On mitigating the analytical limitations of finely stratified experiments

Colin B. Fogarty [*]

Abstract

While attractive from a theoretical perspective, finely stratified experiments such as paired designs suffer from certain analytical limitations not present in block-randomized experiments with multiple treated and control individuals in each block. In short, when using an appropriately weighted difference-in-means to estimated the sample average treatment effect, the traditional variance estimator in a paired experiment is conservative unless the pairwise average treatment effects are constant across pairs; however, in more coarsely stratified experiments, the corresponding variance estimator is unbiased if treatment effects are constant within blocks, even if they vary across blocks. Using insights from classical least squares theory, we present an improved variance estimator appropriate in finely stratified experiments. The variance estimator is still conservative in expectation for the true variance of the difference-in-means estimator, but is asymptotically no larger than the classical variance estimator under mild conditions. The improvements stem from the exploitation of effect modification, and thus the magnitude of the improvement depends upon on the extent to which effect heterogeneity can be explained by observed covariates. Aided by these estimators, a new test for the null hypothesis of a constant treatment effect is proposed. These findings extend to some, but not all, super-population models, depending on whether or not the covariates are viewed as fixed across samples in the super-population formulation under consideration.

1 Introduction

1.1 The analytical limitations of finely stratified experiment

When considering competing experimental designs, both theoretical and practical concerns must be taken into account. While the advice stemming from theoretical derivations is often in harmony with advice addressing issues of implementation, discordant recommendations can be encountered in the literature. As an illustration, consider the choice of granularity of stratification in a randomized experiment as it pertains to the variance of the resulting difference-in-means estimator of the average treatment effect. Imbens (2011) demonstrates that when considering, *ex ante*, whether one should use a completely randomized experiment or a block-randomized experiment, the classical difference-in-means estimator for the average treatment effect in block-randomized experiment has a variance which cannot be higher than that of the estimator from a completely randomized experiment; see also Fisher (1935); Cochran and Cox (1957); Cox (1958) and Greevy et al. (2004) among many. By the same logic, a given block can be further broken into substrata while not increasing the estimator's variance. This leads Imai et al. (2009) and Imbens (2011) to prefer paired experiments from a theoretical perspective. Kallus (2013) further notes that from a population perspective,

[*] Operations Research and Statistics Group, MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge MA 02142 (e-mail: cfogarty@mit.edu)

if one believes the response functions under treatment and control are Lipchitz with respect to some distance metric $\delta(\mathbf{x}_i, \mathbf{x}_j)$, then optimal pair matching with respect to $\delta(\mathbf{x}_i, \mathbf{x}_j)$ minimizes the variance of the difference-in-means estimator.

Moving away from designs with *a priori* fixed block sizes, Higgins et al. (2016) present a new experimental design called "threshold blocking" which produces stratifications wherein each block contains *at least* some number, call it k, individuals in each treatment arm. Taking $k = 1$ in a treatment-control experiment then yields a design that is more flexible than pairing. Higgins et al. (2016) present a near-optimal threshold blocking algorithm when one takes minimizing the maximal within-block covariate discrepancy between any two individuals in the same block as the objective. For the classical treatment-control experiment, the optimal stratification is mix of pairs and triplets, as any feasible stratum with four or more individuals can broken down into substrata of sizes two or three without increasing covariate discrepancy. Sävje (2015) illustrates that this additional flexibility from allowing for both pairs and triplets can result in lower estimator variance than a paired design, much in the same way that variable ratio matching tends to outperform fixed ratio matching in observational studies (Hansen, 2004).

We define a *finely stratified* design as one where within each block, there is either exactly one treated individual or exactly one control individual; both paired studies and optimal stratifications returned by threshold blocking satisfy this definition. We contrast these with *coarsely stratified* designs, wherein each block has at least two individuals in each treatment group. Of course in principle this experimental taxonomy is not exhaustive as a treatment-control experiment could have both fine and coarse strata; we ignore this possibility in what follows. The preceding discussion has illustrated the theoretical merits of fine stratifications relative to coarse stratifications; however, finely stratified designs face certain "analytical limitations" avoided by coarsely stratified designs (Klar and Donner, 1997; Imbens, 2011; Sävje, 2015). As is well known, the true variance of difference-in-means estimator for the sample average treatment effect cannot be identified without further assumptions being made on the individual level treatment effects. Following the tradition of Neyman (1923), conventional estimators for this variance exist which are conservative in expectation with respect to the experimental design's randomization distribution; see Gadbury (2001) for an overview. It is when considering the magnitude of conservativeness for different experimental designs' standard variance estimators that the practical issues faced by finely stratified designs come to light.

As will be presented explicitly in §3, the conventional variance estimator for a paired experiment is conservative in expectation unless the average treatment effect is constant across pairs, in which case it is unbiased; however, the typical variance estimator in a coarsely stratified experiment is unbiased so long as the treatment effect is constant *within* blocks, even if the effects are heterogeneous across blocks. The practitioner must conduct hypothesis tests and form confidence intervals for the sample average treatment effect using a variance estimator appropriate for the design at hand. Hence, if the practitioner believes that the blocks in her experiment were formed on the basis of effect modifying covariates, any benefits in precision from employing a finely stratified design may be washed away by the increased conservativeness of the corresponding variance estimator. Klar and Donner (1997) write that "these limitations lead us...to favour stratified designs in which there are at least two [units] in each stratum" (Klar and Donner, 1997, p. 1753). Imbens (2011) similarly notes that "[These limitations are an] important reason to prefer experiments with at least two units of each treatment type in each stratum" (Imbens, 2011, p. 17).

1.2 An insight from classical least squares squares theory

The analytical limitations of finely stratified experiments thus present an unappealing gap between theory and practice. Practical limitations hinder the actualization of theoretical benefits, an issue which we now seek to mitigate. Recent work by Aronow and Middleton (2013); Lin (2013); Fogarty (2016); Bloniarz et al. (2016) and Lu (2016) among others has shown how regression adjustment can be utilized to provide improved estimators for the average treatment effect in various experimental designs. In this work, we will demonstrate how illustrate how regression adjustment can be utilized to yield improved *variance* estimators in finely stratified experiments while using the classical difference-in-means estimator for the average treatment effect, hence preserving the so-called “hands above the table" analysis (Freedman, 2008; Lin, 2013). The key takeaway from this work is that effect modification can be exploited in a finely stratified experiment to yield improved variance estimates even when the model is misspecified. As the potential impact of effect modification is the source of the discrepancy between the variance estimators in finely and coarsely stratified experiments, this serves to close the gap between variance estimators in these respective designs. See Abadie and Imbens (2008); Ding (2016); Abadie et al. (2017) for recent work on the role of effect modification in variance estimation in related contexts.

Before proceeding, let us take a detour into classical least squares theory to provide insight into the improvements which will follow. Suppose we have n responses $\mathbf{y} = (y_1, ..., y_n)^T$, and an $n \times K$ centered matrix of covariates $\tilde{X} = (I - \mathbf{e}\mathbf{e}^T/n)X$, where I is the identity matrix and $\mathbf{e}$ is a vector containing n ones. Consider running two regressions, the first a regression of $\mathbf{y}$ on $\mathbf{e}$ and the second a regression of $\mathbf{y}$ on $\mathbf{e}$ and $\tilde{X}$. By orthogonality, the coefficient on the intercept column, $\hat{\beta}_0$, will equal the sample mean $\bar{y}$ in both regressions. On the other hand, the variance estimators for $\hat{\beta}_0$ will differ between the two regressions. For the regression on the intercept, the classical variance estimator for $\hat{\beta}_0$ is $\text{var}(\hat{\beta}_0) = \sum_{i=1}^n (y_i - \bar{y})^2/(n(n-1))$. For a regression of y on $\mathbf{e}$ and $\tilde{X}$, the classical variance for $\hat{\beta}_0$ is $\text{var}(\hat{\beta}_0 \mid \tilde{X}) = \sum_{i=1}^n (y_i - \bar{y} - \tilde{\mathbf{x}}_i^T(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y))^2/(n(n-K-1))$. As a result, $\text{var}(\hat{\beta}_0 \mid \tilde{X}) \lessapprox \text{var}(\hat{\beta}_0)$. The use of this improved variance estimator, $\text{var}(\hat{\beta}_0 \mid \tilde{X})$, is typically justified by an ancillarity argument: if the assumptions underpinning the regression model are satisfied, then the distribution of X is ancillary for inference on any slope coefficient β_k. The conditionality principle would then support conditioning on X in the inference that follows, hence restricting attention to the relevant subset of the sample space.

Buja et al. (2014) provide an illuminating discussion not only of the classical arguments for conditioning on X, but also of the breakdown of these arguments in the presence on model misspecification. The fundamental issue is that when X is itself considered to be random, X is ancillary for inference on β_k if and only if the model is correctly specified. The framework considered therein is one of a practitioner jointly sampling responses and covariates *iid* from some target population, with the target of inference being the best linear approximation to the response function for this population. In the analysis of randomized experiments, a generative model of this nature is often implausible, as individuals within a given experiment need not constitute a representative sample. As such, inference is performed on local estimands such as the average treatment effect for the individuals in the experiment at hand, with the act of randomization itself provides the basis for inference for these estimands (Neyman, 1923; Fisher, 1935; Rubin, 1974; Imbens and Rubin, 2015). For these local estimands, conditioning on the covariates for the individuals in the experiment is justified without an ancillarity, argument, as the estimands are themselves defined with respect to the sample at hand. As will be illustrated, variance estimators which utilize X will furnish improvements in power while facilitating Neyman-style conservative inference for the sample average treatment effect.

2 The sample average treatment effect

2.1 Notation for a block-randomized experiment

There are B independent blocks. The i^{th} of B blocks contains n_i individuals, of whom n_{1i} receive the treatment and n_{0i} receive the control. There are $N = \sum_{i=1}^{B} n_i$ total individuals in the study. Let Z_{ij} be an indicator of whether or not the j^{th} individual in block i receives the treatment, such that $\sum_{j=1}^{n_i} Z_{ij} = n_{1i}$ and $\sum_{j=1}^{n_i}(1 - Z_{ij}) = n_{0i}$. A finely stratified experiment is then characterized by $\min\{n_{0i}, n_{1i}\} = 1$ for all i, while in a coarsely stratified experiment $\min\{n_{0i}, n_{1i}\} > 1$ for all i. Individual j in block i has a K-dimensional vector of measured covariates $\mathbf{x}_{ij} = (x_{ij1}, ..., x_{ijK})$. Each individual has a potential outcome under treatment, r_{1ij}, and under control, r_{0ij}, $i = 1, ..., B; j = 1, ..., n_i$. The pair of potential outcomes (r_{1ij}, r_{0ij}) is not jointly observable for any individual. Instead, we observe the response $R_{ij} = r_{1ij}Z_{ij} + r_{0ij}(1 - Z_{ij})$ for each individual. As a consequence, the individual level treatment effect $\tau_{ij} = r_{1ij} - r_{0ij}$ is not observable for any individual, nor is the average of the treatment effects in any block i, $\bar{\tau}_i = n_i^{-1} \sum_{j=1}^{n_i}(r_{1ij} - r_{0ij})$ (Neyman, 1923; Rubin, 1974).

