Article

Journal of Educational and Behavioral Statistics
2022, Vol. 47, No. 1, pp. 69–100
DOI: 10.3102/10769986211027240
Article reuse guidelines: sagepub.com/journals-permissions

Block What You Can, Except When You Shouldn't

Nicole E. Pashley
Rutgers University

Luke W. Miratrix
Harvard University

Several branches of the potential outcome causal inference literature have discussed the merits of blocking versus complete randomization. Some have concluded it can never hurt the precision of estimates, and some have concluded it can hurt. In this article, we reconcile these apparently conflicting views, give a more thorough discussion of what guarantees no harm, and discuss how other aspects of a blocked design can cost, all in terms of estimator precision. We discuss how the different findings are due to different sampling models and assumptions of how the blocks were formed. We also connect these ideas to common misconceptions; for instance, we show that analyzing a blocked experiment as if it were completely randomized, a seemingly conservative method, can actually backfire in some cases. Overall, we find that blocking can have a price but that this price is usually small and the potential for gain can be large. It is hard to go too far wrong with blocking.

Keywords: *causal inference; potential outcomes; precision; finite-sample inference; randomization inference; Neymanian inference*

1. Introduction

Many of us may have embraced George Box's famous quote[1] ever since it was thrown at us during an undergraduate statistics course:

> *Block what you can and randomize what you cannot.*

"Blocking" an experiment—the act of grouping similar units together and then randomizing them into treatment and control within each group—is a tool for increasing the precision of an estimate of a treatment impact. Blocking may also be a consequence of how the units were obtained in the first place. Many of us interpret this fragment of Box's advice[2] to suggest that we should block on whatever characteristics and information we have. But is blocking always "worth

it"? Could it ever be a mistake, causing more harm than good? And if so, when? In this article, we unpack these questions, showing that their answers depend on details often left implicit when thinking about blocking, such as whether the blocks are formed by the researcher or are inherent to the context or to how the experimental sample is obtained. We additionally shed light on some common misconceptions regarding comparisons of blocked designs and completely randomized (i.e., unblocked) designs.

While blocking has been investigated from many perspectives (e.g., Snedecor & Cochran, 1989), we focus on the causal inference potential outcomes framework. In this literature, there is, on the face of it, disagreement regarding the guarantees of blocking. Imai (2008) compares the true variance for the matched-pairs design (i.e., blocked experiments where all blocks have two units) to complete randomization for two scenarios: random sampling of pairs for both designs and random sampling of pairs for the matched design but simple random sampling for complete randomization. That paper concludes that "the relative efficiency of the matched-pair design depends on whether matching induces positive or negative correlations regarding potential outcomes within each pair" (Imai, 2008, p. 4865). A similar comment is made on page 101 of Snedecor and Cochran (1989), with both implying that the matched-pairs design may be helpful or harmful. In contrast, Imbens (2011), assuming a stratified sampling superpopulation model, claims that "In experiments with randomization at the unit-level, stratification is superior to complete randomization, and pairing is superior to stratification in terms of precision of estimation" (p. 1). Imai et al. (2008) similarly conclude that the variance under the blocked design (with at least two units assigned to treatment and control within each block) is never higher than under complete randomization for a superpopulation setup. Researchers are in more agreement regarding finite-sample inference, for which blocking can be helpful or harmful. All of the above conclusions are correct: Their differences stem from differences in the specific sampling frameworks used for the respective analyses. While considerations of the finite versus superpopulation distinction and of block size are important, they are not enough: How the blocks are formed and the specific details of any sampling framework being used also matter. This is the work of this article.

In particular, we consider multiple block types and sampling mechanisms. For each combination of these elements considered, we derive expressions comparing the variance of estimates under blocking to those under complete randomization. This allows us to show which overall frameworks guarantee no harm from blocking and which do not. We focus on comparing true variance rather than delving into the problem of variance estimation (see Pashley & Miratrix, 2021, for discussion of variance estimation). For the superpopulation contexts, we also carefully separate out variance due to the sampling of units from variance due to randomized treatment assignment; this gives more precise statements of the benefits of blocking than have been given in prior literature. We generally follow

the taxonomy of block types and sampling mechanisms discussed in Pashley and Miratrix (2020). In particular, certain sampling mechanisms go hand in hand with certain types of blocks, where the type of block captures how the blocks are formed. These frameworks and associated types of blocks are as follows:

(1) Finite sample: This is the case where the units in the experiment are considered fixed and the only source of randomness is the treatment assignment itself. Here, blocks may be formed in any way before randomization using measured baseline covariates, and we consider them fixed with the sample. Consider a psychology researcher evaluating an intervention on a convenience sample of students grouped by baseline ability.

(2) Simple random sampling: Here, we would sample units from a larger population and then divide the units into blocks based on some covariate(s). These *flexible* blocks are a consequence of the units we have such that before sampling we may not even know how many blocks we will have or what size they will be. If our researcher from above viewed their experimental sample of units as randomly drawn from some target population, we would use this setting.

(3) Stratified random sampling: Here, we sample a specified number of units from each of a prespecified set of strata and then randomize within these groups. Consider a researcher recruiting a sample of children from each of a series of ages and blocking by age. The blocks are inherently due to *fixed* aspects of the units themselves, which also define the population strata. This setting is noteworthy as it is often the implicit assumption made when comparing blocking to complete randomization.

(4) Random sampling of strata: In this framework, the population is made up of an infinite number of blocks, and entire blocks are sampled as units. This could be, for example, a random sample of twins. The blocks themselves are *structural* in that they are naturally grouped as a product of the world, not the researcher. This setting is often assumed when comparing the matched-pairs design to complete randomization.

(5) Two-stage sampling: In this framework, we first select a sample of strata from an infinite population of strata, as just above, but then, within each sampled stratum, draw a sample of individuals. We consider the selected strata to be of infinite size, making the sampled individuals within each block a sample drawn from a stratum-specific superpopulation. These superpopulations are not fixed as they are in the stratified random sampling case due to the sampling of strata. Consider a *multisite trial* where a sample of students is obtained for each of a random sample of schools.

In Section 3, we, for each of these contexts, carefully compare blocking versus complete randomization and provide closed-form formulae for how the variances of an estimator under these approaches will differ. We also use these formulae to provide guidance on using blocking across different contexts. We provide the derivations for these formulae in Supplementary Material A.

Once we finish these comparisons, we broaden our scope to consider how blocking is commonly done in practice. Our primary investigation, following

prior literature on this tension, only pertains to the rare case when all blocks manage to have the exact same proportion of units treated. This is not typically the case. If we do not assume equal proportions across blocks, then units in the same treatment arm can be weighted differently, which can reduce the efficiency of blocking regardless of framework. We discuss this cost further in Section 4.

Next, in Section 5, we address two common misconceptions regarding blocking. We first examine the question of whether analyzing a blocked randomized experiment as if it were completely randomized is in fact a safe approach. It is not: For the same reasons that implementing blocking could potentially have a cost, ignoring it can as well. We then discuss whether the variance estimator for a completely randomized experiment is more stable than that of a blocked experiment on the same data. This is not necessarily the case: The variance of the variance estimator of a completely randomized experiment can either be more, or less, stable than the variance estimator for a blocked experiment.

Finally, in Section 6, we move away from the theoretical discussion and present a few illustrative simulations. We first present a range of scenarios from blocking being mildly costly to blocking being very advantageous to underscore how even slight success in the grouping of units easily offsets blocking's cost. We also show that, while a researcher can deliberately form blocks to reduce precision, it is difficult, but not impossible, for a researcher to unintentionally form blocks from a random sample of units in a way that is disadvantageous.

Throughout our article, we show that while claims of "no harm" are often unfounded, the potential harm is generally going to be minimal. Our primary aim is to unpack the tensions at play in order to help guide further work in the field and help to lay to rest this apparent debate. In the end, we advise researchers to not be afraid to block but to not bother blocking on things that are unlikely to be related to one's outcomes. We believe our findings will generalize to related designs intended to reduce imbalance on covariates between the treatment and control groups such as rerandomization (Branson et al., 2016; Li et al., 2018; Morgan & Rubin, 2012, 2015; Schultzberg & Johansson, 2019), but we do not explore that connection here.

2. Setup

Before delving into comparisons of blocking and complete randomization, we first review these two experimental designs, define our notation, and review some standard, well-known, results.

