UPPSALA
UNIVERSITET

Working Paper 2019:4

Department of Statistics

Re-randomization: A complement or substitute for stratification in randomized experiments?

Mårten Schultzberg and Per Johansson

Working Paper 2019:4
April 2019
Department of Statistics
Uppsala University
Box 513
SE-751 20 UPPSALA
SWEDEN

Working papers can be downloaded from www.statistics.uu.se

Title: Re-randomization: A complement or substitute for stratification in randomized experiments?

Author: Mårten Schultzberg and Per Johansson
E-mail: marten.schultzberg@statistik.uu.se

Rerandomization: A complement or substitute for stratification in randomized experiments?[1]

Mårten Schultzberg and Per Johansson

April 23, 2019

[1] **The authors thank Nikolay Angelov, Bength Muthén, Linda Muthén, Mattias Nordin and seminar participants at the Institute for Evaluation of Labour Market and Education Policy (IFAU) and the UppUpp conference, Uppsala University and SLU for helpful comments. Berndt Lundgren is acknowledged for kindly sharing data.**

Abstract

Rerandomization is a strategy for improving balance on observed covariates in randomized control trails. It has been proposed as a complement to traditional stratified (blocked) designs. However, the relationship and differences between stratification, rerandomization, and the combination of the two have not been previously investigated. In this paper, we show that stratified designs can be recreated by rerandomization and explain why, in most cases, stratification on binary covariates followed by rerandomization on continuous covariates is more efficient than rerandomization on all covariates at the same time.

1 Introduction

The most common design used to improve balance a randomized control trials is stratified randomization, which is also called blocked randomization. In this paper we use the term 'stratification'. The idea of stratification is to divide units into strata (i.e. groups/blocks) based on similarity on a set of covariates, and then perform complete randomization within each stratum. In this way, units from all strata are represented in both the treatment and control groups and imbalances on any of these covariates are thereby avoided, see e.g. Imbens and Rubin (2015) for a recent overview.

Another design is rerandomization. Even though suggested already by R. A. Fisher and implemented by researchers for a long time (cf. the references in Morgan and Rubin (2012)), rerandomization as a design for improving inference was first formally proposed in Morgan and Rubin (2012). As the name suggests, rerandomization consists of redoing the randomization until some pre-specified balance criterion on the observed covariates is met. That is, the randomization is restricted to a subset of allocations that fulfill a rerandomization covariate balance criterion. Rerandomization designs are especially useful when covariates are continuous. Stratification requires continuous covariates to be discretized which leads to a loss of information. Rerandomization is computationally demanding compared to stratification. However, with modern computers this is not a limitation in practice, making rerandomization a powerful and flexible design.

In their original rerandomization paper, Morgan and Rubin (2012) do not propose rerandomization as a substitute for stratification. Instead, the motivation for rerandomization is based on an understanding of that, also after blocking, complete randomization within strata can result in imbalances in other covariates. In this situation, Fisher is alleged to have recommended rerandomization (Morgan and Rubin, 2012). This recommendation was summarized in a presentation by D. B. Rubin as 'Block on what you can and rerandomize on what you cannot'. In The Handbook of Economic Field Experiments, Athey and Imbens (2016) recommended researchers to first and foremost take care in the 'original design' to rule out unbalanced assignments instead of relying on rerandomization. This recommendation by the authors may be interpreted that they view rerandomization as a substitute for stratification which may be unfortunate if relevant continuous covariates are available. It is, however, arguably not obvious how or when to combine these strategies. This paper contributes to the literature by comparing the properties and clarifying the relationship between stratification, rerandomization and the combination of both stratification and rerandomization.

Throughout the rest of this paper, we will use *stratified rerandomization* to mean to first stratify on the available binary covariates and then rerandomize on available continuous covariates, and *rerandomization* to mean to rerandomize on all available covariates at the same time. When rerandomization is applied, we use the Mahalanobis distance rerandomization criterion. We show in this paper that the stratified design can be recreated by rerandomizing on the covariates used to form the strata. Utilizing this equivalence, we compare the relative efficiency of stratified rerandomization as compared to rerandomization. The two designs are asymptotically equivalent with respect to efficiency. However, for moderate large N, stratified rerandomization is in general more efficient than rerandomization. With small N, the relationship may be reversed. We focus on exact inference to units of the experiment. The main reason for this is clarity. The aim is to explain the relationship between stratification and rerandomization, and this is more clearly achieved for inference to the units of the experiments. Most of the theoretical results discussed in this paper extend to inferences to a population under random sampling. In addition, as will be shown, the choice of design is more complex in small sample settings where exact inference may be desirable for its lack of distributional assumptions.

The rest of this paper is structured as follows. Section 2 establishes the considered experimental designs. Section 3 contains an overview of Mahalanobis-based rerandomization highlighting the most important theoretical results from Morgan and Rubin (2012, 2015). Section 4 investigates the relationship between stratification and rerandomization. Section 5 presents a Monte Carlo study confirming the theoretical findings, and Section 6 makse use of electricty consumption data as an illustration. Section 7 contains a discussion and concluding remarks.

2 Complete randomization, stratification and rerandomization

Let $Y_i(w)$ denote the potential outcome for unit i given the 'treatments' ($w = 0, 1$), e.g., treatment and control. For a sample of N experimental units, the sample average treatment effect is defined as

$$SATE = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0).$$

An experiment will have N_1 units assigned the treatment for which we observe $W_i = 1$, and N_0 units assigned the control for which $W_i = 0$ is observed. Under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980) the observed Y_i is equal to $Y(W_i)$. Define the mean outcome for the treated and controls

$$\overline{Y}_1 = \frac{1}{N_1}\sum_{i:W_i=1}^{N_1} Y_i \text{ and } \overline{Y}_0 = \frac{1}{N_0}\sum_{i:W_i=0}^{N_0} Y_i.$$

The mean difference

$$\widehat{\tau} = \overline{Y}_1 - \overline{Y}_0 \tag{1}$$

is an unbiased estimator of the SATE (Neyman, 1923, 1990). Assuming that the units are randomly sampled from a population, this estimator is an unbiased estimator also for of the population average treatment effect for that given population (see e.g. Aronow et al. 2014). Let $\mathbf{W}$ be the $N \times \mathcal{N}_A$ matrix of all $\binom{N}{N_1} = \mathcal{N}_A$ possible allocation vectors under complete randomization, and let $\widehat{\tau}^j$ be the *estimate* for allocation j with assignment vector $\mathbf{W}^j \in \mathbf{W}$.

2.1 Stratification

To further the intuition for what we try to achieve with experimental designs, note that the variance of the SATE estimator under complete randomization can formulated in the following way

$$V_{\mathbf{W}}(\widehat{\tau}) = \frac{1}{\mathcal{N}_A}\sum_{j=1}^{\mathcal{N}_A}(\widehat{\tau}_j - SATE)^2.$$

The idea of stratification is to remove allocations with potential large differences in $(\widehat{\tau}_j - SATE)^2, j = 1, ..., \mathcal{N}_A$, by forming strata based on similarity on observed covariates. The variance of the stratified estimator, similarily defined, is then

$$V_{\mathbf{W}^S}(\widehat{\tau}) = \frac{1}{\mathcal{N}_S}\sum_{j=1}^{\mathcal{N}_S}(\widehat{\tau}_j - SATE)^2,$$

where $\mathbf{W}^S \subset \mathbf{W}$, is the set of all possible allocations under stratification, and $\mathcal{N}_S$ is the cardinality of the set $\mathbf{W}^S$.