Let Ω be the set of $\prod_{i=1}^{B} \binom{n_i}{n_{1i}}$ possible values of $\mathbf{Z} = (Z_{11}, Z_{12}, ..., Z_{Bn_B})^T$ under the block-randomized design. Each $z \in \Omega$ has probability $|\Omega|^{-1}$ of being selected, where the notation $|A|$ denotes the cardinality of the set a. Let $\mathcal{Z}$ denote the event $Z \in \Omega$. Quantities dependent on the assignment vector such as $\mathbf{Z}$ and $\mathbf{R} = (R_{11}, R_{12}, ..., R_{Bn_B})^T$ are random, whereas $\mathcal{F} = \{(r_{1ij}, r_{0ij}, \mathbf{x}_{ij}), i = 1, ..., B, j = 1, ..., n_B\}$ contains fixed quantities for the experiment at hand. In a block-randomized experiment, $\text{pr}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \text{pr}(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}) = |\Omega|^{-1} = \left(\prod_{i=1}^{B} \binom{n_i}{n_{1i}}\right)^{-1}$, and $\text{pr}(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = \text{pr}(Z_{ij} = 1 \mid \mathcal{Z}) = n_{1i}/n_i$.

2.2 The estimand and the estimator

The sample average treatment effect, or $SATE$, is defined as

$$\bar{\Delta} = \frac{1}{N}\sum_{i=1}^{B}\sum_{j=1}^{n_i}\tau_{ij} = \frac{1}{B}\sum_{i=1}^{B} w_i\bar{\tau}_i,$$

where $w_i = B(n_i/N)$. The conventional unbiased estimator for $\bar{\tau}_i$, the average treatment effect for individuals in block i, is simply the observed difference-in-means between the treated and control individuals in block i.

$$\hat{\tau}_i = \sum_{j=1}^{n_i}\left(\frac{Z_{ij}r_{1ij}}{n_{1i}} - \frac{(1 - Z_{ij})r_{0ij}}{n_{0i}}\right).$$

The classical unbiased estimator for the overall sample average treatment effect $\bar{\Delta}$ is

$$\hat{\Delta} = B^{-1}\sum_{i=1}^{B} w_i\hat{\tau}_i, \tag{1}$$

i.e. a weighted average of the block-specific estimators with n_i/N serving as weights (Rosenbaum, 2002, Chapter 2).

3 A comparison of standard variance estimators

3.1 Conventional variance estimation in coarsely stratified experiments

For block i, define the block-specific averages of the potential outcomes under treatment and control as $\bar{r}_{1i} = n_i^{-1} \sum_{i=1}^{n_i} r_{1ij}$ and $\bar{r}_{0i} = n_i^{-1} \sum_{i=1}^{n_i} r_{0ij}$. Further, define σ^2_{1i}, σ^2_{0i}, and $\sigma^2_{\tau i}$ by

$$\sigma^2_{1i} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(r_{1ij} - \bar{r}_{1i}\right)^2; \quad \sigma^2_{0i} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(r_{0ij} - \bar{r}_{0i}\right)^2; \quad \sigma^2_{\tau i} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(\tau_{ij} - \bar{\tau}_i\right)^2.$$

The variance of the sample average treatment effect estimator in block i, $\text{var}(\hat{\tau}_i \mid \mathcal{F}, \mathcal{Z})$, can be expressed as (Imbens and Rubin, 2015, Theorem 6.2)

$$\text{var}(\hat{\tau}_i \mid \mathcal{F}, \mathcal{Z}) = \frac{\sigma^2_{1i}}{n_{1i}} + \frac{\sigma^2_{0i}}{n_{0i}} - \frac{\sigma^2_{\tau i}}{n_i}.$$

This immediately yields the following expression for $\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$:

$$\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{B^2} \sum_{i=1}^{B} w_i^2 \left(\frac{\sigma^2_{1i}}{n_{1i}} + \frac{\sigma^2_{0i}}{n_{0i}} - \frac{\sigma^2_{\tau i}}{n_i}\right).$$

This variance is unknown in practice because it depends on the missing potential outcomes. In a coarsely stratified experiment where we have $\min\{n_{1i}, n_{0i}\} \geq 2$ for all i, the conventional estimator for $\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$ is based on an appropriately weighted sum of the sample variances of the treated and control responses in each block. Let $\bar{R}_{1i} = n_{1i}^{-1} \sum_{i=1}^{n_i} Z_{ij} r_{1ij}$ and $\bar{R}_{0i} = n_{0i}^{-1} \sum_{i=1}^{n_i} (1 - Z_{ij}) r_{0ij}$ be the observed averages of responses for the treated and control individuals in block i. Further, let s^2_{1i} and s^2_{0i} be the sample variances for the responses of the treated and control units in block i,

$$s^2_{1i} = \frac{1}{n_{1i} - 1} \sum_{j=1}^{n_i} Z_{ij}(r_{1ij} - \bar{R}_{1i})^2; \quad s^2_{0i} = \frac{1}{n_{0i} - 1} \sum_{j=1}^{n_i} (1 - Z_{ij})(r_{0ij} - \bar{R}_{0i})^2$$

The classical variance estimator in a coarsely stratified experiment takes on the following form:

$$S^2_{CS} = \frac{1}{B^2} \sum_{i=1}^{B} w_i^2 \left(\frac{s^2_{1i}}{n_{1i}} + \frac{s^2_{0i}}{n_{0i}}\right).$$

A well known fact dating back to Neyman (1923) is that this estimator yields conservative inference for the sample average treatment effect, since

$$\mathbb{E}[S^2_{CS} \mid \mathcal{F}, \mathcal{Z}] - \text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{B^2} \sum_{i=1}^{B} w_i^2 \sigma^2_{\tau i}. \tag{2}$$

Hence, the variance estimator S^2_{CS} is an upper bound on $\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$ in expectation unless the treatment effect is constant within each block (i.e. if for each block i, $\tau_{ij} = \bar{\tau}_i$ for $j = 1, ..., n_i$). This thus enables Neyman-style conservative inference on $\bar{\Delta}$ to proceed using S^2_{CS}.

3.2 Classical results on variance estimation in finely stratified experiments

In a finely stratified experiment, at least one of s^2_{1i} and s^2_{0i} will be undefined as $\min\{n_{1i}, n_{0i}\} = 1$. As a result, the estimator S^2_{CS} cannot be employed. To the best of our knowledge there does not exist a "classical" variance estimator for the general class of finely stratified experiments without making assumptions such as additivity of treatment effects or equal variance of potential outcomes (Rosenbaum, 2002; Hansen, 2004; Sävje, 2015). In the particular case of paired designs where $n_{1i} = n_{0i} = 1$ for all strata, the classical variance estimator is simply the sample variance of the observed paired differences divided by the number of pairs,

$$S^2_P = \frac{1}{B(B-1)} \sum_{i=1}^{B} (\hat{\tau}_i - \hat{\Delta})^2. \tag{3}$$

Imai (2008) discusses inference for the sample average treatment effect within a paired design. Proposition 1 of that work illustrates that S^2_P is also an upper bound in expectation for $\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$, and that the degree of the bias is given by

$$\mathbb{E}[S^2_P \mid \mathcal{F}, \mathcal{Z}] - \text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{B(B-1)} \sum_{i=1}^{B} (\bar{\tau}_i - \bar{\Delta})^2. \tag{4}$$

A comparison of bias expressions (2) and (4) reveals the analytical limitations alluded to in §1.1. For a paired design, S^2_P is biased upwards unless the average treatment effects are the same across pairs. In a coarsely stratified design, S^2_{CS} is unbiased if there is additivity within blocks, even if there is effect heterogeneity across blocks. If the blocks were formed using covariates that are thought to be effect modifiers, it may be the case that the coarsely stratified design yields an unbiased estimator for the variance, while the paired design would yield a variance estimator that is substantially biased upwards. Were (3) the only variance estimator available to facilitate inference in a paired experiment, the practitioner in this case may well be justified in preferring the more coarsely stratified design as a means of shrinking confidence intervals and yielding more powerful hypothesis tests.

4 Conservative variance estimators in finely stratified experiments

4.1 Two recipes with projection matrices

Let Q be an arbitrary $B \times L$ matrix with $L < B$, and let $H_Q = Q^T(Q^TQ)^{-1}Q$ be the orthogonal projection of $\mathbb{R}^B$ onto the column space of Q. Let h_{Qij} be the $\{i, j\}$ element of H_Q. Define $y_i = \hat{\tau}_i/\sqrt{1 - h_{Qii}}$ and $\mu_i = \bar{\tau}_i/\sqrt{1 - h_{Qii}}$. Let $\mathbf{y} = (y_1, ..., y_B)^T$, and let the analogous definitions hold for $\boldsymbol{\mu}$, $\hat{\boldsymbol{\tau}}$, and $\bar{\boldsymbol{\tau}}$. Finally, let Ψ_Q be a $B \times B$ diagonal matrix whose $\{i, i\}$ entry equals $1/(1 - h_{Qii})^2$

Let W be a $B \times B$ diagonal matrix whose i^{th} diagonal element contains $w_i = Bn_i/N$. We will now show that the matrix Q can be used to produce two variance estimators which are conservative in expectation for $\text{var}(\widehat{\Delta} \mid \mathcal{F}, \mathcal{Z})$

Define the first of these estimators, $S^2_1(Q)$, as

$$S^2_1(Q) = \frac{1}{B^2}\mathbf{y}^T W(I - H_Q)W. \tag{5}$$

Proposition 1. *If Q is constant across all elements of Ω:*

$$\mathbb{E}[S^2_1(Q) \mid \mathcal{F}, \mathcal{Z}] - var(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{B^2}\boldsymbol{\mu}^T W(I - H_Q)W\boldsymbol{\mu} \geq 0$$

Proof. Define $\boldsymbol{\mu}$ as before, and let Λ be the covariance matrix for y, a diagonal matrix with $\Lambda_{ii} = 1/(1-h_{Qii})\left(\sigma^2_{1i}/n_{1i} + \sigma^2_{0i}/n_{0i} - \sigma^2_{\tau i}/n_i\right)$. Noting that $W(I-H_Q)W$ is symmetric,

$$\begin{aligned} B^2 E[S_1^2(Q) \mid \mathcal{F}, \mathcal{Z}] &= tr(\Lambda W(I-H_Q)W) + \boldsymbol{\mu}^T W(I-H_Q)W\boldsymbol{\mu} \\ &= \sum_{i=1}^{B} w_i^2 \left(\frac{\sigma^2_{1i}}{n_{1i}} + \frac{\sigma^2_{0i}}{n_{0i}} - \frac{\sigma^2_{\tau i}}{n_i}\right) + \boldsymbol{\mu}^T W(I-H_Q)W\boldsymbol{\mu} \end{aligned}$$

Recalling that $\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z}) = B^{-2}\sum_{i=1}^{B} w_i^2 \left(\sigma^2_{1i}/n_{1i} + \sigma^2_{0i}/(n_{0i}) - \sigma^2_{\tau i}/n_i\right)$

$$\mathbb{E}\left[S_1^2(Q) \mid \mathcal{F}, \mathcal{Z}\right] - \text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{B^2}\boldsymbol{\mu}^T W(I-H_Q)W\boldsymbol{\mu} \geq 0,$$

where the last line stems from $(I-H_Q)$ being a projection matrix, and hence positive semi-definite.