2.1. Experimental Designs and Estimands

We use the Neyman–Rubin model of potential outcomes (Rubin, 1974; Splawa-Neyman et al., 1923/1990) and assume the Stable Unit Treatment Value Assumption, that is, no interference and no multiple forms of treatment (Rubin, 1980). In this framework, each unit has a potential outcome under treatment and

a potential outcome under control, denoted $Y_i(t)$ and $Y_i(c)$, respectively, for unit i ($i = 1, \ldots, n$). We consider two experimental designs on a sample of n units: complete randomization and blocked randomization. Under a completely randomized design, $n_t = np$ of the units are randomly assigned to treatment, with the rest of the $n_c = n - n_t$ units assigned to control, for fixed $p \in (0, 1)$. Under a blocked design, the units are split into K blocks in some manner (see Pashley & Miratrix, 2021, for longer discussion of block types), with n_k units in the kth block ($k = 1, \ldots, K$). Within each block, a completely randomized experiment is performed independently of other blocks such that in the kth block, $n_{t,k} = p_k n_k$ are randomly assigned to treatment for fixed $p_k \in (0, 1)$. For most of the article, except where noted, we assume $p_k = p$ for $k = 1, \ldots, K$.

In addition to having two experimental designs, we consider finite-sample inference and superpopulation inference. In finite-sample inference, the units in the sample and their potential outcomes are considered fixed, and randomness comes solely from the treatment assignment mechanism. The estimand for the finite sample is then the sample average treatment effect (SATE) defined as (see, e.g., Imbens & Rubin, 2015, p. 86)

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

Under blocking the SATE within block k, for $k = 1, \ldots, K$, is

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

where $b_i \in \{1, \ldots, K\}$ indicates the block that unit i belongs to.

In the superpopulation setting, we wish to make inference for some (infinite) superpopulation rather than just the units in our experiment. We therefore need to consider the randomness induced by sampling units from the population into the sample. We thus have two sources of randomness: the sampling and the treatment assignment mechanism. We write our estimand in the superpopulation setting, the population average treatment effect (PATE), as

$$\tau = \mathbb{E}[Y_i(t) - Y_i(c)].$$

The unconditioned expectation denotes a direct average for all units in the superpopulation. We can similarly define the PATE within block k as

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

2.2. Estimation and Variance

There is a different standard treatment effect estimator for complete randomization and blocked randomization. However, the estimators are the same for each design whether we are interested in the SATE or PATE (at least in the

settings considered here). Let $Z_i = t$ if unit i is assigned treatment and $Z_i = c$ if unit i is assigned to control so that $Y_i^{obs} = Y_i(Z_i)$ is the outcome we observe for unit i given a specific treatment Z_i. For complete randomization, the estimator is just the simple difference in means between treatment and control units,

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

where $\mathbb{I}_{Z_i=t}$ is the indicator that unit i received treatment. For blocked randomization, we can define this estimator within each block as

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

$k = 1, \ldots, K$. Then, the overall blocked randomization estimator is a weighted average of these simple difference estimators for each block,

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

We will often take the expectation over the randomization of units to treatment for a fixed sample. In particular, we write the expectation of estimator $\widehat{\tau}$ for a given finite sample $\mathcal{S}$ and for some assignment mechanism $\mathbf{P}$, which may be complete randomization or blocked randomization, as $\mathbb{E}[\widehat{\tau}|\mathcal{S}, \mathbf{P}]$. To reduce clutter, we drop the $\mathbf{P}$ and simply write $\mathbb{E}[\widehat{\tau}|\mathcal{S}]$ if the estimator makes the assignment mechanism clear. For superpopulation settings, we condition on the sampling mechanism, block type, and assignment mechanism, for example, $\mathbb{E}[\widehat{\tau}|\mathcal{F}]$, where $\mathcal{F}$ is some framework.

Both treatment effect estimators are generally unbiased (or nearly so, in the case of random sampling of strata and two-stage sampling as discussed in the remark below), with respect to their appropriate design, under the finite-sample and superpopulation frameworks considered here. So, the difference in performance between these two experimental designs and associated estimators comes down to the differences in variance.

To discuss the true variances of our estimators $\widehat{\tau}_{(CR)}$ and $\widehat{\tau}_{(BK)}$, we need some additional notation (we will follow conventions found in Imbens & Rubin, 2015). The sample variance of potential outcomes under treatment z is

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

where

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

is the mean of the potential outcomes for the units in the sample under treatment z. The sample variance of the individual level treatment effects is

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

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

For the finite sample, the variance of the completely randomized estimator is known to be (Splawa-Neyman et al., 1923/1990)

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

We can use this expression within each block to get block-level variances,

$$\text{var}(\widehat{\tau}_k|\mathcal{S}) = \frac{S_k^2(t)}{n_{t,k}} + \frac{S_k^2(c)}{n_{c,k}} - \frac{S_k^2(tc)}{n_k}.$$

Summing these across the independent blocks, with the weights for block sizes, gives an overall variance of

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

For the superpopulation, we can use a variance decomposition to obtain, under superpopulation setting $\mathcal{F}$,

$$\text{var}(\widehat{\tau}|\mathcal{F}) = \mathbb{E}[\text{var}(\widehat{\tau}|\mathcal{S})|\mathcal{F}] + \text{var}(\mathbb{E}[\widehat{\tau}|\mathcal{S}]|\mathcal{F}).$$

We provide simplified formulae in the following sections when possible to do so.

***Remark*:** Under the frameworks of random sampling of strata and two-stage sampling, our treatment effect estimators, for both blocking and complete randomization, are not actually unbiased for τ. This bias is induced by the random number of units in the sample creating a random denominator in our estimator. Under the conditions given in Pashley and Miratrix (2021) to obtain variance estimators in this setting, in particular either conditioning on the block sizes or assuming that block sizes are independent of block treatment effects, the estimators are unbiased for the PATE. With respect to comparing variances, the bias is the same under blocking or complete randomization (which can be shown by a simple application of the law of total expectation). Thus, these comparisons are still relevant.

3. Blocking Versus Complete Randomization

With the tools from the previous section, we now turn to comparing complete randomization to blocking. In the following sections, we systematically explore the difference in variance of our two designs in the finite sample and under a

TABLE 1.
When Blocking Is Guaranteed Not to Hurt in Terms of Precision

	Equal Proportions Treated		Unequal Proportions Treated	
Framework	Blocking can help?	Blocking can hurt?	Blocking can help?	Blocking can hurt?
Finite sample	Yes	Yes	Yes	Yes
Simple random sampling, flexible blocks	Yes	Yes[a]	Yes	Yes
Stratified sampling, fixed blocks	Yes	No	Yes	Yes
Random sampling of strata, structural blocks	Yes	Yes	Yes	Yes
Two-stage sampling, structural blocks	Yes	No	Yes	Yes

Note. Results for equal proportions treated appear in Section 3, and results for unequal proportions treated appear in Section 4.

[a]No harm for making blocks out of irrelevant covariates. Harm is possible if deliberately implementing a self-destructive blocking choice as discussed in the text.

number of superpopulation frameworks. Table 1 gives an overview of which frameworks provide a guarantee that blocking will be as good or better in terms of precision.

3.1. In the Finite-Sample World

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

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

where

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

Whether this quantity is positive or negative depends on whether a particular form of between-block variation is larger than a form of within-block variation. Finite-sample numerical studies in Section 6 show an example where even in the worst case for blocking, when all blocks have the same distribution of potential outcomes, the increase in variance is not too great. A further numerical illustration in Section 7 further suggests that costs are minimal in practice.

The first term is simply how much variation in means we have across groups. The second is driven by the average within-group variation. If the groups are quite variable internally, but are similar in terms of averages, blocking can hurt.

Most prior work state that although the difference in the brackets can be negative, as the sample size grows this difference will go to a nonnegative quantity. However, this statement depends on the type of blocks we have. In particular, if we have structural blocks such that as n grows, the number of blocks K also grows, the difference in the brackets of Equation 1 will not necessarily go to zero or become positive as $n \to \infty$ (Miratrix et al., 2013). Consider, for example, if n_k is fixed, the above formula's second term is simply a constant times the mean variance. Therefore, a "bad" choice of structural blocks, in terms of between-block variance being lower than within-block variance, could have asymptotic consequences. This has ties to the random sampling of strata framework as discussed in Subsection 3.4.

