

Research Article

Mattias Nordin* and Mårten Schultzberg

Properties of restricted randomization with implications for experimental design

<https://doi.org/10.1515/jci-2021-0057>

received October 28, 2021; accepted August 05, 2022

Abstract: Recently, there has been increasing interest in the use of heavily restricted randomization designs which enforce balance on observed covariates in randomized controlled trials. However, when restrictions are strict, there is a risk that the treatment effect estimator will have a very high mean squared error (MSE). In this article, we formalize this risk and propose a novel combinatoric-based approach to describe and address this issue. First, we validate our new approach by re-proving some known properties of complete randomization and restricted randomization. Second, we propose a novel diagnostic measure for restricted designs that only use the information embedded in the combinatorics of the design. Third, we show that the variance of the MSE of the difference-in-means estimator in a randomized experiment is a linear function of this diagnostic measure. Finally, we identify situations in which restricted designs can lead to an increased risk of getting a high MSE and discuss how our diagnostic measure can be used to detect such designs. Our results have implications for any restricted randomization design and can be used to evaluate the trade-off between enforcing balance on observed covariates and avoiding too restrictive designs.

Keywords: experimental design, restricted randomization, rerandomization, computationally intensive methods

MSC 2020: 62C20, 62K10, 62K99

1 Introduction

For a long time, the gold standard of causal inference has been randomized experiments. By randomly selecting one group of individuals to be treated and one group to serve as controls, it is guaranteed that the two groups are comparable *in expectation*. However, for a realized treatment assignment, there is no guarantee – and in fact, it is generally not the case – that the two groups are similar in both observed and unobserved characteristics. Therefore, it is possible that an estimated treatment effect is substantially different from the true treatment effect.

The traditional solution to this problem has been to use stratification, or blocked randomization, where randomization is performed within strata based on observed discrete (or discretized) covariates, thereby ensuring balance on these characteristics. However, with many, possibly continuous, covariates, there has not existed a straightforward solution to the problem of imbalances between the treatment and control groups.

Recently, due to vastly increasing computing power, a number of methods have emerged to tackle this problem. For instance, Morgan and Rubin [1] suggest that it is possible to perform a large number of

* **Corresponding author: Mattias Nordin**, Department of Statistics, Uppsala Center for Fiscal Studies (UCFS) and Urban Lab, Uppsala University, Uppsala, Sweden, e-mail: mattias.nordin@statistics.uu.se

Mårten Schultzberg: Spotify and Department of Statistics, Uppsala University, Uppsala, Sweden, e-mail: mschultzberg@spotify.com

randomizations and pick a treatment assignment with a very small imbalance between the treatment group and control group (rerandomization). Furthermore, there are several papers that have developed frameworks and algorithms for finding “optimal” or “near-optimal” designs [2–4]. Usually, such methods are based on restricting the set of treatment assignments so narrowly that only assignments with minimal imbalances in observed covariates, according to some criterion, are considered in the randomization.

However, by heavily restricting the set of admissible treatment assignments one also takes a greater risk of getting a “bad” design. Efron [5] and Wu [6] show that different versions of complete randomization (the unrestricted design loosely defined as the design where all possible treatment assignments are equally likely to be selected) minimizes the maximum mean squared error (MSE) of the treatment effect estimator that can be reached, the so-called *minimax property of complete randomization*.

At the same time, there exists a growing body of literature discussing the efficiency of various restricted randomization designs, such as rerandomization, and how they compare to methods adjusting for covariate imbalances after the experiment has taken place [7–10]. Several papers have shown that rerandomization is asymptotically efficient and that the relevance of the minimax property of Efron [5] and Wu [6] therefore is limited. For example, Li and Ding [10] shows that asymptotically, neither rerandomization nor regression adjustment can ever hurt the precision of the difference-in-means estimator as compared to complete randomization. Similarly, Zhao and Ding [11] argue that randomization inference after rerandomization “... inherits all guarantees from inference under complete randomization ...,” while Ding [12] finds that rerandomization designs with a strict balance criterion is highly efficient.

Others have been more cautious in their recommendation of using rerandomization. For example, Athey and Imbens [13] point out risks with rerandomization as compared to, e.g., stratified randomization designs, while Li et al. [14] and Morgan and Rubin [1], argue that care should be taken such that the balance criterion in rerandomization designs should not be too strict.

In this article, it is possible to reconcile both these views. The first strand of articles focus on the case when the sample size tends to infinity. In that case, our analysis confirms the finding that a restrictive design does not imply an increased risk. However, for small or moderate-sized experiments, modern methods which can heavily restrict the possible treatment assignments may in some circumstances carry substantial risk. We are able to quantify this risk and give practitioners tools necessary to make informed design decisions in such situations.

Just as Efron [5] and Wu [6], we study the risk of getting a high MSE for different designs, but we do it in a different framework. While the aforementioned papers consider situations with the worst possible data that can occur (i.e., in a game against nature), we work in a framework that is independent of the data.

Instead, we introduce the notion of the experimenter being behind a “veil of ignorance” where data cannot be observed. In such a case, the experimenter can be viewed as randomly selecting a design from a set of indistinguishable designs (a notion that we formalize below). While each design has its own MSE, this information is unobserved to the experimenter behind the veil of ignorance, and we show that the largest possible MSE that can be reached is minimized under complete randomization. This result is analogous to the minimax property from Efron [5] and Wu [6], but is proven in a framework without assumptions on the relation between the outcome under no treatment and the outcome under treatment.

The advantage with our framework is that we can go further and show that not only is complete randomization minimax optimal but also the heavier a design is restricted (in the sense that the design contains fewer assignment vectors), the higher the MSE is in the worst possible case. In doing so, we add to an emerging literature discussing the trade-off between using restricted randomization designs to lower the expected MSE, while at the same time making the design robust against “unlucky outcomes” (see, e.g., Kapelner et al. [15], Kapelner et al. [16], Pashley and Miratrix [17], and Harshaw et al. [18]).

In our framework, a natural measure of risk is the variance of the MSE. We show that not only does the maximum MSE increase when designs become more restricted but that is also the case for the variance of the MSE. However, we also show that for a given restriction, even behind the veil of ignorance it is possible to identify designs with a lower variance of the MSE. Specifically, we introduce the *assignment correlation* – a measure of how correlated different assignment vectors in a design are with each other – and show that the variance of the MSE increases linearly in this measure. Intuitively, a design with a high assignment

correlation contains less unique information, implying an increased risk of getting a large MSE.¹ Fortunately, the assignment correlation is straightforward to calculate as it only depends on the combinatoric relationship of the assignment vectors.

The practical value therefore lies in that the experimenter can readily observe the assignment correlation. We show that for traditional designs, such as block randomization, the assignment correlation, and therefore the variance of the MSE, is bounded and cannot take on extreme values. However, more modern designs – which tries to enforce balance on continuous covariates – have the ability to search through millions of assignment vectors and may also use algorithms that cleverly find a narrow set of admissible assignment vectors, thereby heavily restricting the design. In such cases, there is a real possibility of getting a large assignment correlation.

Through a simple simulation study, we show that such algorithmic designs will in most cases work well in the sense that covariate imbalance is minimized with only small increases in the assignment correlation. However, we also show that there are instances when they break down and yield very large assignment correlations. This phenomenon is most likely to occur when data contain outliers which force some units to always be treated together.

While the theoretical result is shown under the veil of ignorance, we argue that the assignment correlation may be a relevant measure of risk also in situations when the experimenter observes covariates likely to affect the outcome. For instance, with rerandomization, there is no currently any agreed upon guideline for how strictly balance should be enforced (see, e.g., discussions in Morgan and Rubin [1] and Li et al. [14]). On one hand, the stricter balance is enforced, the lower the expected MSE is if covariates are informative in explaining the outcome. At the same time, with very strict designs, the “randomness” of the small set of remaining assignment vectors may be compromised, in the sense that these assignment vectors are very similar to each other. In a simulation study, we show that this trade-off between lower expected MSE and higher variance of the MSE is present when covariates are informative and gets more acute as the assignment correlation increases. By observing a high assignment correlation in such a case, the experimenter is made aware of this potential issue and may take appropriate action. We elaborate more on this point in the discussion at the end of the article.

In the next section, we lay out the framework for restricted randomization that we work in. In Section 2.1, we re-prove the minimax property of complete randomization, whereas in Section 2.2, we introduce the assignment correlation and prove that the variance of the MSE is a linearly increasing function of it. In Section 3, we discuss implications for some common designs and perform some simple simulation studies. Finally, in Section 4, we discuss the practical implications for experimental designs in light of the theoretical concept of “veil of ignorance” and we also consider how the assignment correlation can be used in practice.

2 Theory of restricted randomization designs

We consider an experimental setting where a sample of N units is observed and we want to estimate the effect of some intervention (treatment) relative to some baseline (control). The outcome of interest is denoted Y , with $Y_i(1)$ being the potential outcome for unit i if treated and $Y_i(0)$ the potential outcome for unit i if not treated. This formulation is general in the sense that we allow for heterogeneous treatment effects. The interest is in estimating the sample average treatment effect, $\tau := 1/N \sum_{i=1}^N (Y_i(1) - Y_i(0))$.