As a means of obtaining an understanding of the efficiency gain of stratification in comparison to complete randomization, note that the conditional mean under an additive effect is equal to

$$Y_i(W_i) = \mathbf{d}_i'\beta{+}\tau W_i + \varepsilon_i, \tag{2}$$

where $\mathbf{d}_i$ is the ith row of the $N \times S$ incidence matrix $\mathbf{d}$ of the binary covariates classifying the S strata (see, e.g, Street and Street (1987)), and β is a $S \times 1$ vector of unknown parameters. Under the additive effect the mean difference estimator is equal to

$$\overline{Y}_1 - \overline{Y}_0 = (\overline{\mathbf{d}}_1 - \overline{\mathbf{d}}_0)'\beta + \tau + \overline{\varepsilon}_1 - \overline{\varepsilon}_0, \tag{3}$$

where $\overline{\mathbf{d}}_1 = \frac{1}{N_1}\sum_{i:W_i=1}^{N_1} \mathbf{d}_i$, $\overline{\mathbf{d}}_0 = \frac{1}{N_0}\sum_{i:W_i=0}^{N_0} \mathbf{d}_i$, $\overline{\varepsilon}_1 = \frac{1}{N_1}\sum_{i:W_i=1}^{N_1} \varepsilon_i$ and $\overline{\varepsilon}_0 = \frac{1}{N_0}\sum_{i:W_i=0}^{N_0} \varepsilon_i$. Let $\overline{\varepsilon}_{1j}$ and $\overline{\varepsilon}_{0j}$ be the mean of the residuals of the treated and controls in allocation j, respectively. Under complete randomization within each stratum, it holds that $\frac{1}{\mathcal{N}_S}\sum_{j=1}^{\mathcal{N}_S}(\overline{\varepsilon}_{1j} - \overline{\varepsilon}_{0j}) \equiv 0$. However, in line with the motivation of experimental design in general: In any specific experiment j, $\overline{\varepsilon}_{1j} - \overline{\varepsilon}_{0j} \neq 0$.

Let n^s be the number of individuals in stratum s, and let n_1^s and n_0^s denote the number of treated and controls units in stratum s, respectively. Then

$$\frac{N_1}{N}\overline{\mathbf{d}}_1 - \frac{N_0}{N}\overline{\mathbf{d}}_0 = \mathbf{1}_S(2n_1^s - n^s), \tag{4}$$

where $\mathbf{1}_S$ is a $S \times 1$ vector of ones. This means that when $n_1^s = n^s/2 = n_0^s$ for all s we have overall balance in the covariates, i.e,

$$(\overline{\mathbf{d}}_1 - \overline{\mathbf{d}}_0) \equiv \mathbf{0}.$$

Thus, when $\mathbf{W}^S$ is restricted to allocations where the treatment and control group are balanced in each stratum the OLS estimator on Equation (2) is asymptotically equivalent to the stratified estimator

$$\widehat{\tau}^{str} = \sum_{s=1}^{S} \frac{n^s}{N} \times \widehat{\tau}_s, \tag{5}$$

where $\widehat{\tau}_s$ is the mean differences estimator within each stratum. For a formal proof see Theorem 9.1 in (Imbens and Rubin, 2015).[1]

[1] Imbens and Rubin (2015) show that the OLS estimator converge to $\tau_w = \sum_{s=1}^S \omega_s\tau_s / \sum_{s=1}^S \omega_s$, where $\omega_s = \frac{n^s}{N} \times (\frac{n_1^s}{n^s}(1 - \frac{n_1^s}{n^s}))$ in our notation. As $n_1^s = n^s/2\ \forall s$, $\omega_s = \frac{n^s}{N} \times 0.25$.why $\tau_w = \sum_{s=1}^S \frac{n^s}{N}\tau_s$

Under the assumption of a homogeneous effect and the fact that ε_i and $\mathbf{d}_i$ are independent, $\sigma^2_Y = Var(Y_i|\mathbf{d}_i) + \sigma^2_\varepsilon$, where $\sigma^2_Y = Var(Y_i)$ and $\sigma^2_\varepsilon = Var(\varepsilon_i)$. Let $\mathbf{Y} = (Y_1, ..., Y_N)'$ and let R^2 be squared multiple correlation (or coefficient of determination) between $\mathbf{Y}$ and $\mathbf{d}$, then $\sigma^2_\varepsilon = (1 - R^2)\sigma^2_Y$. Letting $\widehat{\tau}^{CR}$ and $\widehat{\tau}^{S}$ be the estimators (see Equation (1)) under complete randomization and stratification, respectively, $Var(\widehat{\tau}^{S}) = (1 - R^2)Var(\widehat{\tau}^{CR})$, where the variance under complete randomization and homogeneous effect is $Var(\widehat{\tau}^{CR}) = \sigma^2_Y(1/N_1 + 1/N_0)$ (Neyman, 1923, 1990). The percent reduction in sampling variance (PRIV) of the treatment effect under stratification against complete randomization is thus

$$PRIV_S = 100 \times \frac{Var(\widehat{\tau}^{CR}) - Var(\widehat{\tau}^{s})}{Var(\widehat{\tau}^{CR})} = 100 \times R^2. \tag{6}$$

2.2 Rerandomization

Rerandomization is similar to stratification in the sense that certain allocations are excluded and randomization is conducted in a restricted set $\mathbf{W}^\varphi \subset \mathbf{W}$. As in complete randomization and the stratified design, the assignment mechanism is known and it is easy to implement the Fisher randomization test (FRT) (Fisher, 1935) in the set $\mathbf{W}^\varphi$. For the Mahalanobis-based rerandomization, discussed in detail in the next section, the asymptotic properties of the mean difference estimator is also known (Li et al., 2018).

Rerandomization is more general than stratification in the sense that it can easily incorporate different types of covariates. However, the set $\mathbf{W}^\varphi$ must be found by simulations. The main difference is the exclusion criterion. In a rerandomization design, the researcher must first decide on a covariate balance measure, and then a criterion to exclude allocations that are not sufficiently balanced on the covariates. Another difference is regarding the freedom of determining the overall fraction of treated and controls. With N even, one simply lets $N_1 = N_0$ by design in rerandomization. In stratification, complete randomization is performed within each stratum. As a consequence, the treatment and control groups within each stratum are balanced in expectation, but $N_1 = N_0$ is not guaranteed for any given treatment assignment as was seen in equation (4).

3 Mahalanobis-based rerandomization

Due to the well known properties of the Mahalanobis distance, we restrict the analyses to Mahalanobis-based rerandomization designs and discuss the most important results from Morgan and Rubin (2012, 2015).

