Improving Experiments by Optimal Blocking: Minimizing the Maximum Within-block Distance

Michael J. Higgins* Fredrik Sävje† Jasjeet S. Sekhon‡

First Draft: 11/2012

This Draft: 8/26/2014 (02:02)

Abstract

We develop a new method for blocking in randomized experiments that works for an arbitrary number of treatments. We analyze the following problem: given a threshold for the minimum number of units to be contained in a block, and given a distance measure between any two units in the finite population, block the units so that the maximum distance between any two units within a block is minimized. This blocking criterion can minimize covariate imbalance, which is a common goal in experimental design. Finding an optimal blocking is an NP-hard problem. However, using ideas from graph theory, we provide the first polynomial time approximately optimal blocking algorithm for this problem. We derive the variances of estimators for sample average treatment effects under the Neyman-Rubin potential outcomes model for arbitrary blocking assignments and an arbitrary number of treatments.

*Postdoctoral Research Associate, Program for Quantitative and Analytical Political Science, Princeton University. http://www.princeton.edu/~mjh5/

†PhD Candidate, Department of Economics, Uppsala University. http://fredriksavje.com

‡Robson Professor of Political Science and Statistics. UC Berkeley. http://sekhon.berkeley.edu

1 Introduction

Properly executed random treatment assignment in controlled experiments guarantees that one will estimate the true treatment effect—in expectation. For any specific assignment there may, however, be remarkable imbalances on prognostically important covariates. Allowing for such imbalances can lead to grossly inaccurate estimates: viewed unconditionally, the treatment effect estimator will have a high variance; viewed conditionally on the observed imbalances, it will be biased.[1]

Choosing a mode of analysis that is less sensitive to such imbalances—for example post-stratification (Miratrix et al., 2013) or model-based adjustments[2]—it a common way to mitigate this problem. However, the methodological literature largely agrees that imbalances should, if possible, be avoided all together: one should adjust the *design* of the experiment.

There are two common designs that can improve covariate balance in experiments: *blocking*—where one divides the sample into homogeneous groups of units and randomize within those groups—and *re-randomization*—where one randomizes as usual but only accept assignments fulfilling a balance criteria. While the block design has a sturdy theoretical foundation, going back at least to Fisher (1926), it is re-randomization that has arguably seen the most use, although the practice is seldom reported, and it has been given little theoretical attention until recently.

Both designs have their merits. On the one hand, a re-randomization design (Morgan and Rubin, 2012) only requires that the investigator pre-specifies a balance criteria, and it can, thus, easily be applied to highly complex and non-standard experimental settings. However, tractable standard errors require strong distributional assumptions. The

[1] For a deeper discussion see, e.g., Rubin (2008); Keele et al. (2009); Worrall (2010).

[2] The literature on model-based adjustments is vast, a small subset of the discussions can be found in Duflo et al. (2007); Freedman (2008); Rubin and van der Laan (2011); Lin (2012) and the references therein.

investigator is therefore often restricted to permutation based inference when using a re-randomization design.

With a block design, on the other hand, the investigator is forced to make a choice: either only use a crude grouping of units, e.g. based on a small subset of discrete covariates, which may not solve the imbalance problem to an acceptable level; or use computationally demanding methods that are restricted to produce *matched-pairs*—blocks consisting of only two units. Note that, if interest is in the treatment effect in the sample, the second choice is particularly undesirable as the small group sizes makes it impossible to estimate standard errors (Imbens, 2011). Whenever an investigator has access to a reasonably large set of baseline information and is interested in sample treatment effects, or has more than two treatment categories, neither option is very appealing.

In this paper we provide the first method that can, in polynomial time and from an arbitrary set of covariates, produce a blocking satisfying an arbitrary group size requirement with a proven optimality. We thereby make the blocking trade-off that investigators usually face superfluous in many common situations.

We consider the following blocking problem: Choose a set of *block covariates* for which balance is desired. Select a distance metric that measures the similarity of block covariates between any two units, e.g. Mahalanobis or Euclidian distances. Choose a threshold for the minimum *block size*—the least number of units that must be assigned to each block. Block units so that the size of each block is at least this threshold, and so that the *maximum within-block distance*—the maximum distance between any two units within a block—is minimized. We call such a blocking an *optimal blocking*.

Note that this blocking problem differ from the standard problem in two aspects. First, instead of specifying a sharp limit—a size that all blocks must be—this problem specifies a size threshold. Unlike a sharp limit, which require sample sizes that are divisible by the desired block size, a size threshold can accommodate any sample. Furthermore, if there are

natural clusters of units in the sample the sharp limit could be forced to match units from different clusters. With a size threshold, however, the clusters will in general be preserved if there is no suitable way to split them. Of course, the sharp limit blocking is always acceptable under a size threshold implying that the maximum distance is weakly lower with a threshold. Second, instead of an objective function based on the mean, or summed, distances this problem consider their maximum. While this shift is the main cause of the swiftness of our algorithm, in the case of blocking, stratifying, and matching, minimizing the maximum distance is also a natural formalization going back to at least Cochran's observation that a few large distances often are more problematic than many small ones.[3]

Finding an optimal blocking is an NP-hard problem (Hochbaum and Shmoys, 1986) thus not feasible in most settings. Instead we introduce a fast algorithm that can find an *approximately optimal blocking*—a blocking with maximum within-block distance within a factor of four of the smallest distance possible—in polynomial time. This makes this blocking method applicable in most situations. In particular, compared to existing choices our method is well-suited for experiments where the sample is large or when block sizes of more than two are needed (either for standard error estimation or when there are several treatment arms).[4] However, in small experiments where block sizes of two are acceptable other methods, for example nonbipartite matching (Greevy et al., 2004), will often yield better results.

Specifically, our blocking method builds on Rosenbaum's (1989) observation that optimal matching and blocking problems can be thought of as partitioning problems in graph theory. The sample can be represented by graph, where the vertices are the experimental

[3]How these two aspects affect the variance in the end depends on how the covariates are related to the outcome. For example, using a size threshold requires that the covariates, to some degree, are predictive of the outcome—if they are completely uninformative, a sharp limit will always be better even if a threshold produces better balance.

[4]With the R package implementing our algorithm, a sample of 200,000 units can be blocked using a size threshold of four in less than an hour on an ordinary desktop computer.

units and the edges and their length represent the distances between the units. A blocking then corresponds to a subgraph of this graph that only contains edges between units in the same block. Our method relies upon finding subgraphs such that the length of the longest edge included in the subgraph is minimized, mirroring the min-max objective of the blocking problem. Problems of this type are called *bottleneck problems*. Unlike the sum or mean problem, which often has to be solved using heuristic algorithms with unknown level of optimality (e.g., k-way equipartition and k-means clustering Mitchell, 2001; Ji and Mitchell, 2005; Steinley, 2006), the bottleneck problem can be solved more efficiently and with proven optimality.

As a consequence of the higher accuracy granted by the block design, the standard variance estimators, not taking the design into consideration, will generally not be correct and overestimate the uncertainty. An early contribution realizing this is Neyman (1935) that discusses estimation with a block design under his potential outcomes model—now known as the Neyman-Rubin causal model (NRCM, Splawa-Neyman et al., 1990; Rubin, 1974; Holland, 1986). In more recent work, variance for the matched-pairs design under NRCM has been analyzed by Abadie and Imbens (2008, looking at conditional average treatment effects, CATE, as estimands) and Imai (2008, population average treatment effects, PATE). Imbens (2011) provides an overview and also discusses the impossibility of variance estimation of the sample average treatment effect (SATE) with a matched-pairs design.

In the second part of this paper we build upon this literature and discuss estimation of the SATE under block designs with arbitrarily many treatment categories and an arbitrary blocking of units. Specially, we derive variances for two unbiased estimators of the SATE—the difference-in-means estimator and the Horvitz-Thompson estimator—and give conservative estimators of these variances whenever block sizes are at least twice the number of treatment categories. Unlike most other methods, these are closed-form formulas

of the SATE variance under NRCM applicable to a wide range of settings without making additional parametric assumptions.

The paper is organized as follows: Section 2 introduces the blocking problem we consider and forms it as a graph partitioning problem. Section 3 gives, including proofs, a polynomial time algorithm to obtain an approximately optimal blocking. Section 4 discusses estimation of the SATE under block designs. Section 5 applies our blocking method to the (DATASET). Section 6 concludes.

2 Optimal blocking as a graph theory problem

We extend an observation made by Rosenbaum (1989) to view blocking as a graph partitioning problem. Each unit in the experiment can be viewed as a vertex in a graph. For each pair of units there is an edge connecting their corresponding vertices in the graph. All edges are assigned weights intended to measure similarity in the block covariates of the connected units. Mahalanobis distances, or any other distance metric, could be used to construct these weights. If, for some reason, two units are prohibited to be placed in the same block, the corresponding edge weight is set to infinity. After constructing this graph, we then use graph theory machinery to derive algorithms that solve (or approximately solve) our optimal blocking problem.

We will use the following conventions for our notation. Lowercase letters will either denote constants or indices. Parameters of interest will always be denoted by greek letters. Estimates of these parameters are random variables denoted by careted ($\wedge$) letters. Capital letters will either be used for graph theory notation or to denote random variables that are not parameter estimates. Vectors are denoted by bold lowercase letters. Sets are denoted by bold uppercase letters.

Let $G = (V, E)$ denote an undirected graph, where V is a set of n vertices and E is a

set of edges. Suppose that G is *complete*—between any two distinct vertices $i, j \in V$, there is exactly one edge $ij \in E$ joining these vertices.[5] Hence, E contains $n(n-1)/2$ edges. Suppose each edge ij has a weight w_{ij}, and suppose these weights satisfy the triangle inequality:

$$\forall\ ij, j\ell, i\ell \in E,\ w_{ij} + w_{j\ell} \geq w_{i\ell}. \tag{1}$$

Informally, the triangle inequality dictates that if two vertices are connected through an edge, the weight of that edge can at most be the sum of the weights of edges connecting the two vertices through a third vertex—or simply, the direct route between two vertices cannot be longer than a detour. All distance metrics, including Mahalanobis and Euclidean distances, satisfy the triangle inequality by definition.

A *partition* of V is a separation of V into non-empty *blocks* of vertices such that each vertex $i \in V$ belongs to exactly one block. Formally, a partition is a collection of sets of vertices $\mathbf{p} = \{V_1, V_2, \ldots, V_m\}$ satisfying:

1. $\forall\ V_\ell \in \mathbf{p},\ \emptyset \neq V_\ell \subseteq V$,
2. $\forall\ V_\ell, V_{\ell'} \in \mathbf{p},\ (V_\ell \neq V_{\ell'}) \Rightarrow (V_\ell \cap V_{\ell'} = \emptyset)$,
3. $\bigcup_{V_\ell \in \mathbf{p}} V_\ell = V$.

When vertices denote experimental units, each partition of V can be viewed as a blocking of units: experimental units are in the same block if and only if their corresponding vertices are in the same block of the partition.

The *size* of a block V_ℓ, denoted $|V_\ell|$, is the number of vertices contained in that block. Our original optimal blocking problem can be posed as the following optimal partition problem: Let $\mathbf{P}$ denote the set of all partitions, let t^* denote a prespecified *size threshold*,

[5] We will consistently use ij as a shorthand for an edge $\{i, j\}$.

and let:

$$\mathbf{P}^{t^*} \equiv \{\mathbf{p} \in \mathbf{P} : \forall\, V_\ell \in \mathbf{p},\ |V_\ell| \geq t^*\}\,, \tag{2}$$

denote the set of *valid partitions*—partitions that have at least t^* vertices within every block. An optimal partition $\mathbf{p}^\dagger \in \mathbf{P}^{t^*}$ satisfies:

$$\max_{V_\ell \in \mathbf{p}^\dagger} \max_{i,j \in V_\ell} w_{ij} = \min_{\mathbf{p} \in \mathbf{P}^{t^*}} \left(\max_{V_\ell \in \mathbf{p}} \max_{i,j \in V_\ell} w_{ij} \right) \equiv \lambda. \tag{3}$$

That is, across all partitions that contain blocks with t^* or more vertices, an optimal partition minimizes the *maximum within-block edge weight*—the maximum weight of an edge that joins two vertices within the same block of the partition. Optimal partitions are not necessarily unique.

Finding such a partition, $\mathbf{p}^\dagger$, is NP-hard (Hochbaum and Shmoys, 1986). We derive a polynomial-time algorithm that can find an *approximately optimal* partition $\mathbf{p}^* \in \mathbf{P}^{t^*}$ that has a maximum within-block edge weight of at most a factor of four of the optimal maximum weight:

$$\max_{V_\ell \in \mathbf{p}^*} \max_{i,j \in V_\ell} w_{ij} \leq 4\lambda. \tag{4}$$

3 An approximately optimal algorithm for blocking

3.1 Definitions

We here introduce notation and structure that will be used to show approximate optimality of the partition produced by our algorithm. Our approach follows that of Hochbaum and Shmoys (1986).

Let $G = (V, E)$ denote an arbitrary graph:

- Vertices i and j are *adjacent* in G if the edge $ij \in E$.

- A set of vertices $I \subseteq V$ is *independent in* G if no vertices in the set are adjacent to each other:

$$\nexists\, i, j \in I,\ ij \in E. \tag{5}$$

- An independent set of vertices I in G is *maximal* if, for any additional vertex $i \in V$, the set $i \cup I$ is not independent:

$$\forall\, i \in V \setminus I,\ \exists\, j \in I,\ ij \in E. \tag{6}$$

- The *degree in* G of a vertex $i \in V$ is its number of adjacent vertices in G:

$$\deg(G, i) \equiv |\{j \in V : ij \in E\}|. \tag{7}$$

  Note that, for a complete graph on n vertices, each vertex has degree $n - 1$.

- There exists a *walk from* i_1 *to* i_m *of length* m *in* G when one can construct a m-tuple of vertices, with an edge between each adjacent pair, connecting i_1 and i_m:

$$\exists (i_1, i_2, \cdots, i_{m-1}, i_m),\ \forall\, \ell < m,\ i_\ell i_{\ell+1} \in E. \tag{8}$$

  Note that, if $(i_1, i_2, \cdots, i_{m-1}, i_m)$ is a walk of length m from i_1 to i_m and the edge weights of G satisfy the triangle inequality (1), the weight of $i_1 i_m$ (if it exists) fulfills:

$$w_{i_1 i_m} \leq m \max_{\ell < m}(w_{i_\ell i_{\ell+1}}). \tag{9}$$

- The d^{th} *power* of G is a graph $G^d = (V, E^d)$ where an edge $ij \in E^d$ if and only if there is a walk from i to j in G of d or fewer edges.

- Given a partition $\mathbf{p} \in \mathbf{P}$, let $G(\mathbf{p}) = (V, E(\mathbf{p}))$ denote the *subgraph of G generated by* $\mathbf{p}$; the edge ij is in $E(\mathbf{p})$ if and only if vertices i and j are contained in the same block in the partition:

$$E(\mathbf{p}) \equiv \{ij \in E : \exists V_\ell \in \mathbf{p}, i, j \in V_\ell\}. \tag{10}$$

  Note that, for a complete graph G, if a block in the partition $\mathbf{p}$ contains m vertices, then every vertex i in that block has $\deg(G(\mathbf{p}), i) = m - 1$ in the subgraph $G(\mathbf{p})$.