The division of units into treatment and control is made by a $N \times 1$ assignment vector, $\mathbf{w} \in \mathcal{W}$, containing zeros (untreated) and ones (treated). For simplicity, we will only consider assignment vectors where the sample is split into two evenly sized groups (termed a forced balanced procedure by Rosenberger and

¹ Consistent with our result, in a recent paper, Krieger et al. [19] found that the power of the randomization test becomes worse when assignment vectors are highly correlated with each other.

Lachin [20]), which means that the cardinality of \mathcal{W} is $|\mathcal{W}| = \binom{N}{N/2} = N_A$. This simplification means that we, for instance, do not consider Bernoulli trials, or other trials where the number of treated units is stochastic.

Definition of a design. We define a *design* as a set of assignment vectors from which one assignment vector is randomly chosen. A design has to satisfy the mirror property [16,21], which says that if assignment vector \mathbf{w} is included in a given design, then the assignment vector $\mathbf{1} - \mathbf{w}$ must also be included in that design. A given design is denoted $\mathcal{W}_H^k \subseteq \mathcal{W}$, where H is the cardinality of the design and $k = 1, \dots, \binom{N_A/2}{H/2}$ indexes the different designs that are possible for a given H .

With this definition, we make the restriction that all assignment vectors in a design have the same probability of being chosen of $1/H$. We enforce the mirror property to guarantee that the difference-in-means estimator is unbiased. Note that the number of designs possible for a given H is $\binom{N_A/2}{H/2}$ and not $\binom{N_A}{H}$, because of the mirror property. The difference-in-means estimator of τ for an assignment vector \mathbf{w}_j is

$$\hat{\tau}_j = \frac{1}{N/2} (\mathbf{w}_j^T \mathbf{Y}(1) - (\mathbf{1} - \mathbf{w}_j)^T \mathbf{Y}(0)), \quad (1)$$

where $\mathbf{Y}(0)$ is a $N \times 1$ vector of potential outcomes if not treated and $\mathbf{Y}(1)$ the corresponding potential outcomes if treated. For a given design of size H , \mathcal{W}_H^k , there are H estimates that can be obtained.

Let \mathcal{K} be the set of all possible designs and define $\mathcal{K}_H \subseteq \mathcal{K}$ as the set containing all designs of cardinality H . The MSE of the difference-in-means estimator for a design $\mathcal{W}_H^k \in \mathcal{K}_H$ is

$$\text{MSE}(\hat{\tau}_{\{\mathcal{W}_H^k\}}) := \frac{1}{H} \sum_{\mathbf{w}_j \in \mathcal{W}_H^k} (\hat{\tau}_j - \tau)^2. \quad (2)$$

Note that because the estimator is unbiased, the MSE and the variance are identical.

A randomized design for which $H < N_A$ is sometimes called a *restricted randomization design*,² whereas we call the case when $H = N_A$ *complete randomization*.³ Examples of restricted designs include block designs and rerandomization designs, both of which imply a value of $H < N_A$. The purpose of such restrictions is most often to reduce the MSE of the estimator. Restricted randomization designs generally try to achieve this by excluding assignment vectors from \mathcal{W} that have large imbalances between treatment and control groups in observed covariates, where these covariates are thought to be related to the potential outcomes.

At the design phase, the experimenter would typically not know whether the covariates are relevant and it may therefore be relevant for the experimenter to take into consideration what properties the design has in the unlucky event that the covariates are in fact unrelated to the potential outcomes. In such a case, a restricted design can be viewed as being randomly chosen from the set of equally restricted designs. In such a case, we say that the experimenter is behind a “veil of ignorance.” This idea that basing a restricted design on covariates unrelated to the potential outcomes is the same as making a random restriction is also present in the study of Pashley and Miratrix [17] who for block randomization argue that “...blocking on an X independent of outcome is the same as forming random blocks.”

² Frank Yates in his discussion in Anscombe [22] originally used the term *restricted* to refer to restrictions made to latin square designs in the context of agricultural experiments. Later on, Youden [23] used the term *constrained randomization* also in the context of agricultural experiments.

³ In contexts where designs do not have to be balanced, a forced balanced design may also be called a restricted design. The case when treatment assignment is completely random (i.e., the number of treated and control units need not be the same) is sometimes called *completely randomized design* (CRD), whereas the case of a forced balanced design may be called *balanced completely randomized design* (BCRD), see e.g., Wu [6] and Kapelner et al. [16]. For simplicity, because we exclude CRD, we label BCRD as *complete randomization*.

The worst possible case is that data are in such a way that we select the design from the set which has the highest MSE. This “worst case” is similar to what Efron [5] and Wu [6] consider in their derivations of the minimax property of complete randomization, with the difference being that we do not assume any specific data-generating process. In the next section, we show how the minimax property of complete randomization can be proven in our framework.

2.1 Re-proving the minimax property of complete randomization using the combinatoric approach

The main result in this article, Theorem 4, is proved by studying the behavior of the difference-in-means estimator across various restricted designs. In this section, we build intuition for this way of studying restricted designs by re-proving some known properties of restricted randomization designs. Specifically, we provide an alternative proof that complete randomization satisfies the minimax property (i.e., that it has the smallest maximum MSE) and show that the more restricted a design is, the larger the maximum MSE and the variance of the MSE are. In contrast to Efron [5] and Wu [6], we do not need to make assumptions of an additive homogeneous treatment effects. The only assumption we make is the stable unit treatment value assumption (SUTVA, i.e., no interference and the same treatment), which ensures that the observed outcomes are the same as the relevant potential outcomes.

The key to our formulation of restricted randomization designs is to consider various subsets of the full set of designs, \mathcal{K} . Because the number of possible designs grows extremely fast in N , we will show a simple example for $N = 4$ to build intuition.⁴ For $N = 4$ there are $N_A = \binom{4}{2} = 6$ different assignment vectors:

$$\begin{aligned} \mathbf{w}_1 &= [1 \ 1 \ 0 \ 0]^T, \quad \mathbf{w}_2 = [1 \ 0 \ 1 \ 0]^T, \quad \mathbf{w}_3 = [1 \ 0 \ 0 \ 1]^T, \\ \mathbf{w}_4 &= [0 \ 1 \ 1 \ 0]^T, \quad \mathbf{w}_5 = [0 \ 1 \ 0 \ 1]^T, \quad \mathbf{w}_6 = [0 \ 0 \ 1 \ 1]^T. \end{aligned} \quad (3)$$

For each assignment vector, there is an associated estimate of τ , $\hat{\tau}_1, \dots, \hat{\tau}_6$. For $H = 2$ and $H = 4$, there are three different designs each that satisfy the mirror property, whereas for $H = 6$ there is one:

$$\begin{aligned} \mathcal{W}_2^1 &= \{\mathbf{w}_1, \mathbf{w}_6\}, \quad \mathcal{W}_2^2 = \{\mathbf{w}_2, \mathbf{w}_5\}, \quad \mathcal{W}_2^3 = \{\mathbf{w}_3, \mathbf{w}_4\}, \quad \mathcal{W}_4^1 = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_5, \mathbf{w}_6\}, \\ \mathcal{W}_4^2 &= \{\mathbf{w}_1, \mathbf{w}_3, \mathbf{w}_4, \mathbf{w}_6\}, \quad \mathcal{W}_4^3 = \{\mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4, \mathbf{w}_5\}, \quad \mathcal{W}_6^1 = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4, \mathbf{w}_5, \mathbf{w}_6\}. \end{aligned} \quad (4)$$

Each design has an associated MSE of $\hat{\tau}$, $\text{MSE}(\hat{\tau}_{\{\mathcal{W}_H^k\}})$. Suppose we have the following data:

$$\mathbf{Y}(0) = [1 \ 2 \ 3 \ 4]^T, \quad \mathbf{Y}(1) = \mathbf{Y}(0) + [3 \ 4 \ -2 \ 3]^T = [4 \ 6 \ 1 \ 7]^T. \quad (5)$$

The associated MSEs are

$$\begin{aligned} \text{MSE}(\hat{\tau}_{\{\mathcal{W}_2^1\}}) &= \frac{2}{8}, \quad \text{MSE}(\hat{\tau}_{\{\mathcal{W}_2^2\}}) = \frac{50}{8}, \quad \text{MSE}(\hat{\tau}_{\{\mathcal{W}_2^3\}}) = \frac{8}{8}, \quad \text{MSE}(\hat{\tau}_{\{\mathcal{W}_4^1\}}) = \frac{26}{8}, \\ \text{MSE}(\hat{\tau}_{\{\mathcal{W}_4^2\}}) &= \frac{5}{8}, \quad \text{MSE}(\hat{\tau}_{\{\mathcal{W}_4^3\}}) = \frac{29}{8}, \quad \text{MSE}(\hat{\tau}_{\{\mathcal{W}_6^1\}}) = \frac{20}{8}. \end{aligned} \quad (6)$$

In this section, we consider the distribution of MSE for different sets \mathcal{K}_H , i.e., the distribution of MSE for all the ways H assignment vectors can be drawn from the total of $N_A = \binom{N}{N/2}$ assignment vectors. Naturally, there is only one way to select N_A assignment vectors. That is, there is only one design containing all assignment vectors: *complete randomization*. The MSE for this design is

$$\text{MSE}(\hat{\tau}_{\{\mathcal{W}_{N_A}^1\}}) = \frac{1}{N_A} \sum_{j=1}^{N_A} (\hat{\tau}_j - \tau)^2 = \sigma_{CR}^2. \quad (7)$$

⁴ The number of possible designs is $2^{N_A/2} - 1$, which for $N = 2, 4, 6, 8, 10, \dots$ equals 1, 7, 1023, 3.4×10^{10} , $8.5 \times 10^{37} \dots$

For the aforementioned example, $\text{MSE}(\hat{\tau}_{\{\mathcal{W}_H^1\}}) = \sigma_{CR}^2 = 20/8$.