Let $\mathbf{x}$ be the $N \times K_0$ matrix of fixed covariates, and, for simplicity, let $N_1 = N_0$. For a given allocation j the Mahalanobis distance between the covariate mean vectors of the units assigned to treatment and control, respectively, is defined as

$$M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}) = \frac{N}{4}(\overline{\mathbf{X}}_1^j - \overline{\mathbf{X}}_0^j)' cov(\mathbf{x})^{-1}(\overline{\mathbf{X}}_1^j - \overline{\mathbf{X}}_0^j),\ j = 1, ..., \mathcal{N}_A, \tag{7}$$

where $\overline{\mathbf{X}}_1^j - \overline{\mathbf{X}}_0^j$ is the difference in mean vectors which is a $K_0 \times 1$ stochastic vector as it depends on the random allocation. Morgan and Rubin (2012) suggest randomizing within the set $\mathbf{W}^\varphi$ fulfilling

$$\mathbf{W} \supset \mathbf{W}^\varphi : M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}) \leq a_0 \ \forall \ \mathbf{W}^j \in \mathbf{W}^\varphi,$$

where a_0 is a constant, called the *rerandomization criterion*. This means that instead of randomly choosing one of the $\mathcal{N}_A$ possible allocations, a set of allocations with Mahalanobis distances smaller than a_0 is considered in the final randomization. If the covariate means are normally distributed, $M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}) \sim \chi^2(K_0)$. This implies that a_0 can be indirectly determined by setting

$$p_{a_0} = \Pr(\chi^2(K_0) \leq a_0),$$

and that allocations can be sampled from any desired percentile of allocations with the smallest Mahalanobis distances. As the number of rerandomizations needed to draw an allocation that fulfills the criterion is geometrically distributed with expected value $1/p_{a_0}$, the expected number of rerandomizations before drawing a randomization fulfilling the criterion with, e.g., $p_{a_0} = 0.001$, is 1000.

Morgan and Rubin (2015) extend the idea of Morgan and Rubin (2012) by proposing the rerandomization to be done in tiers of covariates. The most important covariates should be placed in the first tier, often using a strict rerandomization criterion, and the second most important covariates in tier two with a slightly less restrictive criterion, etc. All allocations with large imbalances in the covariates of the first tier are excluded. In the second tier only the admissible allocations in the first tier are considered etc.. This means that the number of possible allocations decreases for each tier, until only allocations fulfilling the overal balance criteria remain in the final tier.

By placing all categorical covariates in the first tier and all continuous covariates in the second tier, rerandomization in tiers can be seen as special case of stratified rerandomization. The main difference is that rerandomization in tiers allows for any type of covariate in the first tier and has an explicit rerandomization criterion for this tier, making it possible to put covariates in tiers based on their believed relative importance rather than variable type.

3.1 Efficiency and the exact FRT

To facilitate the understanding of the relative efficiency of the different designs, we restrict the comparison to an additive treatment effect, or the Fisher null. The variance reduction of the treatment effect will be larger if the effect is heterogenous with respect to the included covariates for both stratification (see e.g. Imbens and Rubin, 2015) and with rerandomization (see Li, Ding and Rubin 2018). For this reason we cannot see this restriction as important for the relative efficiency also for asymptotic inferences under the Neyman null.

Consider the linear projection of the outcome on the covariates under an additive effect, that is,

$$Y_i(W_i) = \mathbf{x}_i'\gamma + \tau W_i + \varepsilon_i, \tag{8}$$

where $\mathbf{x}_i$ is the $1 \times K_0$ vector of covariates for unit i. Again, let R^2 be the squared multiple correlation between $\mathbf{Y}$ and $\mathbf{x}$. Under the assumption that ε_i is normally distributed, Morgan and Rubin (2012) show that PRIV of the treatment effect under Mahalanobis based rerandomization against complete randomization is equal to

$$PRIV_1 = 100 \times \frac{Var(\widehat{\tau}^{CR}) - Var(\widehat{\tau}^{RR})}{Var(\widehat{\tau}^{CR})} = 100 \times R^2(1-\nu_{_0}), 0 \leq \nu_0 \leq 1 \tag{9}$$

where

$$\nu_0 = \frac{\Pr(\chi^2(K_0+2) \leq a_0)}{\Pr(\chi^2(K_0) \leq a_0)}, \tag{10}$$

and $\widehat{\tau}^{RR}$ is the estimator (see Equation (1)) under Mahalanobis based rerandomization. This variance reduction can be compared to the variance reduction in the balanced stratified design based on the S strata which is equal to $100R^2$. As the probability of a Chi-square distributed variable to be less than a_0 is decreasing with the degrees of freedom, $\nu_0 \to 0$ as $a_0 \to 0$ (see Morgan and Rubin (2012) for details). The implication is, thus, that with a rerandomization criterion a_0 close to zero, such that $\nu_0 \simeq 0$, stratification and rerandomization should give similar improvements in efficiency.

Note that the assumption of normal errors, needed to derive (9) is for ε_i to be independent of $\mathbf{x}_i$. The error term is by definition uncorrelated with $\mathbf{x}$, however, equation (8) is not generally the conditional mean under the assumption of an additive effect as it is if (8) is saturated in $\mathbf{x}$. That is, the normality assumption of ε can be relaxed when $\mathbf{x}$ is a set of binary covariates and all their interactions.

With T tiers with covariates $\mathbf{x}_t$, for $t = 1, ...T$, such that $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$ and $\mathbf{x}_t$ an $N \times K_t$ matrix, Morgan and Rubin (2015) show that

$$PRIV_T = (1 - \nu_1)R^2_{1:1,y} + \sum_{t=2}^{T}(1 - \nu_t)(R_{1:t,y} - R_{1:(t-1),y}), \tag{11}$$

where $R^2_{1:t,y}$ is squared multiple correlation between $\mathbf{Y}$ and $\mathbf{x}_{1:t} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t)$ and a_t, is the rerandomization criterion for tier t, implying

$$\nu_{_t} = \frac{\Pr(\chi^2(K_t + 2) \leq a_t)}{\Pr(\chi^2(K_t) \leq a_t)}, t = 1, ..., T. \tag{12}$$

3.1.1 The FRT

As long as the randomization mechanism is known, FRT's can be used without making any assumptions (Fisher, 1935; Imbens and Rubin, 2015). This means that the FRT can be conducted in rerandomization designs as long as the randomization is restricted to any known set of allocations. Inference to a population, assuming random sampling to the experiment, is less straight forward with rerandomization as common test statistics are not longer asymptotically normal (Li et al., 2018). However, Schultzberg and Johansson (2019) show that using Mahalanobis-based rerandomization with a balance criterion close to zero, scaling with R^2, standard asymptotic inference can be used.