- The *bottleneck subgraph of G for weight threshold* w is a subgraph $BG_w = (V, BE_w)$ where only edges $ij \in E$ with a weight of at most w is included in BE_w:

$$BE_w \equiv \{ij \in E : w_{ij} \leq w\}. \tag{11}$$

- The *k-nearest-neighbors subgraph of G* is a subgraph $NNG_k = (V, NNE_k)$ where an edge $ij \in NNE_k$ if and only if j is one of the k closest vertices to i or i is one of the k closest vertices to j. Rigorously, for a vertex i, let $i_{(\ell)}$ denote the vertex j that corresponds to the ℓ^{th} smallest value of $\{w_{ij} : ij \in E\}$ (where ties are broken arbitrarily):

$$w_{ii_{(1)}} \leq w_{ii_{(2)}} \leq \cdots \leq w_{ii_{(m)}}, \tag{12}$$

  then:

$$NNE_k \equiv \left\{ij \in E : j \in \{i_{(\ell)}\}_{\ell=1}^{k} \vee i \in \{j_{(\ell)}\}_{\ell=1}^{k}\right\}. \tag{13}$$

  Note that, if all vertices have a degree in G of at least k, each vertex i has $\deg(NNG_k, i) \geq k$.

- A graph $G' = (V, E')$ is a *spanning subgraph* of G if they contain the same set of vertices and $E' \subseteq E$. Note that subgraphs generated by partitions, bottleneck

subgraphs and k-nearest-neighbors subgraphs are spanning subgraphs.

3.2 The algorithm

We here present our algorithm for deriving an approximately optimal partition $\mathbf{p}^* \in \mathbf{P}^{t^*}$ as defined in (4). When $t^* << n$, this algorithm can be performed in $O(n^2)$ time.

An brief outline of the algorithm is as follows: We first construct a $(t^* - 1)$-nearest-neighbor subgraph. An initial set of vertices, or *seeds*, are then selected within this graph from which the blocks of the partition will be grown. The seeds are spaced sufficiently far apart so that they cannot interfere with each another's growth but dense enough so that the growth will span the entire graph quickly. This is done by selecting seeds so that all seeds are at least at a distance of three edges from each other in the $(t^* - 1)$-nearest-neighbor subgraph, while all other vertices are at most at a distance of two edges from a seed. The seeds and their respective adjacent units then form the blocks of the partition.

In more detail:

1. Construct the $(t^* - 1)$-nearest-neighbors subgraph NNG_{t^*-1} from the complete graph describing the experimental units.

2. (*Obtain maximal independent set of seeds*) Find a maximal independent set of vertices $\mathbf{M}$ in the second power of the (t^*-1)-nearest-neighbors subgraph, $(NNG_{t^*-1})^2 = (V, (NNE_{t^*-1})^2)$.

   That is, $\mathbf{M}$ is a set of vertices satisfying the following two conditions:

   (a) for any two vertices $i, j \in \mathbf{M}$, there is no walk in NNG_{t^*-1} from i to j of two or fewer edges,

   (b) for any vertex $i \notin \mathbf{M}$, there is a vertex $j \in \mathbf{M}$ such that there exists a walk from i to j of two or fewer edges in NNG_{t^*-1}.

3. (*Grow from seeds*) For each $i \in \mathbf{M}$, form a block of vertices, V_i^*, comprised of vertex i and all vertices adjacent to i in NNG_{t^*-1}. Since the blocks are formed from the seeds, the number of blocks is equal to $|\mathbf{M}|$.

4. (*Assign remaining vertices*) Some vertices may not be assigned to a block yet. These vertices are a walk of two edges away from at least one vertex $i \in \mathbf{M}$ in the (t^*-1)-nearest-neighbors subgraph—thus adjacent in $(NNE_{t^*-1})^2$. Assign the unassigned vertices to the block which their adjacent seed is in. If there are several adjacent seeds, choose the one that is at the closest distance to the unassigned vertex.

After Step 4, the blocks $\{V_i^*\}_{i\in\mathbf{M}}$, form a partition $\mathbf{p}^*$ which describe an approximately optimal blocking of the experimental units.

3.3 Proof of approximate optimality

Intuitively, the algorithm produce an approximate optimal blocking because no vertex will be at a distance away from any vertex in its block greater than what the distance of a walk of four edges in the (t^*-1)-nearest-neighbors subgraph would be. This is since all vertices are at a distance away from their block's seed of at most a walk of two edges. Due to the triangle inequality (1) this implies that the greatest edge weight in $E(\mathbf{p}^*)$ is at most 4λ.

More in detail, to show approximate optimality we must show that the algorithm produces a partition that is a valid blocking and that the subgraph generated by the partition contain no edge with a weight greater than 4λ. The first theorem will prove validity and the rest of this section will prove the bound.

Theorem 1 *The algorithm described in Section 3.2 produces a valid partition,* $\mathbf{p}^* \in \mathbf{P}^{t^*}$.

Proof: The blocks in $\mathbf{p}^*$ are given by the seeds, thus there exist no empty blocks:

$$\forall\, V_\ell \in \mathbf{p}^*,\ \emptyset \neq V_\ell \subseteq V. \tag{14}$$

Since Step 2 ensures that the shortest walk between any seeds in NNG_{t^*-1} is at least three, the seeds do not share any adjacent vertices in NNG_{t^*-1}. No vertex can, thus, be added to a block twice in Step 3. This, together with that all unassigned vertices are added only to a single block in Step 4, implies that the blocks are disjoint:

$$\forall\, V_\ell, V_{\ell'} \in \mathbf{p}^*,\ (V_\ell \neq V_{\ell'}) \Rightarrow (V_\ell \cap V_{\ell'} = \emptyset). \tag{15}$$

Step 4 ensures that every vertex is assigned to a block:

$$\bigcup_{V_\ell \in p^*} V_\ell = V, \tag{16}$$

which, all together, makes $\mathbf{p}^*$ a partition.

After Step 3, each block will contain its seed and the seed's adjacent vertices in NNG_{t^*-1}, of which there is at least $t^* - 1$. Step 4 can add, but never remove, vertices from the blocks. As a consequence, after Step 4, each block will contain at least t^* vertices. ■

To prove that $\mathbf{p}^*$ is approximately optimal we will first state and prove necessary lemmas. The proofs rely heavily on properties of bottleneck subgraphs of the complete graph describing the experimental units. Specifically, we will bound the largest edge weight in the $(t^* - 1)$-nearest-neighbors subgraph by showing that it can be contained in a specific bottleneck subgraph. Vertices within the same block will be connected through a walk of edges within the $(t^* - 1)$-nearest-neighbor graph and, hence, the bottleneck subgraph. Approximate optimality follows after applying the triangle inequality.

Lemma 2 *If $w \geq \lambda$, where λ is the maximum within-block edge weight of an optimal partition, each vertex i in the bottleneck subgraph BG_w has* $\deg(BG_w, i) \geq t^* - 1$.

Proof: Recall that $\mathbf{p}^\dagger$ denotes an optimal partition. By definition, all edges ij in the set $E(\mathbf{p}^\dagger)$ must have weights $w_{ij} \leq \lambda$. BG_λ contain all edges of weight λ or less, thus it

contains all edges in $E(\mathbf{p}^\dagger)$ (and oftentimes additional edges). It follows that $G(\mathbf{p}^\dagger)$ is a spanning subgraph of BG_λ, $E(\mathbf{p}^\dagger) \subseteq BE_\lambda$.

Moreover, since each block of $\mathbf{p}^\dagger$ has at least t^* vertices, each vertex $i \in V$ must have $\deg(G(\mathbf{p}^\dagger), i) \geq t^* - 1$. As the degree of a vertex is increasing in its number of edges and since $G(\mathbf{p}^\dagger)$ is a spanning subgraph of BG_λ, each vertex must also have $\deg(BG_\lambda, i) \geq t^* - 1$.

If $w \geq \lambda$, then BG_λ is a spanning subgraph of BG_w and, by the previous argument, each vertex i in BG_w must have $\deg(BG_w, i) \geq t^* - 1$. ■

Define λ^- as the smallest weight threshold such that each vertex i in the corresponding bottleneck subgraph BE_{λ^-} has $\deg(BE_{\lambda^-}, i) \geq t^* - 1$:

$$\lambda^- \equiv \min \{w : \forall i \in V, \deg(BE_w, i) \geq t^* - 1\}. \tag{17}$$

We can show that this weight threshold is no larger than the maximum within-block edge weight in the optimal blocking:

Lemma 3 $\lambda^- \leq \lambda$.

Proof: By Lemma 2, for all $w \geq \lambda$ the degree in BG_w for all vertices must be at least $t^* - 1$, thus:

$$\{w : w \geq \lambda\} \subseteq \{w : \forall i \in V, \deg(BG_w, i) \geq t^* - 1\}. \tag{18}$$

Since the minimum in a subset is weakly greater than the minimum in its superset and $\lambda = \min\{w : w \geq \lambda\}$, it follows from the definition of λ^- that $\lambda^- \leq \lambda$. ■

Remark: BG_{λ^-} generally does not describe an optimal blocking as, in most cases, a partition $\mathbf{p} \in \mathbf{P}^{t^*}$ such that $\lambda^- = \max\{w_{ij} : ij \in E(\mathbf{p})\}$ does not exist. This explains why λ^- oftentimes will be strictly smaller than λ.

Moreover, we can connect the weights of edges in a nearest-neighbor subgraph to this

bottleneck subgraph.

Lemma 4 *All edges in the $(t^* - 1)$-nearest-neighbor subgraph will have an edge weight of at most λ^-:*

$$\forall\, ij \in NNE_{t^*-1},\ w_{ij} \leq \lambda^-. \tag{19}$$

Proof: Let $w^+ = \max\{w_{ij} : ij \in NNE_{t^*-1}\}$ and $\mathbf{W} = \{w : \forall i \in V, \deg(BE_w, i) \geq t^* - 1\}$. Recall that each vertex i in NNG_{t^*-1} has $\deg(NNG_{t^*-1}, i) \geq t^* - 1$, therefore $w^+ \in \mathbf{W}$. Notice, further, that there must exist a vertex that have less than $t^* - 1$ edges in E with weights less than w^+:

$$\exists i \in V, |\{ij \in E : w_{ij} < w^+\}| < t^* - 1. \tag{20}$$

If not, one could form a subgraph where all vertices have a degree of at least $t^* - 1$ and all edges' weights are lower than w^+, thus NNG_{t^*-1} could in that case not be a valid $(t^* - 1)$-nearest-neighbors subgraph.

As a consequence of (20) no w' lower than w^+ is a candidate for λ^-:

$$\nexists\, w' < w^+, w' \in \mathbf{W}, \tag{21}$$

and subsequently w^+ must be the lowest element in $\mathbf{W}$. By the definition of λ^-, $w^+ = \lambda^-$, and the lemma follows from that the weights of all edges in NNE_{t^*-1} are at most w^+. ■

Corollary 5 *By Lemma 3 and 4, $\forall\, ij \in NNE_{t^*-1},\ w_{ij} \leq \lambda$.*

Finally the triangle inequality can used to bound the edge weights in powers of nearest-neighbors subgraph:

Lemma 6 *All edges ij in the d^{th} power of the (t^*-1)-nearest-neighbors subgraph $(NNG_{t^*-1})^d = (V, (NNE_{t^*-1})^d)$ have weight $w_{ij} \leq d\lambda$.*

Proof: Recall that edge $ij \in (NNE_{t^*-1})^d$ if and only if there is a walk of d or fewer edges connecting i to j in NNG_{t^*-1}. By Corollary 5, all edges in NNE_{t^*-1} have a weight of at most λ thus no edge's weight in the walk connecting i to j can be greater than λ. It follows immediately from (9) that all edges ij in $(NNE_{t^*-1})^d$ have weights that satisfy $w_{ij} \leq d\lambda$.

■

Theorem 7 *The algorithm described in Section 3.2 produces an approximately optimal partition.*

Proof: From Theorem 1, $\mathbf{p}^* \in \mathbf{P}^{t^*}$, so we must show that $\max_{ij \in E(\mathbf{p}^*)} w_{ij} \leq 4\lambda$.

Note that all vertices assigned to blocks in Step 3 are adjacent to a seed, $i \in \mathbf{M}$, in the $(t^* - 1)$-nearest-neighbor subgraph, and by Corollary 5 the weight of the edge connecting them is at most λ. Further, recall that vertices assigned in Step 4 are adjacent to a seed, $i \in \mathbf{M}$, in $(NNG_{t^*-1})^2$ and by Lemma 6 their edge weight is at most 2λ. Step 4 assigns these vertices to the seed with the lowest edge weight, thus that weight can at most be 2λ since the seed adjacent in $(NNG_{t^*-1})^2$ always is a valid match (but it need not be the closest match). Also note that there are exactly one seed vertex per block in the partition $\mathbf{p}^*$, thus no two seeds are connected and no vertex is connected to two or more seeds. Taken together, by the end of Step 4, for all seed vertices the weights of their edges is at most 2λ:

$$\forall i \in \mathbf{M}, \forall ij \in E(\mathbf{p}^*), w_{ij} \leq 2\lambda. \tag{22}$$

As $\mathbf{p}^*$ is a partition, for any two non-seed vertices in a block, $j, \ell \in V_i^* \setminus i$, there must exist edges so that they are connected with each other and with the block's seed, $j\ell, ji, i\ell \in E(\mathbf{p}^*)$. By (22), $ji, i\ell \leq 2\lambda$, which, together with the triangle inequality (1),

implies:

$$w_{j\ell} \leq w_{ji} + w_{i\ell} \leq 4\lambda. \tag{23}$$

As there cannot be any edges between vertices in different blocks in a partition, all edges in $E(\mathbf{p}^*)$ are bounded either by (22) or (23), and approximate optimality follows. ■

3.4 Proof of time complexity

Lemma 8 *A maximal independent set of an arbitrary graph, $G = (V, E)$, can be found in polynomial time.*

Proof: We will prove this lemma by deriving a polynomial-time procedure that obtains a maximal independent set:

A. (*Initialize*) Initialize the maximal independent set $\mathbf{M} = \emptyset$ and initialize $\mathbf{I} = V$.

B. (*Add vertex to* $\mathbf{M}$) Set i to any vertex in $\mathbf{I}$. Set $\mathbf{M} = \mathbf{M} \cup i$.

C. (*Update* $\mathbf{I}$) Remove i and all vertices adjacent to i in G from $\mathbf{I}$:

$$\mathbf{I} = \mathbf{I} \setminus (i \cup \{j \in V : ij \in E\}) \tag{24}$$

D. (*Terminate*) If $\mathbf{I} = \emptyset$, terminate. In all other cases, go to Step B.

One can, by proof by contradiction, show that $\mathbf{M}$ is a maximal independent set when the procedure described above terminates. Suppose that $\mathbf{M}$ is not a maximal independent set, then either $\mathbf{M}$ is not independent or it is not maximal.