Define the second estimator, $S_2^2(Q)$, as

$$S_2^2(Q) = \frac{1}{B^2}\hat{\boldsymbol{\tau}}^T W(I-H_Q)\Psi_Q(I-H_Q)W\hat{\boldsymbol{\tau}}, \tag{6}$$

Proposition 2. *If Q is constant across all elements of Ω:*

$$\begin{aligned} &\mathbb{E}[S_2^2(Q) \mid \mathcal{F}, \mathcal{Z}] - var(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z}) \\ &= \frac{1}{B^2}\sum_{i=1}^{B} w_i^2 \left(\frac{\sigma^2_{1i}}{n_{1i}} + \frac{\sigma^2_{0i}}{n_{0i}} - \frac{\sigma^2_{\tau i}}{n_i}\right)\sum_{j\neq i}\frac{h^2_{Qij}}{(1-h_{Qjj})^2} + \frac{1}{B^2}\bar{\boldsymbol{\tau}}^T W(I-H_Q)\Psi_Q(I-H_Q)W\bar{\boldsymbol{\tau}} \geq 0 \end{aligned}$$

Proof. Define $\bar{\boldsymbol{\tau}}$ as before, and let Σ be the covariance matrix for $\hat{\boldsymbol{\tau}}$, a diagonal matrix with $\Sigma_{ii} = 1/\left(\sigma^2_{1i}/n_{1i} + \sigma^2_{0i}/n_{0i} - \sigma^2_{\tau i}/n_i\right)$. Noting that $W(I-H_Q)\Psi_Q(I-H_Q)W$ is symmetric,

$$B^2 E[S_2^2(Q) \mid \mathcal{F}, \mathcal{Z}] = tr(\Sigma W(I-H_Q)\Psi_Q(I-H_Q)W) + \bar{\boldsymbol{\tau}}^T W(I-H_Q)\Psi_Q(I-H_Q)W\bar{\boldsymbol{\tau}}.$$

The $\{i,i\}$ element of $\Sigma W(I-H_Q)\Psi_Q(I-H_Q)W$ is given by

$$(\Sigma W(I-H_Q)\Psi_Q(I-H_Q)W)_{ii} = w_i^2\left(\frac{\sigma^2_{1i}}{n_{1i}} + \frac{\sigma^2_{0i}}{n_{0i}} - \frac{\sigma^2_{\tau i}}{n_i}\right)\left(1 + \sum_{j\neq i}\frac{h^2_{Qij}}{(1-h_{Qjj})^2}\right)$$

Recalling the form of $\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$ and noting that $(I-H_Q)\Psi_Q(I-H_Q)$ is positive semidefinite completes the proof.

Propositions 1 and 2 illustrate that for any constant matrix Q with $L < B$, the corresponding projection matrix can be utilized for conservative variance estimation in a finely stratified experiment through the estimators $S_1^2(Q)$ and $S_2^2(Q)$ defined in (5) and (6). We will first illustrate that certain choices of Q recover the standard variance estimator in a paired experiment when using $S_1^2(Q)$, and further suggest two conventional estimators for finely stratified experiments with varying block sizes. We will then show that the form of the bias expressions in Proposition 1 and 2 provides insight into choices for Q which will provide improvements in variance estimation.

4.2 Preliminary conservative variance estimators with equal and unequal block sizes

Initially, let $\tilde{Q}_1 = [\mathbf{e}, W\mathbf{e} - 1]$ to be a $B \times 2$ matrix with a constant column along with a column corresponding to the centered weights (note that $B^{-1}\sum_{i=1}^{B} w_i = 1$). Define $Q_1 = \tilde{Q}_1 I_{2\times rank(\tilde{Q}_1)}$, where $I_{k\times \ell}$ denotes a matrix of dimension $k \times \ell$ with ones on the diagonal and zeroes everywhere else; this removes the column $W\mathbf{e} - \mathbf{e}$ when block sizes are equal to avoid rank deficiency. We will now consider the implications of choosing $Q = Q_1$ in (5) and (6) to define a conservative variance estimator.

When block sizes are equal $Q_1 = \mathbf{e}$, and hence the diagonal elements of the hat matrix associated with Q_1 equal $1/B$ for each observation. The variance estimator then takes on the simplified form

$$S_1^2(Q_1) = \frac{1}{B(B-1)}\sum_{i=1}^{B}(\hat{\tau}_i - \hat{\Delta})^2.$$

In the case of matched pairs, this estimator is simply the sample variance of the observed paired differences divided by the number of pairs, hence recovering the classical variance estimator. Proposition 1 of Imai (2008) for matched pairs can be viewed as a special case of our Proposition 1 with $Q = \mathbf{e}$. This also indicates that an additive treatment effect model implies unbiasedness of the estimator $S^2(Q_1)$ for $\mathrm{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$ in a finely stratified experiments with equal block sizes, even if the design is not paired. With equal block sizes, we have that $S_2^2(Q_1) \geq S_1^2(Q_1)$, meaning that the estimator $S_1^2(Q_1)$ should always be preferred in this case.

With unequal block sizes, the i^{th} diagonal elements of the hat matrix associated with Q_1 is $1/B + (w_i - 1)^2/\sum_{i=1}^{B}(w_i - 1)^2$. Since the diagonal elements of the hat matrix depend on w_i, the estimator $S_1^2(Q_1)$ will be a strict upper bound in expectation for $\mathrm{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$ under an additive treatment effect model for finite samples $\bar{\tau}_i = 0$ for all i. So long as $(w_i - 1)^2/\sum_{i=1}^{B}(w_i - 1)^2 \to 0$ for all i as $B \to \infty$, the estimator $S_1^2(Q_1)$ and $S_2^2(Q_1)$ will both be asymptotically unbiased for $\mathrm{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$ under an additive treatment effect (this condition would hold under the assumption that the block sizes are bounded, for example). In the unequal block case there is no longer a consistent ordering between $S_1^2(Q_1)$ and $S_2^2(Q_1)$, but the discrepancies tend to be minor: as will be demonstrated Theorem 2, appropriately scaled versions of these two estimators converge in probability to the same limit under mild conditions.

4.3 Improved variance estimation through exploiting effect modification

For each block i, let $\bar{\mathbf{x}}_i$ be the vector of length K whose k^{th} entry is the average of the k^{th} covariate for the individuals in block i, i.e. $\bar{x}_{ik} = n_i^{-1}\sum_{j=1}^{n_i} x_{ijk}$. Let $\bar{X}$ be the $B \times K$ matrix whose k^{th} column contains $(\bar{x}_{1k}, \bar{x}_{2k}, ..., \bar{x}_{Bk})^T$ for $k = 1, ..., K$. Let $M = (I - H_{Q_1})W\bar{X}$ be the weighted covariate means adjusted for Q_1. Let $Q_2 = [Q_1, M]$. While the mutual orthogonality of M, $\mathbf{e}$, and $W\mathbf{e} - \mathbf{e}$ within Q_2 is not required at this point, it facilitates forthcoming illustrations and makes clearer certain connections to heteroskedasticity consistent standard errors. Let $S_1^2(Q_2)$ and $S_2^2(Q_2)$ be the variance estimators corresponding to setting $Q = Q_2$ in (5) and (6).

To understand the potential benefits of the variance estimator $S_1^2(Q_2)$, note that from Proposition 1 the bias in BS_1^2 is $B^{-1}\mu^T W(I - H_Q)W\mu$. Under mild regularity conditions described in §4.4, the diagonal elements of the hat matrix associated with Q_2 tend to 0 implying that $\mu_i \approx \bar{\tau}_i$ in sufficiently large samples. We can then think of $B^{-1}\mu^T W(I - H_{Q_2})W\mu$ as, approximately,the mean squared error from a regression of the weighted treatment effects, $W\bar{\boldsymbol{\tau}}$, on the weighted covariates, along with an intercept and a column for the block sizes. If the matrix $W\bar{X}$ contains covariates which are predictive of the treatment effects in different blocks, $S_1^2(Q_2)$ could yield a substantially less conservative estimator for $\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$ than the estimator $S_1^2(Q_1)$, which does not exploit potential effect modification.

For $S_2^2(Q_2)$, there is an additional connection to commonly employed standard error estimators in linear regression. In fact, since Q_2 was constructed such that $\mathbf{e}$ is orthogonal to all other columns of Q_2, $S_2^2(Q_2)$ exactly corresponds to the square of the HC3 heteroskedasticity consistent standard error for the intercept column in a regression of $W\hat{\boldsymbol{\tau}}$ on Q_2 (MacKinnon and White, 1985; Long and Ervin, 2000). The bias term for $BS_2^2(Q_2)$ is then approximately equal to B times the HC3 variance for the intercept column of a regression of $W\bar{\boldsymbol{\tau}}$ on Q_2, which is itself a close approximation to the mean squared error from a regression of the weighted treatment effects $W\bar{\boldsymbol{\tau}}$ on Q_2.

Importantly, Propositons 1 and 2 make no assumption about the truth of the linear model generating the projection matrix H_Q. While the magnitude of the improvement from using $S_\ell^2(Q_2)$ instead of $S_\ell(Q_1)$ for $\ell = 1, 2$ depends on how well the weighted covariate means $W\bar{X}$ predict $W\bar{\boldsymbol{\tau}}$, any choice of Q in (5) or (6) will yield a variance estimator which is conservative in expectation for $\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})$. As will now be shown, under mild conditions $S_\ell^2(Q_2)$ is asymptotically no worse than $S_\ell^2(Q_1)$ for $\ell = 1, 2$ regardless of the functional form describing the relationship between the observed covariates and the stratum-specific treatment effects. Further, both $B(S_1^2(Q_1) - S_2^2(Q_1))$ and $B(S_1^2(Q_2) - S_2^2(Q_2))$ converge in probability to zero.

4.4 Asymptotic performance of variance estimators

We now give sufficient conditions which enable asymptotically valid inference for $\bar{\Delta}$ to proceed using $S_\ell^2(Q_1)$ and $S_\ell^2(Q_2)$ for $\ell = 1, 2$. In so doing, we will also quantify the potential improvements from exploiting effect modification through the variance estimator. The finite population asymptotics presented herein embed a given experiment with B strata within an infinite sequence of experiments with increasingly many blocks. To reflect their changing values along this sequence, quantities such as $\bar{\Delta}$, M and W should be subscripted by B for precision of notation; we omit this, trading precision for readability. Let $H_M = M(M^TM)^{-1}M^T$ be the hat matrix associated with M as defined in the previous section, and consider the following regularity conditions.

Condition 1. *(Bounded Block Sizes)* There exists a $C_1 < \infty$ such that $n_i < C_1$ for all i and all B as $B \to \infty$.