For the MSEs in equation (6), it is the case that for each \mathcal{K}_H (i.e., for $H = 2, 4, 6$), the average MSE is equal to σ_{CR}^2 . This fact is not a coincidence but something that can be generalized under the following condition:

Condition 1. $\tilde{\mathcal{K}}_H = \{\tilde{\mathcal{W}}_H^1, \dots, \tilde{\mathcal{W}}_H^m\} \subseteq \mathcal{K}_H$ is a set of designs such that

$$\sum_{k=1}^m 1[\mathbf{w} \in \tilde{\mathcal{W}}_H^k] = c, \quad \forall \mathbf{w} \in \mathcal{W}, \quad (8)$$

where $1[\cdot]$ is the indicator function (taking the value of one if the statement in brackets is true, and zero otherwise) and c is a constant.

This condition says that the number of times an assignment vector occurs over all designs in $\tilde{\mathcal{K}}_H$ is the same for all assignment vectors in \mathcal{W} . For the aforementioned example, $c = 1, 2, 1$ for $\tilde{\mathcal{K}}_H = \mathcal{K}_2, \mathcal{K}_4, \mathcal{K}_6$. The motivation behind condition 1 is that we are studying the behavior of different designs with no information available about which specific assignment vectors are more likely to produce an estimate close to τ . Therefore, there is no reason to prefer one specific assignment vector over any other assignment vector and so we only consider sets of designs where each assignment vector is equally likely to be selected.

Condition 1 is valid under the veil of ignorance, but cannot be used when covariates are related to the potential outcomes. While our theoretical results only hold under this condition, the insights we derive using this condition may also be of relevance for situations when covariates have a weak relationship with the outcome, a situation which is quite common in practice. We explore this point in a simulation study in Section 3.5.

Under condition 1, we can state the following result:

Theorem 1. Consider a set of designs $\tilde{\mathcal{K}}_H = \{\tilde{\mathcal{W}}_H^1, \dots, \tilde{\mathcal{W}}_H^m\} \subseteq \mathcal{K}_H$ satisfying condition 1. The expected MSE of the difference-in-means estimator for a randomly chosen design in $\tilde{\mathcal{K}}_H$ is equal to the MSE under complete randomization, i.e.,

$$\frac{1}{m} \sum_{k=1}^m \text{MSE}(\hat{\tau}_{\{\tilde{\mathcal{W}}_H^k\}}) = \sigma_{CR}^2. \quad (9)$$

Proof. See the Supporting information. □

The intuition behind this theorem is straightforward. By condition 1, each assignment vector occurs an equal number of times over all the designs in $\tilde{\mathcal{K}}_H$. For any such set, the average MSE over all the designs in this set is therefore always going to be the same. And because each assignment vector occurs an equal number of times under complete randomization, the expected MSE is always equal to the MSE under complete randomization. The fact that the MSE of sets of restricted designs averages out to the MSE of complete randomization shares some similarities with the literature studying conditional inference in randomized experiments, where tests conditional on observed imbalances in uninformative covariates average out to have the same properties as tests that do not condition on observed imbalances (see, e.g., Krieger et al. [19], Hennessy et al. [24], Branson and Miratrix [25], and Johansson and Nordin [26]).

Remark 1. \mathcal{K}_H satisfies condition 1, which means that the expected MSE for a randomly selected design out of all possible designs containing H assignment vectors is equal to the MSE under complete randomization.

The theorem therefore implies that by randomly selecting a design containing H assignment vectors out of all possible such designs, the expected MSE of the difference-in-means estimator is the same as under complete randomization. However, the distribution of the MSEs will be different for different values of H .

Remark 2. By Theorem 1, all sets of designs under the veil of ignorance (i.e., satisfying condition 1) have the same expected MSE, which equals the MSE under complete randomization. Therefore, the maximum MSE for any such set of designs can never be smaller than the MSE for the set containing only complete randomization ($\mathcal{K}_{N_A} = \{\mathcal{W}\}$). Furthermore, all sets other than the set which contains only complete randomization contain more than one design. Therefore, if all difference-in-means estimates are distinct, the inequalities will in general be strict and \mathcal{K}_{N_A} has the uniquely smallest maximum MSE.

Remark 2 implies a version of the minimax property of complete randomization [5,6], which says that complete randomization minimizes the maximum MSE. In addition, we can go further by proving that more restrictive designs have a greater maximum MSE:

Theorem 2. For \mathcal{K}_H and $\mathcal{K}_{H'}$ such that $H < H'$, it is the case that

- (i) the maximum MSE in the set of all designs in \mathcal{K}_H is greater than, or equal to, the maximum MSE in the set of all designs in $\mathcal{K}_{H'}$.
- (ii) the minimum MSE in the set of all designs in \mathcal{K}_H is smaller than, or equal to, the minimum MSE in the set of all designs in $\mathcal{K}_{H'}$.

If all difference-in-means estimates are distinct, the inequalities are strict.

Proof. See the Supporting information. □

Theorem 2 says that the worst possible design, in terms of MSE, gets worse as H decreases. That is, under the veil of ignorance, the more restrictive the design is, the larger the maximum MSE is.

To measure the risk associated with randomly selecting a design, we not only study the maximum possible MSE but also the variance of the MSE of the difference-in-means estimator. We define this variance as

$$\text{Var}(\text{MSE}(\hat{\tau}_{\{\mathcal{W}_H\}}) : \mathcal{W}_H \in \mathcal{K}_H) := \frac{1}{\binom{N_A/2}{H/2}} \sum_{k=1}^{\binom{N_A/2}{H/2}} \left(\frac{1}{H} \sum_{\mathbf{w}_j \in \mathcal{W}_H^k} (\hat{\tau}_j - \tau)^2 - \sigma_{CR}^2 \right)^2. \quad (10)$$

Theorem 3. Under condition 1, the variance of the MSE of the difference-in-means estimator is decreasing in H .

Proof. See the Supporting information. □

Theorem 3 reinforces the lesson from Theorem 2 that the risk of getting a large MSE of the difference-in-means estimator is smaller, the greater H is. Without any knowledge about $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ it therefore seems as if it is always better to increase H so as to decrease the variance of the MSE of the difference-in-means estimator. However, we have only shown that this is true for sets which contain *all* designs of size H (i.e., \mathcal{K}_H). It might seem impossible to select a subset of \mathcal{K}_H without any *a priori* information, but in the following section we show that it is in fact possible to select subsets of \mathcal{K}_H where the variance of the MSE is smaller. It is possible to do so by using the combinatorial relationship between different assignment vectors.

2.2 Combinatorial uniqueness

In this section, we utilize the information in the combinations, what we call the *pairwise uniqueness*, of the assignment vectors in a design. We show that the variance of the MSE of the difference-in-means estimator depends on this uniqueness and that the uniqueness can be used as a source of information about how “risky” a design is behind the veil of ignorance.

Let w_j^i be the i th element of assignment vector \mathbf{w}_j . The pairwise uniqueness between two assignment vectors \mathbf{w}_j and \mathbf{w}' is defined as

$$U_{j,j'} := \sum_{i=1}^N 1[w_j^i = 1] \cdot 1[w_{j'}^i = 0], \quad (11)$$

where $1[\cdot]$, again, is the indicator function taking a value of one if the statement in brackets is true and zero otherwise. This formula implies that uniqueness is the number of units that are assigned to treatment in one of the assignment vectors but not the other. The uniqueness is theoretically bounded between zero (where an assignment vector is compared to itself) and $N/2$ (where the assignment vector is compared to its mirror).

The uniqueness between two assignment vectors does not depend on data, and so it is a measure that is available even behind the veil of ignorance. Our interest is in studying whether it is possible to use this measure to reduce the risk of a large MSE of the difference-in-means estimator.

To study this question, we continue to restrict attention to sets of designs with H assignment vectors, $\tilde{\mathcal{K}}_H \subseteq \mathcal{K}_H$, which satisfies condition 1, i.e., where all assignment vectors are equally likely to occur over all designs in the set. By Theorem 1, we therefore know that the expected MSE for any set of designs under study equals σ_{CR}^2 . We now add a second condition:

Condition 2. $\tilde{\mathcal{K}}_H = \{\tilde{W}_H^1, \dots, \tilde{W}_H^m\} \subseteq \mathcal{K}_H$ is a set of designs such that

$$\sum_{k=1}^m 1[\{\mathbf{w}_j, \mathbf{w}_{j'}\} \subseteq \tilde{W}_H^k] = d_u, \quad \forall \mathbf{w}_j, \mathbf{w}_{j'} \in \mathcal{W} : U_{j,j'} = u, \quad (12)$$

where $1[\cdot]$ is the indicator function and d_u is a constant for a given value of u .

This condition says that two different assignment vectors \mathbf{w}_j and $\mathbf{w}_{j'}$ will occur together in the same number of designs as any other pair of assignment vectors with the same pairwise uniqueness. The intuition behind this condition is that behind the veil of ignorance, for pairs of assignment vectors with the same uniqueness there is no information available which allows us to say that one pair should occur more often than any other pair. Condition 2 therefore requires such pairs to occur the same number of times.