If $\mathbf{M}$ is not independent, there are vertices $i, j \in \mathbf{M}$ such that $ij \in E$. Since vertices are added to $\mathbf{M}$ sequentially there must have existed a state where, at Step D, one of the vertices, i, was in $\mathbf{M}$ while the other, j, was in $\mathbf{I}$. However, in the iteration that i was added

to $\mathbf{M}$, j would have been removed from $\mathbf{I}$ at Step C since they are adjacent. Hence, that state cannot exist and $\mathbf{M}$ must be independent.

If $\mathbf{M}$ is not maximal, there is a vertex $i \in V \setminus \mathbf{M}$ that is not adjacent to a vertex in $\mathbf{M}$. Vertices can only be removed from $\mathbf{I}$ by being added to $\mathbf{M}$ or by being adjacent to a vertex being added to $\mathbf{M}$. If there exist a vertex i in $V \setminus \mathbf{M}$ that is not adjacent to a vertex in $\mathbf{M}$, it must, thus, never been removed from $\mathbf{I}$. However, the procedure only terminates when $\mathbf{I}$ is empty hence $\mathbf{M}$ must be maximal.

As $\mathbf{I}$ is finite, and at least one element of $\mathbf{I}$ is removed at each iteration, the procedure must terminate. Step A takes $O(n)$ time, where n is the number of vertices in G. At each iteration, Steps B and D require $O(1)$ time and Step C takes $O(n)$ time. The number of iterations are $O(n)$, thus the whole procedure requires $O(n^2)$ time. ■

Theorem 9 *The algorithm described in Section 3.2 will terminate in polynomial time.*

Proof: Step 1 of the algorithm can be completed in $O(t^* n \log n)$ time (Vaidya, 1989). From Lemma 8, Step 2 requires $O(n^2)$ time. Steps 3 and 4 can also be completed in $O(n^2)$ time. Thus, when $t^* << n$, the entire algorithm is performed in $O(n^2)$ time. When $t^* \propto n$, the algorithm takes $O(n^2 \log n)$ time. ■

Remark: The improvements discussed in Section 3.5 to Step 2 of the blocking algorithm can be implemented by changing $\mathbf{I}$ to a sequence of vertices in the procedure for finding a maximal independent set, where the specific improvement prescribe the order of the sequence. Vertices are then added to $\mathbf{M}$ in that order. If this sequence of vertices can be constructed in polynomial time, the improved Step 2 and the whole algorithm will, thus, still terminate in polynomial time.

3.5 Refinements

Although the partition produced by our algorithm satisfies (4), additional refinement could in some cases reduce the maximum within-block edge weight further—even if the 4λ-bound cannot be reduced.

The maximal independent set of vertices $\mathbf{M}$ derived in Step 2 will, in general, not be unique. While our proof of approximate optimality is valid for any maximal independent set, a more deliberate choice could in many cases result in a lower maximum distance. For example, by selecting seeds that are endpoints of the edges with the greatest weights one can ensure that the most far-away vertices never will be assigned to the same block. While this procedure would eliminate the most extreme blockings, the algorithm will in general not assign far-away vertices to the same block even if they are not separated as seeds.

By recognizing that in most cases small sized blocks (subject to t^* and natural clusters) are desirable, a more efficient heuristic can be devised. Selecting seeds with few adjacent vertices in $(NNG_{t^*-1})^2$ will require more seeds to make $\mathbf{M}$ a *maximum* independent set. As each seed will form a block, this approach will lead to more blocks and thus smaller blocks.[6] Since there will be many vertices with the same number of adjacent vertices in $(NNG_{t^*-1})^2$, clever tie breaking can further improve this heuristic. One such tiebreaker is selecting seeds for which their adjacent vertices themselves have many adjacent vertices. This will ensure that undesirable vertices, from perspective of the first criteria, cannot become seeds. Another tiebreaker can be constructed by noting that in a final blocking any vertex' edge weights bound all other vertices weights in the block through the triangle inequality (1). That is, for a vertex i and any j, ℓ is the same block:

$$w_{j\ell} \leq w_{ji} + w_{i\ell}. \tag{25}$$

[6] An alternative approach, which in our simulations has proven less efficient, is to select seeds by their corresponding units' distance from the sample covariate mean.

By selecting vertices with low edge weights in NNG_{t^*-1} as seeds one thereby effectively bound most other edge weights in the that block and, in many cases, forces the maximum edge weight below the approximately optimal bound.

When the remaining unassigned vertices are assigned to blocks in Step 4, the algorithm only considers blocks which contain seeds that are exactly a walk of two from the vertex in the $(t^* - 1)$-nearest-neighbor subgraph. In some instances the seed at the closest distance to an unassigned vertex need not be the one at the shortest walk to the vertex. By allowing the unassigned vertices to be match to any seed one could therefore reduce the maximum distance in these cases. Approximate optimality is maintained as the seed at a walk of two always is an option. The seed that is at the closest distance can therefore not be at greater distance than the distance to the seed that is a walk of two away. While not all vertices in a block need to be adjacent in $(NNG_{t^*-1})^4$, as before, all vertices will be at a distance of at most 2λ from its block's seed. By the triangle inequality the distance between any two vertices in a block must, thus, be at most 4λ.

Due to the block size threshold, the algorithm will occasionally produce partitions with block sizes much larger than t^*—for example, when there is tight clusters of experimental units. Whenever a block in a partition formed by the algorithm contain at least $2t^*$ vertices, that block can safely be split into two or more blocks each with at least t^* vertices (by any procedure) and the resulting partition will also be approximately optimal (and generally closer to the true optimum). In a similar vein, swapping vertices between blocks will maintain approximate optimality if the maximum edge weight in the new partition is at most as large as in the old partition (Kernighan and Lin, 1970; Hochbaum and Pathria, 1996).

4 Inference of blocking estimators under Neyman-Rubin model

4.1 Notation and preliminaries

There are n units, numbered 1 through n. There are r treatments, numbered 1 through r. Each unit is assigned to exactly one treatment. Each unit i has a vector of block covariates $\mathbf{x_i}$. A distance between block covariates (such as the Mahalanobis distance) can be computed between each pair of distinct covariates.

Suppose the units are partitioned into b blocks (for example, by our algorithm), numbered 1 through b, with each block containing at least t^* units, with $t^* \geq r$. Let n_c denote the number of units in block c. Assume that the units within each block c are ordered in some way: let (k, c) denote the k^{th} unit in block c. Let z denote the remainder of n/r, and let z_c denote the remainder of n_c/r.

4.1.1 Balanced complete and block randomization

Treatment assignment is *balanced* if z treatments are replicated $\lfloor n/r \rfloor + 1$ times, and $r - z$ of the treatments are replicated $\lfloor n/r \rfloor$ times. A balanced treatment assignment is *completely randomized* if each of the

$$\binom{r}{z} \prod_{i=0}^{z-1} \binom{n - i(\lfloor n/r \rfloor + 1)}{\lfloor n/r \rfloor + 1} \prod_{i=0}^{r-z-1} \binom{n - z(\lfloor n/r \rfloor + 1) - i\lfloor n/r \rfloor}{\lfloor n/r \rfloor} \tag{26}$$

possible treatment assignments are equally likely. Treatment is *balanced block randomized* if treatment is balanced and completely randomized within each block and treatment is assigned independently across blocks.

Let T_{kcs} denote treatment indicators for each unit (k, c):

$$T_{kcs} = \begin{cases} 1, & \text{unit } (k,c) \text{ receives treatment } s, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Let $\#T_{cs} = \sum_{k=1}^{n_c} T_{kcs}$ denote the number of units in block c that receive treatment s, and let $\#T_s = \sum_{c=1}^{b} \#T_{cs}$ denote the number of units in total assigned to s. Let z_c denote the remainder of n_c/r.

Under balanced complete randomization, $\#T_s$ has distribution

$$\#T_s = \begin{cases} \lfloor n/r \rfloor + 1 & \text{with probability } z/r, \\ \lfloor n/r \rfloor & \text{with probability } (r-z)/r. \end{cases} \tag{28}$$

Under balanced block randomization, $\#T_{cs}$ has distribution

$$\#T_{cs} = \begin{cases} \lfloor n_c/r \rfloor + 1 & \text{with probability } z_c/r, \\ \lfloor n_c/r \rfloor & \text{with probability } (r-z_c)/r. \end{cases} \tag{29}$$

Since $t^* \geq r$, it follows that $\#T_s \geq \#T_{cs} \geq 1$.

4.1.2 Model for response: the Neyman-Rubin Causal Model

We assume responses follow the Neyman-Rubin Causal Model (NRCM) (Splawa-Neyman et al., 1990; Rubin, 1974; Holland, 1986). Let y_{kcs} denote the *potential outcome* of unit (k, c) given treatment s—the hypothetical observed value of unit (k, c) had that unit received treatment s. Under the NRCM, the potential outcome y_{kcs} is non-random, and the value of this outcome is observed if and only if (k, c) receives treatment s; exactly one of

$\{y_{kcs}\}_{s=1}^{r}$ is observed. The observed response is:

$$Y_{kc} \equiv y_{kc1}T_{kc1} + y_{kc2}T_{kc2} + \cdots + y_{kcr}T_{kcr}. \tag{30}$$

Inherent in this equation is the *stable-unit treatment value assumption* (SUTVA): the observed Y_{kc} only depends on which treatment is assigned to unit (k, c), and is not affected by the treatment assignment of any other unit (k', c').

4.1.3 Common parameters and estimates under the Neyman-Rubin Causal Model

The domain-level mean and variance of potential outcomes for treatment s are:

$$\mu_s \equiv \frac{1}{n}\sum_{c=1}^{b}\sum_{k=1}^{n_c} y_{kcs}, \tag{31}$$

$$\sigma_s^2 \equiv \sum_{c=1}^{b}\sum_{k=1}^{n_c}\frac{(y_{kcs}-\mu_s)^2}{n} = \sum_{c=1}^{b}\sum_{k=1}^{n_c}\frac{y_{kcs}^2}{n} - \left(\sum_{c=1}^{b}\sum_{k=1}^{n_c}\frac{y_{kcs}}{n}\right)^2, \tag{32}$$

and the domain-level covariance between potential outcomes for treatment s and treatment t is:

$$\begin{aligned}\gamma_{st} &\equiv \sum_{c=1}^{b}\sum_{k=1}^{n_c}\frac{(y_{kcs}-\mu_s)(y_{kct}-\mu_t)}{n} \\ &= \sum_{c=1}^{b}\sum_{k=1}^{n_c}\frac{y_{kcs}y_{kct}}{n} - \left(\sum_{c=1}^{b}\sum_{k=1}^{n_c}\frac{y_{kcs}}{n}\right)\left(\sum_{c=1}^{b}\sum_{k=1}^{n_c}\frac{y_{kct}}{n}\right).\end{aligned} \tag{33}$$

Two estimators for μ_s are the sample mean and the Horvitz-Thompson estimator (Horvitz

and Thompson, 1952):

$$\hat{\mu}_{s,\text{samp}} \equiv \sum_{c=1}^{b} \sum_{k=1}^{n_c} \frac{y_{kcs} T_{kcs}}{\#T_s}, \tag{34}$$

$$\hat{\mu}_{s,\text{HT}} \equiv \sum_{c=1}^{b} \sum_{k=1}^{n_c} \frac{y_{kcs} T_{kcs}}{n/r}. \tag{35}$$

Two estimators for σ_s^2 are the sample variance and the Horvitz-Thompson estimate of the variance:

$$\hat{\sigma}^2_{s,\text{samp}} \equiv \frac{n-1}{n} \sum_{c=1}^{b} \sum_{k=1}^{n_c} \frac{T_{kcs} \left(y_{kcs} - \sum_{c=1}^{b} \sum_{k=1}^{n_c} \frac{y_{kcs} T_{kcs}}{\#T_s} \right)^2}{\#T_s - 1}. \tag{36}$$

$$\begin{aligned} \hat{\sigma}^2_{s,\text{HT}} \equiv & \frac{(n-1)r}{n^2} \sum_{c=1}^{b} \sum_{k=1}^{n_c} y_{kcs}^2 T_{kcs} \\ & - \frac{(n-1)r^2}{n^2(n-r) + nz(r-z)} \sum_{(k,c) \neq (k',c')} y_{kcs} y_{k'c's} T_{kcs} T_{k'c's} \end{aligned} \tag{37}$$

The sample estimators weight observations by the inverse of the number of observations receiving treatment s, and the Horvitz-Thompson estimators weight observations by the inverse of the probability of being assigned treatment s. Block-level parameters μ_{cs}, σ_{cs}, and γ_{cst}, and block-level estimators $\hat{\mu}_{cs,\text{samp}}$, $\hat{\mu}_{cs,\text{HT}}$, $\hat{\sigma}^2_{cs,\text{samp}}$, and $\hat{\sigma}^2_{cs,\text{HT}}$ are defined as above except that sums range over only units in block c.

When treatment is balanced and completely randomized, the domain-level estimators satisfy the following properties:

Lemma 10 *Under balanced and completely randomized treatment assignment, for any*

treatments s and t with $s \neq t$,

$$\begin{aligned}
\mathbb{E}(\hat{\mu}_{s,samp}) &= \mu_s, && (38)\\
Var(\hat{\mu}_{s,samp}) &= \frac{r-1}{n-1}\sigma_s^2 + \frac{rz(r-z)}{(n-1)(n-z)(n+r-z)}\sigma_s^2, && (39)\\
cov(\hat{\mu}_{s,samp}, \hat{\mu}_{t,samp}) &= \frac{-\gamma_{st}}{n-1}. && (40)
\end{aligned}$$

Lemma 11 *Under balanced and completely randomized treatment assignment, for any treatments s and t with $s \neq t$,*

$$\begin{aligned}
\mathbb{E}(\hat{\mu}_{s,HT}) &= \mu_s, && (41)\\
Var(\hat{\mu}_{s,HT}) &= \frac{r-1}{n-1}\sigma_s^2 + \frac{z(r-z)}{n^3(n-1)} \sum_{(k,c)\neq(k',c')} y_{kcs}y_{k'c's}, && (42)\\
cov(\hat{\mu}_{s,HT}, \hat{\mu}_{t,HT}) &= \frac{-\gamma_{st}}{n-1} - \frac{z(r-z)}{(r-1)n^3(n-1)} \sum_{(k,c)\neq(k',c')} y_{kcs}y_{k'ct}. && (43)
\end{aligned}$$

Lemma 12 *Under balanced and completely randomized treatment assignment, for any treatment s,*

$$\mathbb{E}(\hat{\sigma}^2_{s,samp}) = \sigma_s^2, \quad \mathbb{E}(\hat{\sigma}^2_{s,HT}) = \sigma_s^2. \tag{44}$$

Recall that, in balanced block randomized designs, treatment is balanced and completely randomized within each block. Thus, analogous properties for block-level estimators hold under balanced block randomization. Lemmas 10 and 11 are proven in Appendix A, and Lemma 12 is proven in Appendix B.