Condition 2. *(Bounded Fourth Moments).* There exists a $C_2 < \infty$ such that, for all B,
$B^{-1}\sum_{i=1}^B \sum_{j=1}^{n_i} w_i^4 r_{1ij}^4/n_i < C_2$, $B^{-1}\sum_{i=1} w_i^4 r_{0ij}^4/n_i < C_2$, $B^{-1}\sum_{i=1} w_i^4 \bar{\tau}_i^4 < C_2$ and
$B^{-1}\sum_{i=1}^B \sum_{j=1}^{n_i} w_i^4 x_{ijk}^4/n_i < C_2$ for $k = 1, .., K$.

Condition 3. *(Existence of Population Moments).*

- $B^{-1}\sum_{i=1}^B w_i\bar{\tau}_i$, $B^{-1}\sum_{i=1}^B w_i^2\bar{\tau}_i$, $B^{-1}\sum_{i=1}^B w_i^2\bar{\tau}_i^2$ and $B^{-1}\sum_{i=1}^B w_i^2(\sigma_{1i}^2/n_{1i} + \sigma_{0i}^2/n_{0i} - \sigma_{\tau i}^2/n_i)$ converge to finite limits as $B \to \infty$.
- $B^{-1}\sum_{i=1}^B w_i\bar{\tau}_i m_{ik}$ converges to a finite limit for $k = 1, ..., K$ as $B \to \infty$. Let $\boldsymbol{\eta}_M$ be the vector of length k containing these limits, i.e. $\eta_{Mk} = \lim_{B\to\infty} B^{-1}\sum_{i=1}^B w_i\bar{\tau}_i m_{ik}$.
- $B^{-1}M^TM$ converges to a finite, invertible matrix as $B \to \infty$. Call this limit Σ_M.

Let $\boldsymbol{\beta}_M = \Sigma_M^{-1}\boldsymbol{\eta}_M$. The following theorems illustrate that $S_\ell^2(Q_1)$ and $S_\ell^2(Q_2)$ for $\ell = 1, 2$ can all be used to conduct asymptotically conservative inference for the sample average treatment effect, $\bar{\Delta}$. After establishing asymptotic normality, we demonstrate that inference using $S_\ell^2(Q_2)$ will be no less powerful than that conducted using $S_\ell^2(Q_1)$ for $\ell = 1, 2$.

Theorem 1. *Under Conditions 1-3 and conditional on $\mathcal{F}$ and $\mathcal{Z}$,*

$$\sqrt{B}(\hat{\Delta} - \bar{\Delta}) \xrightarrow{d} \mathcal{N}\left(0, B^{-1}\sum_{i=1}^{B} w_i^2 \left(\frac{\sigma_{1i}^2}{n_{1i}} + \frac{\sigma_{0i}^2}{n_{0i}} - \frac{\sigma_{\tau i}^2}{n_i}\right)\right).$$

Theorem 2. *Under Conditions 1-3 and conditional on $\mathcal{F}$ and $\mathcal{Z}$, then for $\ell = 1, 2$,*

$$\begin{aligned} BS_\ell^2(Q_1) - var(\sqrt{B}\hat{\tau} \mid \mathcal{F}, \mathcal{Z}) &\xrightarrow{p} \lim_{B\to\infty} \frac{1}{B} \bar{\boldsymbol{\tau}}^T W (I - H_{Q_1}) W \bar{\boldsymbol{\tau}}; \\ BS_\ell^2(Q_2) - var(\sqrt{B}\hat{\tau} \mid \mathcal{F}, \mathcal{Z}) &\xrightarrow{p} \lim_{B\to\infty} \frac{1}{B} \bar{\boldsymbol{\tau}}^T W (I - H_{Q_2}) W \bar{\boldsymbol{\tau}} \\ &= \lim_{B\to\infty} \frac{1}{B} \bar{\boldsymbol{\tau}}^T W (I - H_{Q_1}) W \bar{\boldsymbol{\tau}} - \beta_M^T \Sigma_M \beta. \end{aligned}$$

Corollary 1. *For $\ell = 1, 2$,*

$$BS_\ell^2(Q_1) - BS_\ell^2(Q_2) \xrightarrow{p} \boldsymbol{\beta}_M^T \Sigma_M \boldsymbol{\beta}_M \geq 0.$$

The proofs are deferred to the appendix. The above results, in concert with Propositions 1 and 2, justify multiple means by which inference can be conducted for the sample average treatment effect, $\bar{\Delta}$, in finely stratified experiments. The results validate new standard error estimators for inference on the $SATE$ in finely stratified experiments while using classical weighted difference-in-mean estimator. Furthermore, these results highlight how effect modification can be leveraged to reduce the degree of conservativeness of the performed inference. As Corollary 1 demonstrates, standard errors derived by including suitably weighted average values for covariates within blocks are, asymptotically, never worse than those derived without including covariate information.

5 Consonant and dissonant super-population formulations

5.1 Population-level causal estimands

The preceding results make no assumptions about the manner by which individuals were selected for inclusion into the block-randomized experiment in the first place; that is, they neither require nor postulate the existence of a larger population from which individuals were drawn. The target of estimation, the sample average treatment effect, attests merely to the treatment effect for individuals in the sample at hand, and the act of randomization provides a reasoned basis for making probabilistic statements (Fisher, 1935). That being said, it is sometimes desired to postulate that individuals in the study at hand were in fact draws from a super-population, and to perform inference on the average treatment effect within that super-population.

5.2 Conditional average treatment effect (CATE)

As an initial super-population extension, suppose we consider the covariates $\mathbf{x}_{ij}$ and the block sizes $\{n_1, ..., n_B\}$ as fixed and consider the pairs of potential outcomes (r_{1ij}, r_{0ij}) as having arisen through the following sampling mechanism.

$$(r_{1ij}, r_{0ij}) = (f_{1i}(\mathbf{x}_{ij}), f_{0i}(\mathbf{x}_{ij})) + (\epsilon_{1ij}, \epsilon_{0ij}),$$

where $(\epsilon_{1ij}, \epsilon_{0ij})$ are drawn from an arbitrary distribution with mean $(0, 0)$ and block-specific variance-covariance matrix $\Sigma_{i\epsilon}$. Let $f_{1ij} = f_{1i}(\mathbf{x}_{ij})$, and let $f_{0ij} = f_{0i}(\mathbf{x}_{ij})$. Let $\mathcal{C} = \{\mathbf{x}_{ij}\}$ be the set containing the covariates for all individuals. Within this super-population abstraction, the conditional average treatment effect, or $CATE$, in a finely stratified experiment is defined as

$$\bar{\Delta}^{(C)} = \frac{1}{N}\sum_{i=1}^{B}\sum_{j=1}^{n_i}(f_{1ij} - f_{0ij}) \tag{7}$$

Let $\bar{f}_i = n_i^{-1}\sum_{j=1}^{n_i}(f_{1ij} - f_{0ij})$, and let $\bar{\mathbf{f}} = (\bar{f}_1, ..., \bar{f}_B)^T$. Note that (7) reflects the view of the covariates as fixed, in much the same way that conventional least squares theory operates under the assumption of fixed covariates. The classical unbiased estimator for the overall conditional average treatment effect remains the weighted difference-in-means estimator given in (1). The true variance for this estimator is inflated, as unlike with the sample average treatment effect we no longer condition on the potential outcomes in each block. Nonetheless, we now demonstrate the variance estimators $S_1^2(Q)$ given in (5) and $S_2^2(Q)$ given in (6) remain conservative estimators in expectation for $\text{var}(\hat{\Delta} \mid \mathcal{C}, \mathcal{Z})$.

Proposition 3. *If Q is constant across all elements of Ω:*

$$\mathbb{E}[S_1^2(Q) \mid \mathcal{C}, \mathcal{Z}] - var(\hat{\Delta} \mid \mathcal{C}, \mathcal{Z}) = \frac{1}{B^2}\mathbf{g}^T W(I - H_Q)W\mathbf{g} \geq 0,$$

where $\mathbf{g}$ *is a vector of length B with $g_i = (1 - h_{Qii})^{-1/2}\bar{f}_i$. Further,*

$$\begin{aligned}&\mathbb{E}[S_2^2(Q) \mid \mathcal{C}, \mathcal{Z}] - var(\hat{\Delta} \mid \mathcal{C}, \mathcal{Z})\\ &= \frac{1}{B^2}\sum_{i=1}^{B} w_i^2 var(\hat{\tau}_i \mid \mathcal{C}, \mathcal{Z})\sum_{j\neq i}\frac{h_{Qij}^2}{(1 - h_{Qjj})^2} + \frac{1}{B^2}\mathbf{f}^T W(I - H_Q)\Psi_Q(I - H_Q)W\mathbf{f} \geq 0\end{aligned}$$

The proof is analogous to that of Propositions 1 and 2. The insights from Theorem 2 similarly extend variance estimation for the conditional average treatment effect: through using regression adjustments on the average *level* of the covariates in a given block results in less conservative variance estimators, with the degree of improvement now dependent on the extent to which the average of the weighted covariates in a given block are able to predict $w_i f_i$, the weighted conditional average treatment effect in a block given the covariate values.

In the case of equal block sizes, if the stratum-level treatment effects are homoskedastic (i.e. $\text{var}(\hat{\tau}_i \mid \mathcal{C}, \mathcal{Z})$ is constant across all blocks), then we are also entitled to an additional variance estimator connected to $HC2$ standard errors. Let $\tilde{\Psi}_Q$ be a diagonal matrix whose i^{th} diagonal element is $\tilde{\Psi}_{Qii} = 1/(1 - h_{Qii})$, and define $S_3^2(Q)$ as

$$S_3^2(Q) = \frac{1}{B^2}\hat{\boldsymbol{\tau}}^T W(I - H_Q)\tilde{\Psi}_Q(I - H_Q)W\hat{\boldsymbol{\tau}}, \tag{8}$$

Proposition 4. *If Q is constant across all elements of Ω, block sizes are equal (such that $W = I$), and $var(\hat{\tau}_i \mid \mathcal{C}, \mathcal{Z})$ is constant across blocks:*

$$\mathbb{E}[S_3^2(Q) \mid \mathcal{C}, \mathcal{Z}] - var(\hat{\Delta} \mid \mathcal{C}, \mathcal{Z}) = \frac{1}{B^2}\bar{\mathbf{f}}^T(I - H_Q)\tilde{\Psi}_Q(I - H_Q)\bar{\mathbf{f}} \geq 0.$$

The proof is deferred to the appendix. In the general case with across block heteroskedasticity, unequal block sizes, or when conducting inference on the the sample average treatment effect $S_3^2(Q)$ need not be conservative in expectation. It does, however, converge in probability to the same limiting value as $S_1^2(Q)$ and $S_2^2(Q)$, indicating that the prospect of anticonservative inference through $S_3^2(Q)$ may only be a realistic concern in small samples.

These developments demonstrate that the modes of inference presented for the sample average treatment effect in §4 yield harmonious extensions to inference on the conditional average treatment effect. That is, hypothesis tests and confidence intervals for the sample average treatment can also be interpreted as hypothesis tests and confidence intervals for the conditional average treatment effect should the practitioner deem the super-population formulation.