It is a good practice to formulate the risk of falsely rejecting the null hypothesis before any experiment is conducted. Restricting the focus of inference to the units of the sample or SATE, all stochastic variation in an experiment stems from the randomization, which means that the lowest level of risk in a FRT is determined by the number of possible treatment allocation in the randomized experiment. This means that if the lowest risk of making a false decision is chosen to be 5%, the number of possible allocations in a two-sided test needs to be at least 40. For 1%, the corresponding number of allocations is 200. That is, the lowest possible risk in a double sided exact test can be viewed as the resolution, r, of the exact p-value in the FRT. Let $\mathcal{N}_R$ be the cardinality of the set $\mathbf{W}^{\varphi}$, then $r = 2/\mathcal{N}_R$.

For most sample sizes, $\mathcal{N}_R$ is large for any rerandomization criteria. Therefore, a too large r is not usually an issue. A too large number of allocations, is on the other hand, a problem as it makes it intractable to calculate the exact p-value. This issue was pointed out by Athey and Imbens (2016), and used as an argument for recommending researchers to *not* use rerandomization. The rapid growth of combinations is, however, a potential problem for using exact tests under any randomized design, not only rerandomization. For example, in pair-wise stratification (each stratum is of size two), which implies that $S = N/2$ and that the number of allocations is

$$\mathcal{N}_S = \binom{2}{1}^{N/2}.$$

Already for $N = 60$, there are $1.073\,7 \times 10^9$ possible allocations which is hardly manageable for an average computer. One common solution is to approximate the exact p-value by Monte Carlo simulation. An alternative, suggested in Johansson and Schultzberg (2018), is to do an exact test on a limited set of allocations of 'optimal' allocations found by rerandomization.

Before discussing this algorithm, it is useful to first provide the intuition for the unbiasedness of SATE estimator after randomization (in the sets $\mathbf{W}, \mathbf{W}^S$ or $\mathbf{W}^{\varphi}$). The intuition is simply symmetry of the allocations. Under the sharp null or a homogeneous treatment effect, for a given random allocation $\mathbf{W}^j$ with an estimate $\widehat{\tau}^j$ the estimate of the 'mirror allocation' $\mathbf{W}^{j'} = \mathbf{1} - \mathbf{W}^j$ is simply $-\widehat{\tau}^j$. Thus, any set $\mathbf{W}^{\varphi}$ with cardinality two or larger containing only mirror allocations will be unbiased. As the Mahalanobis distance is affinely invariant it is also the case that $M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}) \equiv M(\mathbf{x}, \mathbf{1} - \mathbf{W}^j|\mathbf{W})$. This also gives intuition for why the SATE estimator is unbiased under Mahalanobis-based rerandomization. The implication for any algorithm with an aim of finding a set of allocations with cardinality $\mathcal{H}$ is that, by only sampling allocations from the first half of the lexicographically ordered allocations, only $\mathcal{H}/2$ allocations fulfilling the criterion must be found after which the corresponding mirror allocations are added to the set.

The implied smallest number of allocations needed for inference is then $\mathcal{H} = 2/r$. The set $\mathbf{W}^{\varphi *}$ is the optimal set of allocations, conditional on the Mahalanobis distance balance measure, if it contains the allocations with the $\mathcal{H}/2$ first unique order statistics of the Mahalanobis distance across all $\mathcal{N}_A$ allocations, i.e., $\{M^{[1]}, ..., M^{[H/2]}|\mathcal{N}_A\}$ as we include also the mirror allocations. For normally distributed covariates or large sample settings, this corresponds to using the rerandomization criterion $a_0 = M^{[H]}$ in $p_{a_0} = \Pr(\chi^2(K_0) \leq a_0)$, which implies that $\mathcal{H} = \mathcal{N}_A p_{a_0}$.

When N is large, the Mahalanobis distance cannot within a reasonable time limits be calculated over all of the $\mathcal{N}_A/2$ allocations due to the rapid growth of $\binom{N}{N_1}$. In this situation, the algorithm sequentially keeps the $\mathcal{H}/2$ allocations with the smallest Mahalanobis distance over subsets until a total number of $\mathcal{I}$ allocations from the original $\mathcal{N}_A/2$ has been drawn. When the algorithm is finished, the mirror allocations are included to give $\mathcal{H}$ in total. The final $\mathcal{H}$ allocations with the smallest Mahalanobis distances are then used as $\mathbf{W}^{\varphi *}$. This procedure differs from the procedure suggested in Morgan and Rubin (2012) in the sense that p_{a_0} is not set arbitrarily before the rerandomization. Instead p_{a_0} is a function of $\mathcal{I}$. The only restriction is computational time and the implied a_0 can be probabilistically bounded by setting $\mathcal{I}$ accordingly (Johansson and Schultzberg, 2018). Within each set $\mathcal{I}$, the Mahalanobis distances are Chi square-distributed, with the implication that the probability for an allocation to be accepted will depend on $\mathcal{I}$, that is $p_{a_0}^{\mathcal{I}} = \Pr(\chi^2(K_0) \leq a_0|\mathcal{I})$ where $\lim_{\mathcal{I}\to\mathcal{N}_A} p_{a_0}^{\mathcal{I}} = p_{a_0}$. In expectation $p_{a_0}^{\mathcal{I}} = \mathcal{H}/\mathcal{I}$.

4 The relationship between stratification and Mahalanobis-based rerandomization

Let $\mathbf{x}_1$ the $N \times K_1$ matrix of binary covariates and $\mathbf{x}$ be the $N \times K_0$ matrix containing the K_1 binary covariates in $\mathbf{x}_1$ and all their interactions, implying $K_0 = \sum_{i=1}^{K_1} \binom{K_1}{i} = S - 1$.

Theorem 4.1. *Let $\mathbf{W}^o$ be the set of allocations minimizing the conditional variance of the outcome under the sharp null. Then this set can be obtained as*

$$\mathbf{W}^o = \mathbf{W}^\varphi : M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}^\varphi) = \min_{\mathbf{W}^j \in \mathbf{W}} M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}) \Longleftrightarrow \mathbf{W}^j \in \mathbf{W}^\varphi$$

Proof. *As $\mathbf{x}$ contains only binary covariates including all interactions, the linear projection of $\mathbf{Y}$ on $\mathbf{x}$ (cf. Equation 8) is fully saturated and therefore equals the conditional expectation of $\mathbf{Y}$ under the sharp null.*

This, together with the fact that the Mahalanobis distance is affinely invariant, implies that minimizing the variance in $\mathbf{x}$ *will directly minimize* $Var(\mathbf{Y}|\mathbf{x})$ *for all* $\boldsymbol{\gamma}$ *in Equation 8.The if and only if part ensures that* $\mathbf{W}^{\varphi}$ *is not a proper subset of* $\mathbf{W}^{o}$*.* □

Theorem 4.1 implies that rerandomizing on only binary covariates and their interactions, using the strictest possible Mahalanobis criterion, always randomizes within the set of allocations with the minimum variance in the outcome under the sharp null.