Under the NRCM, the covariance γ_{st} is not directly estimable; such an would require estimate requires knowledge of potential outcomes under both treatment s and treatment t within a single unit. However, when blocks contain several replications of each treatment, and when potential outcomes satisfy some smoothness conditions with respect to the block covariates, good estimates of the block-level covariances γ_{cst} may be obtained. For details, see Abadie and Imbens (2008); Imbens (2011).

4.2 Estimating the sample average treatment effect

Given any two treatments s and t, we wish to estimate the *sample average treatment effect of s relative to t* (SATE$_{st}$), denoted δ_{st}. The SATE$_{st}$ is a sum of differences of potential outcomes:

$$\delta_{st} \equiv \frac{1}{n}\sum_{c=1}^{b}\sum_{k=1}^{n_c}(y_{kcs} - y_{kct}) = \sum_{c=1}^{b}\frac{n_c}{n}\sum_{k=1}^{n_c}(\mu_{cs} - \mu_{ct})\,. \tag{45}$$

We consider two estimators for δ_{st}: the difference-in-means estimator:

$$\hat{\delta}_{st,\text{diff}} \equiv \sum_{c=1}^{b}\frac{n_c}{n}\sum_{k=1}^{n_c}\left(\frac{y_{kcs}T_{kcs}}{\#T_{cs}} - \frac{y_{kct}T_{kct}}{\#T_{ct}}\right) = \sum_{c=1}^{b}\frac{n_c}{n}\sum_{k=1}^{n_c}(\hat{\mu}_{cs,\text{samp}} - \hat{\mu}_{ct,\text{samp}}) \tag{46}$$

and the Horvitz-Thompson estimator (Horvitz and Thompson, 1952):

$$\hat{\delta}_{st,\text{HT}} \equiv \sum_{c=1}^{b}\frac{n_c}{n}\sum_{k=1}^{n_c}\left(\frac{y_{kcs}T_{kcs}}{n_c/r} - \frac{y_{kct}T_{kct}}{n_c/r}\right) = \sum_{c=1}^{b}\frac{n_c}{n}\sum_{k=1}^{n_c}(\hat{\mu}_{cs,\text{HT}} - \hat{\mu}_{ct,\text{HT}})\,. \tag{47}$$

These estimators are shown to be unbiased under balanced block randomization in Theorems 13 and 14.

Properties of these estimators are most easily seen by analyzing the block-level terms.

Consider first the difference-in-means estimator. By linearity of expectations:

$$\mathbb{E}(\hat{\delta}_{st,\text{diff}}) = \sum_{c=1}^{b} \frac{n_c}{n}(\mathbb{E}(\hat{\mu}_{cs,\text{samp}}) - \mathbb{E}(\hat{\mu}_{ct,\text{samp}})). \tag{48}$$

When treatment is balanced block randomized, by independence of treatment assignment across blocks:

$$\text{Var}(\hat{\delta}_{st,\text{diff}}) = \sum_{c=1}^{b} \frac{n_c^2}{n^2} \left[\text{Var}(\hat{\mu}_{cs,\text{samp}}) + \text{Var}(\hat{\mu}_{ct,\text{samp}}) - 2\text{cov}(\hat{\mu}_{cs,\text{samp}}, \hat{\mu}_{ct,\text{samp}})\right]. \tag{49}$$

Linearity of expectations and independence across blocks can also be exploited to obtain similar expressions hold for the Horvitz-Thompson estimator.

From Lemmas 10 and 11, and using (48) and (49), we can show that both the difference-in-means estimator and the Horvitz-Thompson estimator for the SATE_{st} are unbiased, and we can compute the variance of these estimates.

Theorem 13 *Under balanced block randomization, for any treatments s and t with $s \neq t$:*

$$\begin{aligned}
\mathbb{E}(\hat{\delta}_{st,diff}) &= \delta_{st}, & (50)\\
\textit{Var}(\hat{\delta}_{st,diff}) &= \sum_{c=1}^{b} \frac{n_c^2}{n^2}\left(\frac{r-1}{n_c-1}(\sigma_{cs}^2 + \sigma_{ct}^2) + 2\frac{\gamma_{cst}}{n_c-1}\right) \\
&\quad + \sum_{c=1}^{b} \frac{n_c^2}{n^2}\left(\frac{r z_c(r-z_c)}{(n_c-1)(n_c-z_c)(n_c+r-z_c)}(\sigma_{cs}^2 + \sigma_{ct}^2)\right). & (51)
\end{aligned}$$

Theorem 14 *Under balanced block randomization, for any treatments s and t with $s \neq t$:*

$$\mathbb{E}(\hat{\delta}_{st,HT}) = \delta_{st}, \tag{52}$$

$$\begin{aligned} Var(\hat{\delta}_{st,HT}) &= \sum_{c=1}^{b} \frac{n_c^2}{n^2} \left(\frac{r-1}{n_c - 1}(\sigma_{cs}^2 + \sigma_{ct}^2) + 2\frac{\gamma_{cst}}{n_c - 1} \right) \\ &\quad + \sum_{c=1}^{b} \frac{z_c(r - z_c)}{n_c^3(r-1)} \sum_{k=1}^{n_c} \sum_{\ell \neq k} \left(\frac{r-1}{n_c - 1} (y_{kcs} y_{\ell cs} + y_{kct} y_{\ell ct}) + 2\frac{y_{kcs} y_{\ell ct}}{n_c - 1} \right) \end{aligned} \tag{53}$$

Note that, when r divides each n_c, then $\hat{\delta}_{st,\text{diff}} = \hat{\delta}_{st,\text{HT}}$ and

$$\text{Var}(\hat{\delta}_{st,\text{diff}}) = \text{Var}(\hat{\delta}_{st,\text{HT}}) = \sum_{c=1}^{b} \frac{n_c^2}{n^2} \left(\frac{r-1}{n_c - 1}(\sigma_{cs}^2 + \sigma_{ct}^2) + 2\frac{\gamma_{cst}}{n_c - 1} \right). \tag{54}$$

When r does not divide each n_c, simulation results (not presented) seem to suggest that the difference-in-means estimator has a smaller variance, especially when block sizes are small.

4.3 Estimating the variance

As discussed in Section 4.1.3, estimation of the variance for both the difference-in-means and Horvitz-Thompson estimators is complicated by γ_{cst} terms, which cannot be estimated without making assumptions about the distribution of potential outcomes. We give conservative estimates (in expectation) for these variances by first deriving unbiased estimators for the block-level variances $\text{Var}(\hat{\mu}_{cs,\text{diff}})$ and $\text{Var}(\hat{\mu}_{cs,\text{HT}})$ and bounding the total variance using the Cauchy-Schwarz inequality and the arithmetic mean/geometric mean (AM-GM) inequality (Hardy et al., 1952) on the covariance terms γ. These conservative variances make no distributional assumptions on the potential outcomes.

4.3.1 Block-level variance estimates

The variance for the block-level estimators (as derived in Lemmas 10 and 11) can be estimated unbiasedly several ways. We consider the following variance estimators.

Lemma 15 *Define:*

$$\widehat{Var}(\hat{\mu}_{cs,samp}) \equiv \left(\frac{r-1}{n_c-1} + \frac{rz_c(r-z_c)}{(n_c-1)(n_c-z_c)(n_c+r-z_c)}\right)\hat{\sigma}^2_{cs,samp}, \tag{55}$$

$$\begin{aligned}\widehat{Var}(\hat{\mu}_{cs,HT}) \equiv\ & \frac{r-1}{n_c-1}\hat{\sigma}^2_{cs,HT} \\ & + \frac{r^2 z_c(r-z_c)}{n_c^3(n_c-r)+n_c^2 z_c(r-z_c)} \sum_{k=1}^{n_c}\sum_{\ell\neq k} y_{kcs}y_{\ell cs}T_{kcs}T_{\ell cs},\end{aligned} \tag{56}$$

Under balanced block randomization, for any treatment s:

$$\mathbb{E}\left[\widehat{Var}(\hat{\mu}_{cs,samp})\right] = Var(\hat{\mu}_{cs,samp}), \quad \mathbb{E}\left[\widehat{Var}(\hat{\mu}_{cs,HT}))\right] = Var(\hat{\mu}_{cs,HT}). \tag{57}$$

This Lemma is proven in Appendix B.

4.3.2 Conservative variance estimates for SATE$_{st}$ estimators

Define the following variance estimators:

$$\widehat{\text{Var}}(\hat{\delta}_{st,\text{diff}}) \equiv \sum_{c=1}^{b} \frac{n_c^2}{n^2} \left[\left(\frac{r}{n_c - 1} + \frac{r z_c (r - z_c)}{(n_c - 1)(n_c - z_c)(n_c + r - z_c)} \right) (\hat{\sigma}^2_{cs,\text{samp}} + \hat{\sigma}^2_{ct,\text{samp}}) \right], \tag{58}$$

$$\begin{aligned} \widehat{\text{Var}}(\hat{\delta}_{st,\text{HT}}) \equiv \sum_{c=1}^{b} \frac{n_c^2}{n^2} \Bigg[& \frac{r}{n_c - 1} (\hat{\sigma}^2_{cs,\text{HT}} + \hat{\sigma}^2_{ct,\text{HT}}) \\ & + \frac{r^2 z_c (r - z_c)}{n_c^3 (n_c - r) + n_c^2 z_c (r - z_c)} \sum_{k=1}^{n_c} \sum_{\ell \neq k} (y_{kcs} y_{\ell cs} T_{kcs} T_{\ell cs} + y_{kct} y_{\ell ct} T_{kct} T_{\ell ct}) \\ & + \frac{2 r^2 z_c (r - z_c)}{n_c^4 (r - 1) - n_c^2 z_c (r - z_c)} \sum_{k=1}^{n_c} \sum_{\ell \neq k} y_{kcs} y_{\ell ct} T_{kcs} T_{\ell ct} \Bigg]. \end{aligned} \tag{59}$$

We now show that these estimators are conservative (in expectation). First, we begin with the following lemma:

Lemma 16 *Under balanced block randomization, for any treatments s and t with $s \neq t$:*

$$\begin{aligned} & \mathbb{E} \left(\frac{2 r^2 z_c (r - z_c)}{n_c^4 (r - 1) - n_c^2 z (r - z)} \sum_{k=1}^{n_c} \sum_{\ell \neq k} y_{kcs} y_{\ell ct} T_{kcs} T_{\ell ct} \right) \\ & = \frac{2 z_c (r - z_c)}{n_c^3 (n_c - 1)(r - 1)} \sum_{k=1}^{n_c} \sum_{\ell \neq k} y_{kcs} y_{\ell ct}. \end{aligned} \tag{60}$$

This lemma is proved in Appendix B.

Also note, by the Cauchy-Schwarz and the AM-GM inequalities respectively, that:

$$\gamma_{cst} \leq \sqrt{\sigma^2_{cs} \sigma^2_{ct}} \leq \frac{\sigma^2_{cs} + \sigma^2_{ct}}{2}. \tag{61}$$

The first two terms are equal if and only if there exists constants a and b such that, for all $k \in \{1, \ldots, n_c\}$, $y_{kcs} = a + b y_{kct}$. The last two terms are equal if and only if $\sigma^2_{cs} = \sigma^2_{ct}$.

Hence, (61) is satisfied with equality if and only if there exists a constant a such that, for all $k \in \{1, \ldots, n_c\}$, $y_{kcs} = a + y_{kct}$; that is, if and only if treatment shifts the value of the potential outcomes by a constant for all units within block c.

Theorem 17 *Under balanced block randomization, for any treatments s and t with $s \neq t$:*

$$\mathbb{E}(\widehat{Var}(\hat{\delta}_{st,diff})) \geq Var(\hat{\delta}_{st,diff}), \quad \mathbb{E}(\widehat{Var}(\hat{\delta}_{st,HT})) \geq Var(\hat{\delta}_{st,HT}). \tag{62}$$

with equality if and only if, for each block c, there is a constant a_c such that, for all $k \in \{1, \ldots, n_c\}$, $y_{kcs} = a_c + y_{kct}$. $\sigma^2_{cs} = \sigma^2_{ct}$.

Proof: Define:

$$\begin{aligned}
\text{Var}^*_{st,\text{diff}} &\equiv \sum_{c=1}^{b} \frac{n_c^2}{n^2} \left(\frac{r}{n_c - 1} (\sigma^2_{cs} + \sigma^2_{ct}) \right) \\
&\quad + \sum_{c=1}^{b} \frac{n_c^2}{n^2} \left(\frac{r z_c (r - z_c)}{(n_c - 1)(n_c - z_c)(n_c + r - z_c)} (\sigma^2_{cs} + \sigma^2_{ct}) \right) \qquad (63) \\
\text{Var}^*_{st,\text{HT}} &\equiv \sum_{c=1}^{b} \frac{n_c^2}{n^2} \left(\frac{r}{n_c - 1} (\sigma^2_{cs} + \sigma^2_{ct}) \right) \\
&\quad + \sum_{c=1}^{b} \frac{z_c (r - z_c)}{n_c^3 (r - 1)} \sum_{k=1}^{n_c} \sum_{\ell \neq k} \left(\frac{r - 1}{n_c - 1} (y_{kcs} y_{\ell cs} + y_{kct} y_{\ell ct}) + 2 \frac{y_{kcs} y_{\ell ct}}{n_c - 1} \right) \qquad (64)
\end{aligned}$$

By Theorems 13 and 14, and by equation (61), it follows that:

$$\text{Var}^*_{st,\text{diff}} \geq \text{Var}(\hat{\delta}_{st,\text{diff}}), \quad \text{Var}^*_{st,\text{HT}} \geq \text{Var}(\hat{\delta}_{st,\text{HT}}), \tag{65}$$

with equality if and only if, for each block c, there is a constant a_c such that, for all $k \in \{1, \ldots, n_c\}$, $y_{kcs} = a_c + y_{kct}$. Moreover, by Lemmas 15 and 16, and by linearity

of expectations, we have that:

$$\mathbb{E}(\widehat{\text{Var}}(\hat{\delta}_{st,\text{diff}})) = \text{Var}^*_{st,\text{diff}}, \quad \mathbb{E}(\widehat{\text{Var}}(\hat{\delta}_{st,\text{HT}})) = \text{Var}^*_{st,\text{HT}}. \tag{66}$$

The theorem immediately follows. ■

4.4 Comparing block randomization and complete randomization

We now describe conditions under which the variance of SATE_{st} estimates under balanced block randomization is smaller than those under balanced complete randomization (without blocking). We then show that these conditions are met (in expectation) when the assignment of units into blocks of fixed size is completely randomized. Thus, unless block covariates are worse than random chance at predicting potential outcomes, blocking will only improve precision of SATE_{st} estimates. These results are a generalization of those found in Imai (2008). To make the mathematics more tractable, we only consider the case where r divides each n_c.