5.3 Population average treatment effect (PATE)

As an alternative super-population formulation, suppose we now consider the block sizes $\{n_1, ..., n_B\}$ as fixed, but the covariates within a given block, $\{\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i}\}$ as random. We now consider the pair of potential outcomes $\{r_{1ij}, r_{0ij}\}$ as having arisen through the following model:

$$\begin{aligned} \mathbf{x}_{ij} &= \boldsymbol{\zeta}_i + \varepsilon_{ij} \\ (r_{1ij}, r_{0ij}) \mid \mathbf{x}_{ij} &= (f_{1i}(\mathbf{x}_{ij}), f_{0i}(\mathbf{x}_{ij})) + (\epsilon_{1ij}, \epsilon_{0ij}), \end{aligned}$$

where $\boldsymbol{\zeta}_i$ are block-specific fixed effects, ε_{ij} are *iid* from some mean zero, finite variance distribution G, and the $(\epsilon_{1ij}, \epsilon_{0ij})$ are drawn *iid* from an arbitrary distribution F with mean $(0, 0)$ and block-specific variance-covariance matrix $\Sigma_{i\epsilon}$. Within this super-population abstraction, the *population* average treatment effect, or $PATE$, in a finely stratified experiment is defined as.

$$\bar{\Delta}^{(P)} = \sum_{i=1}^{B} (n_i/N) \int (f_{1i}(\boldsymbol{\zeta}_i + \varepsilon_{ij}) - f_{0i}(\boldsymbol{\zeta}_i + \varepsilon_{ij}))dG(\varepsilon_{ij}) \tag{9}$$

The classical weighted difference-in-means estimator $\hat{\Delta}$ remains an unbiased estimator for the population average treatment effect. Imai (2008) consider this model in a paired experiment with $\boldsymbol{\zeta}_i = \boldsymbol{\zeta}_0$ for all i. Therein, they demonstrate not only that the average of the paired differences yields an unbiased estimator for the average population average treatment effect, but that the classical variance estimator for the difference-in-means, S^2_P, is an *unbiased* estimator for $\text{var}(\hat{\Delta}|\mathcal{Z})$ regardless of whether or not the underlying treatment effect is additive.

It is here that we see the potential incongruity between inferential methods for the sample average treatment effect and for the population average treatment effect appear. The improvements presented herein empower the practitioner to use the average level of the covariates within a given block as a means to improve variance estimation when the sample or conditional average treatment effects are the targets of estimation. If the target is instead the population average treatment effect as formulated in this section, randomness in $\{\mathbf{x}_{ij}\}$ renders these conclusions inapplicable. As an illustration, consider the expectation of $S^2_1(Q_2)$ within this super-population formulation.

$$
\begin{aligned}
&\mathbb{E}[S_1^2(Q_2) \mid \mathcal{Z}] = \mathbb{E}[\mathbb{E}[S_1^2(Q_2) \mid \mathcal{Z}, \mathcal{C}]] \\
&= \mathbb{E}[\text{var}(\hat{\Delta} \mid \mathcal{Z}, \mathcal{C})] + \frac{1}{B^2}\mathbb{E}[\mathbf{g}^T W(I - H_{Q_2})W\mathbf{g} \mid \mathcal{Z}] \\
&= \text{var}(\hat{\Delta} \mid \mathcal{Z}) - \text{var}\left(\frac{1}{N}\sum_{i=1}^{B}\sum_{j=1}^{n_i}(f_{1i} - f_{0i}) \mid \mathcal{Z}\right) + \frac{1}{B^2}\mathbb{E}[\mathbf{g}^T W(I - H_{Q_2})W\mathbf{g} \mid \mathcal{Z}] \\
&= \text{var}(\hat{\Delta} \mid \mathcal{Z}) - \text{var}\left(\frac{1}{N}\sum_{i=1}^{B}\sum_{j=1}^{n_i}(f_{1i} - f_{0i}) \mid \mathcal{Z}\right) + \frac{1}{B^2}\mathbb{E}[\mathbf{g}^T W(I - H_{Q_1})W\mathbf{g} \mid \mathcal{Z}] \\
&- \frac{1}{B^2}\mathbb{E}[\mathbf{g}^T W H_M W\mathbf{g} \mid \mathcal{Z}] \\
&\approx \text{var}(\hat{\Delta} \mid \mathcal{Z}) + \frac{1}{B^2}\mathbb{E}[\mathbf{g} \mid \mathcal{Z}]^T W(I - H_{Q_1})W\mathbb{E}[\mathbf{g} \mid \mathcal{Z}] - \frac{1}{B^2}\mathbb{E}[\mathbf{g}^T W H_M W\mathbf{g} \mid \mathcal{Z}],
\end{aligned}
$$

where the approximation stems from ignoring the division by $\sqrt{1 - h_{Q_2ii}}$ in $g_i = \bar{f}_i/\sqrt{1 - h_{Q_2ii}}$ in the term $\mathbb{E}[\mathbf{g}^T W(I - H_{Q_1})W\mathbf{g} \mid \mathcal{Z}]$, a safe approximation in large samples. The last line need not be greater than $\text{var}(\hat{\Delta} \mid \mathcal{Z})$. For example, in the case where $\boldsymbol{\zeta}_i = \boldsymbol{\zeta}_0$ for all i, it will approximately equal $\text{var}(\hat{\Delta} \mid \mathcal{F}) - B^{-2}\mathbb{E}[\mathbf{g}^T W H_M W\mathbf{g} \mid \mathcal{Z}]$, meaning that it provides an underestimate. In short, the implications of this derivation are that effect modification cannot be safely exploited in variance estimation when conducting inference for the $PATE$ as defined in (9), while it can be exploited for inference on the $SATE$ and $CATE$. Valid inference for the sample average treatment effect using the developments in §4 may, or may not, be anti-conservative for inference on the population average treatment effect depending on whether or not the corresponding variance estimator was constructed using M, the average level of the covariates. As a consequence, $S_1^2(Q_2)$ and $S_2^2(Q_2)$ cannot be relied upon to yield valid inference for the $PATE$.

The above derivation also illustrates that if the covariate means $\bar{X}$ were not present, the issue with potentially anti-conservative variance estimation disappears. That is, $S_1^2(Q_1)$ and $S_2^2(Q_1)$ can safely be utilized when conducting inference on $PATE$ since these estimators do not exploit effect modification. Similar dissonance for variance estimates for population average treatment effects versus sample average treatment effects is also observed when conducting inference after regression adjustment in completely randomized experiments; compare, for example the suggested variance estimator of Pitkin et al. (2013) and Berk et al. (2013) to that of Lin (2013).

6 An exact test for additivity with power under linear effect modification

The developments of the previous sections naturally lend themselves to a new test of the null hypothesis of an additive treatment effect model when the researcher suspects the presence of effect modification on the basis of observed covariates. Suppose we want to test the null hypothesis of an additive treatment effect model against the alternative that there is effect heterogeneity,

$$
\begin{aligned}
&\mathbf{H}_o : \tau_{ij} = \bar{\Delta} \;\text{ for some } \bar{\Delta}, \;\text{ for all }\; i = 1, .., B; j = 1, .., n_i \\
&\mathbf{H}_a : \tau_{ij} \neq \tau_{i'j'}. \;\text{ for some }\; i, i', j, j'
\end{aligned}
$$

Let $F(\mathbf{Z})$ be the F-ratio for a partial F-test comparing a regression of $W\hat{\boldsymbol{\tau}}$ on Q_1 to one on $Q_2 = [Q_1, M] = [Q_1, (I - H_{Q_1})W\bar{X}]$ with observed treatment allocation $\mathbf{Z}$,

$$F(\mathbf{Z}) = \left(\frac{\hat{\boldsymbol{\tau}}^T W(I - H_{Q_1})W\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}^T W(I - H_{Q_2})W\hat{\boldsymbol{\tau}}}{\hat{\boldsymbol{\tau}}^T W(I - H_{Q_2})W\hat{\boldsymbol{\tau}}}\right)\frac{B - rank(Q2)}{K}$$
$$= \left(\frac{\hat{\boldsymbol{\tau}}^T W H_M W\hat{\boldsymbol{\tau}}}{\hat{\boldsymbol{\tau}}^T W(I - H_{Q_2})W\hat{\boldsymbol{\tau}}}\right)\frac{B - rank(Q2)}{K},$$

where the second line stems from orthogonality of M and Q_1. Small values for this ratio indicate that the reduction in residual variation from using Q_2 was modest relative to the model only containing Q_1. Large values for this ratio indicate substantial reduction in residual variation from exploiting effect modification through Q_2.

Note that while the null hypothesis specifies that the treatment effect is additive, it does not specify the value of the additive treatment effect. That is, in general the true value of the additive effect, call it $\bar{\Delta}_0$, is a nuisance parameter for the desired inference. Fortunately, our choice of test statistic eschews this dependence.

Proposition 5. *$F(\mathbf{Z})$ is a pivotal statistic for testing the null of an additive treatment effect in a finely stratified experiment.*

Proof. Suppose the null hypothesis was true and that the additive treatment effect equaled $\bar{\Delta}_0$. Note then that in the i^{th} block, $w_i\hat{\tau}_i$ can be written as

$$w_i\hat{\tau}_i = w_i\sum_{j=1}^{n_i}\left(Z_{ij}(r_{0ij} + \bar{\Delta}_0)/n_{1i} - (1 - Z_{ij})r_{0ij}/n_{0i}\right)$$
$$= w_i\bar{\Delta}_0 + w_i\sum_{j=1}^{n_i}\left(Z_{ij}r_{0ij}/n_{1i} - (1 - Z_{ij})r_{0ij}/n_{0i}\right),$$

Hence, the vector $W\hat{\boldsymbol{\tau}}_i$ can be broken into the sum of two vectors, one of which is a mean zero random variable, and the other being the deterministic vector $\bar{\Delta}_0 W\mathbf{e}$. To complete the proof, simply note that $W\mathbf{e}$ is in the columnspace of both Q_1 and Q_2, such that the term $\bar{\Delta}_0 W\mathbf{e}$ drops out of both the numerator and denominator of $F(\mathbf{Z})$.

Let t be the observed value of $F(\mathbf{Z})$ in the sample at hand. To compute a p-value corresponding to t, we can simply choose an arbitrary value for the additive treatment effect, say $\bar{\Delta}_0 = 0$, and compute the randomization distribution of $F(\mathbf{Z})$ which is entirely specified under the null,

$$p_{val} = \frac{1}{|\Omega|}\sum_{\mathbf{z}\in\Omega}\chi\left\{F(\mathbf{z}) \geq t \mid \mathcal{F}, \mathcal{Z}, \tau_{ij} = 0 \;\; \forall \;\; i, j\right\}, \tag{10}$$

where $\chi\{A\}$ is an indicator that the event A occurred.

Among alternatives to strict additivity, the test will be more powerful when there exists effect modification that is well modeled by a regression on Q_2. The test will not be particularly powerful when there exists heterogeneity that is not well modeled as a linear function of the covariates which compose Q_2. Regardless, the test will maintain the desired size even in finite samples as, for each fixed value of $\bar{\Delta}_0$, the distribution of $T(\mathbf{Z})$ can be computed exactly under the null of additivity.