We can see the structure of Equation 1 more starkly if we consider the case of K equal-sized blocks with $n_k = n/K$ units in each, $p = 1/2$, and no treatment effect, with $Y_i(c) = Y_i(t)$ for all units. In this case, we obtain

$$\text{var}\big(\widehat{\tau}_{(CR)}|\mathcal{S}\big) - \text{var}\big(\widehat{\tau}_{(BK)}|\mathcal{S}\big) = \frac{1}{n-1}\left[4\text{Var}_k(\bar{Y}_k(c)) - \frac{4(K-1)}{n}\frac{1}{K}\sum_{k=1}^{K} S_k^2(c)\right].$$

This simplification makes the comparison of within- versus between-block variability more clear.

3.2. With Simple Random Sampling and Flexible Blocks

Given a simple random sample of units, the variance for the completely randomized design yields the following clean and well-known result (see Imbens & Rubin, 2015):

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

where $\sigma^2(z)$ is the variance of potential outcomes in the (infinite) superpopulation under treatment $z \in \{t, c\}$.

Flexible blocking in this context would be to divide the units into groups using any baseline covariate information we might like after the sample of units is obtained. This common setting is what might happen in, for example, experiments that have an initial recruitment drive, with the researchers then tailoring

the design of their experiment to the obtained sample. In these cases, we might group by some categorical covariate, aggregating those types of units too few in number to make their own blocks into a single "leftover" block. Or we might form blocks out of a continuous covariate by clustering units together as best we can based on the observed sample covariate distribution. We denote this population framework by conditioning on $\mathcal{F}_{\text{SRS}}$.

The difference of variances between blocking and complete randomization is the expectation over Equation 1 with respect to the simple random sampling. In this context, to understand how blocking compares to complete randomization, we need to specify the mechanism of how the blocks are formed. Assume our blocking procedure uses observed baseline covariate information X to group units into blocks and is such that, given a fixed constellation of X values, we will always end up with the same number of blocks with the same values of X within each block. Further assume that any units with the same values of X, indistinguishable to the blocking algorithm, are interchangeable.[3] Then, the expectation of performance over the simple random sampling would also capture the process of making the blocks based on X as a function of the random set of X. The overall difference between blocking and complete randomization will then, in essence, be the difference between how much variation we manage to keep across blocks (as represented by block means) and how much variation is left over within blocks, all averaged across the possible samples. In other words, Equation 1 shows that a good blocking algorithm should reduce heterogeneity in the $Y_i(z)$ within blocks and maximize variation of $\bar{Y}_k(z)$ across blocks.

To explore the possible harm of flexible blocking, we first consider a scenario where blocking would usually be considered a "bad idea": building blocks out of a covariate that is independent from the outcomes.

> **Theorem 3.1** (Simple random sampling and flexible blocking with independent covariates): If we have a fixed blocking algorithm and the covariates used to form blocks are independent of potential outcomes in the superpopulation, then $\text{var}(\widehat{\tau}_{(CR)}|\mathcal{F}_{\text{SRS}}) = \text{var}(\widehat{\tau}_{(BK)}|\mathcal{F}_{\text{SRS}})$.

See A.2 of the Supplementary Materials for proof. The core idea is that blocking on an X independent of outcome is the same as forming random blocks, which is equivalent to no blocking, that is, complete randomization. In these cases, the small amount of random separation of the block means perfectly offsets the blocking cost (these are the two terms in Equation 1).

Interestingly, however, it is possible to do worse when the independence of Theorem 3.1 does not hold: One way to do this is to group units so their group means of X tend to be similar while leaving the within-group variances of X high. If X is highly predictive of Y, Equation 1 suggests that doing this grouping based on X will result in the same grouping structure on Y, producing high within-group variability and low between-group variability in outcomes. If we can do this type

of grouping well enough for most samples we draw, we would get overall inferior performance under this blocked design. This is in fact possible, as we illustrate in Subsection 6.2. Blocking in this manner would happen due to, for example, the occasional misconception that blocks, rather than the units within the blocks, should be made as similar to each other as possible.

In principle, systematic blocking worsening precision could also happen inadvertently: even if we are systematically reducing the variation of X within blocks, if the systematic relationship of X and Y happens to be exactly, perversely, wrong, we could still fail. We illustrate this as well in the latter part of our simulation Subsection 6.2, where we designed a relationship where the outcome depended on whether the covariate X was even or odd and blocked in a manner to ensure balance on this across blocks. This blocking mechanism substantially reduces within-block variance in our covariate X, suggesting gains, but still results in poor performance. Such inadvertent causing of harm seems unlikely to happen in practice.

These two simulations show that, without further assumptions on the structural relationship between X and the outcomes, we cannot guarantee that blocking is not harmful. That being said, as our examples suggest, if a blocking procedure tends to group like with like, in terms of X, then it is hard to imagine a case that could be worse than blocking on something irrelevant. That is, it is hard to imagine a case when blocking would cause any harm in this setting. In our view, the "Independence" case used in Theorem 3.1 is a reasonable worst case for blocking approaches that are not explicitly blocking to minimize cross-block heterogeneity.[4]

We next turn to simple random sampling for the other types of blocks. We first note that structural blocks cannot, by design, be used in a simple random sampling setting; for example, we cannot randomly sample n individuals who are twins from an infinite population of twins and expect to find any complete pairs of twins in our final sample. Block type is essential in determining which superpopulation framework makes sense for a given experiment and therefore what guarantees the researcher can realistically expect.

Simple random sampling with fixed blocks is similarly ill-defined: What happens if a singleton, or no, units from a given block are sampled due to random chance? We could extend this case to flexible blocks if we have simple rules for how these partially sampled blocks are combined with the others; if the fixed blocks are few in number, relative to the overall sample size, this adjustment is likely to be only a small number of units, implying that we would obtain substantively similar results. That being said, we consider the fixed blocks case more carefully in the next section.

3.3. With Stratified Sampling and Fixed Blocks

In stratified sampling, there is a superpopulation that contains fixed strata, and we sample a fixed number of units from each stratum and then randomize the

units within each stratum independently. This is the typical type of framework associated with a predefined categorical covariate used for blocking. For example, a medical trial may recruit a set number of individuals in groups defined by preset age ranges and gender categories. We denote this population framework by conditioning on $\mathcal{F}_{\text{strat}}$.

Similar to complete randomization under simple random sampling, the variance of the blocking design simplifies to the following result in this setting (see Imbens, 2011):

$$\text{var}\left(\hat{\tau}_{(BK)}|\mathcal{F}_{\text{strat}}\right) = \sum_{k=1}^{K} \frac{n_k^2}{n^2}\left(\frac{\sigma_k^2(t)}{n_{t,k}} + \frac{\sigma_k^2(c)}{n_{c,k}}\right),$$

where $\sigma_k^2(z)$ is the population variance of potential outcomes under treatment $z \in \{t, c\}$ for units within stratum k.

In this context, it is not possible for blocking to be harmful (assuming equal proportion treated in all blocks):

Theorem 3.2 (Variance comparison under stratified sampling): The difference in variance between complete randomization and blocked randomization under the stratified sampling framework is

$$\text{var}\left(\hat{\tau}_{(CR)}|\mathcal{F}_{\text{strat}}\right) - \text{var}\left(\hat{\tau}_{(BK)}|\mathcal{F}_{\text{strat}}\right) = \frac{1}{n-1}\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right) \geq 0,$$

where $\mu_k(z)$ is the population mean of potential outcomes under treatment z in stratum k and $\text{Var}_k(\cdot)$ is defined as in Equation 2.

The above expression is very similar to the positive term in Equation 1 for the finite-sample framework. Now, however, we no longer have the negative term.

Remark: Interestingly, when comparing blocking to complete randomization in an infinite population setting, researchers have typically evaluated the completely randomized design under the simpler sampling mechanism of simple random sampling and analyzed the blocked design under the stratified sampling framework (e.g., see Imai et al., 2008; Imbens, 2011). The found difference between the two estimators is therefore an agglomeration of differences in the characteristics of the samples under the two different sampling regimes as well as the difference in doing a blocked experiment versus a completely randomized experiment. Specifically, the difference is

$$\text{var}(\hat{\tau}_{(CR)}|\mathcal{F}_{\text{SRS}}) - \text{var}(\hat{\tau}_{(BK)}|\mathcal{F}_{\text{strat}}) = \frac{1}{n_c}\sum_{k=1}^{K}\frac{n_k}{n}(\mu_k(c) - \mu(c))^2 + \frac{1}{n_t}\sum_{k=1}^{K}\frac{n_k}{n}(\mu_k(t) - \mu(t))^2.$$

We see that this difference is also positive, which is where the results claiming that blocking does not cause harm under a superpopulation setting typically comes from. However, in general, the difference $\text{var}(\hat{\tau}_{(CR)}|\mathcal{F}_{\text{SRS}}) - \text{var}\left(\hat{\tau}_{(CR)}|\mathcal{F}_{\text{strat}}\right)$ may be positive or negative, showing that the traditional estimates

of the benefits of blocking can either be under- or overstated in this context. The result in Theorem 3.2 compares blocking to stratified sampling followed by complete randomization, *not* simple random sampling. For further discussion, see Supplementary Material A.4.