Consider the case when either a stratification or rerandomization design will be chosen with the aim of estimating the SATE or PATE. We investigate the relation between these two designs by comparing rerandomization based on $\mathbf{x}$ and stratification. We start by studying the Mahalanobis-based randomization based on the $N \times S$ incidence matrix $\mathbf{d}$. Again, $\mathbf{d} = (\mathbf{d}_1, ..., \mathbf{d}_S)$ where $\mathbf{d}_s \; \forall \; s = 1, ..., S$, are $N \times 1$ vectors with one for units belonging to stratum s and zero otherwise. Thus

$$\mathbf{N}^S = \mathbf{d}'\mathbf{1}_N = (n^1, n^2, ...n^S)',$$

is the $S \times 1$ vector of the number of units in each stratum. As $n^S = N - \sum_{s=1}^{S-1} n^s$, we can drop the last column and let $\mathbf{q} = (\mathbf{d}_1, ..., \mathbf{d}_{S-1})$ with $cov(\mathbf{q})$ positive definite. Based on $\mathbf{q}$, the mean-difference vector of treated and control group for allocation j is

$$\overline{\mathbf{Q}}_1^j - \overline{\mathbf{Q}}_0^j = \frac{\mathbf{N}_{1j}^S}{N_1} - \frac{\mathbf{N}_{0j}^S}{N_0} = \frac{2\mathbf{N}_{1j}^S - \mathbf{N}^S}{N/2},$$

where $\mathbf{N}_{1j}^S = (n_{1j}^1, n_{1j}^2, ...n_{1j}^S)'$ and $\mathbf{N}_{0j}^S = (n_{0j}^1, n_{0j}^2, ...n_{0j}^S)'$ are the vectors of number of treated and controls in each stratum for allocation j, respectively. With the objective to obtain $N_1 = N_0$ and to have all strata be of even sample size, that is, $n^s \bmod 2 = 0$, for all s, $\overline{\mathbf{Q}}_1^j - \overline{\mathbf{Q}}_0^j \equiv \mathbf{0}$ for j where $2\mathbf{N}_{1j}^S = \mathbf{N}^S$. As $cov(\mathbf{q})$ is positive definite this implies

$$M(\mathbf{q}, \mathbf{W}^j|\mathbf{W}) \begin{cases} = 0 \;\forall\; \mathbf{W}^j \in \mathbf{W}^S \\ \\ \neq 0 \;\forall\; \mathbf{W}^j \notin \mathbf{W}^S \end{cases} . \tag{13}$$

This says that, Mahalanobis-based rerandomization with $a_0 = 0$, based on $\mathbf{q}$, randomizes within the same set of allocations as in stratification

Corollary 4.1. *If* $n^s \bmod 2 = 0 \;\forall\; s = 1, ..., S$*, it follows that*

$$\mathbf{W}^o = \mathbf{W}^S = \mathbf{W}^{\varphi} : M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}) = 0 \Longleftrightarrow \mathbf{W}^j \in \mathbf{W}^{\varphi}.$$

Proof. *Since the Mahalanobis distance is affinely invariant, it holds that*

$$M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}) = M(\mathbf{q}, \mathbf{W}^j|\mathbf{W}),$$

and the proof follows directly from Theorem 4.1 and from equation 13. The if and only if part ensures that $\mathbf{W}^{\varphi}$ *is not a proper subset of* $\mathbf{W}^S$*.* □

A detailed example illustrating theorem 4.1 and Corollary 4.1 is given in the Appendix. Corollary 4.1 shows that, in the case when all strata have even sample size, stratification gives exactly the same design as rerandomization with the rerandomization criterion zero.[2] The rerandomization criterion recreating the stratified design is in accordance with Equations 9, and 6 on the PRIV for the two designs. That is, if the two designs are identical they should give identical variance reductions and this is only the case if $\nu_0 = 0$, which is obtained by letting $a_0 = 0$. In other words, randomization on continuous covariates with close-to-zero criteria corresponds approximately to stratifying on these covariates but without any information loss due to discretization.

When $n^{s'} \bmod 2 \neq 0$ for $s' \in (1, ...S)$, and with the aim of letting $N_1 = N_0$, the researcher would randomly assign $n^{s'}/2 - \frac{1}{2}$ or $n_1^{s'}/2 + \frac{1}{2}$ to be treated. In the first case $n_1^{s'} - n_0^{s'} = -1$ and in the second $n_1^{s'} - n_0^{s'} = 1$. In this situation $\mathbf{W}^S$ cannot be shown to be equal to $\mathbf{W}^o$. Restricting randomization to $\mathbf{W}^S$ removes imbalances within each stratum but does not guarantee that the covariates are balanced over the full sample. This is not a problem for the SATE estimator (Equation 5) as the within strata estimators are all unbiased. With the set $\mathbf{W}^o$ we are guaranteed to obtain an estimator with an overall minimum variance, under the null, which one is more efficient will depend on the context.

4.1 Stratified rerandomization and rerandomization: efficiency and computational time

The recommendation to 'Block on what you can and rerandomize on what you cannot' was originally given at the time when the computational cost was high. As the cost for computations has fallen, it is of interest to compare the performance of the SATE estimator under stratified rerandomization and

[2] Note that this equivalence is also true with heterogenous treatment effects, thus the fact the we restrict the comparison to the situation with an additive effect is no restriction when all strata are of even size.

rerandomization. Since Mahalanobis-based rerandomization on categorical covariates with interactions can be made equivalent to stratification, it follows that the relative performance can be investigated using the framework of rerandomization in tiers. As strict equivalence is possible only under Corollary 4.1; we restrict the comparison to the completely balanced experiment accordingly.

Let $\mathbf{x}_1$ be the set of K_1 binary covariates (including all interactions), $\mathbf{x}_2$ be a set of K_2 continuous covariates, and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ be the $N \times K_0$ fixed matrix. The PRIV for the SATE estimator under rerandomization and rerandomization in tiers (used for stratified rerandomization) are given in Equations 9 and 11, but repeated here for convenience:

$$PRIV_1 = 100 \times (1 - \nu_{_0})R^2$$

and

$$PRIV_2 = (1 - \nu_1)R_1^2 + (1 - \nu_2)(R^2 - R_1^2),$$

where R_1^2 is squared multiple correlation between $\mathbf{Y}$ and $\mathbf{x}_1$, and

$$\nu_t = \frac{Pr(\chi^2(K_t + 2) \leq a_t)}{Pr(\chi^2(K_t) \leq a_t)} \quad , \; t = 0, 1, 2. \tag{14}$$

Due to the balanced strata, rerandomizing on $\mathbf{x}_1$ with the zero criteria (stratifying) implies $a_1 = p_{a_1} = \nu_1 = 0$. That is, for stratified rerandomization, all the variance in $\widehat{\tau}$ from $\mathbf{x}_1$ is controlled for and the number of remaining allocations is $\mathcal{N}_S$.