When treatment assignment is balanced and completely randomized, the following estimator is always unbiased for the SATE_{st}:

$$\hat{\delta}_{st,\text{cr}} \equiv \frac{1}{n/r} \sum_{c=1}^{b} \sum_{k=1}^{n_c} y_{kcs} T_{kcs} - y_{kct} T_{kct}. \tag{67}$$

When r divides each n_c (and hence, r divides n), this estimator is the same as $\hat{\delta}_{st,\text{diff}}$ and $\hat{\delta}_{st,\text{HT}}$. Hence, under these assumptions, this estimator has variance:

$$\begin{aligned} \text{Var}(\hat{\delta}_{st,\text{cr}}) &= \frac{r-1}{n-1}(\sigma_s^2 + \sigma_t^2) + 2\frac{\gamma_{st}}{n-1} \\ &= \sum_{c=1}^{b} \frac{n_c^2}{\sum_c n_c^2} \left(\frac{r-1}{n-1}(\sigma_s^2 + \sigma_t^2) + 2\frac{\gamma_{st}}{n-1} \right). \end{aligned} \tag{68}$$

Proofs of (67) and (68) follow those in Appendix A.

Suppose an experimenter has already partitioned experimental units into blocks and is deciding between completely randomizing treatment and block randomizing treatment. By (54) and (68), when r divides each n_c, the variance of SATE_{st} estimators under block randomization will be as small or smaller than that under complete randomization precisely when

$$\sum_{c=1}^{b} \frac{n_c^2}{\sum n_c^2} \left(\frac{r-1}{n-1}(\sigma_s^2 + \sigma_t^2) + 2\frac{\gamma_{st}}{n-1} \right) - \frac{n_c^2}{n^2} \left(\frac{r-1}{n_c-1}(\sigma_{cs}^2 + \sigma_{ct}^2) + 2\frac{\gamma_{cst}}{n_c-1} \right) \geq 0. \tag{69}$$

We can write this condition in terms of a comparison between block-level variances and sample-level variances. For all units (k, c), define $y_{kc(s+t)} \equiv y_{kcs} + y_{kct}$. Let σ_{s+t}^2 and $\sigma_{c(s+t)}^2$ denote the domain-level and block-level variance of these $y_{kc(s+t)}$ as defined in (32). It follows that the variance under block randomization will be as small or smaller than that under complete randomization if and only if

$$\delta_{\text{cr,blk}} \equiv \sum_{c=1}^{b} n_c^2 \left(\frac{(r-2)(\sigma_s^2 + \sigma_t^2) + \sigma_{s+t}^2}{(n-1)\sum_c n_c^2} - \frac{(r-2)(\sigma_{cs}^2 + \sigma_{ct}^2) + \sigma_{c(s+t)}^2}{(n_c-1)n^2} \right) \geq 0 \tag{70}$$

This formula gives some insight as to what properties of a blocking are helpful in reducing variance. Terms of $\delta_{\text{cr,bl}}$ will be positive (and thus, will favor estimates under block randomization) if and only if

$$\frac{(r-2)(\sigma_{cs}^2 + \sigma_{ct}^2) + \sigma_{c(s+t)}^2}{(r-2)(\sigma_s^2 + \sigma_t^2) + \sigma_{s+t}^2} \leq \frac{(n-1)\sum_c n_c^2}{(n_c-1)n^2} \tag{71}$$

Since the right-hand-side fraction gets smaller as n_c gets larger, it follows that blocking

helps most when the block-level variances in the largest-sized blocks is small.

We now show that, when units are randomly assigned to blocks, the precision of estimates of the SATE_{st} will be the same in expectation under either complete randomization or block randomization. More formally, we say that an assignment of n units into blocks of sizes $\mathbf{n} = (n_1, \ldots, n_b)$ is a *completely randomized blocking with block sizes* $\mathbf{n}$ if each possible blocking with those block sizes is equally likely. Under completely randomized blocking, the block-level variances σ^2_{cs}, σ^2_{ct}, and $\sigma^2_{c(s+t)}$ are random variables; sample-level variances σ^2_s, σ^2_t, and σ^2_{s+t} and block sizes are constants. In Appendix C, we prove the following theorem.

Theorem 18 *Under completely randomized blocking, when r divides each n_c,*

$$\mathbb{E}(\delta_{cr,blk}) = 0. \tag{72}$$

That is, even when units are assigned to blocks randomly, the variance of an estimate of the SATE_{st} under block randomization be no larger than that under complete randomization. When block covariates predict potential outcomes better than at random, blocking guarantees an increase the precision of SATE_{st} estimates.

5 Results

6 Discussion

Recently, there has been a renewed interest in developing new methods for blocking. Greevy et al. (2004) provided a blocking method for the case of two treatments that is optimal in the sense of minimizing a global measure of distance across matched-pairs. A

number of other blocking methods have been proposed that are not formally optimal, but which may perform well in practice (e.g., Moore, 2012, 2014).

We build on this literature by offering a new method for blocking that is approximately optimal at minimizing the maximum within-block distance. It is the first algorithm to be (approximately) optimal at minimizing this loss function, and it produces good balance within blocks and not just globally across blocks like previous methods. Our method works for an arbitrary number of treatments, and it is an polynomial time algorithm. Our method produces blocks with at least as many observations as there are treatment categories. But the analyst may create blocks with more than the number of treatment categories. Larger block sizes are helpful as they avoid analytical difficulties when one wants to estimate conditional variances (Imbens, 2011), although at the cost of increasing the mean square error of the estimator.

We show that unless the blocking covariates are worse than random chance at predicting the potential outcomes, blocking cannot harm the precision of the estimate of the sample average treatment effect. But note that in a given sample, blocking covariates could be worse. An alternative formalization is that the sample is a random sample from an infinite population, with the expectation in the expected-squared-error taken over this population. In this case, blocking cannot harm the precision of the estimator (Imbens, 2011). Again, there can be a harm in blocking for a given sample realization. Moreover, although blocking cannot decrease the precision of the estimator, it may increase the estimated variance, if the blocking covariates are uninformative. But this cost depends on the variance estimator used and is not a general issue. For example, as Imbens (2011) notes, if one uses the randomization distribution to conduct the hypothesis test, there is no cost to blocking in expectation.

As an alternative to blocking, some advocate re-randomization when a given randomization results in poor balance in observed covariates (Hayes and Moulton, 2009; Morgan

and Rubin, 2012). Re-randomization restricts the randomization scheme, as assignments with poor balance are ruled out. If the rule for which randomizations are acceptable and which are not is precise and set *a priori*, randomization inference is well defined.

Far more common than blocking or re-randomization are *ex-post* methods of adjusting experimental data such as post-stratification or using a model based estimator that incorporates covariate information. Such methods work can well. For example, Miratrix et al. (2013) show that post-stratification is nearly as efficient as blocking: the difference in their variances is on the order of $1/n^2$, with a constant depending on treatment proportion. However, in finite samples, post-stratification can increase variance if the number of strata is large and the strata are poorly chosen. Lin (2012) shows the regression adjustment can provide significant gains in precision, and Rosenblum and van der Laan (2009) show that hypothesis tests can be asymptotically valid even when the adjustment model is incorrectly specified. However, like post-stratification, regression adjustment may increase the finite sample variance, and will do so on average for any sample size, if the covariates are not informative.

An argument in favor of blocking as opposed to *ex-post* adjustment is that by building covariate adjustment into the experimental design one is increasing the transparency of the analysis. The results cited regarding post-stratification assume that the analyst didn't pick the strata as a function of the realized treatment assignment. The regression adjustment results assume that the analyst did not pick the adjustment model based on the realized outcomes. One assumes that the analyst does not run a number of adjustment models, and then only report one. Human nature being what it is, this assumption may be optimistic. A major benefit of randomized experiments aside from the randomization is that the design stage is separated from the analysis stage by construction. The less that there is to do at the analysis stage, the less likely it is that the analyst can or will fish for particular results, unconsciously or not.

A Proof of Lemmas 10 and 11

The following proofs use methods found in Cochran (1977) and Lohr (1999). Additionally, the variance calculations for the sample mean follow Miratrix et al. (2013) closely. To help the reader, we refer each unit by a single index.

For any distinct units i and j, and distinct treatments s and t, the following expectations hold under complete randomization:

$$\begin{aligned}
\mathbb{E}\left(\frac{T_{is}}{\#T_s}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{T_{is}}{\#T_s}\middle|\#T_s\right)\right] \\
&= \mathbb{E}\left(\frac{\frac{\#T_s}{n}}{\#T_s}\right) = \mathbb{E}\left(\frac{1}{n}\right) = \frac{1}{n}, && (73)\\
\mathbb{E}\left(\frac{T_{is}}{(\#T_s)^2}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{T_{is}}{(\#T_s)^2}\middle|\#T_{is}\right)\right] \\
&= \mathbb{E}\left(\frac{\frac{\#T_s}{n}}{(\#T_s)^2}\right) = \mathbb{E}\left(\frac{1}{n}\frac{1}{\#T_s}\right) = \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s}\right), && (74)\\
\mathbb{E}\left(\frac{T_{is}T_{js}}{(\#T_s)^2}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{T_{is}T_{js}}{(\#T_s)^2}\middle|\#T_{is}\right)\right] \\
&= \mathbb{E}\left(\frac{\frac{\#T_s}{n}\frac{\#T_s-1}{n-1}}{(\#T_s)^2}\right) = \mathbb{E}\left(\frac{(\#T_s)^2-\#T_s}{n(n-1)(\#T_s)^2}\right) \\
&= \frac{1}{n(n-1)}\mathbb{E}\left(1-\frac{1}{\#T_s}\right) \\
&= \frac{1}{n(n-1)} - \frac{1}{n(n-1)}\mathbb{E}\left(\frac{1}{\#T_s}\right), && (75)\\
\mathbb{E}\left(\frac{T_{is}T_{jt}}{\#T_s\#T_t}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{T_{is}T_{jt}}{\#T_s\#T_t}\middle|\#T_s,\#T_t\right)\right] \\
&= \mathbb{E}\left(\frac{\frac{\#T_s}{n}\frac{\#T_t}{n-1}}{\#T_s\#T_t}\right) = \mathbb{E}\left(\frac{1}{n(n-1)}\right) = \frac{1}{n(n-1)}. && (76)
\end{aligned}$$

We first compute the expectation of the block-level estimator $\hat{\mu}_{s,\text{samp}}$. By 73,

$$\mathbb{E}(\hat{\mu}_{s,\text{samp}}) = \mathbb{E}\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right) = \sum_{i=1}^{n} y_{is}\mathbb{E}\left(\frac{T_{is}}{\#T_s}\right) = \sum_{i=1}^{n}\frac{y_{is}}{n} = \mu_s. \quad (77)$$

We now derive the variance of this estimator. Observe that, by (74) and (75):

$$\begin{aligned}
&\mathbb{E}\left(\hat{\mu}_{s,\text{diff}}^2\right) = \mathbb{E}\left[\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2\right] \\
= \;& \mathbb{E}\left[\sum_{i=1}^{n}\frac{y_{is}^2T_{is}^2}{(\#T_s)^2}+\sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{js}T_{is}T_{js}}{(\#T_s)^2}\right] = \mathbb{E}\left[\sum_{i=1}^{n}\frac{y_{is}^2T_{is}}{(\#T_s)^2}+\sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{js}T_{is}T_{js}}{(\#T_s)^2}\right] \\
= \;& \sum_{i=1}^{n}y_{is}^2\mathbb{E}\left(\frac{T_{is}}{(\#T_s)^2}\right)+\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{js}\mathbb{E}\left(\frac{T_{is}T_{js}}{(\#T_s)^2}\right) \\
= \;& \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s}\right)\sum_{i=1}^{n}y_{is}^2+\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{js}\left(\frac{1}{n(n-1)}-\frac{1}{n(n-1)}\mathbb{E}\left(\frac{1}{\#T_s}\right)\right) \\
= \;& \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s}\right)\sum_{i=1}^{n}y_{is}^2+\left(\left(\sum_{i=1}^{n}y_{is}\right)^2-\sum_{i=1}^{n}y_{is}^2\right)\left(\frac{1}{n(n-1)}-\frac{1}{n(n-1)}\mathbb{E}\left(\frac{1}{\#T_s}\right)\right) \\
= \;& \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}y_{is}\right)^2-\frac{1}{n(n-1)}\sum_{i=1}^{n}y_{is}^2 \\
& +\mathbb{E}\left(\frac{1}{\#T_s}\right)\left(\left(\frac{1}{n}+\frac{1}{n(n-1)}\right)\sum_{i=1}^{n}y_{is}^2-\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}y_{is}\right)^2\right). \qquad (78)
\end{aligned}$$

We can simplify the last term in parentheses.

$$\begin{aligned}
& \left(\frac{1}{n}+\frac{1}{n(n-1)}\right)\sum_{i=1}^{n}y_{is}^2-\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}y_{is}\right)^2 \\
= \;& \frac{1}{n-1}\sum_{i=1}^{n}y_{is}^2-\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}y_{is}\right)^2 \\
= \;& \frac{1}{n-1}\sum_{i=1}^{n}y_{is}^2-\frac{n}{n-1}\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 \\
= \;& \frac{n}{n-1}\sum_{i=1}^{n}\frac{y_{is}^2}{n}-\frac{n}{n-1}\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 \\
= \;& \frac{n}{n-1}\left(\sum_{i=1}^{n}\frac{y_{is}^2}{n}-\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2\right)=\frac{n}{n-1}\sigma_s^2. \qquad (79)
\end{aligned}$$

The last equality is obtained by applying (32).