Note that, in principle, other estimation procedures beyond linear regression could be used to compare sums of squared errors including and not including the observed covariates. In general, the corresponding test statistics will not be pivotal, meaning that their distribution could depend on the value of the additive treatment effect. This can be accommodated through the technique of Berger and Boos (1994) by first finding $1-\gamma$ confidence interval for the value of $\bar{\Delta}$ through inversion of randomization tests under the assumption of additivity (Rosenbaum, 2002), and finding the maximal p-value for the test statistic for values of $\bar{\Delta}_0$ within the confidence interval, and adding γ to the result. See Ding et al. (2015) for a recent application of this idea to testing for effect variation in completely randomized experiments.

7 Illustrations and simulations

7.1 Variance estimation in finely stratified experiments

We now explore the improvements in inference for the $SATE$ and $CATE$ that can be attained by exploiting effect modification, and illustrate our caveat about these benefits not extending to inference on the $PATE$. There are B blocks, $0.4B$ of which are triplets and $0.6B$ of which are pairs. The blocks are formed by taking an *iid* sample of a $k = 10$ dimensional vector of covariates $\mathbf{x}_i$, where each component is *iid* uniform on the interval [0,1]. Modifying the function utilized in the simulation study of Friedman (1991) to remove linear terms, for each block-level covariate vector $\mathbf{x}_i$ we then sample potential outcomes under treatment and control from the following distribution:

$$\begin{aligned} r_{1ij} &= a\left(10\sin(\pi x_{i1}x_{i2}) + 20(x_{i3}-1/2)^2 + 10\exp(x_{i4}) + 5(x_{i5}-1/2)^3\right) + b\epsilon_{ij} \qquad (11)\\ r_{0ij} &= 10\sin(\pi x_{i1}x_{i2}) + 20(x_{i3}-1/2)^2 + 10\exp(x_{i4}) + 5(x_{i5}-1/2)^3 + \epsilon_{ij}\\ \epsilon_{ij} &\overset{iid}{\sim} \mathcal{N}(0,1) \end{aligned}$$

For the simulations in this subsection, we set $B = 100$ for the number of strata, and $a = 2$ and $b = 2$ in (11). Under this specification the average treatment effect at the population level, $\bar{\Delta}^{(P)}$, is roughly 24.1, and is fixed as an estimand across samples. Further, var$(\hat{\Delta} \mid \mathcal{Z} = 0.437$. The sample average treatment effect, $\bar{\Delta}$, and the conditional average treatment effect, $\bar{\Delta}^{(C)}$, vary with each realization, as their definitions depend on $\mathcal{F}$ and $\mathcal{C}$ respectively. There is effect heterogeneity present, as $\mathbb{E}[\tau_{ij} \mid \mathcal{C}] = 10\sin(\pi x_{i1}x_{i2}) + 20(x_{i3}-1/2)^2 + 10\exp(x_{i4}) + 5(x_{i5}-1/2)^3$. In this generative model, $\mathbb{E}[\text{var}(\hat{\Delta} \mid \mathcal{C}, \mathcal{Z})] = 0.443$, and $\mathbb{E}[\text{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})] = 0.401$. In each simulation, we

1. Simulate covariates $\mathbf{x}_i$, $i = 1, ..., 100$ and potential responses (r_{1ij}, r_{0ij}), $j = 1, .., n_i$, setting $n_i = 3$ for 40 blocks and $n_i = 2$ for 60 blocks
2. Compute the variance estimators $S_1^2(Q)$, $S_2^2(Q)$, and $S_3^2(Q)$

We form the matrix Q used to compute the variance estimators in three ways,

1. *None.* Only including a constant column and the stratum weights .
2. *Correct.* Including a constant column, stratum weights, and weighted transformed covariates $w_i \sin(\pi x_{i1}x_{i2})$, $w_i(x_{i3}-1/2)^2$, $w_i \exp(x_{i4})$, $w_i(x_{i5}-1/2)^3$ (*Correct*).
3. *Incorrect.* Including a constant column, stratum weights, and weighted values for the original 10 covariates (without transformation), $w_i x_{ik}$, $k = 1, ..., 10$.

The functional form for effect modification is thus correctly specified within the second form, and incorrectly specified in the third form.

Table 1: Expectations for variance estimators for various matrices Q. Target expectations for valid inference on the $SATE$, $CATE$, and $PATE$ are 0.0401, 0.0443, and 0.437 respectively.

	Covariates in Q		
	None	Correct	Incorrect
$S_1^2(Q)$	0.437	0.0460	0.108
$S_2^2(Q)$	0.447	0.0474	0.126
$S_3^2(Q)$	0.437	0.0443	0.110

Table 1 shows the results of this simulation. We see that, as Propositions 1-4 guarantee, $S_1^2(\cdot)$ and $S_2^2(\cdot)$ remained conservative in expectation for the variances for estimating the $SATE$ and $CATE$ for all choices of Q. Using the correctly specified covariates allows the expectations to come closest to the true values for the variances, while the incorrect specification still performs substantially better than ignoring the covariates altogether. For inference on the $PATE$, we see that only the choice of Q which ignores the covariates yields a valid estimator for the variance; the choices incorporating the covariates would produce substantially anticonservative inference for the population average treatment effect. While not guaranteed to be as such in this simulation, we see that $S_3^2(\cdot)$ produced estimators which remained conservative in expectation for inference on the $SATE$ and $CATE$ for all three choices of Q, and for the $PATE$ through the choice of Q ignoring the covariates.

7.2 Testing for effect heterogeneity

We now demonstrate the test for effect modification proposed in §6. We use a similar generative model for the covariates as was employed in §7.1, but we instead set $b = 1$ and conduct the test for effect heterogeneity for increasing values of a in (11). At $a = 1$, the null hypothesis of additivity is true; all values $a \neq 1$ imply that the null is false. We also set $B = 20$ as a means of illustrating the exactness of the test. As in the previous section, we conduct the test utilizing both the correct and incorrect specification for the functional form of the covariates in forming the matrix Q_2. Hence, in each iteration we

1. Simulate covariates $\mathbf{x}_i$, $i = 1, ..., 20$ and potential responses (r_{1ij}, r_{0ij}), $j = 1, .., n_i$, setting $n_i = 3$ for 8 blocks and $n_i = 2$ for 12 blocks

2. Randomly allocate individuals to treatment or control in accordance with the finely stratified design, recording the observed outcomes and the value for the test statistic $F(\mathbf{Z})$

3. Estimate the permutation p-value in (10) through Monte Carlo simulation.

We set $\alpha = 0.05$ for this study. Figure 1 shows the power of our testing procedure as a function of a, under both correct and incorrect specifications for Q_2. Note that at $a = 0$ our tests have the correct size. As a increases, the power increases for both choices of Q_2, but more rapidly for the correct choice of Q_2. Hence, while the specification of Q_2 affects the power of the test, it does not affect its validity in terms of maintaining the desired Type I error rate.

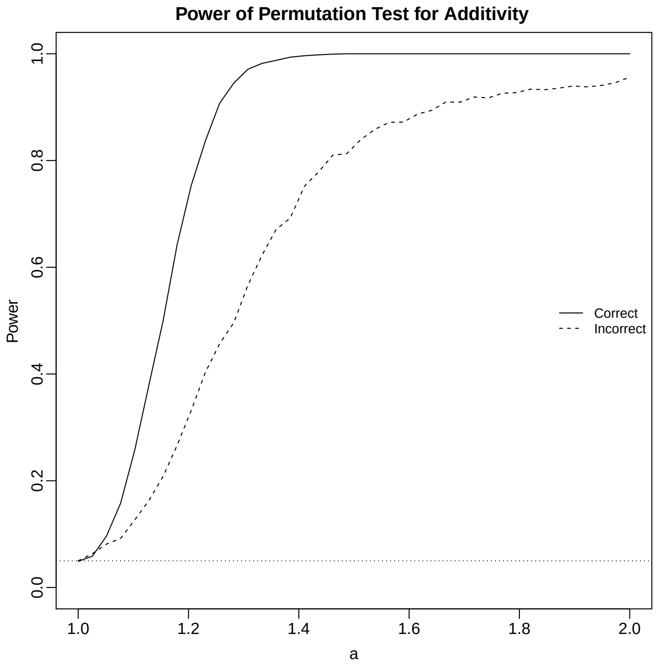

Figure 1: The power of the permutation test for effect modification under both correct and incorrect covariate specification as a function of a, which controls the departure from additivity. $a = 1$ corresponds to the null being true. The horizontal line corresponds to $\alpha = 0.05$, the desired size in the simulation.

7.3 Pairs or quartets?

In this study, we explore the extent to which the analytical limitations of paired experiments described in the introduction are mitigated by the methods presented herein. In our simulation study, we have $N = 80$ total individuals. We have a single covariate for each individual. In our study, we fix the observed values for the covariate at $\mathbf{x} = (0.25, 0.25, 0.5, .0.5, ..., 10, 10)$. Hence, there are 40 pairs of individuals who share the same value for the covariate, hence forming natural pairs. Due to concerns over the analytical limitations of paired experiments, the practitioner may instead choose to create 20 quartets of individuals, namely those taking on values $\{0.25, 0.25, 0.5, 0.5\}$, $\{0.75, 0.75, 1, 1\}$,....$\{9.75, 9.75, 10, 10\}$.

For each experiment, we generate potential outcomes for individual j as

$$r_{1j} = 100 + 30x_j + \epsilon_j, \quad r_{0j} = 20x_j + \epsilon_j, \quad \epsilon_j \overset{iid}{\sim} \mathcal{N}(0, 10^2).$$

x_j is thus an effect modifier, as the treatment effect for individual j is $100 + 10x_j$. We first consider two situations: one in which the practitioner has access to $\mathbf{x}$ itself, and one in which the practitioner instead has access to $b_j = \exp(x_j/3)$ for each individual. Let Q_x be the $B \times 2$ matrix with $\mathbf{e}$ in the first column and the value of x_j defining each pair in the second column, and let Q_b be the analogous for the incorrectly specified covariate b_j . We imagine the target of inference is the conditional average treatment effect, and hence seek to estimate $\text{var}(\hat{\Delta} \mid \mathcal{C}, \mathcal{Z})$ for both the paired and quartet design. Under this specification, we simply utilize the bias formulae in Propositions 1, 2 and 4 to calculate the expectation for the variance estimators S^2_P, $S^2_\ell(Q_x)$ and $S^2_\ell(Q_b)$ for $\ell = 1, 2, 3$. For the quartet design, we simply compute the true value of the variance, along with the bias of the conventional estimator S^2_{CS}.