3.4. With Random Sampling of Structural Blocks

In this context, we sample complete blocks and then randomize the individuals within the blocks into treatment and control. Here, the blocks are *structural*: They are a consequence of the world, not the researcher's choices. Examples include twin studies or a multisite trial with random sampling of sites where each site is then a small randomized trial. These multisite trials are common in education and medical research where sites may be schools or hospitals. For instance, with schools, we may select schools from an "infinite" number of schools and then randomize treatment to classrooms within each school. We denote this population framework by conditioning on $\mathcal{F}_{\text{site}}$.

Here, the difference between the variances is again the expectation over Equation 1 with respect to sampling of blocks:

$$\begin{aligned}&\text{var}\big(\widehat{\tau}_{(CR)}|\mathcal{F}_{\text{site}}\big) - \text{var}\big(\widehat{\tau}_{(BK)}|\mathcal{F}_{\text{site}}\big)\\ &= E\left[\frac{1}{n-1}\left[\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\bar{Y}_k(c) + \sqrt{\frac{1-p}{p}}\bar{Y}_k(t)\right) - \sum_{k=1}^{K}\frac{n_k}{n}\frac{n-n_k}{n}\text{var}(\widehat{\tau}_k|\mathcal{S})\right]\middle|\mathcal{F}_{\text{site}}\right]\end{aligned} \tag{3}$$

As the blocks themselves are sampled with block membership fixed, the expectation can be thought of as over all blocks in the population. As in the finite framework, because the strata themselves are finite, it is possible that blocking could result in higher variance. In particular, it is possible to have systematically poor blocks if the block means do not vary. For example, if we use elementary school classrooms as blocks, we may find that schools break up students into classes such that the classrooms all look similar to each other, in that they have similar proportions of high- and low-achieving students but by the same token have higher within-classroom variability.

This framework is typically adopted under comparisons of matched pairs (a blocked experiment with blocks of size two) and complete randomization. Therefore, researchers studying matched-pairs designs often note that the design can hurt in the superpopulation setting (e.g., Imai, 2008). However, we make clear here that this harm is not due to the size of the blocks but rather the block types and sampling scheme.

3.5. With Two-Stage Sampling

We next extend our prior setting by letting the sampled strata be themselves infinite in size. The two-stage sampling scheme then works as follows: first randomly select K blocks, then randomly select n_k units within the kth selected block. We generally allow the n_k to depend on the block selected such that we may not know the total sample size beforehand. For instance, we might imagine first selecting schools from an "infinite" population of schools and then selecting students from an "infinite" population of students within each school. This sampling scheme is popular in education research and has close ties to multilevel modeling and how multisite experiments are evaluated. We denote this model with $\mathcal{F}_{2-\text{stage}}$.

If we condition on which blocks were chosen in the first stage, we can draw on results from stratified sampling (Subsection 3.3), giving our variance difference of

$$\begin{aligned}
&\text{var}\big(\widehat{\tau}_{(CR)}|\mathcal{F}_{2-\text{stage}}\big) - \text{var}\big(\widehat{\tau}_{(BK)}|\mathcal{F}_{2-\text{stage}}\big) \\
&= \mathbb{E}\left[\frac{1}{n-1}\text{Var}_k\left(\sqrt{\frac{p}{1-p}}\mu_k(c) + \sqrt{\frac{1-p}{p}}\mu_k(t)\right)\middle|\mathcal{F}_{2-\text{stage}}\right] \\
&\geq 0.
\end{aligned}$$

The sampling of units within strata, as compared to taking the entire structural block, guarantees that blocking will not be harmful with respect to variance as compared to complete randomization. Of course, in order to obtain this guarantee, the sampling mechanism and population model need to be valid for the study at hand. That is, we must have large strata with two-stage sampling to rely on this result. For studies where entire strata are included in the experiment, the setting of Subsection 3.4 is the one that should be assumed.

4. The Consequences of Variable Proportion of Blocks Treated

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

With varying p_k, the units within each block are weighted differently than they would be in a complete randomization when calculating a treatment effect estimate. That is, in a complete randomization, the treated units are all weighted proportional to $1/p$, but here, the treated units in each block get weighted instead by $1/p_k$, meaning units with low probability of treatment will "count more" toward the overall treatment mean and their variability will have greater

relevance for the overall variance of the estimator. This can be seen from the estimator formulation:

$$\begin{aligned}\widehat{\tau}_{(BK)} &= \sum_{k=1}^{K} \frac{n_k}{n} \widehat{\tau}_k \\ &= \sum_{k=1}^{K} \frac{n_k}{n} \left(\frac{1}{p_k n_k} \sum_{i:b_i=k} \mathbb{I}_{Z_i=t} Y_i(t) - \frac{1}{(1-p_k)n_k} \sum_{i:b_i=k} (1 - \mathbb{I}_{Z_i=t}) Y_i(c) \right) \\ &= \sum_{k=1}^{K} \left(\frac{1}{n_t} \frac{p}{p_k} \sum_{i:b_i=k} \mathbb{I}_{Z_i=t} Y_i(t) - \frac{1}{n_c} \frac{1-p}{1-p_k} \sum_{i:b_i=k} (1 - \mathbb{I}_{Z_i=t}) Y_i(c) \right).\end{aligned}$$

Sävje (2015) also discussed the effect of variable proportions treated on variance, and Higgins et al. (2015) explored estimators for blocked designs with possibly unequal treatment proportions but also multiple treatments. The costs here are similar to the costs of variable selection probabilities in survey sampling (see Särndal et al., 2003).

When different blocks have different proportions of units treated, it is possible to systematically have blocks and treatment groups with more variance to also have more weight, which could cause blocking to be harmful even in the stratified sampling setting of Subsection 3.3, where we usually have guarantees on the benefits of blocking. In the face of unequal proportions treated, we have the following result for the stratified sampling context:

Theorem 4.1 (Variance comparison with unequal treatment proportions):

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

The first term in the above result is exactly the usual difference in the stratified sampling setting with equal proportions, as shown in Theorem 3.2, and is always nonnegative. The second term, however, can be positive or negative depending upon the proportion treated within each block and the variability of potential outcomes under treatment and control within each block. With unequal proportions treated across blocks, we do not have a setting with simple guarantees on the benefits of blocking.

Consider the following simplification when we have constant additive treatment effects within each block such that $\sigma_k^2(c) = \sigma_k^2(t) = \sigma_k^2$:

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

The second term can be negative if, for example, either $1 - p_k < p < p_k$ or $p_k < p < 1 - p_k$ holds for each of the blocks. This could occur, for instance, if $p = 0.5$, but in some blocks, the researcher is unable to treat exactly half the units (e.g., because of odd number block sizes). In this circumstance, if our blocks are such that $\mu_k(c) = \mu(c)$ and $\mu_k(t) = \mu(t)$ for all k, the whole expression will be negative. Without the simplification that $\sigma_k^2(c) = \sigma_k^2(t) = \sigma_k^2$, the effect of the varying proportion treated can be mitigated or exacerbated depending upon whether systematically more or fewer units are allocated to the more variable treatment condition within the blocks.

In our simulations in Section 6, we in fact see degradation in the benefits of blocking on weakly predictive covariates when the proportion treated is only approximately equal, rather than precisely equal, for the finite sample. This suggests that in many realistic scenarios, one might not want to block on covariates that are only weakly predictive.

5. Misconceptions on the Comparison of Blocking and Complete Randomization

In this section, we explore two misconceptions we have encountered regarding comparisons of blocking and complete randomization. First, there is a common belief that one can simply ignore blocking that was done and analyze a blocked experiment as a completely randomized one without consequence. This belief is a misconception in some contexts: Implementing a blocked design and then ignoring the blocking when calculating the variance estimator will not necessarily be conservative for the variance of $\widehat{\tau}_{(BK)}$. Second, there is a belief that the completely randomized variance estimator (under a completely randomized design) is guaranteed to have lower variability than the typical blocking variance estimator (under a blocked design). This also does not hold in some contexts.