The insight that the pairwise uniqueness can provide additional information about the variability of the MSE leads us to the following result:

Theorem 4. Consider a set of designs $\tilde{\mathcal{K}}_H = \{\tilde{W}_H^1, \dots, \tilde{W}_H^m\} \subseteq \mathcal{K}_H$ satisfying conditions 1 and 2. The variance of the MSE of the difference-in-means estimator can be written as

$$\text{Var}(\text{MSE}(\hat{\tau}_{\{\mathcal{W}_H\}}) : \mathcal{W}_H \in \tilde{\mathcal{K}}_H) = \psi \frac{4}{N^2} \left(2 \frac{N_A - H}{HN_A} + \frac{H - 2}{H} \phi(\tilde{\mathcal{K}}_H) - \frac{N_A - 2}{N_A} \phi(\mathcal{K}) \right), \quad (13)$$

where $\psi = \psi(\mathbf{Y}(0), \mathbf{Y}(1))$ is a function of the data (related to the kurtosis of the data), $\phi(\tilde{\mathcal{K}}_H)$ is the expected value of $(4/N)^2(u - N/4)^2$ over all designs in $\tilde{\mathcal{K}}_H$, and $\phi(\mathcal{K})$ is the expected value of $(4/N)^2(u - N/4)^2$ over all $N_A(N_A - 2)$ possible pairwise combinations of two distinct assignment vectors (excluding mirrors).

Proof. See the Supporting information. □

The parameter ψ can, roughly speaking, be considered a function of the kurtosis of the data, with the explicit form given in the supporting information in the proof of Theorem 4.⁵ The parameter $\phi(\tilde{\mathcal{K}}_H)$ is written as

⁵ Over random sample and i.i.d data, the expectation of ψ simplifies to $E(\psi) = (\text{Var}(Y(0))^2 + \text{Var}(Y(1))^2 + 2\text{Var}(Y(0))\text{Var}(Y(1)) + 4\text{Var}(Y(0))\text{Cov}(Y(0), Y(1)) + 4\text{Var}(Y(1))\text{Cov}(Y(0), Y(1)) + 4\text{Cov}(Y(0), Y(1))^2)/2$. For homogeneous treatment effects, this in turn simplifies to $E(\psi) = 8\text{Var}(Y(0))^2$.

$$\phi(\tilde{\mathcal{K}}_H) := \left(\frac{4}{N}\right)^{2N/2-1} \sum_{u=1} v_u(\tilde{\mathcal{K}}_H)(u - N/4)^2, \quad (14)$$

where u is the *uniqueness*, i.e., the number of uniquely treated units in one, but not the other, assignment vector in a pairwise comparison of assignment vector. $v_u(\tilde{\mathcal{K}}_H)$ is the proportion of pairwise uniqueness values (excluding mirrors) in the designs in $\tilde{\mathcal{K}}_H$ with uniqueness $U = u$ and $(4/N)^2$ is a normalizing constant. For much of the discussion below, it will be useful to discuss the corresponding parameter for a given design, \mathcal{W}_H , defined as

$$\varphi(\mathcal{W}_H) := \left(\frac{4}{N}\right)^{2N/2-1} \sum_{u=1} \mu_u(\mathcal{W}_H)(u - N/4)^2, \quad (15)$$

where $\mu_u(\mathcal{W}_H)$ is the proportion of pairwise uniqueness values (excluding mirrors) in design \mathcal{W}_H . It is straightforward to show that $\phi(\tilde{\mathcal{K}}_H) = \frac{1}{m} \sum_{k=1}^m \varphi(\tilde{\mathcal{W}}_H^k)$, i.e., that $\phi(\tilde{\mathcal{K}}_H)$ is the average value of $\varphi(\mathcal{W}_H)$ over the designs in $\tilde{\mathcal{K}}_H$.

The parameter φ is a key parameter bounded between zero (if $U = N/4$ for all pairs) and one.⁶ As φ increases, the less unique information is available in each assignment vector in the design. At first glance, one might have expected that information should be increasing in uniqueness. However, because mirrors are always included, this is not the case. For two assignment vectors $\mathbf{w}_j, \mathbf{w}_{j'}$ with $U_{j,j'} = u$, the mirrors are also included with associated values of $U_{N_{A+1}-j, N_{A+1}-j'} = u$ and $U_{j, N_{A+1}-j'} = U_{N_{A+1}-j, j'} = N/2 - u$. The pairwise uniquenesses for a design will therefore always be symmetrically distributed around $N/4$.

A large value of φ implies that the assignment vectors in the design are highly correlated with each other, which also means that the associated treatment effect estimates will correlate with each other. On the other hand, with a small value of φ , the assignment vectors are less correlated, which also means that the estimates of τ have less correlation. We refer to φ as the *assignment correlation* for a design and ϕ as the average assignment correlation for a set of designs.

The amount of information available is maximized when the assignment correlation is zero (i.e., when $u = N/4$). The key insight from Theorem 4 is that the variance of the MSE of the difference-in-means estimator is – under the veil of ignorance – linearly increasing in the average assignment correlation, i.e., the average value of $(u - N/4)^2$ over all pairs in all designs in $\tilde{\mathcal{K}}_H$. Theorem 1 says that the expected MSE from a randomly selected design satisfying condition 1 is equal to σ_{CR}^2 . But for the given design actually selected, the MSE is in general something different. With a large average assignment correlation, the risk that the MSE will be much larger than σ_{CR}^2 is larger than that with a small average assignment correlation. Theorem 4 also shows that the data, $(\mathbf{Y}(0), \mathbf{Y}(1))$, only enter multiplicatively in one place through ψ , which means that the relative variance of the MSE for two different sets of designs is independent of the data. Instead, what determines the relative variances are the values of H and $\phi(\tilde{\mathcal{K}}_H)$ for each set of designs.

It is worth noting that as N increases, the variance of the MSE goes to zero at the rate N^2 , which means that for large sample sizes, the variability of the MSE should be small. However, as we show in Section 3, for smaller experiments such as $N = 50$, the variance of the MSE can be substantial. If $H = N_A$ (complete randomization), the variance of the MSE is zero, because $\phi(\mathcal{K}) = \phi(\mathcal{K}_{N_A})$.

By observing the assignment correlation for a design, the experimenter has an easily accessible diagnostic tool which, behind the veil of ignorance, is informative of the risk of getting a design with a large MSE. In such a case, we can view \mathcal{W}_H as being randomly sampled from the set of all designs with the given values of φ and H . We denote that set $\mathcal{K}_{H,\varphi}$. That is, it is the case that

$$\phi(\mathcal{K}_{H,\varphi}) = \varphi(\mathcal{W}_H), \quad \forall \mathcal{W}_H \in \mathcal{K}_{H,\varphi}. \quad (16)$$

In the Supplementary material, we show that the assignment correlation is bounded between zero and one and that under complete randomization it converges quickly to $1/(N-1)$ as $N \rightarrow \infty$. It is straightforward to calculate φ . The computational complexity is $O(NH^2)$, which means that for small H it is a trivial calculation,

⁶ Technically, the upper bound is $\left(\frac{4}{N}\right)^2 (N/4 - 1)^2 = 1 - \frac{8}{N} + \frac{16}{N^2} < 1$, which happens if U equals 1 or $N/2 - 1$ for all pairs.

but for large H , it might not be computationally feasible to calculate. In such cases, φ can instead be Monte Carlo approximated. In our experience, a calculation of the assignment correlation for $H = 10,000$ does not take much time on a standard desktop computer, but beyond that it starts to be computationally intensive. The value of N does not matter much for the computational time.

The assignment correlation can be viewed as a measure of randomness in a design, where a higher correlation indicates less randomness. Indeed, it is strongly related to, but distinct from, the entropy metric and standard deviation metric discussed by Krieger et al. [4].⁷ In addition, the assignment correlation is, after a linear transformation, identical to what Krieger et al. [19] call the pairwise allocation correlation.

3 Implications for the design of experiments

In this section, we use the result in the previous section to discuss the small-sample properties of the MSE. The key to our discussion is the assignment correlation. Different designs have implications for what values the assignment correlation can take and it is therefore helpful to go through some common designs.

3.1 Block randomization

The perhaps most common design aside from complete randomization is block randomization. In cases with only one block covariate with two equal-sized groups (such as gender), it can be shown that the assignment correlation for such a design is

$$\sum_{u=1}^{N/2-1} \frac{\sum_{i=0}^u \binom{N/4}{N/4-i}^2 \binom{N/4}{N/4-(u-i)}^2}{\binom{N/2}{N/4}^2 - 2} \left(\frac{4}{N}\right)^2 (u - N/4)^2. \quad (17)$$

This assignment correlation converges quickly to $1/(N - 2)$ as $N \rightarrow \infty$ and can never take extreme values. A block design with two blocks is therefore a “safe” design in the sense that there are clear limitations to how large the variance of the MSE can be.

