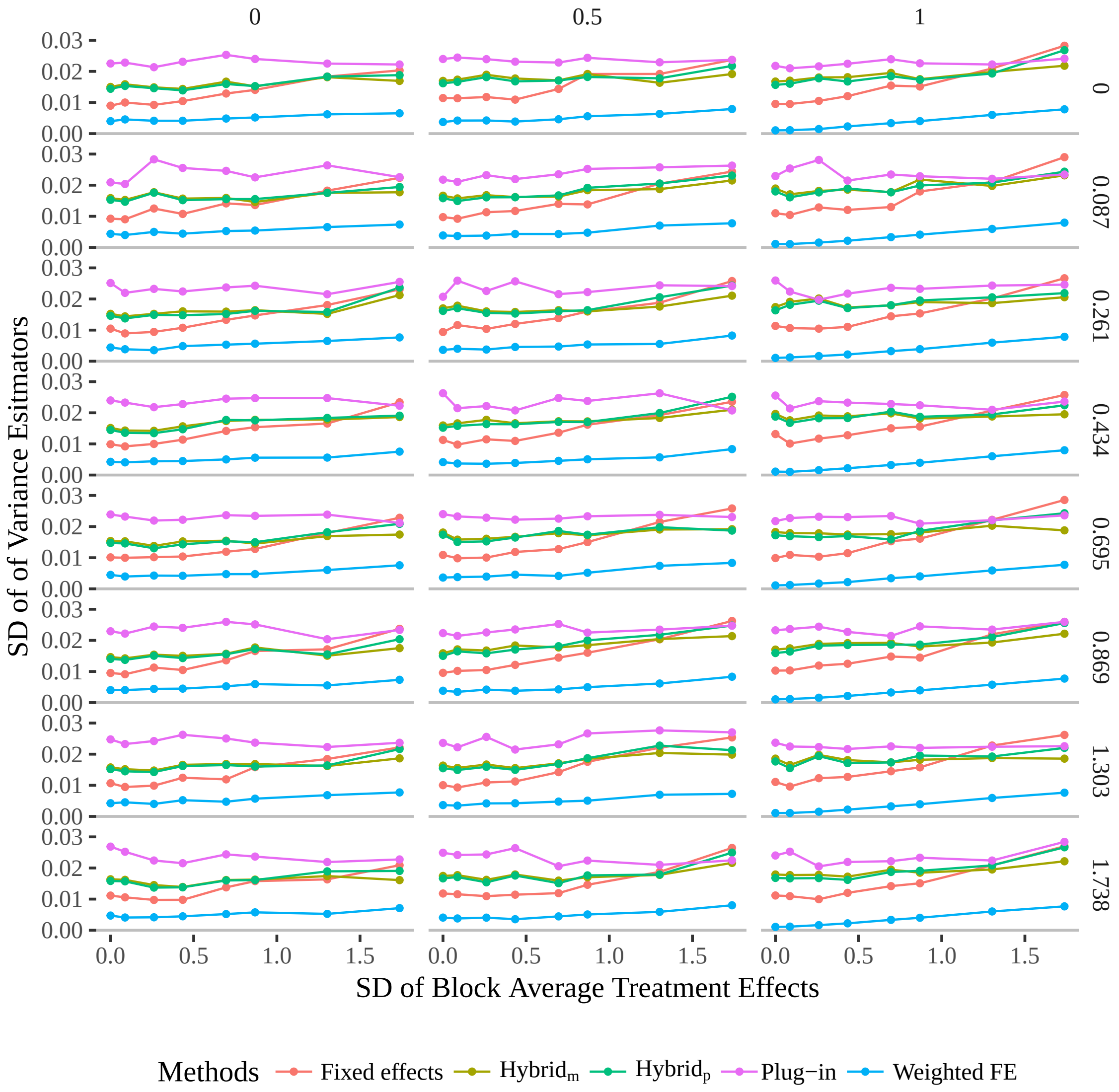

Figure 3: Simulations to assess variance estimators' variance. Each column represents a different value of ρ, with values denoted at the top of the graph. Each row shows the standard deviation of block average control potential outcomes. The x-axis shows the standard deviation of block average treatment effects. FE stands for fixed effects.