Before exploring these misconceptions in the next two sections, we first need to introduce the standard variance estimators. For a completely randomized design, the standard variance estimator is

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

where $s^2(z)$ is the estimated sampling variance for units assigned to treatment $z \in \{t, c\}$. Under complete randomization, this estimator is conservative for the finite sample, with bias $S^2(tc)/n$, and is unbiased under simple random sampling (see, e.g., Imbens & Rubin, 2015).

Variance estimation for the blocked design is a bit more complicated. However, the usual variance estimation strategy for blocked experiments with at least two treated and two control units in each block is to estimate the variance within each block separately, as in the completely randomized design, with

$$\widehat{\sigma}_k^2 = \widehat{\text{var}}(\widehat{\tau}_k) = \frac{s_k^2(c)}{n_{c,k}} + \frac{s_k^2(t)}{n_{t,k}}.$$

Here, the $s_k^2(z)$ are analogous to $s^2(z)$ within each block k. These variance estimators can then be combined into an overall variance estimator of

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

Extending results for the completely randomized variance estimator, it is well-known (see, e.g., Imbens, 2011) that $\widehat{\sigma}_{(BK)}^2$ is conservative in the finite sample and unbiased under stratified random sampling. See Pashley and Miratrix (2021) for a full discussion of variance estimation under the blocked design, including variance estimators for designs that include blocks with a singleton treatment or control unit, and the resulting bias in estimation under different population frameworks.

5.1. Misconception I: Completely Randomized Variance Estimators Are Conservative for the Blocked Design

What happens if we ignore blocking done in the design stage when analyzing the data? First, if we do not have $p_k = p$ for all k, then it is possible that ignoring the blocking structure can cause bias, with $\mathbb{E}\left[\widehat{\tau}_{(CR)} | \mathcal{S}, \mathbf{P}_{blk}\right] \neq \tau_{\mathcal{S}}$. $\widehat{\tau}_{(CR)}$ could be biased, even under a constant treatment effect assumption, because the completely randomized estimator will effectively give higher or lower weight to some units than the blocked estimator. In this case, variance comparison is less relevant.

When $p_k = p$, although bias in the treatment effect estimator is no longer a concern, using the standard completely randomized variance estimator under a blocked design can cause issues. First, the correct analysis follows from the experimental design. Therefore, if blocking was implemented, it should be taken into account in the analysis. If standard errors based on the completely randomized design are used instead, we are using standard errors that do not actually reflect uncertainty in the design we ran. In other words, the standard errors for the completely randomized design are irrelevant for the blocked experiment we actually ran!

Now we turn to a more technical discussion of why researchers may believe the blocks can be ignored and why this belief is flawed. In particular, the estimated standard errors for a blocked design, when there are many blocks and/or few units, can sometimes be unstable. This instability is separate from the true performance of the blocked estimator; the instability is in estimating the uncertainty, not the uncertainty itself. A researcher might think, therefore, to implement blocking to realize its gains, but then perform the analysis as a completely

randomized experiment to avoid these concerns. Unfortunately, this strategy is not guaranteed to be a good choice.

Regarding this first misconception, we have the following:

> **Theorem 5.1** (Completely randomized variance estimator under blocking: Finite sample): In the finite-sample setting, analyzing a blocked experiment as if it were completely randomized could give anticonservative estimators for variance.

That is, it is possible to have $\mathbb{E}\left[\widehat{\sigma}^2_{(CR)}|\mathcal{S}, \mathbf{P}_{blk}\right] \leq \text{var}\left(\widehat{\tau}_{(BK)}|\mathcal{S}, \mathbf{P}_{blk}\right)$, where $\mathbf{P}_{blk}$ is a blocked randomization assignment mechanism. See Supplementary Material B.1 for a derivation that proves this result (assuming $p_k = p$ for all k and with a positive correlation of potential outcomes). Subsection 6.3 illustrates this result with a simulation.

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

> **Corollary 5.1** (Completely randomized variance estimator under blocking: Stratified sampling): Analyzing a blocked experiment as if it were completely randomized will not give anticonservative estimators for variance if we are analyzing for a superpopulation with fixed blocks and stratified random sampling.

In the context of stratified random sampling, ignoring the blocking in variance estimation is "safe." See Supplementary Material B.2 for more on this result (assuming $p_k = p$ for all k).

5.2. Misconception II: Completely Randomized Variance Estimators Have Lower Variance

Comparisons of blocking to complete randomization often include a discussion on the performance of the variance estimators in terms of their own variance under each design. There is a misconception that the variance of the blocking variance estimator will always be larger than that of the complete randomization variance estimator. We next show that misconception is not necessarily true. Assume that $n_{z,k} \geq 2$ for all blocks k and all treatment assignments z. We focus on the superpopulation framework with stratified random sampling. The question is, do we have guarantees that one variance estimator will have lower variance?

If $\text{var}(s^2(c)|\mathcal{F}_{\text{strat}}, \mathbf{P}_{cr}) \leq \text{var}(s^2_k(c)|\mathcal{F}_{\text{strat}}, \mathbf{P}_{blk})$ and $\text{var}(s^2(t)|\mathcal{F}_{\text{strat}}, \mathbf{P}_{cr}) \leq \text{var}(s^2_k(t)|\mathcal{F}_{\text{strat}}, \mathbf{P}_{blk})$ for most $k = 1, \ldots, K$, then the completely randomized variance estimator will have lower variability than the blocking variance estimator. Imbens (2011) gives such an example.[5] In Imbens's example, the potential outcomes under control have no variation ($\sigma^2_k(c) = 0$ for all $k = 1, \ldots, K$), and the distribution of the potential outcomes under treatment is the same in all of the strata ($\sigma^2_k(t) = \sigma^2(t)$ for all $k = 1, \ldots, K$). Then, Imbens argues, because $s^2(t)$

is a less noisy estimator of $\sigma^2(t)$ than any of the $s_k^2(t)$, the variance of the variance estimator would be smaller under the completely randomized design than under the blocked design. The notion that the variance estimator for complete randomization is less noisy because it is using more data may be true in many situations.

However, this result does not hold in general. For instance, consider a population with four strata. Within each stratum, there is zero treatment effect, and all units are identical. Between strata, however, the potential outcomes differ. For convenience, say in Stratum 1 $Y_i(c) = Y_i(t) = 1$, in Stratum 2 $Y_i(c) = Y_i(t) = 2$, in Stratum 3 $Y_i(c) = Y_i(t) = 3$, and in Stratum 4 $Y_i(c) = Y_i(t) = 4$. Now assume that four units are sampled from each of the strata. In a blocked design, our variance estimate would always be 0. But in a completely randomized design, the variance estimate would change based on which units were assigned to treatment and control. Thus, the blocking variance estimator would have 0 variance, whereas the completely randomized variance estimator would have nonzero variance.

To further explore this question, we compare the variances of the standard variance estimators (under their respective designs) in a simulation in Subsection 6.3. We find that as the blocking estimator gets more precise, relative to complete randomization, the precision of the associated variance estimator also improves. In the case where blocking is not beneficial, we do see a slight increase in the instability of the variance estimator. Overall, we see the relative uncertainty of the blocking estimator is proportional to the true variance of the blocking estimator, and so, when the true variance goes down, the uncertainty of estimating that variance goes down as well. In general, we do not find this additional instability, when blocking is ineffectual, to be a concern; more serious, perhaps, would be the impact of degrees of freedom adjustment when doing inference in the case of experiments with a small number of small blocks.

6. Simulations

We underscore the findings and arguments in our theoretical work with a few illustrative simulations. Our first set of simulations illustrate the general value of blocking along with its potential cost as a function of how successful the researcher is in separating out relatively homogenous sets of units. The second set of simulations unpacks the difficulties in obtaining overall theoretical results on flexible blocking by showing a range of scenarios including one that might deceptively look beneficial, but where flexible blocking can hurt. The third set of simulations illustrate the misconceptions discussed in the previous section.

6.1. Cost and Benefit of Blocking as a Function of Block Variation

Following our theoretical results, we compare actual variances of the treatment effect estimators to avoid complications of any costs or differences in estimating these variances. We examine a series of scenarios ranging from a collection of blocks where there is little variation from block to block (causing blocking to be less beneficial) to scenarios where the blocks are well separated and blocking is critical for controlling variation. In our first numerical study, we treat 20% of units in all of the blocks, with specific block sizes of 10, 15, and 20. We had a total of eight blocks. In the second numerical study, we keep the sizes and number of blocks the same but allow the proportion treated to vary from block to block, from 0.1 to 0.3, with no block having exactly 20% treated. We keep the overall proportion treated the same so that the completely randomized design was the same in both numerical studies. Because the number of treated units is forced to be a positive integer, the small blocks of size 10 have more disparate proportions (.1 or .3), the blocks of size 15 have slightly less disparate proportions (2/15 or 4/15), and the blocks of size 20 have the least disparate proportions (0.15 or 0.25). Different setups for these unequal proportions may yield different results.