This result is for a block design with the fewest possible number of blocks: two. On the other end of the spectrum, there can be a block design with $N/2$ blocks with two units in each block. In that case, the assignment correlation for the set of all such designs equals

$$\sum_{u=1}^{N/2-1} \frac{\binom{N/2}{u}}{2^{N/2} - 2} \left(\frac{4}{N}\right)^2 (u - N/4)^2. \quad (18)$$

In this case, the assignment correlation converges quickly to $1/(N - N/2) = 2/N$ as $N \rightarrow \infty$. More generally, for any block design with N/b blocks with exactly b units in each block, the assignment correlation converges to $1/(N - N/b)$ as $N \rightarrow \infty$.

Overall, any given block design implies a specific assignment correlation and, different from the strategies discussed below, there is a relatively tight bound for how large it can be.

3.2 Rerandomization

Morgan and Rubin [1] formalize the idea of a *rerandomization design* as a restricted design where only assignment vectors which satisfy a balance criterion based on some observed covariates, such as the

⁷ The entropy metric and standard deviation metric count how often two units are treated together over all assignment vectors in the design, whereas the assignment correlation counts how many units are uniquely treated for each pairwise comparison of assignment vectors, square these, and then average.

Mahalanobis distance, are part of the design. Let $M(\mathbf{X}, \mathbf{w})$ be the Mahalanobis distance of the mean differences of covariates \mathbf{X} of treated and controls given assignment vector \mathbf{w} . Define a threshold value a where $\Pr(M(\mathbf{X}, \mathbf{w}) \leq a) = p_a$. The set of all admissible assignment vectors in a rerandomization design is $\mathcal{W}_H^{RR} = \{\mathbf{w} \in \mathcal{W} : M(\mathbf{X}, \mathbf{w}) \leq a\}$, where $H = N_A \cdot p_a$.

As pointed out in Schultzberg and Johansson [27], block designs can be seen as a special case of rerandomization where the observed covariates on which to rerandomize are categorical. However, with rerandomization using continuous covariates, we can no longer be certain that the assignment correlation will stay reasonably close to $1/(N - 1)$, as was the case for block randomization. Instead, the assignment correlation depends on the distribution of the covariates. This fact does not imply that rerandomization designs will always produce large assignment correlations. With moderate p_a and well-behaved covariates, many rerandomization designs will have assignment correlations which are close to the assignment correlation under complete randomization (see Section 3.4). Rather, the point is that there is no guarantee for a relatively tight upper limit of the assignment correlation, such as is the case for block randomization.⁸

If covariates are informative in explaining the outcome, a small p_a is desirable. At the same time, when choosing p_a , Morgan and Rubin [1] argue that care should be taken such that p_a is large enough for randomization inference to be possible, whereas Li et al. [14] warn against setting p_a too small and suggest that $p_a = 0.001$ is reasonable. They conclude that “How to choose p_a is an open problem” (p. 9162). In Section 4, we discuss how the assignment correlation can be used to guide the experimenter in the choice of p_a .

3.3 Algorithmic designs

Rerandomization provides a straightforward and intuitive way of finding balanced assignment vectors for the experiment. However, with informative covariates, to achieve a given variance reduction with a rerandomization design, the number of expected rerandomizations needed increases exponentially with the number of covariates, making rerandomization computationally very demanding when there are many covariates that need to be balanced.

In response to this issue, a number of different algorithms have been suggested which more efficiently find balanced assignment vectors, making it possible to balance a larger number of covariates (see, for instance, Lauretto et al. [2], Kallus [3], and Krieger et al. [4]). At the same time, for certain type of data, especially containing outliers, there may only be a few ways that treatment can be assigned which enforce balance. In that case, the assignment vectors from using algorithms may be very similar to each other, compromising the randomness of the design. With the assignment correlation, this is something that is directly observable to the experimenter. We illustrate this point below using the pair-switching algorithm of Krieger et al. [4].

The pair-switching algorithm works by randomly choosing an assignment vector. Then it goes through all possible combinations of pair switches, i.e., where one treated and one control unit switch treatment status. The switch that results in the largest drop in Mahalanobis distance is then made, resulting in a new assignment vector, and the algorithm tries all switches again. This process continues until no switch lower the Mahalanobis distance anymore. To obtain a set of assignment vectors, the algorithm is run with different randomly chosen assignment vectors to start the algorithm with.

3.4 Simulation of assignment correlations

For block randomization with equal-sized blocks, each design implies a specific assignment correlation that (asymptotically) is bounded between $1/(N - 2)$ and $1/(N - N/2)$ (see Section 3.1). For rerandomization and

⁸ As mentioned above, there exists an upper bound for the assignment correlation smaller than one, but it is far from as tight as for block randomization.

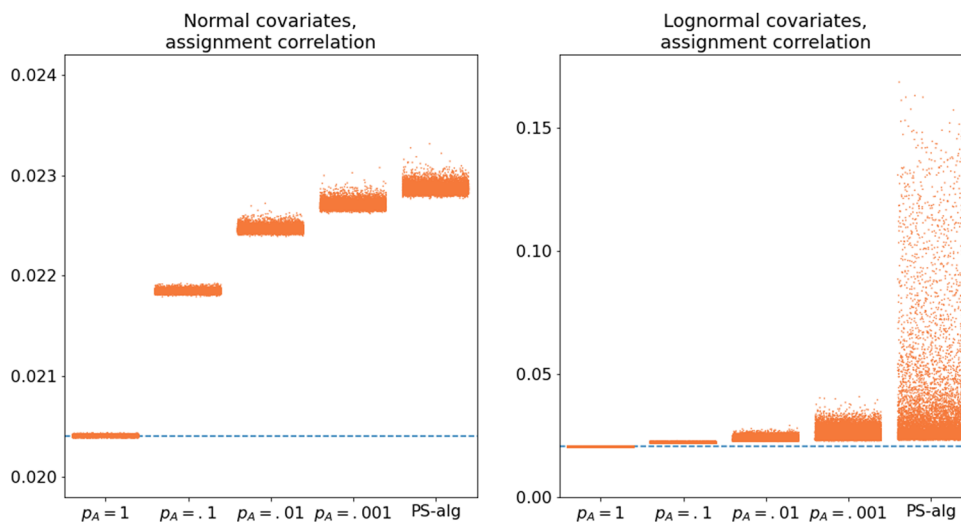


Figure 1: Distribution of assignment correlation for different designs. Note: The figure shows the distribution of the assignment correlation for different designs on normal (left graph) and lognormal (right graph) covariates with the dashed horizontal line indicating the assignment correlation for complete randomization. $N = 50$ and for each sample, the assignment correlations are approximated with 10,000 assignment vectors (including mirrors). A total of 10,000 random samples are drawn. Note that the scale on the y-axis is different for each graph.

the algorithmic designs, there is no way to analytically derive the values of the assignment correlation; rather, they depend on the data. Instead, we perform a simple simulation study to analyze what values the assignment correlation can take, where we study different rerandomization strategies, as well as the pair-switching algorithm of Krieger et al. [4].

We set the sample size to $N = 50$ and balance on five standard normal covariates (corresponding results for $N = 100$ are shown in the Supplementary material). We consider rerandomization with $p_a = \{1, 0.1, 0.01, 0.001\}$. For these designs, $H = p_a \times N_A$, which makes it computationally infeasible to exactly calculate the assignment correlation. Instead, we make a Monte Carlo approximation by randomly sampling 5,000 assignment vectors. With $p_a = 1$, the design is complete randomization. For that case, we know what the exact value of the assignment correlation should be (see the Supplementary material; it is almost exactly $1/(50 - 1)$) and we can therefore see the sampling error due to the Monte Carlo approximation. For the pair-switching algorithm, we randomly choose 5,000 assignment vectors and apply the algorithm, resulting in 5,000 new assignment vectors. For each strategy, no duplicated assignment vector is allowed: if a duplicate is found, it is discarded and a new vector is drawn. With the mirrors added, each design contains 10,000 assignment vectors.

The left graph of Figure 1 plots the estimated assignment correlations for 10,000 different samples with Table 1 providing some key statistics from the simulation. The dashed horizontal line in the figure indicates the true value of the assignment correlation under complete randomization. Beginning from the left, we see that for complete randomization ($p_a = 1$), the Monte Carlo approximated values are very close to the true value, suggesting that 10,000 assignment vectors are sufficient to obtain a good estimate of the assignment correlation.

Turning to the rerandomization designs, we see that for these designs, the more restrictive the design is, the larger the assignment correlation becomes, but even for a fairly unrestrictive rerandomization design ($p_a = 0.1$), there is a clear difference to complete randomization. The spread of the estimated assignment correlations is larger compared to complete randomization because they depend on the data: A design with $p_a = 0.1$ implies a different assignment correlation for each sample, which is different for complete randomization when the assignment correlation is fixed. Finally, the pair-switching algorithm gives the largest assignment correlation. This result is expected as it is the most restrictive design with the smallest associated Mahalanobis distances.⁹

⁹ The maximum Mahalanobis distance for each design is, on average over the 10,000 replications, 21.0 ($p_a = 1$), 1.70 ($p_a = 0.1$), 0.59 ($p_a = 0.01$), 0.23 ($p_a = 0.001$), and 0.11 (pair-switching algorithm).