Table 2: A comparison of variance estimation in the paired and quartet designs

	Pairs					Quartets	
Covariates	$\text{var}(\widehat{\Delta})$	S^2_P	S^2_1	S^2_2	S^2_3	$\text{var}(\widehat{\Delta})$	S^2_{CS}
	5.00	21.35	-	-	-	5.65	5.68
Correct, Linear	-	-	5.09	5.26	5.00	-	-
Incorrect, Linear	-	-	7.12	8.76	8.25	-	-
Correct, Cubic	-	-	5.52	5.59	5.00	-	-
Incorrect, Cubic	-	-	7.23	6.06	5.24	-	-

Table 2 contains relevant numerical information for comparing estimation under the two designs. We first see that the true variance under the paired design is smaller than that under the quartet design (5.00 vs 5.65), such that if we had access to this true variance the paired design would undoubtedly be preferred; however, the classical variance estimator in a paired experiment has an expectation of 21.35, while the classical variance estimator in the quartet experiment has an expectation of 5.68. This comparison of conventional variance estimators highlights the motivation for recommending quartet experiments over paired experiments within the literature. In the rows labeled “Correct, Linear" and “Incorrect, Linear" we see the improvements in variance estimation both under proper specification and misspecification of the response function. When the response function is properly specified both $S^2_1(Q_x)$ and $S^2_2(Q_x)$ are lower in expectation than S^2_{CS}, showing that proper modeling of the treatment effect heterogeneity provides variance estimators whose expectations are smaller than that of the coarsely stratified experiments. When the heterogeneity is not properly modeled, improvements in the variance estimator are still attained; however, the expectations for the variance estimators now exceed that from the quartet experiment.

Recall once again that the findings of Propositions 1 and 2 facilitate conservative estimation of the variance for any fixed matrix Q. This allows us, before conducting the experiment, to decide to include polynomial terms in the matrix Q to more flexibly model the relationship between the covariates at hand and the treatment effects. Suppose we now add quadratic and cubic terms of $\mathbf{x}$ and $\mathbf{b}$, calling the corresponding matrices Q_{x3} and Q_{b3} respectively. In the last two rows of 2, we consider the performance of $S^2_1(Q_{x3}), S^2_1(Q_{b3})$, $S^2_2(Q_{x3})$, and $S^2_2(Q_{b3})$. We see that both $S^2_1(Q_{x3})$ and $S^2_2(Q_{x3})$ have a larger expectation than what was attained when we omitted the polynomial terms. By adding two more predictor variables, the sum of diagonals of the correspond hat matrix increases from 2 to 4, hence resulting in additional inflation of residuals. For $S^2_1(Q_{b3})$, we see that this inflation has also swamped any benefit from flexibility in modeling as its expectation is larger than that of $S^2_1(Q_b)$. For $S^2_2(Q_{b3})$, we see that the additional flexibility has been beneficial, and the expectation for the variance estimator has decreased relative to $S^2_2(Q_{b3})$, although not enough to fall below the level of S^2_{CS}.

A component of the remaining conservativeness of the estimators $S^2_1(Q_{b3})$ and $S^2_2(Q_{b3})$ stems from our variance estimators being unbiased in expectation regardless of the degree of heteroskedasticity across blocks, and hence having to be inflated in the presence of high leverage points. If one is willing to do away with the requirement, $S^2_3(Q_{b3})$ becomes an appealing estimator. This estimator combines elements of $S^2_1(Q)$, essentially adjusting the variance estimator by $1/(1-h_{Qii})$ instead of $1/(1-h_{Qii})^2$, and $S^2_2(Q)$, by adjusting residuals instead of responses to account for influential points. The column labeled $S^2_3(\cdot)$ corresponds to this estimator. In the setting considered herein, it is exactly unbiased with Q_x and Q_{b3}, owing to the fact that $\text{var}(\hat{\tau}_i \mid \mathcal{C}, \mathcal{Z})$ is constant across pairs. It is necessarily less conservative than $S^2_2(\cdot)$ for all four choices of Q considered, and it is less conservative than $S^2_1(\cdot)$ for all choices of Q except for Q_b. As B decreases the estimators all converge to the same limit, yet here we see the potential benefits of using the estimator $S_3^2(\cdot)$.

8 An example: The Children's Television Workshop Experiment

Ball et al. (1973) designed an experiment to evaluate an educational television program which sought to improve reading skills for young children. §10.7 of Imbens and Rubin (2015) examined a subset of the experiment conducted in Youngstown, Ohio with $B = 8$ primary schools. In each school, a pair of first-grade classes was selected, with one class in each pair assigned to watch the show during reading class and the other class assigned to continuing with the usual curriculum. Each class has a pre-test score assessing average reading ability, x_{ij} in our notation, along with a post-test after the experiment, $R_{ij} = Z_{ij}r_{1ij} + (1 - Z_{ij})r_{0ij}$, where Z_{ij} is 1 if the class was shown the educational program *The Electric Company*. $\hat{\tau}_i$ is the difference between the observed treatment and control scores on the post-test in the ith pair.

In this data set, the conventional difference-in-means estimator was $\hat{\tau} = 13.4$, with an observed value of the conventional standard error of $S_P = 4.6$. We now consider using the estimators developed herein to improve upon this standard error estimate. We include linear and quadratic terms in the covariates, defining $Q_2 = [\mathbf{e}, (I - \mathbf{e}\mathbf{e}^T/B)\bar{X}]$, where the $\{i, 1\}$ entry of $\bar{X}$ is $\bar{x}_{i1} = (x_{i1} + x_{i2})/2$, and the $\{i, 2\}$ entry of $\bar{X}$ is $\bar{x}_{i2} = (x_{i1}^2 + x_{i2}^2)/2$. The values for $S_1(Q_2)$, $S_2(Q_2)$, and $S_3(Q_2)$ are 4.2, 4.34, and 3.57 respectively. All three estimators would thus facilitate the construction of narrower confidence intervals than the ones constructed using S_P while maintaining the conclusion that the treatment was effective at $\alpha = 0.05$. As noted, $S_3^2(Q_2)$ is not in general unbiased for the variance when the target of inference is the sample average treatment effect, so the discrepancy between this estimator and the other two may well stem from downwards bias. This concern is not relevant for the other two estimators, a reason to prefer them particularly in small samples.

9 Discussion

When the target of estimation is either the sample or the conditional average treatment effect, the developments presented in this work facilitate improved variance estimation for finely stratified experiments for inference conducted based on both the conventional difference-in-means estimator and estimators utilizing regression adjustment. As the simulation study in §7.3 illustrated, these have certainly mitigated the analytical limitations of finely stratified experiments by providing more powerful inference than that available through classical variance estimators, yet the analytical issues have not been entirely resolved. If the regression model is grossly misspecified, the variance estimators presented herein may not provide an improvement over that of an experiment with blocks of size four.

One direction for future research is investigating the extent to which the improvements presented in this work can be employed in the super-population setting considered by van der Laan et al. (2012) wherein rather than pairs being drawn *iid*, individuals are drawn *iid* and then optimally paired after being selected into the study. More generally, the nature of these improvements raise additional questions about the extent to which inference on local estimands in randomized experiments should be transferable to population-level estimands in popular super-population formulations. With respect to the conventional variance estimator in a completely randomized experiment Imbens and Rubin (2015) describe the consonance between finite-population and super-population inference through this variance estimator as an "attractive property." (Imbens and Rubin, 2015, §6.7, p.101). Yet as was noted in §6.3, variance estimators which exploit effect heterogeneity can yield anticonservative inference at the level of the $PATE$ as defined in Imai (2008). Our perspective is that rather than detracting from the appeal of these new estimators, this dissonance forces the researcher to critically assess the question, "to whom does the inference apply?" The answer is often left ambiguous in the analysis of randomized experiments, and *iid* assumptions are often made vacuously, without consideration of the true nature of the process by which the data came to be and the corresponding ramifications for the integrity of the performed inference.

A Lemmas

Lemma 1. *Under Conditions 1-3,* $B^{-1}\sum_{i=1}^{B} w_i\hat{\tau}_i m_{ik}$ *converges in probability to* $\lim\limits_{n\to\infty} B^{-1}\sum_{i=1}^{n} w_i\bar{\tau}_i m_{ik}$ *for any* $k = 1, ..., rank(M)$*. Further, and* $B^{-1}\sum_{i=1}^{B} w_i^2\hat{\tau}_i$ *and* $B^{-1}\sum_{i=1}^{B} w_i\hat{\tau}_i$ *converges in probability to* $\lim\limits_{n\to\infty} B^{-1}\sum_{i=1}^{n} w_i^2\bar{\tau}_i$ *and* $\lim\limits_{n\to\infty} B^{-1}\sum_{i=1}^{n} w_i\bar{\tau}_i$ *respectively.*

Proof. We prove the result for $B^{-1}\sum_{i=1}^{B} w_i\hat{\tau}_i m_{ik}$; the proof for the remaining two weighted sums are analogous. For any k, $\mathbb{E}[B^{-1}\sum_{i=1}^{B} w_i\hat{\tau}_i m_{ik} \mid \mathcal{F}, \mathcal{Z}] = B^{-1}\sum_{i=1}^{B} w_i\bar{\tau}_i m_{ik}$, which has a finite limit by Condition 3. We now show that $\text{var}(B^{-1}\sum_{i=1}^{n} w_i\hat{\tau}_i m_{ik} \mid \mathcal{F}, \mathcal{Z})$ converges to zero.

$$\begin{aligned}
\text{var}\left(B^{-1}\sum_{i=1}^{n} w_i\hat{\tau}_i m_{ik} \mid \mathcal{F}, \mathcal{Z}\right) &= B^{-2}\sum_{i=1}^{B} w_i^2\left(\sigma_{1i}^2/n_{1i} + \sigma_{0i}^2/(n_{0i}) - \sigma_{\tau i}^2/n_i\right)(m_{ik})^2 \\
&\leq B^{-2}\left\{\sum_{i=1}^{B}\left(\sum_{j=1}^{n_i} w_i^2(r_{1ij}^2 + r_{0ij}^2)\right)^2\right\}^{1/2}\left\{\sum_{i=1}^{B} m_{ik}^4\right\}^{1/2} \\
&\leq B^{-2}\left\{\sum_{i=1}^{B} n_i^2\left(\sum_{j=1}^{n_i} w_i^4(r_{1ij}^4/n_i + r_{0ij}^4/n_i)\right)\right\}^{1/2}\left\{\sum_{i=1}^{B} m_{ik}^4\right\}^{1/2} \\
&\leq C_1C_2/B
\end{aligned}$$

by Conditions 2 and 3, which tends to zero as $B \to \infty$. Chebyshev's inequality and Condition 3 complete the proof.

Lemma 2. *Under Condition 2,* $h_{ii} \to 0$.