Our data-generating mechanism generally follows the one presented in Pashley and Miratrix (2021). The simulations are for the finite sample. Data were generated once from normal distributions for each setting in a manner to ensure that the empirical averages and variances for each block match the theoretical values set for that simulation. This avoids possible pitfalls of odd behavior from a single random finite sample. The block means and treatment effects were set such that they had a negative correlation with block size; larger blocks had lower potential outcomes and smaller treatment effects than small blocks. Through the simulation, we varied how close the blocks were in terms of control means and treatment effects. We also varied the correlation of potential outcomes within blocks as $\rho = 0,\ 0.5,$ and 1. Correlation of 1 corresponds to additive treatment effects within each block. The variances (and variance ratios) were calculated based on the full schedule of potential outcomes using the formulas presented in Subsection 2.2.

The first numerical study corresponds to the mathematical argument presented in Subsection 3.1 and examines how much blocking can hurt. The second numerical study illustrates what changes when proportions treated are not the same across blocks as discussed in Section 4.

We see on the x-axis of Figure 1 an R^2-like measure of how predictive blocks are of the outcome, calculated for each finite data set investigated. The R^2 measure was varied by manipulating the spread of block means under control and the spread of block treatment effects. That is, the x-axis tells us how "good" our blocking is. If we were blocking based on a covariate value, this would measure how predictive that covariate is of the outcome (higher R^2 means more

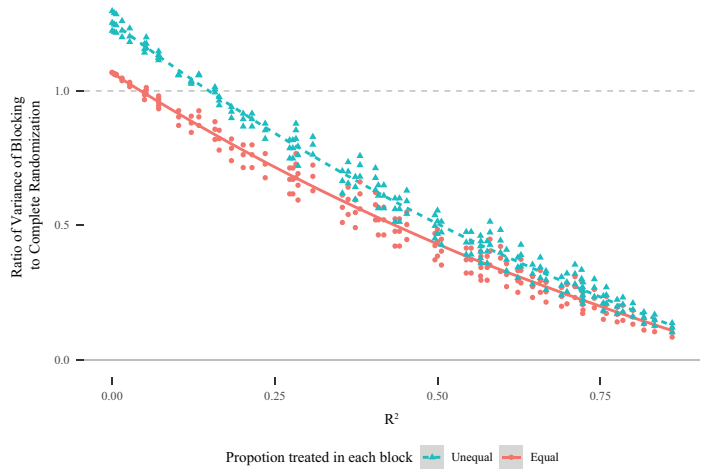


FIGURE 1. *Numerical study to assess completely randomized versus blocked design in finite-sample context when* $\mathrm{p_k} = 0.2$ *(equal proportions) or unequal proportions across blocks. The* y*-axis is* $Var(\widehat{\tau}_{(\mathrm{BK})}|\mathrm{S})/Var(\widehat{\tau}_{(\mathrm{CR})}|\mathrm{S})$.

predictive). The *y*-axis is the ratio of variances of blocking versus complete randomization; values above 1 (dashed horizontal line) indicate a cost to blocking, and values below 1 indicate a benefit to blocking.

Generally, as expected, in most scenarios, we find blocking to be helpful. We see large gains in blocking for moderate-to-large R^2 and only a slight penalty to blocking when the R^2 is relatively small. That is, even in the extreme case where blocks are formed in a (unrealistically) poor manner such that all blocks have exactly the same average potential outcomes but there is variability within in each block, we observed at most an increase in the variance of about 7% for using blocking with equal proportions. On the other hand, the benefits of blocking go to an almost 92% reduction in variance in the most extreme scenario considered.

When we vary the proportions treated, the gains of blocking are muted. The maximum variance increase using blocking in this case is 30%, which occurs in the "worst-case" blocking scenario. This additional cost of blocking is due to the inability to weight all units equally because of the variable proportions treated within each block. This is analogous to the additional cost of incorporating weights in, for example, survey experiments (Miratrix et al., 2018). The maximum benefit of blocking was similar to the equal proportion case, with an approximately 89.8% reduction in variance using blocking.

These results provide some practical guidance: If our blocking factor is even moderately predictive of outcomes, we expect blocking to be beneficial. With equal proportions treated across all blocks, even if we have poor blocks, the cost of blocking will not be too large. When we move away from equal proportions

being treated, the story gets a bit murkier in terms of exactly how harmful blocking can be, but we still expect gains from blocking as long as our blocking factor is reasonably predictive of outcomes.

6.2. Costs and Perils of Flexible Blocking

To illustrate the benefits and perils of flexible blocking, we designed three data-generating scenarios, "linear," "indep," and "odd," dictating the relationship between X and Y. In all the three, X ranges as an integer from 1 to 16. For linear, Y is linearly related to X, making X a good thing to block on. For indep, Y is independent of X, making X a useless thing to block on. Finally, for odd, Y is high if X is odd and low if X is even; blocking on X in this case has unclear benefit.

We also examined three methods for blocking. Our most principled, "flex," divides the units up by X, after sorting X, in a manner that ensures each block has an even number of units. "Interleave" makes blocks by interleaving units, trying to make a collection of blocks that are as similar to each other as possible; this is doing blocking the exact wrong way. Finally, "peevish" is our perverse, existence-proof blocking method where we group units by similar values of X but also ensure there is the same number of odd and even units in each block.

For each of these nine combinations, we repeatedly (10,000 times) generated data and analyzed it via a simulated blocked experiment and a simulated completely randomized experiment. Thus, this corresponds to the simple random sampling setting. We impose a strict zero treatment effect for all units in the experiment, meaning we can measure within- versus between-block variance on outcomes without worrying about the treatment assignment (as the variance is equal in the two treatment groups). We then looked at the relative standard error of the blocking approach compared to complete randomization. Numbers above 100% mean blocking performs worse than doing nothing. Numbers below 100% show blocking to be beneficial.

Results are given in Table 2. When X is independent, how we block does not matter. Blocking neither harms nor helps. Interleaving (deliberately making heterogeneous blocks with respect to X) can cause harm if X matters: When X is linearly related to Y, we have a variance increase. And finally, as a proof of concept, we show that for our peevish blocking approach, even though we are reducing the within-block variance of X, we are doing it in a manner that exactly fails, given the odd data-generating process (DGP) we consider: While this seems unlikely to happen in practice, this demonstrates that, in principle, one could get hurt by a blocking algorithm even while it looks like things are improved. This combined with our odd DGP shows that we cannot get any mathematical guarantees on blocking causing no harm without further assumptions on the DGP or restrictions on the blocking variance estimator approach.

TABLE 2.
Flexible Blocking Simulation Results

	Relative Variance			Variance Ratios			
Blocking Method	Indep	Linear	Odd	X	Y (Indep)	Y (Linear)	Y (Odd)
Flex	100.5	8.2	98.9	0.5	76.4	0.5	75.4
Interleave	99.3	110.4	99.7	93.8	76.1	93.8	76.2
Peevish	100.9	31.4	110.4	7.6	76.2	7.6	93.6

Note. For first three columns, numbers are relative percent of the average size of the standard error of the blocking approach versus complete randomization. Each column corresponds to a data-generating process dictating the relationship of X to Y. Each row corresponds to a form of blocking. The final four columns show the ratio of average variance of X within block to overall and the ratio of average variance of Y within block to overall for each DGP considered. DGP, data-generating process; Indep = independent.

6.3. Illustration of the Misconceptions of Section 5

We next extend the analysis of our simulations to briefly illustrate our two misconceptions in Section 5. The simulation setting is the same as Subsection 6.1, but we only look at the case with equal proportions treated across blocks. Because both of the misconceptions dealt with variance estimation, for each finite population, we simulated 5,000 randomizations and calculated the various variance estimators for each randomization. For the first misconception considered in Section 5, we explore the results of using the completely randomized variance estimator as an estimator for the blocked design. For the second misconception considered in Subsection 5.2, we compare the variance of the standard Neyman variance estimator used with the completely randomized design to the standard extension of the Neyman variance estimator to blocking with the blocked design. Note that we have at least two units in each block with equal proportions treated, allowing the usual blocking variance estimator to be used here.