Continuing from (78) and applying (79), we find that:

$$\begin{aligned}
\mathbb{E}\left(\hat{\mu}^2_{s,\text{diff}}\right) &= \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} y_{is}\right)^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n} y_{is}^2 \\
&\quad + \mathbb{E}\left(\frac{1}{\#T_s}\right)\left(\left(\frac{1}{n}+\frac{1}{n(n-1)}\right)\sum_{i=1}^{n} y_{is}^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} y_{is}\right)^2\right) \\
&= \frac{n}{n-1}\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 - \frac{1}{n-1}\sum_{i=1}^{n}\frac{y_{is}^2}{n} + \mathbb{E}\left(\frac{1}{\#T_s}\right)\frac{n}{n-1}\sigma_s^2. \qquad (80)
\end{aligned}$$

Since $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$, it follows from (80) and (32) that:

$$\begin{aligned}
\text{Var}(\hat{\mu}_{s,\text{diff}}) &= \mathbb{E}(\hat{\mu}_s^2) - (\mathbb{E}(\hat{\mu}_s))^2 \\
&= \frac{n}{n-1}\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 - \frac{1}{n-1}\sum_{i=1}^{n}\frac{y_{is}^2}{n} + \mathbb{E}\left(\frac{1}{\#T_s}\right)\frac{n}{n-1}\sigma_s^2 - \left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 - \frac{1}{n-1}\sum_{i=1}^{n}\frac{y_{is}^2}{n} + \mathbb{E}\left(\frac{1}{\#T_s}\right)\frac{n}{n-1}\sigma_s^2 \\
&= \frac{-1}{n-1}\left(\frac{1}{n-1}\sum_{i=1}^{n}\frac{y_{is}^2}{n} - \left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2\right) + \mathbb{E}\left(\frac{1}{\#T_s}\right)\frac{n}{n-1}\sigma_s^2 \\
&= \frac{-1}{n-1}\sigma_s^2 + \mathbb{E}\left(\frac{1}{\#T_s}\right)\frac{n}{n-1}\sigma_s^2 = \frac{n}{n-1}\left(\mathbb{E}\left(\frac{1}{\#T_s}\right) - \frac{1}{n}\right)\sigma_s^2. \qquad (81)
\end{aligned}$$

Note that $\lfloor n/r \rfloor = n/r - z/r$. Thus, under complete randomization:

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{\#T_s}\right) &= \frac{z}{r}\left(\frac{1}{\lfloor n/r \rfloor + 1}\right) + \left(1 - \frac{z}{r}\right)\left(\frac{1}{\lfloor n/r \rfloor}\right) \\
&= \frac{z\lfloor n/r \rfloor}{r(\lfloor n/r \rfloor)(\lfloor n/r \rfloor + 1)} + \frac{(r-z)(\lfloor n/r \rfloor + 1)}{r(\lfloor n/r \rfloor)(\lfloor n/r \rfloor + 1)} \\
&= \frac{z(n/r - z/r) + (r-z)(n/r - z/r + 1)}{r(n/r - z/r)(n/r - z/r + 1)} \\
&= \frac{(1/r)z(n-z) + (1/r)(r-z)(n-z+r)}{(1/r)(n-z)(n+r-z)} \\
&= \frac{z(n-z) + (r-z)(n-z+r)}{(n-z)(n+r-z)} \\
&= \frac{r(n-z) + r^2 - zr}{(n-z)(n+r-z)} = \frac{r(nr + r^2 - 2rz)}{(n-z)(n+r-z)}. \qquad (82)
\end{aligned}$$

It follows that:

$$\begin{aligned}
\text{Var}(\hat{\mu}_{s,\text{diff}}) &= \frac{n}{n-1}\sigma_s^2\left(\mathbb{E}\left(\frac{1}{\#T_s}\right) - \frac{1}{n}\right) \\
&= \frac{n}{n-1}\sigma_s^2\left(\frac{nr + r^2 - 2rz}{(n-z)(n+r-z)} - \frac{1}{n}\right) \\
&= \frac{n}{n-1}\frac{n^2r + nr^2 - 2nrz - (n-z)(n+r-z)}{n(n-z)(n+r-z)}\sigma_s^2 \\
&= \frac{nr(n+r-z) - nrz - (n-z)(n+r-z)}{(n-1)(n-z)(n+r-z)}\sigma_s^2 \\
&= \frac{nr(n+r-z) - rz(n+r-z) - (n-z)(n+r-z) + rz(r-z)}{(n-1)(n-z)(n+r-z)}\sigma_s^2 \\
&= \frac{(nr - rz - n + z)(n+r-z) + rz(r-z)}{(n-1)(n-z)(n+r-z)}\sigma_s^2 \\
&= \frac{(r-1)(n-z)(n+r-z) + rz(r-z)}{(n-1)(n-z)(n+r-z)}\sigma_s^2 \\
&= \left(\frac{r-1}{n-1} + \frac{rz(r-z)}{(n-1)(n-z)(n+r-z)}\right)\sigma_s^2. \qquad (83)
\end{aligned}$$

We now derive covariances of this estimator. Note that:

$$\begin{aligned}
& \mathbb{E}\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\sum_{i=1}^{n}\frac{y_{it}T_{it}}{\#T_t}\right) \\
= \; & \mathbb{E}\left(\sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}T_{is}T_{jt}}{\#T_s\#T_t}\right)+\mathbb{E}\left(\sum_{i=1}^{n}\frac{y_{is}y_{it}T_{is}T_{it}}{\#T_s\#T_t}\right) \\
= \; & \sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{jt}\mathbb{E}\left(\frac{T_{is}T_{jt}}{\#T_s\#T_t}\right)+\sum_{i=1}^{n}y_{is}y_{it}\mathbb{E}\left(\frac{T_{is}T_{it}}{\#T_s\#T_t}\right) \\
= \; & \sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{jt}\frac{1}{n(n-1)}+0=\sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}}{n(n-1)}. \qquad (84)
\end{aligned}$$

Recall $\text{cov}(X,Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$. It follows that:

$$\begin{aligned}
\text{cov}(\hat{\mu}_{s,\text{diff}},\hat{\mu}_{t,\text{diff}}) &= \mathbb{E}\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_{is}}\sum_{i=1}^{n}\frac{y_{it}T_{it}}{\#T_{it}}\right)-\sum_{i=1}^{n}\frac{y_{is}}{n}\sum_{i=1}^{n}\frac{y_{it}}{n} \\
&= \sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}}{n(n-1)}-\sum_{i=1}^{n}\frac{y_{is}}{n}\sum_{i=1}^{n}\frac{y_{it}}{n} \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}y_{is}\sum_{i=1}^{n}y_{it}-\frac{1}{n(n-1)}\sum_{i=1}^{n}y_{is}y_{it}-\frac{1}{n^2}\sum_{i=1}^{n}y_{is}\sum_{i=1}^{n}y_{it} \\
&= \frac{1}{n^2(n-1)}\sum_{i=1}^{n}y_{is}\sum_{i=1}^{n}y_{it}-\frac{1}{n(n-1)}\sum_{i=1}^{n}y_{is}y_{it} \\
&= \frac{-1}{n-1}\left(\sum_{i=1}^{n}\frac{y_{is}y_{it}}{n}-\sum_{i=1}^{n}\frac{y_{is}}{n}\sum_{i=1}^{n}\frac{y_{it}}{n}\right)=\frac{-\gamma_{st}}{n-1}. \qquad (85)
\end{aligned}$$

Our derivation of the variance and covariance of the sample mean show that the variance and covariate expressions derived in Miratrix et al. (2013) are incorrect by a factor of $\frac{n}{n-1}$.

We now turn our attention to the Horvitz-Thompson estimator. For any distinct units i and j, and distinct treatments s and t, the following expectations hold under complete

randomization:

$$\begin{aligned}
\mathbb{E}\left(\#T_s\right) &= \mathbb{E}\left(\sum_{i=1}^{n} T_{is}\right) = \sum_{i=1}^{n} \mathbb{E}(T_{is}) = \sum_{i=1}^{n} 1/r = n/r, && (86)\\
\mathbb{E}\left((\#T_s)^2\right) &= \frac{z}{r}\left(\lfloor n/r \rfloor + 1\right)^2 + \left(1 - \frac{z}{r}\right)\left(\lfloor n/r \rfloor\right)^2 \\
&= \frac{z}{r}\left((\lfloor n/r \rfloor)^2 + 2\lfloor n/r \rfloor + 1\right) + \left(1 - \frac{z}{r}\right)\left(\lfloor n/r \rfloor\right)^2 \\
&= (\lfloor n/r \rfloor)^2 + \frac{2z}{r}\lfloor n/r \rfloor + \frac{z}{r} = (\lfloor n/r \rfloor + z/r)^2 + (z/r - (z/r)^2) \\
&= (n/r)^2 + z/r(1 - z/r), && (87)\\
\mathbb{E}\left(\#T_s \#T_t\right) &= \frac{z(z-1)}{r(r-1)}\left(\lfloor n/r \rfloor + 1\right)^2 + \frac{(r-z)(r-z-1)}{r(r-1)}\left(\lfloor n/r \rfloor\right)^2 \\
&\quad + \frac{2z(r-z)}{r(r-1)}\left(\lfloor n/r \rfloor + 1\right)\left(\lfloor n/r \rfloor\right) \\
&= \frac{1}{r(r-1)}\begin{pmatrix} z(z-1)((\lfloor n/r \rfloor)^2 + 2\lfloor n/r \rfloor + 1) \\ +(r-z)(r-z-1)(\lfloor n/r \rfloor)^2 \\ +2z(r-z)((\lfloor n/r \rfloor)^2 + \lfloor n/r \rfloor) \end{pmatrix} \\
&= \frac{1}{r(r-1)}\begin{pmatrix} (z(z-1) + (r-z)(r-z-1) + 2z(r-z))(\lfloor n/r \rfloor)^2 \\ +(2z(z-1) + 2z(r-z))\lfloor n/r \rfloor \\ +z(z-1) \end{pmatrix} \\
&= \frac{1}{r(r-1)}\begin{pmatrix} (z(z-1) + (r-z)(r+z-1))(\lfloor n/r \rfloor)^2 \\ +(2z(r-1))\lfloor n/r \rfloor + z(z-1) \end{pmatrix} \\
&= \frac{1}{r(r-1)}\begin{pmatrix} (z^2 - z + r^2 - z^2 - r + z)(n/r - z/r)^2 \\ +(2z(r-1))(n/r - z/r) + z(z-1) \end{pmatrix} \\
&= \frac{1}{r^3(r-1)}\left((r^2 - r)(n-z)^2 + r(r-1)2z(n-z) + r^2 z(z-1)\right) \\
&= \frac{1}{r^3(r-1)}\left(r(r-1)\left((n-z)^2 + 2z(n-z)\right) + r^2 z(z-1)\right) \\
&= \frac{1}{r^2(r-1)}\left((r-1)\left((n-z)^2 + 2z(n-z) + z^2\right) - z^2(r-1) + rz(z-1)\right) \\
&= \frac{1}{r^2(r-1)}\left((r-1)(n-z+z)^2 - rz + z^2\right) \\
&= \frac{n^2(r-1) - z(r-z)}{r^2(r-1)}. && (88)
\end{aligned}$$

Using these expressions, we can compute the following expectations under complete randomization, assuming distinct treatments s and t and distinct units i and j:

$$\begin{aligned}
\mathbb{E}(T_{is}T_{js}) &= = \mathbb{E}\left(\mathbb{E}(T_{is}T_{js}|\#T_s)\right) = \mathbb{E}\left(\frac{\#T_s(\#T_s-1)}{n(n-1)}\right) \\
&= \frac{\mathbb{E}[(\#T_s)^2] - \mathbb{E}[\#T_s]}{n(n-1)} = \frac{(n/r)^2 + z/r(1-z/r) - n/r}{n(n-1)} \\
&= \frac{(n/r)^2 - (z/r)^2 - (n/r - z/r)}{n(n-1)} \\
&= \frac{n^2 - z^2 - (nr - zr)}{n(n-1)r^2} \\
&= \frac{(n-z)(n+z) - (nr-zr)}{n(n-1)r^2} \\
&= \frac{(n-z)(n-r+z)}{n(n-1r^2)} = \frac{n(n-r)+z(r-z)}{n(n-1)r^2}, \qquad (89) \\
\mathbb{E}(T_{is}T_{jt}) &= \mathbb{E}[\mathbb{E}(T_{is}T_{jt}|\#T_s, \#T_t)] \\
&= \mathbb{E}\left(\frac{\#T_s\#T_t}{n(n-1)}\right) = \frac{\mathbb{E}(\#T_s\#T_t)}{n(n-1)} \\
&= \frac{n^2(r-1) - z(r-z)}{n(n-1)r^2(r-1)}. \qquad (90)
\end{aligned}$$

Under complete randomization, the expectation of the Horvitz-Thompson estimator is:

$$\mathbb{E}(\hat{\mu}_{s,\text{HT}}) = \mathbb{E}\left(\sum_{i=1}^{n} \frac{y_{is}T_{is}}{n/r}\right) = \sum_{i=1}^{n} \frac{y_{is}\mathbb{E}(T_{is})}{n/r} = \sum_{i=1}^{n} \frac{y_{is}(1/r)}{n/r} = \sum_{i=1}^{n} \frac{y_{is}}{n} = \mu_s. \qquad (91)$$

The variance of this estimator is derived as follows. By (89):

$$
\begin{aligned}
\mathbb{E}(\hat{\mu}^2_{s,\mathrm{HT}}) &= \mathbb{E}\left(\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{n/r}\right)^2\right) \\
&= \left(\frac{r^2}{n^2}\right)\left(\mathbb{E}\left(\sum_{i=1}^{n}y_{is}^2T_{is}^2\right)+\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{js}T_{is}T_{js}\right)\right) \\
&= \left(\frac{r^2}{n^2}\right)\left(\mathbb{E}\left(\sum_{i=1}^{n}y_{is}^2T_{is}\right)+\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{js}T_{is}T_{js}\right)\right) \\
&= \left(\frac{r^2}{n^2}\right)\left(\sum_{i=1}^{n}y_{is}^2\mathbb{E}(T_{is})+\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{js}\mathbb{E}(T_{is}T_{js})\right) \\
&= \left(\frac{r^2}{n^2}\right)\left(\sum_{i=1}^{n}\frac{y_{is}^2}{r}+\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{js}\frac{n(n-r)+z(r-z)}{r^2n(n-1)}\right) \\
&= \left(\frac{r^2}{n^2}\right)\left[\sum_{i=1}^{n}\frac{y_{is}^2}{r}+\left(\frac{n(n-r)+z(r-z)}{r^2n(n-1)}\right)\left(\left(\sum_{i=1}^{n}y_{is}\right)^2-\sum_{i=1}^{n}y_{is}^2\right)\right] \\
&= \left(\frac{r}{n^2}-\frac{n(n-r)+z(r-z)}{n^3(n-1)}\right)\sum_{i=1}^{n}y_{is}^2 \\
&\quad +\left(\frac{n(n-r)+z(r-z)}{n^3(n-1)}\right)\left(\sum_{i=1}^{n}y_{is}\right)^2 \\
&= \left(\frac{rn(n-1)-n(n-r)-z(r-z)}{n^2(n-1)}\right)\sum_{i=1}^{n}\frac{y_{is}^2}{n} \\
&\quad +\left(\frac{n(n-r)+z(r-z)}{n(n-1)}\right)\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 \\
&= \left(\frac{n^2r-n^2-z(r-z)}{n^2(n-1)}\right)\sum_{i=1}^{n}\frac{y_{is}^2}{n}+\left(\frac{n(n-r)+z(r-z)}{n(n-1)}\right)\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 \\
&= \left(\frac{n^2(r-1)-z(r-z)}{n^2(n-1)}\right)\sum_{i=1}^{n}\frac{y_{is}^2}{n}+\left(\frac{n(n-r)+z(r-z)}{n(n-1)}\right)\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2.
\end{aligned}
\tag{92}
$$