As a tool for comparing the two designs, we use the ratio of the PRIV's of the two designs

$$\begin{aligned} RPRIV &= \frac{PRIV_2}{PRIV_1} \\ &= \frac{(1 - \nu_2)R^2 + \nu_2 R_{_1}^2}{(1 - \nu_0)R^2} \\ &= \frac{1 - \nu_2}{1 - \nu_0} + \frac{\nu_2}{1 - \nu_0}\frac{R_1^2}{R^2}. \end{aligned} \tag{15}$$

From this equation, it is clear that the two methods for any $R_{_1}^2$, $R^2 > 0$, and with $\mathcal{H}$ fixed, gives the same variance reduction asymptotically as

$$\frac{1 - \nu_2}{1 - \nu_0} + \frac{\nu_2}{1 - \nu_0}\frac{R_1^2}{R^2} \xrightarrow{p} 1 \text{ as } N \to \infty. \tag{16}$$

This follows from the fact that both ν_0 and ν_2 converge to zero with N for a fixed $\mathcal{H}$.

To get an understanding of the behavior and limitations of the two designs, we evaluate the relative efficiency of the two procedures, for the more realistic situation with N fixed in the next section.

4.1.1 The relative efficiency for fixed N

Given a level of resolution r in the exact p-value of a two-sided hypothesis test, the minimum number of allocations that must remain after stratification is $\mathcal{H} = 2/r$. Under the assumption

$$M(\mathbf{x}_2, \mathbf{W}^j|\mathbf{W}^S) \sim \chi^2(K_2),$$

it follows that $\mathcal{H} = \mathcal{N}_S p_{a_2}$, which implies that rerandomization criterion in the second tier of the stratified rerandomization is bounded

$$\frac{\mathcal{H}}{\mathcal{N}_S} \leq p_{a_2} \leq 1. \tag{17}$$

Thus, when $\mathcal{N}_S \simeq \mathcal{H}$, $p_{a_2} = \nu_{t_2} \simeq 1$. Putting it differently, if only a few allocations can be discarded under the rerandomization based on the second tier, one cannot expect substantial variance reductions in the covariates in the second tier. This suggests that, if $\mathcal{N}_S$ is close to $\mathcal{H}$, and $\mathbf{x}_2$ is believed to explain a lot of variation in the outcome, it might be a bad idea to stratify first.

In order for the stratified rerandomization to have larger percentage reduction in variance than rerandomization, it must hold that

$$\frac{1-\nu_2}{1-\nu_0} + \frac{(R_1^2/R^2)\nu_2}{1-\nu_0} > 1, 0 \leq \nu_0, \nu_2 \leq 1.$$

$$\iff$$

$$(1 - \frac{R_1^2}{R^2}) < \frac{\nu_0}{\nu_2}, 0 \leq \nu_0, \nu_2 \leq 1. \tag{18}$$

As $0 \leq R_1^2/R^2 \leq 1$, this means that whenever $\frac{\nu_0}{\nu_2} > 1$ stratified rerandomization is more efficient that rerandomization in expectation.

Let $Q(r, K)$ be the quantile function of the Chi-square distribution such that $Q(p_{a_t}, K_t) = a_t$, for $t = 0, 1, 2$. For a given r, the optimal criteria for the stratified rerandomization and rerandomization are $a_2 = Q(r, K_2)$ and $a_0 = Q(r, K_0)$, respectively. With $r = \mathcal{H}/\mathcal{N}_S = \mathcal{H}/\mathcal{N}_A$ for stratified rerandomization and rerandomization, respectively, we get

$$\nu_0/\nu_2 = \frac{\Pr\left(\chi^2(K_0+2) \leq Q(\frac{\mathcal{H}}{\mathcal{N}_A}, K_0)\right)}{\Pr\left(\chi^2(K_0) \leq Q(\frac{\mathcal{H}}{\mathcal{N}_A}, K_0)\right)} \Big/ \frac{\Pr\left(\chi^2(K_2+2) \leq Q(\frac{\mathcal{H}}{\mathcal{N}_S}, K_2)\right)}{\Pr\left(\chi^2(K_2) \leq Q(\frac{\mathcal{H}}{\mathcal{N}_S}, K_2)\right)}. \tag{19}$$

From this expression it becomes clear that the relative efficiency of the SATE estimator under stratified rerandomization and rerandomization depends on the degrees of freedom in the Chi-square distribution of the Mahalanobis distances of the allocations and the number of possible allocations in the rerandomization step of the two designs (i.e. $\mathcal{N}_A$ and $\mathcal{N}_S$). The efficiency of the rerandomization is decreasing in the degrees of freedom and increasing in the number of allocations. The stratification reduces both the degrees of freedom (from K_0 to K_2) and the number of allocations (from $\mathcal{N}_A$ to $\mathcal{N}_S$) in the second tier rerandomization.

4.1.1.1 Variance-reduction evaluation for pairwise stratification To illustrate how the relative efficiency of stratified rerandomization compared to rerandomization depends on R_1^2/R^2, we consider the case with $N/2$ strata of size 2 for $N = 12, 14, 16, 18, 20, 22, 24$, and one continuous covariate. With $K_1 = N/2 - 1$ binary and one continuous covariate the total number of covariates is $K_0 = N/2$. We vary the lowest level of risk to be 5%,1% and 0.1%, which means that $\mathcal{H} = 40, 200$ and 2000.

From Figure 1 we can see that with a 5% level of risk, rerandomization is preferable to stratified reranomization, i.e. has larger expected PRIV, only when $R_1^2/R^2 \leq 0.20$ and $N = 12$. When the level of risk is set to 1%, $\mathcal{N}_S < \mathcal{H}$ for $N \leq 14$, which implies that stratified rerandomization is not an option for $N \leq 14$. For $N = 16$, rerandomization is preferable to stratified rerandomization when $R_1^2/R^2 \leq 0.53$. For larger experiments, stratified rerandomization is preferable for all R_1^2/R^2. With the level of risk set to 0.1%, $\mathcal{N}_S < \mathcal{H}$ for $N \leq 20$, which means that stratified rerandomization is not an option in these cases. For $N = 22$ rerandomization is preferable to stratified rerandomization when $R_1^2/R^2 \leq 0.78$, and for $N = 24$, stratified rerandomization is preferable for all R_1^2/R^2.

Figure 1 clearly shows the trade off between the number of remaining allocations after stratification and the ‘cost’ of increasing the degrees of freedom in the Chi-square distribution of the Mahalanobis distance for the Mahalanobis-based rerandomization. By first stratifying, the number of degrees of freedom in the Chi-square distribution of the Mahalanobis distances in the remaining rerandomization is reduced from $N/2$ to one. Lower degrees of freedom means less diffused Mahalanobis distances which provides better precision in the rerandomization in the second tier. On the other hand, if there are few allocations left after the stratification, the rerandomization on the remaining covariates becomes restricted, as $\mathcal{H}$ allocation must be kept for inference based on the pre experiment choice of the maximum level of risk. It is important to understand that the pairwise stratification is a 'worst case' scenario for rerandomization as the difference between K_0 and K_2 is maximized for each N in this design. If the number of binary covariates is fixed over N, stratified rerandomization and rerandomization would give more similar PRIV in accordance with Equation 16.