Table 1: Distribution of assignment correlation for different designs

	Mean	Min	Max	Quantiles			
				0.5	0.75	0.975	0.999
Five standard normal covariates							
$p_A = 1$	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
$p_A = 0.1$	0.0218	0.0218	0.0219	0.0218	0.0219	0.0219	0.0219
$p_A = 0.01$	0.0225	0.0224	0.0227	0.0225	0.0225	0.0225	0.0226
$p_A = 0.001$	0.0227	0.0226	0.0231	0.0227	0.0227	0.0228	0.0229
PS-alg	0.0229	0.0228	0.0233	0.0229	0.0229	0.0229	0.0231
Five log-normal covariates							
$p_A = 1$	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
$p_A = 0.1$	0.0222	0.0218	0.0229	0.0222	0.0223	0.0224	0.0227
$p_A = 0.01$	0.0239	0.0224	0.0291	0.0238	0.0245	0.0254	0.0272
$p_A = 0.001$	0.0258	0.0227	0.0408	0.0250	0.0273	0.0307	0.0373
PS-alg	0.0341	0.0229	0.1687	0.0262	0.0323	0.0820	0.1516

Note: The table presents data from the same simulated distributions of the assignment correlation as is shown in Figure 1. $N = 50$ and for each sample, the assignment correlations are approximated with 10,000 assignment vectors (including mirrors). A total of 10,000 random samples are drawn.

With normally distributed covariates, data are well-behaved whereas real data may contain covariates which are skewed and contain outliers. We therefore perform the same simulation study, but where covariates instead follow a lognormal distribution with the distribution of the assignment correlation being shown in the right graph of Figure 1, with key statistics in Table 1.

We now see quite a different picture where more restrictive designs having a relatively larger assignment correlation on average, but where the distribution also has a large right tail. Indeed, for the pair-switching algorithm, the assignment correlation can become quite extreme with values up to eight times as large as the expectation.

What are the implications of such extreme assignment correlations? To see that, note that if the outcome is independent and identically distributed with unit variance and homogeneous treatment effects, then $\sigma_{CR}^2 \approx 4/N$ and $\psi \approx 8$, which means for large H , a standard deviation increase in the MSE relative to its expectation equals

$$\frac{\sqrt{\text{Var}(\text{MSE}(\hat{\tau}_{\{W_H\}}) : W_H \in \tilde{\mathcal{K}}_{H,\phi})}}{\sigma_{CR}^2} = \sqrt{2\left(\phi(\tilde{\mathcal{K}}_H) - \frac{1}{N-1}\right)}. \tag{19}$$

From this formula, we can see that a value of $\phi = 0.0229$ (the mean value of ϕ for the pair-switching algorithm with normal covariates) implies that the standard deviation of the MSE relative to its expectation is around 0.07: a 7% increase in the MSE for a standard deviation increase relative to its expectation. For $\phi = 0.0341$ (the mean value of ϕ for the pair-switching algorithm with lognormal covariates), the corresponding number is 17%, and for the most extreme value of $\phi = 0.1687$, it is 54%.

Equation (19) is useful in that it is possible, by only observing ϕ , to say how variable the MSE is under the veil of ignorance. In the design phase, covariates are added to hopefully lower the MSE. However, in the unlucky event that covariates are in fact not informative of the outcome, then the experimenter can decide that the MSE should have a standard deviation that differs by at most $b \times 100\%$ from the MSE under complete randomization. Using equation (19), that is equivalent to saying that

$$\phi \leq \frac{b^2}{2} + \frac{1}{N-1}. \tag{20}$$

For $b = 0.1$ (the standard deviation of the MSE should be at most 10% of expected MSE), it would imply $\phi \leq \{0.0254; 0.0151; 0.0100; 0.0070; 0.0060\}$ for $N = \{50; 100; 200; 500; 1,000\}$. Another reasonable criterion would be to say that the values that the assignment correlation take should not exceed the value of

the assignment correlation for the most extreme block design (the design with $N/2$ groups with two units in each group). As discussed in Section 3.1, in such a case the assignment correlation equals $\frac{2}{N}$. In the case with $N = 50$, we see that it would admit all rerandomization designs (and the pair-switching design) with normal covariates, but that it would rule out some designs when covariates are drawn from a lognormal distribution. As with any simple heuristic, such cutoff values for the assignment correlation should only be viewed as an indication that it may be worth considering whether the design should be adjusted. A more thorough analysis would require the experimenter to make guesses on how likely it is that the covariates are uninformative and what the risk preferences are. We return to this point in the concluding discussion.

To further illustrate the variability of the MSE depending on the different designs used – and to extend the analysis to the situation where covariates are informative – we turn to another Monte Carlo simulation.

3.5 Informative covariates

Theorem 4 is derived under the veil of ignorance, equivalent to using a restricted design on covariates which do not explain the outcome. Of course, in practice, restricted designs are often used to balance covariates in the treatment and control groups when they are expected to explain the outcome. To study such a situation, we focus on the case in Morgan and Rubin [1] with homogeneous treatment effects, $Y_i(1) = Y_i(0) + \tau$. We can write the potential outcome under no treatment as a linear projection

$$Y_i(0) = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad (21)$$

where this formulation is nonrestrictive in the sense that the true relationship between $Y(0)$ and \mathbf{x} may be nonlinear.

At the design phase, \mathbf{x}_i is observed to the experimenter, but ε_i is unobserved. For a given design, the MSE of the design is given by \mathbf{x}_i and ε_i . As discussed above, if the covariates do not explain anything of the outcome, we can think of the chosen design as randomly chosen from the set of all designs with given cardinality H and assignment correlation φ , $\mathcal{K}_{H,\varphi}$. An analogous way of thinking about this is that for a given design, \mathbf{x} is fixed, but ε can be viewed as stochastic (since it is unobserved). The advantage of this latter way of thinking is that we can perform analyses when covariates are informative.

Specifically, we want to study the distribution of MSE for various designs (with different associated assignment correlations) depending on how informative the covariates are. To do so, we set up a simple simulation study with \mathbf{X} either being normally or lognormally distributed covariates in the same way as in Section 3.4 and $\varepsilon \sim N(0, c^2)$, with c equaling

$$c = \begin{cases} \sqrt{\text{Var}(\mathbf{X}\boldsymbol{\beta}) \frac{1-R^2}{R^2}}, & \text{for } R^2 > 0, \\ 1, & \text{for } R^2 = 0, \end{cases} \quad (22)$$

where R^2 is the R -squared from the regression of $Y(0)$ on \mathbf{x} .¹⁰ $\boldsymbol{\beta}$ is a 5×1 vector of ones and for $R^2 > 0$ and a vector of zeros for $R^2 = 0$. To make results comparable between realizations of ε , we normalize the outcome as $Y^*(0) = Y(0)/\text{Var}(Y(0))$, where $\text{Var}(Y(0))$ is the sample variance of $Y(0)$ for a given realization of ε . We compare rerandomization designs with $p_a = \{1; 0.5; 0.25; 0.1; 0.025; 0.01; 0.005; 0.0025; 0.001\}$, as well as the pair-switching design of Krieger et al. [4].

Because we condition on observing \mathbf{X} , it is not randomly sampled. Instead, we perform one simulation with a specific realization of \mathbf{X} from a multivariate normal and one from a multivariate lognormal distribution (both taken from a realization in the simulation in Section 3.4). The results are not sensitive to the specific realization of \mathbf{X} , other than by virtue of different \mathbf{X} resulting in designs giving different assignment correlations. For all designs, we fix the number of assignment vectors in each design to $H = 10,000$

¹⁰ This R -squared is the population R -squared (over random sampling of ε), corresponding to the asymptotic results about variance reduction proportional to R -squared in Morgan and Rubin [1].