It follows by (92) and (32) that:

$$
\begin{aligned}
& \mathrm{Var}(\hat{\mu}_{s,\mathrm{HT}}) = \mathbb{E}(\hat{\mu}_{s,\mathrm{HT}}^2) - (\mathbb{E}(\hat{\mu}_{s,\mathrm{HT}}))^2 \\
= \; & \left(\frac{n^2(r-1) - z(r-z)}{n^2(n-1)}\right) \sum_{i=1}^{n} \frac{y_{is}^2}{n} \\
& + \left(\frac{n(n-r) + z(r-z)}{n(n-1)}\right) \left(\sum_{i=1}^{n} \frac{y_{is}}{n}\right)^2 - \left(\sum_{i=1}^{n} \frac{y_{is}}{n}\right)^2 \\
= \; & \left(\frac{n^2(r-1) - z(r-z)}{n^2(n-1)}\right) \sum_{i=1}^{n} \frac{y_{is}^2}{n} \\
& + \left(\frac{n(n-r) + z(r-z) - n(n-1)}{n(n-1)}\right) \left(\sum_{i=1}^{n} \frac{y_{is}}{n}\right)^2 \\
= \; & \left(\frac{n^2(r-1) - z(r-z)}{n^2(n-1)}\right) \sum_{i=1}^{n} \frac{y_{is}^2}{n} + \left(\frac{n(1-r) + z(r-z)}{n(n-1)}\right) \left(\sum_{i=1}^{n} \frac{y_{is}}{n}\right)^2 \\
= \; & \left(\frac{n^2(r-1) - z(r-z)}{n^2(n-1)}\right) \sum_{i=1}^{n} \frac{y_{is}^2}{n} - \left(\frac{n(r-1) - z(r-z)}{n(n-1)}\right) \left(\sum_{i=1}^{n} \frac{y_{is}}{n}\right)^2 \\
= \; & \frac{r-1}{n-1} \left(\sum_{i=1}^{n} \frac{y_{is}^2}{n} - \left(\sum_{i=1}^{n} \frac{y_{is}}{n}\right)^2\right) - \frac{z(r-z)}{n^3(n-1)} \left(\sum_{i=1}^{n} y_{is}^2 - \left(\sum_{i=1}^{n} y_{is}\right)^2\right) \\
= \; & \frac{r-1}{n-1}\sigma_s^2 - \frac{z(r-z)}{n^3(n-1)} \left(-\sum_{i=1}^{n} \sum_{j \neq i} y_{is} y_{js}\right) \\
= \; & \frac{r-1}{n-1}\sigma_s^2 + \frac{z(r-z)}{n^3(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} y_{is} y_{js}. \qquad (93)
\end{aligned}
$$

Now we derive the covariance. Note that:

$$\begin{aligned}
& \mathbb{E}\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{n/r}\sum_{i=1}^{n}\frac{y_{it}T_{it}}{n/r}\right) \\
= & \mathbb{E}\left(\sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}T_{is}T_{jt}}{(n/r)^2}\right)+\mathbb{E}\left(\sum_{i=1}^{n}\frac{y_{is}y_{it}T_{is}T_{it}}{(n/r)^2}\right) \\
= & \sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}}{(n/r)^2}\mathbb{E}\left(T_{is}T_{jt}\right)+\sum_{i=1}^{n}\frac{y_{is}y_{it}}{(n/r)^2}\mathbb{E}\left(T_{is}T_{it}\right) \\
= & \sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}}{(n/r)^2}\frac{n^2(r-1)-z(r-z)}{n(n-1)r^2(r-1)}+0 \\
= & \sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}(n^2(r-1)-z(r-z))}{n^3(n-1)(r-1)} \\
= & \sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}}{n(n-1)}-\sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}z(r-z)}{n^3(n-1)(r-1)}. \qquad (94)
\end{aligned}$$

Thus, using $\text{cov}(X,Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$ and applying (94) and (33), we have:

$$\begin{aligned}
\text{cov}(\hat{\mu}_{s,\text{HT}},\hat{\mu}_{t,\text{HT}}) &= \mathbb{E}(\hat{\mu}_{s,\text{HT}}\hat{\mu}_{t,\text{HT}})-\mathbb{E}(\hat{\mu}_{s,\text{HT}})\mathbb{E}(\hat{\mu}_{t,\text{HT}}) \\
&= \mathbb{E}\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{n/r}\sum_{i=1}^{n}\frac{y_{it}T_{it}}{n/r}\right)-\sum_{i=1}^{n}\frac{y_{is}}{n}\sum_{i=1}^{n}\frac{y_{it}}{n} \\
&= \sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}}{n(n-1)}-\sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}z(r-z)}{n^3(n-1)(r-1)}-\sum_{i=1}^{n}\frac{y_{is}}{n}\sum_{i=1}^{n}\frac{y_{it}}{n} \\
&= \frac{-\gamma_{st}}{n-1}-\sum_{i=1}^{n}\sum_{j\neq i}\frac{y_{is}y_{jt}z(r-z)}{n^3(n-1)(r-1)}. \qquad (95)
\end{aligned}$$

This proves the two lemmas.

B Proof of Lemmas 15 and 16

To help the reader, we suppress the block index in the following derivations, identifying units by a single index.

Note that, under complete randomization and for distinct treatments s and t and distinct units i and j, the following expectations hold:

$$\begin{aligned}
\mathbb{E}\left(\frac{T_{is}}{\#T_s - 1}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{T_{is}}{\#T_s - 1}\middle| \#T_s\right)\right] \\
&= \mathbb{E}\left(\frac{\frac{\#T_s}{n}}{\#T_s - 1}\right) = \frac{1}{n}\mathbb{E}\left(\frac{\#T_s}{\#T_s - 1}\right) \\
&= \frac{1}{n}\mathbb{E}\left(\frac{\#T_s - 1}{\#T_s - 1}\right) + \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s - 1}\right) \\
&= \frac{1}{n} + \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s - 1}\right), \qquad (96) \\
\mathbb{E}\left(\frac{T_{is}}{\#T_s(\#T_s - 1)}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{T_{is}}{\#T_s(\#T_s - 1)}\middle| \#T_s\right)\right] \\
&= \mathbb{E}\left(\frac{\frac{\#T_s}{n}}{\#T_s(\#T_s - 1)}\right) = \frac{1}{n}\mathbb{E}\left(\frac{\#T_s}{\#T_s(\#T_s - 1)}\right) \\
&= \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s - 1}\right), \qquad (97) \\
\mathbb{E}\left(\frac{T_{is}T_{js}}{\#T_s(\#T_s - 1)}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{T_{is}T_{js}}{\#T_s(\#T_s - 1)}\middle| \#T_s\right)\right] \\
&= \mathbb{E}\left(\frac{\frac{\#T_s}{n}\frac{\#T_s - 1}{n-1}}{\#T_s(\#T_s - 1)}\right) \\
&= \frac{1}{n(n-1)}\mathbb{E}\left(\frac{\#T_s(\#T_s - 1)}{\#T_s(\#T_s - 1)}\right) = \frac{1}{n(n-1)}. \qquad (98)
\end{aligned}$$

We show that $\mathbb{E}(\hat{\sigma}^2_{s,\text{samp}}) = \sigma^2_s$. The fact that $\mathbb{E}\left[\widehat{\text{Var}}(\hat{\mu}_{s,\text{samp}})\right] = \text{Var}(\hat{\mu}_{s,\text{samp}})$ follows immediately.

First, note that:

$$\begin{aligned}
&\sum_{i=1}^{n} T_{is}\left(y_{is}T_{is} - \sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2 \\
= &\sum_{i=1}^{n} T_{is}(y_{is}T_{is})^2 - 2\sum_{i=1}^{n} T_{is}\left(y_{is}T_{is}\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right) + \sum_{i=1}^{n} T_{is}\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2 \\
= &\sum_{i=1}^{n} y_{is}^2 T_{is} - 2\sum_{i=1}^{n}\left(y_{is}T_{is}\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right) + \sum_{i=1}^{n} T_{is}\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2 \\
= &\sum_{i=1}^{n} y_{is}^2 T_{is} - 2\#T_s\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2 + \#T_s\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2 \\
= &\sum_{i=1}^{n} y_{is}^2 T_{is} - \#T_s\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2. \qquad (99)
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2_{s,\text{diff}}) &= \mathbb{E}\left(\frac{n-1}{n}\sum_{i=1}^{n}\frac{T_{is}\left(y_{is} - \sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2}{\#T_s - 1}\right) \\
&= \frac{n-1}{n}\mathbb{E}\left(\frac{\sum_{i=1}^{n} y_{is}^2 T_{is} - \#T_s\left(\sum_{i=1}^{n}\frac{y_{is}T_{is}}{\#T_s}\right)^2}{\#T_s - 1}\right) \\
&= \frac{n-1}{n}\left(\sum_{i=1}^{n} y_{is}^2\mathbb{E}\left(\frac{T_{is}}{\#T_s - 1}\right) - \mathbb{E}\left(\frac{\left(\sum_{i=1}^{n} y_{is}T_{is}\right)^2}{\#T_{is}(\#T_{is}-1)}\right)\right). \qquad (100)
\end{aligned}$$

By (96):

$$\sum_{i=1}^{n} y_{is}^2\mathbb{E}\left(\frac{T_{is}}{\#T_s - 1}\right) = \frac{1}{n}\sum_{i=1}^{n} y_{is}^2 + \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s - 1}\right)\sum_{i=1}^{n} y_{is}^2. \qquad (101)$$

By (97) and (98):

$$\begin{aligned}
& \mathbb{E}\left(\frac{\left(\sum_{i=1}^{n} y_{is}T_{is}\right)^2}{\#T_{is}(\#T_{is}-1)}\right) \\
= \; & \mathbb{E}\left(\frac{\sum_{i=1}^{n}(y_{is}T_{is})^2}{\#T_{is}(\#T_{is}-1)}\right) + \mathbb{E}\left(\frac{\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}T_{is}T_{js}}{\#T_{is}(\#T_{is}-1)}\right) \\
= \; & \sum_{i=1}^{n} y_{is}^2 \mathbb{E}\left(\frac{T_{is}}{\#T_{is}(\#T_{is}-1)}\right) + \sum_{j\neq i} y_{is}y_{js}\mathbb{E}\left(\frac{T_{is}T_{js}}{\#T_{is}(\#T_{is}-1)}\right) \\
= \; & \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s-1}\right)\sum_{i=1}^{n} y_{is}^2 + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js} \\
= \; & \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s-1}\right)\sum_{i=1}^{n} y_{is}^2 + \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} y_{is}\right)^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n} y_{is}^2.
\end{aligned} \tag{102}$$

Thus, by (101), (102), and (32), it follows that:

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2_{s,\text{diff}}) \; = \; & \frac{n-1}{n}\left(\sum_{i=1}^{n} y_{is}^2\mathbb{E}\left(\frac{T_{is}}{\#T_s-1}\right) - \mathbb{E}\left(\frac{\left(\sum_{i=1}^{n} y_{is}T_{is}\right)^2}{\#T_{is}(\#T_{is}-1)}\right)\right) \\
= \; & \frac{n-1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} y_{is}^2 + \frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s-1}\right)\sum_{i=1}^{n} y_{is}^2\right) \\
& -\frac{n-1}{n}\left(\frac{1}{n}\mathbb{E}\left(\frac{1}{\#T_s-1}\right)\sum_{i=1}^{n} y_{is}^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} y_{is}\right)^2 + \frac{1}{n(n-1)}\sum_{i=1}^{n} y_{is}^2\right) \\
= \; & \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n} y_{is}^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} y_{is}\right)^2\right) \\
= \; & \frac{n-1}{n}\left(\frac{n}{n-1}\sum_{i=1}^{n}\frac{y_{is}^2}{n} - \frac{n}{n-1}\left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2\right) \\
= \; & \frac{n-1}{n}\left(\frac{n}{n-1}\sigma_s^2\right) = \sigma_s^2.
\end{aligned} \tag{103}$$

We now focus on estimating the variance of Horvitz-Thompson estimators. By (89),

for any treatment s, the following expectation holds under complete randomization.

$$\begin{aligned}
& \mathbb{E}\left(\frac{n(n-1)r^2}{n(n-r)+z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}T_{is}T_{js}\right) \\
= & \frac{n(n-1)r^2}{n(n-r)+z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}\mathbb{E}(T_{is}T_{js}) \\
= & \frac{n(n-1)r^2}{n(n-r)+z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}\frac{n(n-r)+z(r-z)}{n(n-1)r^2} \\
= & \sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}, \qquad (104)
\end{aligned}$$

Applying (104), we show unbiasedness of the Horvitz-Thompson variance estimator under complete randomization:

$$\begin{aligned}
\mathbb{E}\left(\hat{\sigma}^2_{s,\text{HT}}\right) = & \mathbb{E}\left(\frac{(n-1)r}{n^2}\sum_{i=1}^{n} y_{is}^2 T_{is}\right) \\
& -\mathbb{E}\left(\frac{(n-1)r^2}{n^2(n-r)+nz(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}T_{is}T_{js}\right) \\
= & \frac{(n-1)r}{n^2}\sum_{i=1}^{n} y_{is}^2\mathbb{E}(T_{is}) \\
& -\frac{1}{n^2}\mathbb{E}\left(\frac{n(n-1)r^2}{n(n-r)+z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}T_{is}T_{js}\right) \\
= & \frac{(n-1)r}{n^2}\sum_{i=1}^{n} y_{is}^2(1/r) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js} \\
= & \left(\frac{1}{n}-\frac{1}{n^2}\right)\sum_{i=1}^{n} y_{is}^2 - \frac{1}{n^2}\left(\left(\sum_{i=1}^{n} y_{is}\right)^2 - \sum_{i=1}^{n} y_{is}^2\right) \\
= & \frac{1}{n}\sum_{i=1}^{n} y_{is}^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} y_{is}\right)^2 = \sum_{i=1}^{n}\frac{y_{is}^2}{n} - \left(\sum_{i=1}^{n}\frac{y_{is}}{n}\right)^2 = \sigma_s^2. \qquad (105)
\end{aligned}$$

Thus, by (89) and (105), we show unbiasedness of the variance estimator $\widehat{\text{Var}}(\hat{\mu}_{s,\text{HT}})$ under complete randomization:

$$\begin{aligned}
\mathbb{E}(\widehat{\text{Var}}(\hat{\mu}_{s,\text{HT}})) &= \mathbb{E}\left(\frac{r-1}{n-1}\hat{\sigma}^2_{s,\text{HT}} + \frac{r^2 z(r-z)}{n^3(n-r)+n^2 z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}T_{is}T_{js}\right) \\
&= \frac{r-1}{n-1}\mathbb{E}\left(\hat{\sigma}^2_{s,\text{HT}}\right) + \mathbb{E}\left(\frac{r^2 z(r-z)}{n^3(n-r)+n^2 z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}T_{is}T_{js}\right) \\
&= \frac{r-1}{n-1}\sigma^2_s + \mathbb{E}\left(\frac{z(r-z)}{n^3(n-1)}\frac{n(n-1)r^2}{n(n-r)+z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}T_{is}T_{js}\right) \\
&= \frac{r-1}{n-1}\sigma^2_s + \frac{z(r-z)}{n^3(n-1)}\mathbb{E}\left(\frac{n(n-1)r^2}{n(n-r)+z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js}T_{is}T_{js}\right) \\
&= \frac{r-1}{n-1}\sigma^2_s + \frac{z(r-z)}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{js} = \text{Var}(\hat{\mu}_{s,\text{HT}}). \qquad (106)
\end{aligned}$$

From (90), it follows that:

$$\begin{aligned}
&\mathbb{E}\left(\frac{n(n-1)r^2(r-1)}{n^2(r-1)-z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{jt}T_{is}T_{jt}\right) \\
= &\frac{n(n-1)r^2(r-1)}{n^2(r-1)-z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{jt}\mathbb{E}(T_{is}T_{jt}) \\
= &\frac{n(n-1)r^2(r-1)}{n^2(r-1)-z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{jt}\frac{n^2(r-1)-z(r-z)}{n(n-1)r^2(r-1)} \\
= &\sum_{i=1}^{n}\sum_{j\neq i} y_{is}y_{jt} \qquad (107)
\end{aligned}$$