If there is no a priori information of the relative importance of $\mathbf{x}_1$ and $\mathbf{x}_2$ in explaining the outcome, it is reasonable to assume that $\mathbf{x}_1$ explains as much as $\mathbf{x}_2$, i.e., $R^2_{_1}/R^2 = 0.5$. With $R^2_{_1}/R^2 = 0.5$ the Mahalanobis-based rerandomization is in expectation more efficient than Mahalanobis-based rerandomization on $\mathbf{x}$ whenever $\nu_0/\nu_2 > 0.5$ (see Equation 18). For $N = 16$ we saw that exact inference is possible under both stratified rerandomization and rerandomization with $\alpha = 1\%$. As rerandomization was more efficient for all $R^2_{_1}/R^2 \leq 0.53$ this implies that rerandomization is preferable with an agnostic assumption on the importance of the two types of covariates dependence with the outcome in this case. However, as these results build on asymptotic properties they should be interpreted with caution. Finite sample properties in this case will be presented in Section 5.

4.1.2 Computational time as a function of the number of covariates

In the previous section it was shown that, for sample sizes smaller than 24, the number of remaining allocations after stratification on $\mathbf{x}_1$ restricts the subsequent rerandomizations on $\mathbf{x}_2$ so much that rerandomization on $\mathbf{x}$ may be preferable. However, this restriction is only a concern for these small experiments.

Even with moderately large experiments, it is intractable to go trough all allocations (also after stratification) wherefore it is important to take the computational time it takes to find H acceptable allocation into account when comparing designs using rerandomization. The expected number of considered allocations needed to find one acceptable allocation for a given criterion a_t in tier t using rerandomization is $1/p_{a_t}$. This means that, on average, $\mathcal{H}/p_{a_t}$ allocations need to be sampled to obtain $\mathcal{H}$ acceptable allocations. For $\mathcal{H} = 40$, this means that with $p_{a_t} = 0.00001$, 4 million allocations needs to be sampled to obtain 40 allocations that fulfills the criterion on average. However, it is ν_t and not p_{a_t} that determines the efficiency gain from the design as, for a given R^2, the PRIV is only a function of ν_t. As ν_t increases with K_t, the variance reduction from the rerandomization decreases in K_t for a fixed p_{a_t}. This means that in order to achieve the same variance reduction from a large set as for a small set of covariates, the criterion p_{a_t} needs to be reduced, and, therefore the number of sampled allocations needs to be increased. An alternative to searching for the $\mathcal{H}$ optimal allocations among all $\mathcal{N}_A$ allocations, the procedure discussed in section 3.1.1 can be used. There, the idea was to let $p^{\mathcal{I}}_{a_t} = \Pr(\chi^2(K_t) \leq a_t|\mathcal{I})$, where $\mathcal{I}$ is the number of allocations in a randomly drawn subset of $\mathbf{W}$, $\mathbf{W}^s$ or $\mathbf{W}^\varphi$. In expectation, $p^{\mathcal{I}}_{a_t} = \mathcal{H}/\mathcal{I}$. Using this procedure, we can analyse the computational time for the hypothetical experiment with an increasing N by calculating the PRIV for different number of covariates (K_t) for a fixed $\mathcal{H}$, letting the number of considered allocations $\mathcal{I}$ increase or, in other words, by decreasing $p^{\mathcal{I}}_{at}$.

Figure 2 displays the expected PRIV in the rerandomization covariates, where $\mathcal{I}$ is varied between 20 thousand and 40 billions for $K_t = 1, ..., 32$. It is apparent from the figure that the relation between the expected variance reduction and the number of sampled allocations is non-linear in K_t. For $K_t \leq 3$ the number of considered allocations that remove all of the variation is small. For K_t larger than five, the number of considered allocations needed to reduce all the variation is extreme. For $K_t \geq 11$ not even 40 billion allocations is enough to get 99% reduction in PRIV. This illustrates the potential benefits of reducing the number of degrees of freedom in the second tier by using stratified rerandomization, and why rerandomizing in tiers is a good idea in general.

These results have great importance for the comparison of stratified rerandomization and rerandomization. That is, since $K_2 < K_0$ by construction, the computational time of the rerandomization step in stratified rerandomization may be substantially smaller than in rerandomization, especially if K_1 is large.

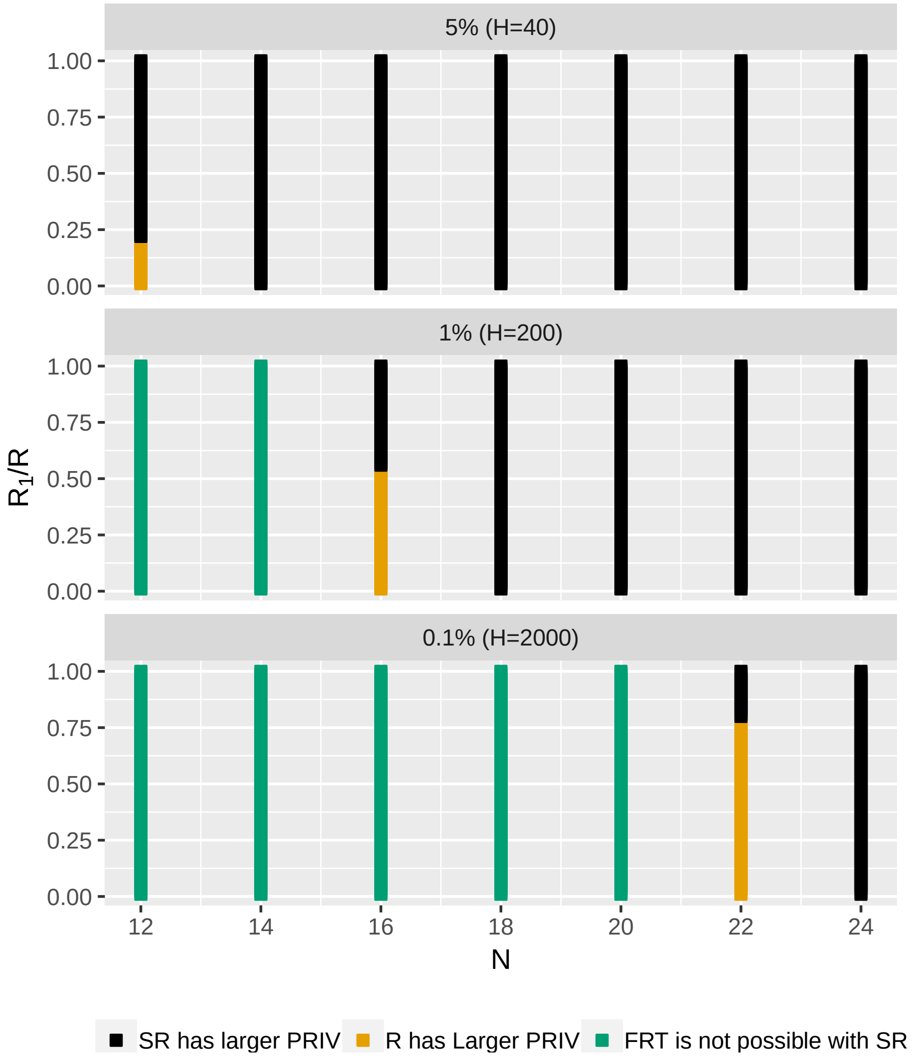

Figure 1: Comparison of expected PRIV under optimal stratified rerandomization (SR) and rerandomization (R) designs for sample sizes beteen 12 and 22, for a test of level 5%, 1% and 0.1%.