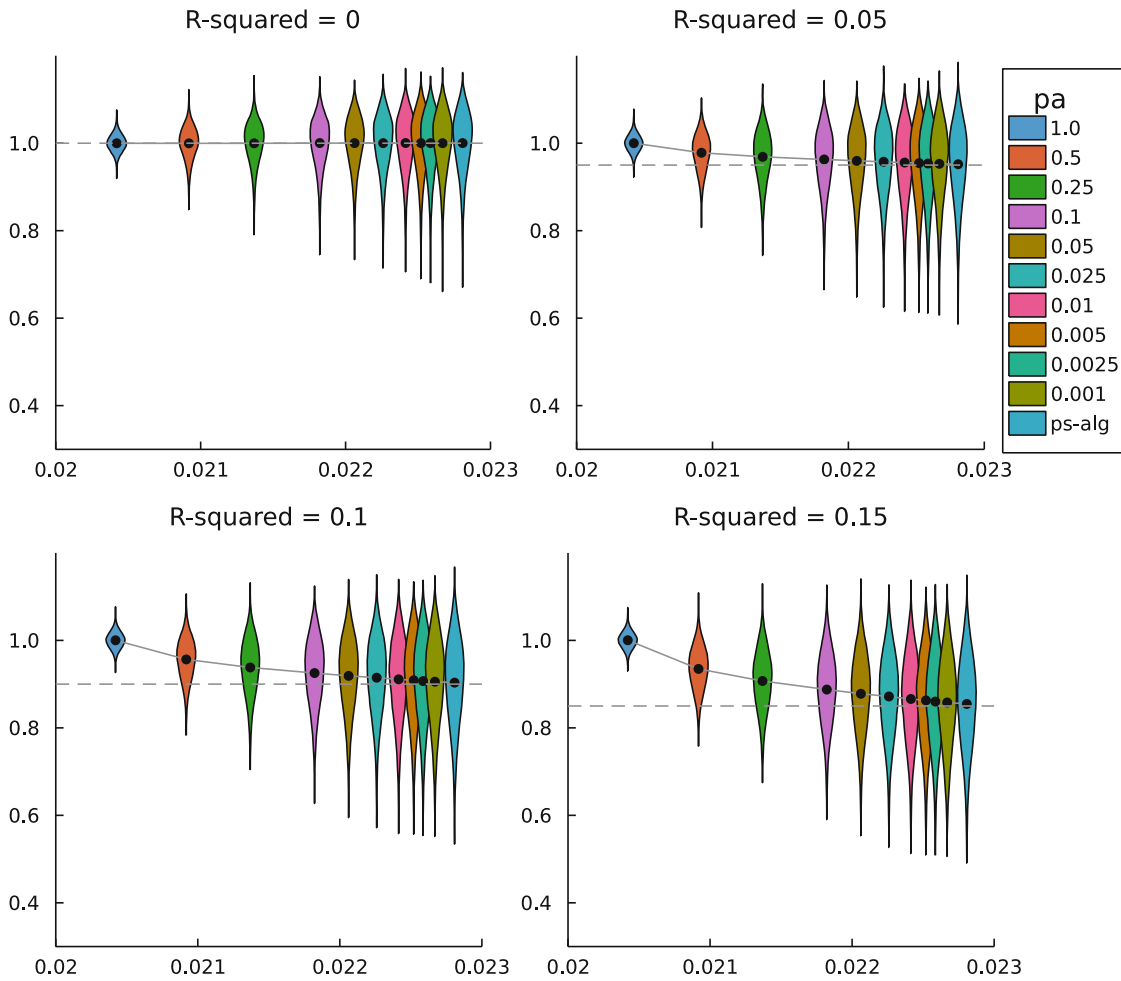


Figure 2: Distribution of MSE (y-axis) against assignment correlation (x-axis) for different designs, normal covariates. Note: The plots are violin plots indicating the distribution of the MSE for different designs. The dashed line shows the expected MSE if covariates are perfectly balanced ($1 - R^2$). The black dots indicate the mean of the distribution of MSE for each design, connected with a gray line.

(including mirrors) and we perform 10,000 replications. Because the designs are only based on a fixed \mathbf{X} , each replication uses the same set of assignment vectors for each design (for the normal and lognormal covariates, respectively).

Figure 2 displays the distribution of MSE for the normal covariates with $R^2 = \{0;0.05;0.1;0.15\}$ for the different designs. Because $Y^*(0)$ is normalized, $\sigma_{CR}^2 = 1$, and the y-axis gives the MSE relative to σ_{CR}^2 . On the x-axis is the assignment correlation and we see that it takes values from 0.0204 (for $p_a = 1$) to 0.0228 (for the pair-switching design) with more restrictive designs having higher assignment correlations; results which are very much in line with those in Figure 1.

Beginning in the top left panel where $R^2 = 0$, we see that the expected MSE equals σ_{CR}^2 , as it should, for all designs. As the assignment correlation increases, the variance of the MSE increases exactly in the way described in equation (13) (something that we show in the Supplementary material where the empirical variances are compared to the theoretical expectation). Note that the design with $p_a = 1$ is not equivalent to complete randomization because we have set $H = 10,000$ to make results comparable with all other designs.

Turning to the other three panels, we see that the more restrictive designs have lower average MSE, which approaches the theoretical minimum expected MSE of $1 - R^2$ (i.e., when covariates are perfectly balanced in a rerandomization design). However, we also see that the variance of the MSE is larger when the assignment correlation is larger, implying that there exists a trade-off between lowering the expected MSE and lowering the variability of the MSE.

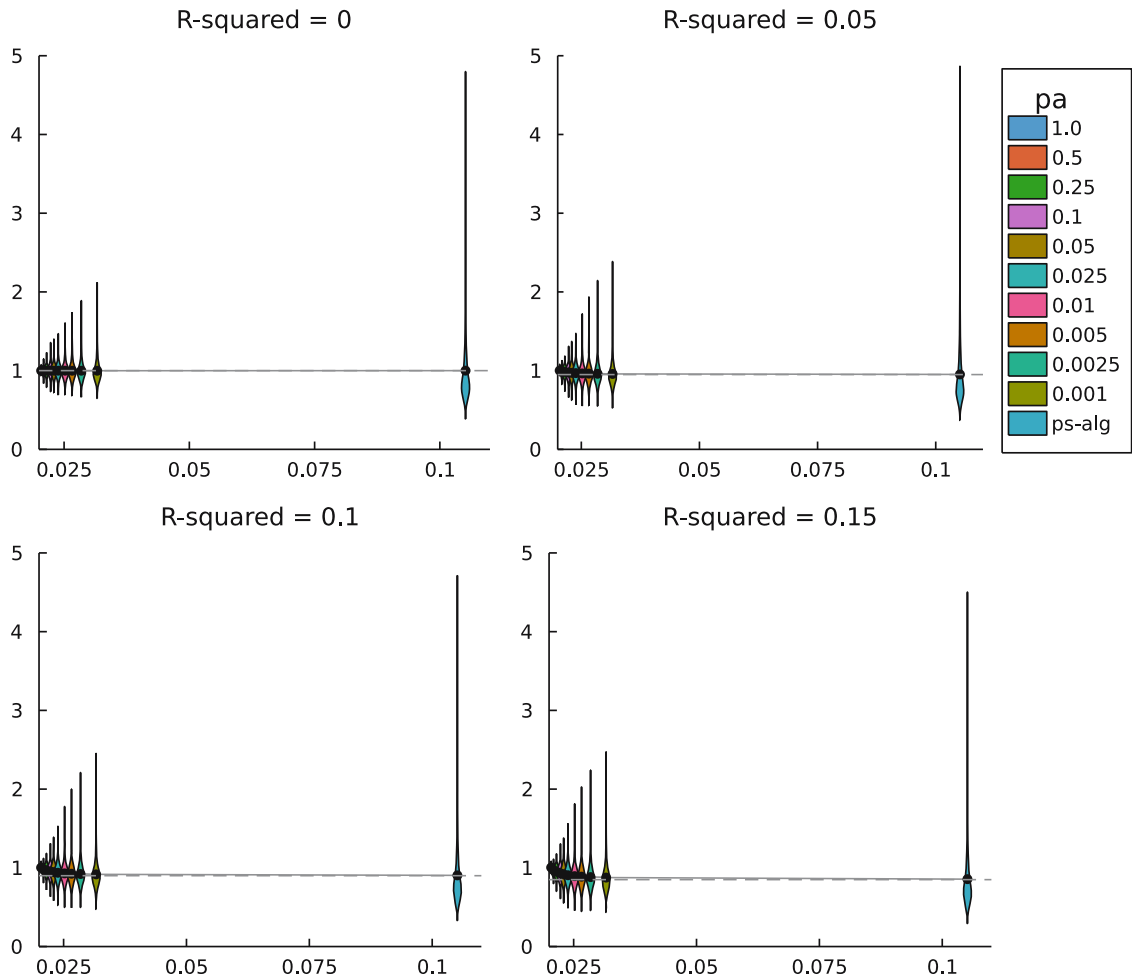


Figure 3: Distribution of MSE (y -axis) against assignment correlation (x -axis) for different designs, lognormal covariates. Note: The plots are violin plots indicating the distribution of the MSE for different designs. The dashed line shows the expected MSE if covariates are perfectly balanced ($1 - R^2$). The black dots indicate the mean of the distribution of MSE for each design, connected with a gray line.

Figure 3 displays the distribution of MSE where \mathbf{X} comes from a lognormal distribution containing substantial outliers. Here, we see that the assignment correlation can take on much larger values, especially for the pair-switching design where it equals 0.105. Again, for $R^2 = 0$, the variances of the MSE are in line with equation (13) (see Supplementary material). The trade-off between expected MSE and variability of MSE is much stronger here. For instance, the average MSE for the pair-switching design when $R^2 = 0.1$ is 0.904 (a 9.6% drop compared to complete randomization) with a standard deviation of 0.27. This can be compared, for instance, to the rerandomization design with $p_a = 0.001$ where the average MSE is 0.911 with a standard deviation of 0.13. The reason for this large discrepancy in standard deviations is that the assignment correlation is 0.105 in the former case compared to 0.0316 in the latter case. The experimenter has to choose whether the additional savings in expected MSE of 0.7% points is worth the risk of a much larger variability in the MSE. In fact, a risk-averse experimenter may even prefer the rerandomization design of $p_a = 0.1$ with an average of MSE of 0.934 and a standard deviation of 0.08, or even complete randomization where the MSE is fixed at one.¹¹

¹¹ To be clear, the MSE here is calculated for $H = 10,000$. The true variability for the different designs is slightly lower as H is much larger for the different rerandomization designs. However, it is not computationally feasible to calculate such MSEs, and

4 Discussion

In this article, we study inference to the sample in balanced experiments with two groups. We show that, under the veil of ignorance, more restrictive design is more “risky,” in the sense that they have a more variable MSE with the least restrictive design, complete randomization, being the least risky design. Importantly, we also show that this risk is increasing in the *assignment correlation* – an easily observable measure of the “degree of randomness” in a design. Intuitively, more restrictive designs put restrictions on which assignment vectors are admissible, which may imply that only assignment vectors which are very similar to each other are included in a design.