Thus, by (107):

$$\begin{aligned}
& \mathbb{E}\left(\frac{2r^2z(r-z)}{n^4(r-1)-n^2z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{jt}T_{is}T_{jt}\right) \\
= \; & \mathbb{E}\left(\frac{2z(r-z)}{n^2}\frac{r^2}{n^2(r-1)-z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{jt}T_{is}T_{jt}\right) \\
= \; & \mathbb{E}\left(\frac{2z(r-z)}{n^3(n-1)(r-1)}\frac{n(n-1)r^2(r-1)}{n^2(r-1)-z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{jt}T_{is}T_{jt}\right) \\
= \; & \frac{2z(r-z)}{n^3(n-1)(r-1)}\mathbb{E}\left(\frac{n(n-1)r^2(r-1)}{n^2(r-1)-z(r-z)}\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{jt}T_{is}T_{jt}\right) \\
= \; & \frac{2z(r-z)}{n^3(n-1)(r-1)}\sum_{i=1}^{n}\sum_{j\neq i}y_{is}y_{jt}. \qquad (108)
\end{aligned}$$

C Proof of Theorem 18

Consider a block of size n_c. Let $\mathbf{S}_{n_c}$ denote an arbitrary subset of $\{1,\ldots,n\}$ of size $|\mathbf{S}_{n_c}| = n_c$. Let $\mathbf{1}(i \in S_{n_c})$ denote an indicator function: $\mathbf{1}(i \in \mathbf{S}_{n_c}) = 1$ if $i \in \mathbf{S}_{n_c}$; otherwise, $\mathbf{1}(i \in \mathbf{S}_{n_c}) = 0$. Possible values of the within-block variance σ^2_{cs} are

$$\frac{1}{n_c}\sum_{i\in\mathbf{S}_{n_c}}y_{is}^2\mathbf{1}(i\in\mathbf{S}_{n_c}) - \frac{1}{n_c^2}\left(\sum_{i\in\mathbf{S}_{n_c}}y_{is}\mathbf{1}(i\in\mathbf{S}_{n_c})\right)^2 \qquad (109)$$

and possible values of the variance $\sigma^2_{c(s+t)}$ are

$$\begin{aligned}
& \frac{1}{n_c}\sum_{i\in\mathbf{S}_{n_c}}y_{is}^2\mathbf{1}(i\in\mathbf{S}_{n_c}) + y_{it}^2\mathbf{1}(i\in\mathbf{S}_{n_c}) \\
& -\frac{1}{n_c^2}\left(\sum_{i\in\mathbf{S}_{n_c}}y_{is}\mathbf{1}(i\in\mathbf{S}_{n_c}) + y_{it}\mathbf{1}(i\in\mathbf{S}_{n_c})\right)^2. \qquad (110)
\end{aligned}$$

Under completely randomized blocking, the probability that block c is comprised of the units in $\mathbf{S}_{n_c}$ is $\binom{n}{n_c}$. Thus, the expectation of σ_{cs} is

$$\begin{aligned}
\mathbb{E}(\sigma_{cs}^2) &= \binom{n}{n_c}^{-1} \sum_{\mathbf{S}_{n_c}} \left(\frac{1}{n_c} \sum_{i \in \mathbf{S}_{n_c}} y_{is}^2 \mathbf{1}(i \in \mathbf{S}_{n_c}) - \frac{1}{n_c^2} \left(\sum_{i \in \mathbf{S}_{n_c}} y_{is} \mathbf{1}(i \in \mathbf{S}_{n_c}) \right)^2 \right) \\
&= \binom{n}{n_c}^{-1} \sum_{\mathbf{S}_{n_c}} \left(\frac{1}{n_c} \sum_{i \in \mathbf{S}_{n_c}} y_{is}^2 \mathbf{1}(i \in \mathbf{S}_{n_c}) - \frac{1}{n_c^2} \sum_{i \in \mathbf{S}_{n_c}} y_{is}^2 \mathbf{1}(i \in \mathbf{S}_{n_c}) \right) \\
&\quad - \binom{n}{n_c}^{-1} \sum_{\mathbf{S}_{n_c}} \frac{1}{n_c^2} \sum_{i \neq j \in \mathbf{S}_{n_c}} y_{is} y_{js} \mathbf{1}(i \in \mathbf{S}_{n_c}) \mathbf{1}(j \in \mathbf{S}_{n_c}) \\
&= \binom{n}{n_c}^{-1} \frac{n_c - 1}{n_c^2} \sum_{\mathbf{S}_{n_c}} \sum_{i \in \mathbf{S}_{n_c}} y_{is}^2 \mathbf{1}(i \in \mathbf{S}_{n_c}) \\
&\quad - \binom{n}{n_c}^{-1} \frac{1}{n_c^2} \sum_{\mathbf{S}_{n_c}} \sum_{i \neq j \in \mathbf{S}_{n_c}} y_{is} y_{js} \mathbf{1}(i \in \mathbf{S}_{n_c}) \mathbf{1}(j \in \mathbf{S}_{n_c}) \\
&= \binom{n}{n_c}^{-1} \frac{n_c - 1}{n_c^2} \sum_{i=1}^{n} \sum_{\mathbf{S}_{n_c}} y_{is}^2 \mathbf{1}(i \in \mathbf{S}_{n_c}) \\
&\quad - \binom{n}{n_c}^{-1} \frac{1}{n_c^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{\mathbf{S}_{n_c}} y_{is} y_{js} \mathbf{1}(i \in \mathbf{S}_{n_c}) \mathbf{1}(j \in \mathbf{S}_{n_c}) \\
&= \binom{n}{n_c}^{-1} \frac{n_c - 1}{n_c^2} \sum_{i=1}^{n} y_{is}^2 \binom{n-1}{n_c - 1} - \binom{n}{n_c}^{-1} \frac{1}{n_c^2} \sum_{i=1}^{n} \sum_{j \neq i} y_{is} y_{js} \binom{n-2}{n_c - 2}.
\end{aligned} \tag{111}$$

Now since

$$\binom{n-1}{n_c - 1} \binom{n}{n_c}^{-1} = \frac{(n-1)! n_c! (n - n_c)!}{(n_c - 1)! (n - n_c)! n!} = \frac{n_c}{n} \tag{112}$$

$$\binom{n-2}{n_c - 2} \binom{n}{n_c}^{-1} = \frac{(n-2)! n_c! (n - n_c)!}{(n_c - 2)! (n - n_c)! n!} = \frac{n_c (n_c - 1)}{n(n-1)}. \tag{113}$$

It follows that

$$\begin{aligned}
\sigma^2_{cs} &= \binom{n}{n_c}^{-1} \frac{n_c - 1}{n_c^2} \sum_{i=1}^{n} y_{is}^2 \binom{n-1}{n_c - 1} - \binom{n}{n_c}^{-1} \frac{1}{n_c^2} \sum_{i=1}^{n} \sum_{j \neq i} y_{is} y_{js} \binom{n-2}{n_c - 2} \\
&= \frac{n_c}{n} \frac{n_c - 1}{n_c^2} \sum_{i=1}^{n} y_{is}^2 - \frac{n_c(n_c - 1)}{n(n-1)} \frac{1}{n_c^2} \sum_{i=1}^{n} \sum_{j \neq i} y_{is} y_{js} \\
&= \frac{n_c - 1}{n(n_c)} \sum_{i=1}^{n} y_{is}^2 - \frac{n_c - 1}{n(n-1)n_c} \sum_{i=1}^{n} \sum_{j \neq i} y_{is} y_{js} \\
&= \frac{n_c - 1}{n(n_c)} \sum_{i=1}^{n} y_{is}^2 - \frac{n(n_c - 1)}{(n-1)n_c} \sum_{i=1}^{n} \sum_{j \neq i} \frac{y_{is} y_{js}}{n^2} \\
&= \frac{n_c - 1}{n(n_c)} \sum_{i=1}^{n} y_{is}^2 - \frac{n(n_c - 1)}{(n-1)n_c} \left(\sum_{i=1}^{n} \frac{y_{is}}{n} \right)^2 + \frac{n_c - 1}{n(n-1)n_c} \sum_{i=1}^{n} y_{is}^2 \\
&= \frac{n_c - 1}{n(n_c)} \sum_{i=1}^{n} y_{is}^2 - \frac{n(n_c - 1)}{(n-1)n_c} \left(\sum_{i=1}^{n} \frac{y_{is}}{n} \right)^2 + \frac{n_c - 1}{n(n-1)n_c} \sum_{i=1}^{n} y_{is}^2 \\
&= \frac{(n-1)(n_c - 1)}{n(n-1)n_c} \sum_{i=1}^{n} y_{is}^2 - \frac{n(n_c - 1)}{(n-1)n_c} \left(\sum_{i=1}^{n} \frac{y_{is}}{n} \right)^2 + \frac{n_c - 1}{n(n-1)n_c} \sum_{i=1}^{n} y_{is}^2 \\
&= \frac{n(n_c - 1)}{n(n-1)n_c} \sum_{i=1}^{n} y_{is}^2 - \frac{n(n_c - 1)}{(n-1)n_c} \left(\sum_{i=1}^{n} \frac{y_{is}}{n} \right)^2 \\
&= \frac{n(n_c - 1)}{(n-1)n_c} \sum_{i=1}^{n} \frac{y_{is}^2}{n} - \frac{n_c(n_c - 1)}{n(n-1)} \left(\sum_{i=1}^{n} y_{is} \right)^2 = \frac{n(n_c - 1)}{(n-1)n_c} \sigma_s^2. \qquad (114)
\end{aligned}$$

Likewise, substituting $y_{is} + y_{it}$ in for y_{is}, we obtain:

$$\sigma^2_{c(s+t)} = \frac{n(n_c - 1)}{(n-1)n_c} \sigma^2_{s+t}. \qquad (115)$$

Therefore

$$
\begin{aligned}
& \mathbb{E}\left(\frac{n_c^2}{(n_c-1)n^2}[(r-2)(\sigma_{cs}^2+\sigma_{ct}^2)+\sigma_{c(s+t)}^2]\right) \\
= & \frac{n_c^2}{(n_c-1)n^2}[(r-2)(\mathbb{E}(\sigma_{cs}^2)+\mathbb{E}(\sigma_{ct}^2))+\mathbb{E}(\sigma_{c(s+t)}^2)] \\
= & \frac{n_c^2}{(n_c-1)n^2}\left[(r-2)\left(\frac{n(n_c-1)}{(n-1)n_c}\sigma_s^2+\frac{n(n_c-1)}{(n-1)n_c}\sigma_t^2\right)+\frac{n(n_c-1)}{(n-1)n_c}\sigma_{s+t}^2\right] \\
= & \frac{n_c^2}{(n_c-1)n^2}\frac{n(n_c-1)}{(n-1)n_c}\left[(r-2)(\sigma_s^2+\sigma_t^2)+\sigma_{s+t}^2\right] \\
= & \frac{n_c}{n(n-1)}\left[(r-2)(\sigma_s^2+\sigma_t^2)+\sigma_{s+t}^2\right]. \qquad (116)
\end{aligned}
$$

Finally, it follows that the expected difference in variances is:

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{c=1}^{b}\frac{n_c^2}{(n-1)\sum n_c^2}[(r-2)(\sigma_s^2+\sigma_t^2)+\sigma_{s+t}^2]\right] \\
& -\mathbb{E}\left[\sum_{c=1}^{b}\frac{n_c^2}{(n_c-1)n^2}[(r-2)(\sigma_{cs}^2+\sigma_{ct}^2)+\sigma_{c(s+t)}^2]\right] \\
= & \sum_{c=1}^{b}\frac{n_c^2}{(n-1)\sum n_c^2}[(r-2)(\sigma_s^2+\sigma_t^2)+\sigma_{s+t}^2] \\
& -\sum_{c=1}^{b}\frac{n_c}{n(n-1)}\left[(r-2)(\sigma_s^2+\sigma_t^2)+\sigma_{s+t}^2\right] \\
= & \frac{(r-2)(\sigma_s^2+\sigma_t^2)+\sigma_{s+t}^2}{n-1}\left(\frac{\sum n_c^2}{\sum n_c^2}-\frac{\sum n_c}{n}\right) \\
= & \frac{(r-2)(\sigma_s^2+\sigma_t^2)+\sigma_{s+t}^2}{n-1}(0)=0. \qquad (117)
\end{aligned}
$$

That is, in expectation, the variance of estimates of the SATE under block randomization with random blocks be the same as those under complete randomization.

References

Abadie, A. and Imbens, G. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, No. 91-92:175–187.

Cochran, W. (1977). *Sampling techniques*. Wiley, New York, NY.

Duflo, E., Glennerster, R., and Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, 4:3895 – 3962.

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

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

Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, 5(4):263–275.

Hardy, G., Littlewood, J., and Pólya, G. (1952). *Inequalities*. Cambridge University Press, second edition.

Hayes, R. J. and Moulton, L. H. (2009). *Cluster randomised trials*. CRC press London.

Hochbaum, D. and Shmoys, D. (1986). A unified approach to approximation algorithms for bottleneck problems. *Journal of the ACM (JACM)*, 33(3):533–550.

Hochbaum, D. S. and Pathria, A. (1996). The bottleneck graph partition problem. *Networks*, 28(4):221–225.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

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

Imbens, G. W. (2011). Experimental design for unit and cluster randomized trials. Working Paper.

Ji, X. and Mitchell, J. (2005). Finding optimal realignments in sports leagues using a branch-and-cut-and-price approach. *International Journal of Operational Research*, 1(1):101–122.

Keele, L., McConnaughy, C., White, I., List, P., and Bailey, D. (2009). Adjusting experimental data. *Experiments in Political Science*.

Kernighan, B. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307.

Lin, W. (2012). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Annals of Applied Statistics*.

Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society, Series B*, 75(2):369–396.

Mitchell, J. (2001). Branch-and-cut for the k-way equipartition problem. Technical report, Citeseer.

Moore, R. T. (2012). Multivariate continuous blocking to improve political science experiments. *Political Analysis*, 20(4):460–479.

Moore, R. T. (2014). Genetic algorithms for experimental design. Working Paper.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282.

Neyman, J. (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement of Journal of the Royal Statistical Society*, 2:107 180.

Rosenbaum, P. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

Rosenblum, M. and van der Laan, M. J. (2009). Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3):937–945.

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

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484).

Rubin, D. B. and van der Laan, M. J. (2011). Targeted ancova estimator in rcts. In *Targeted Learning*, Springer Series in Statistics, pages 201–215. Springer New York.

Splawa-Neyman, J., Dabrowska, D., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34.

Vaidya, P. M. (1989). An $o(n \log n)$ algorithm for the all-nearest-neighbors problem. *Discrete & Computational Geometry*, 4(1):101–115.

Worrall, J. (2010). Evidence: philosophy of science meets medicine. *Journal of evaluation in clinical practice*, 16(2):356–362.