The results are shown in Figure 2. The graphs are restricted to low values of R^2 to showcase where the more interesting results occur. In the left graph, we see that, for some scenarios considered with a very low R^2 value, using the completely randomized variance estimator with a blocked experiment can result in underestimation (in expectation) of the true variance of the blocked experiment; this is illustrated by points falling below the line at 1. This demonstrates that it is possible for this approach to be anticonservative (although only slightly so). More seriously, this estimator rapidly becomes substantially *conservative*; we do not advise analyzing a blocked experiment as if it were completely randomized. For comparison, the standard blocking variance estimator is also shown. The blocking variance estimator is conservative if there are additive effects

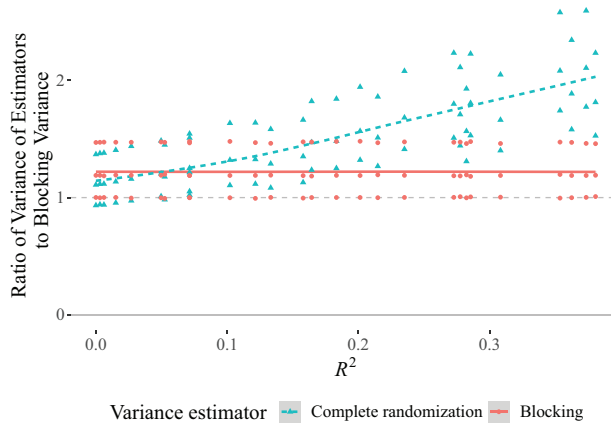
within each block, but the completely randomized variance estimator quickly becomes more conservative at an R^2 less than 0.1.

On the right-hand side of Figure 2, we compare the variability of the standard complete randomization and blocking variance estimators for a series of experiments. We see that while it is possible, when blocks are relatively homogeneous, for the complete randomization estimator to have less instability, it is generally substantially more unstable. The relative instability of the completely randomized variance estimator increases as R^2 increases, and thus, the blocking variance estimator has lower true variance.

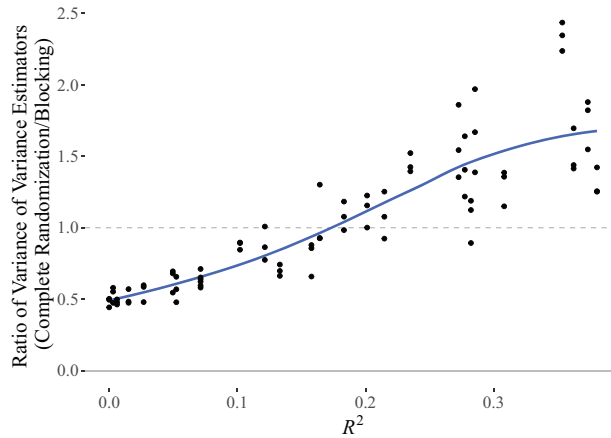
7. Teacher Professional Development, an Applied Example

To further explore the benefits and costs of blocking in a more natural context, we used a data set from a previous randomized trial of a teacher professional development program (Heller et al., 2012) to explore a few hypothetical scenarios of how blocking might work or fail to work in practice. In the actual experiment, the blocks of teachers were defined by geographic area and were thus outside the control of the researcher. We have 15 cohorts (blocks) of teachers with 52%–87% of units treated. Cohort sizes ranged from eight to 29 teachers. Here, cohort membership explained only about 10% of the variance in the baseline test. Furthermore, due to different pragmatic considerations, the proportion treated in each block varied considerably.[6] This is a good example of the worst sort of blocking one might have in a natural context (the blocking was externally forced, hence the nonideal design).

This experiment provides an opportunity in that three tests—a pretest and two follow-up tests—were given to the teachers; we use these outcomes to investigate how different hypothetical blocking strategies may have worked differently. In these investigations, we evaluate how well blocking would work for different hypothetical experiments on units with the same covariates and outcomes as the real units. We initially use the second test as the baseline and the third test as a final outcome and assume no actual treatment impact, using the third test as both the potential outcome for control as well as treatment. This gives us a fully "observed" set of potential outcomes on which we can calculate the impacts of different hypothetical design decisions. That is, we ignore the treatment assignment of units that actually occurred, merely using these data to provide realistic values for an outcome–covariate relationship we might see in practice. Thus, this is a numerical illustration based on practical data values rather than a valid reanalysis of the original study. We consider several different (hypothetical) blocking decisions, constraining ourselves by the block sizes and treatment proportions actually used. We then use the formulae from this work to calculate what the standard errors would be under the different blocking strategies considered, holding the schedule of potential outcomes (the third test) and our assumed baseline test (the second test) fixed.



(a) Ratio of variance estimates to the actual blocking variance.



(b) Variance of variance estimators.

FIGURE 2. *Relationship of estimated variance to actual and precision in the estimated variance. Panel A: Ratio of estimated variance estimators to the actual blocking variance. Panel B: Variance of variance estimators.* Note. *At top, we see the variance estimator for complete randomization can be lower, on average, than the true blocking variance in some circumstances. At bottom, we see the complete randomization variance estimator can be less or more variable than the blocking variance estimator.*

The different explored blocking strategies, and the relative costs and benefits of those strategies, are given in Table 3. The second and third sets of columns show the same exercise but using different tests as baseline and outcome. We report on the

second and third tests, as in this case, there is no actual treatment delivered between tests, arguably making the relationship between baseline and outcome more natural. That being said, the trends are nearly identical for alternate configurations.

We first compare the existing design (blocking on geography) versus a hypothetical complete randomization across all units. Under the finite-sample context, we find a modest 4% increase in the standard error due to having blocked in this haphazard fashion.

In the positive direction, for an experiment where we use the same set of block sizes and proportions treated as our original experiment, but group the teachers into these blocks by their similarity on their baseline test, we find massive benefits to blocking, with the blocked estimator having standard errors 67% of the size of complete randomization; this type of blocking (on a baseline test) would be a natural choice for experimenters with control over their design. For reference, the R^2 measure of our baseline test on the outcome was about 60%. Near the limit, if we were somehow able to obtain and block on a baseline test that was perfectly correlated with the outcome, our standard errors fall to a quarter of the size of complete randomization.

In the other direction, blocking randomly, using the given pattern of blocks and proportions treated, gives around a 5% increase in the standard errors compared to complete randomization, with some variation depending on how lucky or unlucky we are in grouping similar units by chance. The 99th percentile worst allocation, out of 1,000 random allocations tried, was an 8% increase. Identifying the worst of a collection of blocking schemes is similar to the minimax analysis of Nordin and Schultzberg (2020), where they investigate how various restrictions on randomization can lead to the risk of higher variance experiments when one has no covariates predictive of outcome.

Differential rates of treatment can have a cost. To explore this, we examine the case of using the blocks as given, randomizing as equal a proportion of units to treatment as possible across blocks (here 71% with some variation due to needing to treat integer numbers of units). This case gives a slight benefit to the original blocking structure, with standard errors around a half percentage points smaller. Even in this nonideal context, the cohorts were still different enough from each other to offset the potential penalty of blocking. Randomly blocking with equal assignment can still be worse than no blocking but not much worse; even the 99th percentile of 1,000 trials was only a 2.5% increase.

Overall, for the hypothetical experiments motivated by these data, the worst blocking choices were not so bad, and the best were quite good. This reinforces our overall findings: While it is not true that blocking is always beneficial (here for a finite-sample context), it appears difficult to make it substantially harmful. The main concern appears to be when the harm of uninformative blocking is amplified by substantially varying proportions assigned to treatment within blocks. With (roughly) equal proportions of units treated across blocks, the harm of blocking was minimal despite blocks that had little relationship to the

TABLE 3.
Applied Example: Numerical Illustration Results

Scenario Scenario	Relative Standard Error					
	T3 (T2)		T2 (T1)		T1 (T3)	
Actual design	104.0%		102.1		99.5	
Balanced assignment proportions	99.6		99.6		98.0	
Randomly blocking units	104.8	(108.4)	104.8	(108.0)	104.9	(109.1)
Random with balanced assignment	100.9	(102.5)	100.9	(102.6)	100.8	(102.6)
Blocked by baseline test	67.1		93.7		88.9	
Blocked by max inform baseline test	25.4		18.5		27.7	

Note. Relative size of standard error for blocked vs. complete randomization under a variety of blocking strategies applied to the Heller et al. (2012) example data. Numbers in parenthesis are the upper 99% of 1,000 random assignments. First pair of columns used Test 2 for baseline values and Test 3 for hypothetical outcome, second pair used Test 1 for baseline and Test 2 for outcome, and third pair used Test 3 as baseline and Test 1 as outcome.

outcome. We note that these explorations are all scenarios with no impact on any unit. More generally, unequal proportions alone would not necessarily be problematic; as discussed in Section 4, with treatment effect heterogeneity, unequal proportions treated could actually help if we happen to assign more units to more variable treatment arms.