Simulation results indicate that with designs that balance on well-behaved covariates – and where the designs are not too restrictive – the assignment correlation tends to be quite modest, implying that the variance of the MSE, our measure of risk, is not that high. However, in some extreme cases, the assignment correlation can increase substantially, causing concerns that the design is volatile.

The theoretical results are valid under the veil of ignorance where covariates carry no information about the outcome, with the expected MSE being the same for all designs (Theorem 1). In reality, restricted designs such as rerandomization are used when covariates are expected to be informative in explaining the outcome, thereby decreasing the MSE. It is therefore natural to ask what relevance the preceding analysis holds in applied settings.

There are several reasons for why we consider our approach to be relevant. First, before an experiment has taken place, it is in general unknown whether covariates explain the outcome and researchers may be inclined to include covariates in designs even when there is no strong *a priori* reason for their inclusion. Second, and more importantly, even when covariates can explain the outcome, in many cases the question arise of just *how restricted* a design should be, with no satisfactory answer given in the literature [14]. In a simulation study, we show that when covariates can explain the outcome, the average MSE tends to be lower for more restrictive designs (with higher associated assignment correlations), but the MSE is also more variable.

Furthermore, rerandomization is a fairly moderate design which even for small values of p_a generally contains millions of admissible assignment vectors. However, with modern algorithmic designs, it is possible for designs to include extremely few assignment vectors (see Johansson et al. [28] and Kallus [29]), and we would expect even more such designs to emerge with clever algorithms to search high-dimensional spaces. In such cases, we expect the assignment correlation to be especially useful. In the design stage, the experimenter may choose whether it is worth it to choose an even more restrictive design to squeeze out the last gains in lower MSE or, whether to be concerned that such an approach compromises the “randomness” of the design. In such a case, the experimenter can simply look at the assignment correlation. If it does not change much when further restricting the design, then it is not particularly risky to choose the restrictive design. However, if the assignment correlation balloons, such as it did for the lognormal covariates in the simulation study in Section 3.5, then the experimenter should question whether it is appropriate with such a restrictive design.

With rerandomization, an algorithmic way of deciding p_a (i.e., just how restricted the design should be) would be to set a threshold value for the largest admissible assignment correlation. Such a threshold could be given, for instance, by deciding the maximum value the standard deviation of the MSE can take relative to the expected MSE, or by limiting the assignment correlation to the assignment correlation of the most restrictive block design (see Section 3.4). One can then continuously rerandomize, gradually lowering p_a and calculate the assignment correlation. p_a is then chosen based on when the assignment correlation exceeds the threshold value or when the allocated time for rerandomization runs out. In this way, the experimenter can get as low expected MSE as possible, while at the same time feel confident that the

since the variance of the MSE is much smaller for $p_a = 1$ compared to the other designs, the approximation error due to a smaller H is of second-order importance.

resulting design is not too narrow and retains sufficient variation in the assignments. Similar algorithms can be used for various algorithmic designs.

If a design implies a large assignment correlation, it is a sign that the design is being driven by outliers, forcing a certain configuration of treated and control to ensure balance. In such a case, the experimenter may consider trimming, transforming, or excluding offending covariates.

Just how large is an assignment correlation that is “too large”? The threshold values discussed above are shortcuts, but are not guaranteed to be relevant in every situation. Unfortunately, it is not possible to give an unambiguous answer to that question. Rather, the answer depends on (i) how large the expected R -squared is (i.e., how much of the variation in the outcome the covariates are expected to explain) and (ii) what the risk-preference of the experimenter is. For the first question, for a hypothesized R -squared it is possible – at least with Mahalanobis-based rerandomization – to calculate the expected MSE reduction, as well as what the variance of the MSE would be if the covariates turned out to be uninformative, for various designs. For the second question, in most cases it seems likely that the experimenter is risk-averse, meaning that a large assignment correlation should be avoided, all else equal. A risk-neutral experimenter on the other hand would be indifferent between different assignment correlations whereas if they are risk-loving, a large assignment correlation may even be preferable. It is also possible that the experimenter’s risk preference depends on whether the study is well-powered or not (for a recent study of how power of randomization tests relates to the correlation between assignment vectors, see Krieger et al. [19]). A complete analysis of this issue would be well-studied in a decision-theoretic framework, something that is beyond the scope of this article, but which would be an interesting avenue for future research.

Acknowledgments: The authors are grateful for helpful comments and suggestions by Peng Ding, Guido Imbens, Per Johansson, two anonymous reviewers, and seminar participants at Uppsala University.

Author contributions: Both authors have contributed equally to all aspects of the article.

Conflict of interest: Mårten Schultzberg is currently employed by Spotify. Other than that, the authors have no conflict of interest.

Data availability statement: Code for reproducing all simulation results in the article and the Supplementary material is available at <https://github.com/mattiasnordin/Properties-of-restricted-randomization>.

References

- [1] Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. *Annals Statist.* 2012;40(2):1263–82.
- [2] Lauretto MS, Stern RB, Morgan KL, Clark MH, Stern JM. Haphazard intentional allocation and rerandomization to improve covariate balance in experiments. In: *AIP Conference Proceedings*. Vol. 1853. AIP Publishing LLC; 2017. p. 050003.
- [3] Kallus N. Optimal a priori balance in the design of controlled experiments. *J R Statist Soc B (Statist Methodol)*. 2018;80(1):85–112.
- [4] Krieger AM, Azriel D, Kapelner A. Nearly random designs with greatly improved balance. *Biometrika*. 2019;106(3):695–701.
- [5] Efron B. Forcing a sequential experiment to be balanced. *Biometrika*. 1971;58(3):403–17.
- [6] Wu C-F. On the robustness and efficiency of some randomized designs. *Annals Statist.* 1981;9(6):1168–77.
- [7] Freedman DA. On regression adjustments to experimental data. *Adv Appl Math.* 2008;40(2):180–93.
- [8] Lin W. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique. *Annals Appl Statist.* 2013;7(1):295–318.
- [9] Middleton JA. A unified theory of regression adjustment for design-based inference. 2018. arXiv: <http://arXiv.org/abs/arXiv:1803.06011>.
- [10] Li X, Ding P. Rerandomization and regression adjustment. *J R Statist Soc Ser B (Statist Methodol)*. 2020;82(1):241–68.
- [11] Zhao A, Ding P. Covariate adjustment in multi-armed, possibly factorial experiments. 2021. arXiv: <http://arXiv.org/abs/arXiv:2112.10557>.

- [12] Ding P. Two seemingly paradoxical results in linear models: the variance inflation factor and the analysis of covariance. *J Causal Infer.* 2021;9(1):1–8.
- [13] Athey S, Imbens GW. The econometrics of randomized experiments. In: *Handbook of economic field experiments*. Amsterdam, Netherlands: Elsevier; 2017. vol. 1. p. 73–140.
- [14] Li X, Ding P, Rubin DB. Asymptotic theory of rerandomization in treatment-control experiments. *Proc Nat Acad Sci.* 2018;115(37):9157–62.
- [15] Kapelner A, Krieger AM, Sklar M, Azriel D. Optimal rerandomization via a criterion that provides insurance against failed experiments. 2019. arXiv:1905.03337.
- [16] Kapelner A, Krieger AM, Sklar M, Shalit U, Azriel D. Harmonizing optimized designs with classic randomization in experiments. *Am Statistic.* 2021;75(2):195–206.
- [17] Pashley NE, Miratrix LW. Block What You Can, Except When You shouldn't. *J Educat Behav Statist.* 2022;47(1):69–100.
- [18] Harshaw C, Sävje F, Spielman D, Zhang P. Balancing covariates in randomized experiments with the Gram-Schmidt Walk design. 2022. arXiv:1911.03071v5.
- [19] Krieger AM, Azriel D, Sklar M, Kapelner A. Improving the power of the randomization test. 2020. arXiv:2008.05980.
- [20] Rosenberger WF, Lachin JM. *Randomization in clinical trials: theory and practice*. New York, United States: John Wiley & Sons, Incorporated; 2015.
- [21] Johansson P, Schultzberg M. Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. *J Comput Graph Statistics.* 2020;29(4):798–813.
- [22] Anscombe FJ. The validity of comparative experiments (discussion by Yates). *J R Statist Soc Ser A (General)* 1948;111(3):181–211.
- [23] Youden WJ. Randomization and experimentation. *Technometrics.* 1972;14(1):13–22.
- [24] Hennessy J, Dasgupta T, Miratrix L, Pattanayak C, Sarkar P. A conditional randomization test to account for covariate imbalance in randomized experiments. *J Causal Infer.* 2016;4(1):61–80.
- [25] Branson Z, Miratrix LW. Randomization tests that condition on non-categorical covariate balance. *J Causal Infer.* 2019;7(1):1–29
- [26] Johansson P, Nordin M. Inference in experiments conditional on observed imbalances in covariates. *American Statistician.* 2022;1–11. <https://doi.org/10.1080/00031305.2022.2054859>.
- [27] Schultzberg M, Johansson P. Rerandomization: a complement or substitute for stratification in randomized experiments? *J Statist Planning Infer.* 2022;218:43–58.
- [28] Johansson P, Rubin D, Schultzberg M. On optimal rerandomization designs. *J R Statist Soc Ser B (Statist Methodol).* 2021;83(2):395–403.
- [29] Kallus N. On the optimality of randomization in experimental design: How to randomize for minimax variance and design-based inference. *J R Statist Soc Ser B (Statist Methodol).* 2020;83(2):404–9.