Proof. From Condition 2, we have that $B^{-1}M^TM$ converges to a finite, invertible matrix; let $\Lambda = (\lim_{B\to\infty} B^{-1}M^TM)^{-1}$. Note that $h_{ii} = m_i^T(M^TM)^{-1}m_i = B^{-1}(m_i)^T(B^{-1}M^TM)^{-1}(m_i)$

$$\lim_{B\to\infty} h_{ii} = \lim_{B\to\infty} B^{-1}m_i^T\Lambda m_i = 0$$

Lemma 3. *Under Conditions 1-3 and conditional on* $\mathcal{F}, \mathcal{Z}$,

$$B^{-1}\sum_{i=1}^{B} w_i^2\hat{\tau}_i^2 \xrightarrow{p} \lim_{B\to\infty} B^{-1}\sum_{i=1}^{B} w_i^2\left(\bar{\tau}_i^2 + \sigma_{1i}^2/n_{1i} + \sigma_{0i}^2/(n_{0i}) - \sigma_{\tau i}^2/n_i\right),$$

Proof. We have that $\mathbb{E}[B^{-1}\sum_{i=1}^{B} w_i^2\hat{\tau}_i^2 \mid \mathcal{F},\mathcal{Z}] = B^{-1}\sum_{i=1}^{B} w_i^2(\bar{\tau}_i^2 + \sigma_{1i}^2/n_{1i} + \sigma_{0i}^2/(n_{0i})\ -\sigma_{\tau i}^2/n_i)$. It suffices to show that $\mathrm{var}(B^{-1}\sum_{i=1}^{B} w_i^2\hat{\tau}_i^2 \mid \mathcal{F},\mathcal{Z})$ converges to zero.

$$\begin{aligned}
&\mathrm{var}\left(B^{-1}\sum_{i=1}^{B} w_i^2\hat{\tau}_i^2 \mid \mathcal{F},\mathcal{Z}\right)\\
&= B^{-2}\sum_{i=1}^{B}\mathrm{var}(w_i^2\hat{\tau}_i^2 \mid \mathcal{F},\mathcal{Z}) \le B^{-2}\sum_{i=1}^{B}\mathbb{E}[w_i^4\hat{\tau}_i^4 \mid \mathcal{F},\mathcal{Z}]\\
&\le B^{-2}\sum_{i=1}^{B} n_i^2 w_i^4\mathbb{E}\left[\left(\sum_{j=1}^{n_i}(Z_{ij}r_{1ij'}^2/n_{1i}^2 - (1-Z_{ij})r_{0ij}^2/n_{0i}^2)\right)^2\right]\\
&\le B^{-2}\sum_{i=1}^{B} n_i^2 w_i^4\mathbb{E}\left[\left(\sum_{j=1}^{n_i} Z_{ij}r_{1ij}^2/n_{1i}^2\right)^2 + \left(\sum_{j=1}^{n_i}(1-Z_{ij})r_{0ij}^2/n_{0i}^2\right)^2\right]\\
&\le B^{-2}\sum_{i=1}^{B} n_i^3 w_i^4\sum_{j=1}^{n_i} r_{1ij}^4/n_{1i}^4 + B^{-2}\sum_{i=1}^{B} n_i^3 w_i^4\sum_{j=1}^{n_i} r_{0ij}^4/n_{0i}^4\\
&\le 2C_1^4C_2/B,
\end{aligned}$$

which tends to zero as $B\to\infty$.

B Proof of Theorem 1

Noting that the random variables $w_i\hat{\tau}_i$ are independent, it suffices to show that the triangular array version of Lyapunov's condition is satisfied. Let $s_B^2 = \sum_{i=1}^{B}\mathrm{var}(w_i\hat{\tau}_i \mid \mathcal{F},\mathcal{Z})$. As was demonstrated in the proof of Lemma 3, $B^{-1}\sum_{i=1}^{B}\mathbb{E}[w_i^4\hat{\tau}_i^4 \mid \mathcal{F},\mathcal{Z}] \le C_1^*$ for a constant C_1^*. By Condition 2, $B^{-1}\sum_{i=1}^{B} w_i^4\bar{\tau}_i^4 \le C_2$ for a constant C_2. Further, by Condition 3 we have that $s_B^2/B = B^{-1}\sum_{i=1}^{B} w_i^2\left(\sigma_{1i}^2/n_{1i} + \sigma_{0i}^2/(n_{0i}) - \sigma_{\tau i}^2/n_i\right)$ has a finite limit as $B\to\infty$, call it L^*. Hence, using a standard moment inequality,

$$\begin{aligned}
\lim_{B\to\infty}\frac{1}{s_B^4}\sum_{i=1}^{B}\mathbb{E}[|w_i\hat{\tau}_i - w_i\bar{\tau}_i|^4 \mid \mathcal{F},\mathcal{Z}] &\le \lim_{B\to\infty}\frac{8}{B^2(s_B^2/B)^2}B\sum_{i=1}^{B}\mathbb{E}[w_i^4\hat{\tau}_i^4 \mid \mathcal{F},\mathcal{Z}]/B + w_i^4\bar{\tau}_i^4/B\\
&\le \lim_{B\to\infty}\frac{8}{B^2L^*}B(C_1^* + C_2) = 0.
\end{aligned}$$

The conditions for Lyapunov's Central Limit Theorem are thus satisfied for $s_B^{-1}\left(\sum_{i=1}^{B} w_i(\hat{\tau}_i - \bar{\tau}_i)\right)$.

C Proof of Theorem 2

We prove the result for $S_1^2(Q_2)$ in the case of unequal block sizes. Let $\boldsymbol{\eta}_{Q_1} = [\bar{\Delta}, B^{-1}\lim_{B\to\infty}(w_i - 1)w_i\hat{\tau}_i]$. Let Σ_{Q_1} be a 2×2 diagonal matrix with $\Sigma_{Q_1 11} = 1$ and $\Sigma_{Q_1 22} = \lim_{B\to\infty}B^{-1}\sum_{i=1}^{B}(w_i - 1)^2$. Let $\boldsymbol{\beta}_{Q_1} = \Sigma_{Q_1}^{-1}\boldsymbol{\eta}_{Q_1}$. Recalling that Q_1 and M are orthogonal, we decompose $BS_1^2(Q_2)$ as

$$\begin{aligned}
BS_2^2(Q_2) &= B^{-1}(y^TW(I - H_{Q_2})Wy)\\
&= B^{-1}\left(y^TWWy - y^TWQ_1(Q_1^TQ_1)^{-1}Q_1^TWy - y^TWM(M^TM)^{-1}M^TWy\right)
\end{aligned}$$

$B^{-1}y^T WWy$ converges in probability to $\lim B^{-1}(\sum_{i=1}^B w_i^2(\bar{\tau}_i^2 + \sigma_{1i}^2/n_{1i} + \sigma_{0i}^2/(n_{0i}) - \sigma_{\tau i}^2/n_i))$ by Lemmas 1 and 3. By Lemmas 2 and 3, $B^{-1}y^T WQ_1(Q_1^T Q_1)^{-1}Q_1^T Wy$ converges in probability to $\boldsymbol{\beta}_{Q_1}^T \Sigma_{Q_1} \boldsymbol{\beta}_{Q_1}$, and $B^{-1}y^T WM(M^T M)^{-1}M^T Wy$ converges in probability to $\boldsymbol{\beta}_M^T \Sigma_M \boldsymbol{\beta}_M$. Hence,

$$B\left(S_1^2(Q_2) - \mathrm{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z})\right) \xrightarrow{p} B^{-1}\bar{\boldsymbol{\tau}}^T W(I - H_{Q_2})W\bar{\boldsymbol{\tau}}$$

as desired. The proof for $S_1^2(Q_1)$ simply follows by eliminating the terms above pertaining to the matrix M. The proofs for $S_2^2(Q_1)$ and $S_2^2(Q_2)$ are analogous.

D Proof of Proposition 4

Proof. Define $\bar{\mathbf{f}}$ as before, and let $\tilde{\Sigma}$ be the covariance matrix for $\hat{\boldsymbol{\tau}} \mid \mathcal{C}$, which by assumption homoskedasticity and equal block sizes has constant diagonal elements, call them ν.

$$B^2 E[S_3^2(Q) \mid \mathcal{F}, \mathcal{Z}] = tr(\tilde{\Sigma}(I - H_Q)\tilde{\Psi}_Q(I - H_Q)) + \bar{\mathbf{f}}^T(I - H_Q)\tilde{\Psi}_Q(I - H_Q)\bar{\mathbf{f}}.$$

The trace of $\tilde{\Sigma}(I - H_Q)\tilde{\Psi}_Q(I - H_Q)$ is given by

$$\begin{aligned}
tr(\tilde{\Sigma}W(I - H_Q)\tilde{\Psi}_Q(I - H_Q)W) &= \nu \sum_{i=1}^B \left((1 - h_{Qii}) + \sum_{j \neq i} \frac{h_{Qij}^2}{1 - h_{Qjj}}\right) \\
&= \nu \sum_{i=1}^B \left((1 - h_{Qii}) + \sum_{j \neq i} \frac{h_{Qij}^2}{1 - h_{Qii}}\right) \\
&= \nu \sum_{i=1}^B (1 - h_{Qii} + h_{Qii}) = B\nu
\end{aligned}$$

The second line utilizes symmetry of $I - H_Q$, while the third utilizes idempotence of H_Q, implying that $\sum_{j \neq i} h_{Qij}^2 = h_{Qii}(1 - h_{Qii})$. Noting that $\mathrm{var}(\hat{\Delta} \mid \mathcal{F}, \mathcal{Z}) = \nu/B$ under the assumptions of the proposition and that $(I - H_Q)\tilde{\Psi}_Q(I - H_Q)$ is positive semidefinite completes the proof.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2017). Sampling-based vs. design-based uncertainty in regression analysis. *arXiv preprint arXiv:1706.01778.*

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

Aronow, P. M. and Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1):135–154.

Ball, S., Bogatz, G., Rubin, D., and Beaton, A. (1973). *Reading with Television: An Evaluation of The Electric Company. A Report to the Children's Television Workshop. Volumes 1 and 2.*, volume 1 and 2. ERIC, Princeton, NJ.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016.

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., and Zhao, L. (2013). Covariance adjustments for the analysis of randomized field experiments. *Evaluation review*, 37(3-4):170–196.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhao, L., and Zhang, K. (2014). Models as approximations, part i: A conspiracy of nonlinearity and random regressors in linear regression. *arXiv preprint arXiv:1404.1578*.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*. c. Wiley.

Cox, D. R. (1958). *Planning of experiments*. Wiley, New York.

Ding, P. (2016). A paradox from randomization-based causal inference. *Statistical Science*, to appear.

Ding, P., Feller, A., and Miratrix, L. (2015). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd.

Fogarty, C. B. (2016). Regression assisted inference for the average treatment effect in paired experiments. *arXiv preprint arXiv:1612.05179*.

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

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, pages 1–67.

Gadbury, G. L. (2001). Randomization inference and bias of standard errors. *The American Statistician*, 55(4):310–313.

Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618.

Higgins, M. J., Sävje, F., and Sekhon, J. S. (2016). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences*, 113(27):7369–7376.

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

Imai, K., King, G., Nall, C., et al. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, 24(1):29–53.

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

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kallus, N. (2013). Optimal a priori balance in the design of controlled experiments. *arXiv preprint arXiv:1312.0531*.

Klar, N. and Donner, A. (1997). The merits of matching in community intervention trials: a cautionary tale. *Statistics in medicine*, 16(15):1753–1764.

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

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Lu, J. (2016). Covariate adjustment in randomization-based causal inference for 2k factorial designs. *Statistics & Probability Letters*, 119:11–20.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (in Polish). *Roczniki Nauk Roiniczych*, X:1–51. Reprinted in Statistical Science, 1990, 5(4):463-480.

Pitkin, E., Berk, R., Brown, L., Buja, A., George, E., Zhang, K., and Zhao, L. (2013). Improved precision in estimating average treatment effects. *arXiv preprint arXiv:1311.0291*.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York.

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

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

van der Laan, M. J., Balzer, L. B., and Petersen, M. L. (2012). Adaptive matching in randomized trials and observational studies. *Journal of statistical research*, 46(2):113.