8. Conclusion

Different types of blocks and sampling frameworks can change the answer to the question "Is blocking always beneficial in terms of the precision of my estimators?" We argue that these varying factors are why the current literature can seem confusing and contradictory. Overall, our answer is that blocking is often beneficial, but there are many nuances.

We carefully compared complete randomization to blocking, identifying that prior literature has often collapsed the sampling step and randomization step. Overall, we show that blocking often, but not always, improves precision and that guarantees about blocking depend on the framework adopted. Blocking will not reduce precision, compared to a complete randomization, when working in the stratified sampling framework with equal proportion of units treated across blocks, no matter how small the blocks are or how poorly they separate the units. Similarly, we find that in the simple random sampling setting, given a fixed algorithm for creating flexible blocks, if one makes blocks out of covariates independent of the potential outcomes (the "bad idea" scenario), the blocking estimator will also be no worse than complete randomization. In the other two main frameworks considered, however, blocking is not guaranteed to reduce

variance. We also show that the variance estimators for blocked experiments do not, as is sometimes believed, generally have higher instability than with complete randomization.

These results assume that the blocks have equal proportions of units treated; if the proportions treated differ, we lose all guarantees that blocking will reduce variance regardless of framework. That being said, the simulations and numerical example show the potential for large gains of blocking even with unequal proportions. While the cost of blocking on a weakly predictive covariate in this context can be larger, we still did not see substantial losses of precision.

Even when blocking is unlikely to be helpful in terms of precision, there are several reasons an experimenter might block. First and foremost, an experimenter may simply be forced to block, given the context or constraints of the experiment; our work suggests that little sleep should be lost when this occurs. As the numerical example shows, the negative impact is likely to be small. The simulations further show that blocking can lead to very large gains and only a small potential cost in finite-sample settings with equal proportion treated. Blocking can also guarantee that one has a good balance on those covariates used to make one's blocks; this can increase the credibility of an experiment regardless of any documented relationship between the covariates used and outcomes. In general, experiments with observed systematic differences in the treatment and control group are viewed with greater skepticism. Blocking is good insurance against this concern.

Overall, we advocate for the advice "thoughtfully block when you can" to emphasize that blocking is usually beneficial but must be applied with some thought to avoid edge cases such as inadvertently creating blocks that are equal in distribution. Where possible, researchers should form blocks out of covariates predictive of outcome. They might consider blocking on multiple such covariates to increase the likelihood of obtaining a beneficial blocking. Unless one can predict how the variation in the treated and control units will differ for different blocks, we advise keeping the proportion of units treated similar across blocks. In terms of analysis, we show that one should analyze as a blocked experiment if blocking was done: Completely randomized variance estimators are not necessarily conservative for the blocked design.

In future investigations, it would be useful to assess other practical concerns with blocked designs. Two such concerns are (1) degrees of freedom concerns due to the larger number of parameters that need estimation and (2) further assessment of the stability of variance estimators, which we touched on in Subsection 5.2. The trade-off between decreased precision and reduced degrees of freedom by using blocked or matched-pairs designs has been noted by others (e.g., Box et al., 2005, p. 93; Imbens, 2011; Snedecor & Cochran, 1989, p. 101) and is an important practical limitation to consider when using these designs. Future work should investigate how the real costs of degrees of freedom loss and

instability in variance estimation depend on the experimental design within these frameworks.

Authors' Note

Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

Acknowledgments

The authors thank Guillaume Basse, Avi Feller, Colin Fogarty, Michael Higgins, Luke Keele, and Lo-Hua Yuan for their comments and edits on an earlier version of this article. We also thank members of Luke Miratrix's and Donald B. Rubin's research labs for their useful feedback on the project and Peter Schochet and Kosuke Imai for insightful discussion of this material. Special thanks to Joan Heller of Heller Research Associates for access to the data for the numerical example. Thanks to anonymous reviewers for many comments including ideas regarding the justifications of blocking in practice.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D150040. This material is also based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303.

Notes

1. Typically, the quote is attributed to Box and appears on page 103 of Box et al. (1978).
2. Box actually provided more nuanced advice regarding the advantages and disadvantages of blocking than is shown in this popular quote.
3. We believe algorithms that will end up with random sizes of blocks, even for the same array of units, should follow this argument with a further conditioning step. Formal justification of this is beyond the scope of this work.
4. Sävje (2015) came to similar conclusions in an investigation of a specific form of flexible blocking called threshold blocking (see Higgins et al., 2016). In particular, Sävje (2015) discussed how the blocking may help or harm, depending on the true relationship between the covariates used to block and the outcome and provided a useful decomposition of the true variance in terms of aspects of the blocking algorithm. However, threshold blocking allows for unequal proportions of units treated within each block, which causes the additional complications we consider in Section 4. Sävje (2015) showed no

harm of another form of flexible blocking that does have equal proportions treated across blocks, fixed-size blocking (which here would be forming matched pairs), in a specific setting with a single binary covariate and blocking in a sensible way to match units with the same covariate values together, to the extent possible.

5. We note, however, that we disagree with the generalization made in that paper about the variability of the variance estimators as explained in this discussion.
6. The original experiment was also a trial of multiple versions of treatment that varied by site, which we have collapsed, creating further imbalance in the block sizes and proportions treated. We also imputed some missing values.

References

Box, G. E., Hunter, J. S., & Hunter, W. G. (2005). Statistics for experimenters: Design, innovation, and discovery, 2nd edition. In D. J. Balding, P. Bloomfield, N. A. C. Cressie, N. I. Fisher, I. M. Johnstone, J. B. Kadane, L. M. Ryan, D. W. Scott, A. F. M. Smith, & J. L. Teugels (eds.) *Wiley series in probability and statistics* (p. 93). John Wiley.

Box, G. E., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building. Wiley series in probability and mathematical statistics: Applied probability and statistics*. John Wiley.

Branson, Z., Dasgupta, T., & Rubin, D. B. (2016). Improving covariate balance in 2k factorial designs via rerandomization with an application to a New York City Department of Education high school study. *The Annals of Applied Statistics*, *10*(4), 1958–1976.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333–362.

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

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

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

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society. Series A*, *171*(2), 481–502.

Imbens, G. W. (2011). *Experimental design for unit and cluster randomized trials* [Conference session]. Conference International Initiative for Impact Evaluation, Cuernavaca, Mexico.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Li, X., Ding, P., & Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, *115*(37), 9157–9162.

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

Miratrix, L. W., Sekhon, J. S., & 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.

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

Morgan, K. L., & Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, *110*(512), 1412–1421.

Nordin, M., & Schultzberg, M. (2020). *Properties of restricted randomization with implications for experimental design*. arXiv preprint arXiv:2006.14888. https://arxiv.org/abs/2006.14888

Pashley, N. E., & Miratrix, L. W. (2021). Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, *46*(3), 271–296.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591–593.

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

Sävje, F. (2015). *The performance and efficiency of threshold blocking*. arXiv preprint arXiv:1506.02824. https://arxiv.org/abs/1506.02824

Schultzberg, M., & Johansson, P. (2019). *Re-randomization: A complement or substitute for stratification in randomized experiments?* https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1306368&dswid=-1295

Snedecor, G., & Cochran, W. (1989). *Statistical methods* (8th ed.). Iowa State University Press.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. (1990). On the application of probability theory to agricultural experiments: Essay on principles: Section 9. *Statistical Science*, *5*(4), 465–472. (Original work published 1923)

Authors

NICOLE E. PASHLEY is an assistant professor in the Department of Statistics at Rutgers University, 501 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA; email: nicole.pashley@rutgers.edu; np755@stat.rutgers.edu. Her research interests are primarily focused on causal inference, particularly design-based methodology.

LUKE W. MIRATRIX is an associate professor at the Harvard Graduate School of Education, 14 Appian Way, Cambridge, MA 02138, USA; email: luke_miratrix@gse.harvard.edu. His research interests are primarily pertaining to causality with a focus on developing the methodology to assess and characterize treatment effect heterogeneity in randomized clinical trials and observational studies.

Manuscript received October 26, 2020
First revision received March 10, 2021
Second revision received May 27, 2021
Accepted May 28, 2021