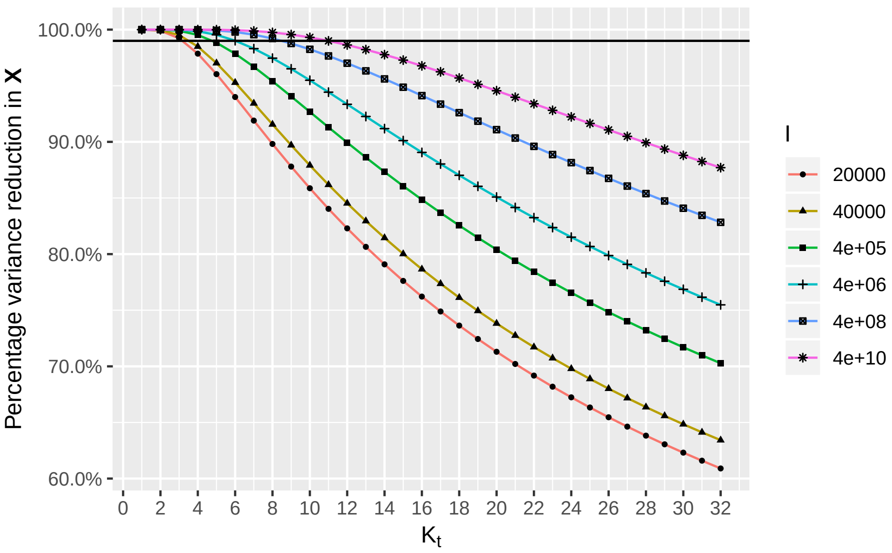

Figure 2: Expected percentage variance reduction in the rerandomization covariates for various numbers of considered allocations as a function of degrees of freedom in the rerandomization. Values above the solid line implies that more than 99% of the variation in the covariates is removed by the rerandomization.

4.1.2.1 A perspective on computational time The seemingly large numbers of considered allocations in Figure 2 are in fact a very small part of the total number of allocations for moderately large samples. For example, if the sample size is 50, 4×10^{10} constitutes $100(4 \times 10^{10}/\binom{50}{25}) = 0.03\%$ of all possible allocations.

The implication of the growth of combinations is difficult to comprehend but to give a perspective, we exemplify by predicting the computational time for finding the globally best allocation for $N = 50, 60, 70, 80, 90, 100$. A decently fast software implementation can go through around 1,000,000 allocations per second [3], depending on the sample size and the number of covariates. Table 1 displays the estimated time consumption for calculating the Mahalanobis distance for all possible allocations. Already for $N = 50$ it takes 4 years, and for $N = 60$ the time consumption is more than 3000 years, indicating the complexity of finding the allocations with the $\mathcal{H}$ globally smallest Mahalanobis distances. This problem can be fully parallelized, and there are likely software implementation that can speed up the calculations by some factor. However, given the rapid increase, going through all allocations for samples sizes such as $N = 100$ is still completely intractable with current hardware and software.

Table 1: Estimated time consumption (years) for calculating the Mahalanobis distance for all $\binom{N}{N/2}$ allocations, assuming that 1,000,000 allocations can be considered each second.

N	50	60	70	80	90	100
Expected time (years)	4	3750	3.56×10^6	3.41×10^9	3.29×10^{12}	3.20×10^{15}

5 Monte Carlo simulations

Three Monte Carlo (MC) simulations are conducted to study the power of the SATE estimator in an exact FRT under stratification (S), stratified rerandomization (SR), Mahalanobis based rerandomization (MR_0), and complete randomization (C).

Data are generated as

$$Y_i(0) = \mathbf{x}_{1i}\boldsymbol{\beta}_1 + \beta_2 x_{2i} + \epsilon_i \tag{20}$$

[3] The figure 1,000,000 comes from timing an implementation in the programming language Julia v1.1.0. The corresponding figure for an implementation in base R v3.5.3 is around 30,000.

where $\mathbf{x}_{1i}$ is the $K_1 \times 1$ vector of binary covariates including their interactions, and x_{2i} and ϵ_i are both i.i.d. exponentially distributed variables: $x_{2i} \sim \exp(\lambda_1)$ and $\epsilon_i \sim \exp(\lambda_2)$. The alternative for which the power is studied is set to $Y_i(1) = Y_i(0) + 0.6$. The parameters are chosen such that in expectation $R^2 = 50\%$ in all settings. The number of replications in each cell is 5,000 and the exact p-value is defined as

$$\pi_{mj} = \frac{1}{\mathcal{H}} \sum_{r=1}^{\mathcal{H}} \mathbf{1}(|\widehat{\tau}(\mathbf{W}_r^*, \mathbf{Y}_j)| \geq |\widehat{\tau}^j|), j = 1, ..., \mathcal{H},\ m = 1, ...5,000, \tag{21}$$

were $\widehat{\tau}(\mathbf{W}_r^*, \mathbf{Y}_j)$ is the distribution of estimates over all allocations r given allocation j (for any of the sets $\mathbf{W}^* = \mathbf{W}, \mathbf{W}^* = \mathbf{W}^S$ or $\mathbf{W}^* = \mathbf{W}^{\varphi *}$) in replication m. The power in replicate m is calculated as

$$P_m = \frac{1}{\mathcal{H}} \sum_{j=1}^{\mathcal{H}} \mathbf{1}(\pi_{mj} \leq 0.05). \tag{22}$$

The designs of the first two Monte Carlo simulations are stimulated by the theoretical results in Section 4.1.1 and 4.1.2 where $K_1 = N/2 - 1$. The first case (Section 5.1) studies if the stratification on $\mathbf{x}_1$ in the stratified rerandomization may prohibit more efficient inference that could be achieved by rerandomization on $\mathbf{x}$ for $N = 16$. The second simulation (Section 5.2) considers the situation when $N = 64$ and $\beta_2 = 0$. Because of the large number of degrees of freedom in the Chi-square distribution and because x_2 does not contribute to this design, it can be seen as a 'worst case' scenario for the rerandomization design. Finally, the third simulation (Section 5.3) compares the designs in a moderately large sample size, $N = 28$, with $K_1 = K_2 = 1$ and shows how the power is affected by increasing $\mathcal{I}$, or decrasing $p_{a_0}^{\mathcal{I}}$.

We have also conducted the same MC simulations with $x_2 \sim N(0, 0.25)$, $\epsilon_i \sim N(0, 0.5)$ and with heterogenous effects with a mean of 0.6. The results from these Monte Carlo simulations are very similar to the ones discussed below and can be obtained upon request.

5.1 MC simulation 1

The study aims at examining the power of the FRT when $\mathcal{H}$ approaches $\mathcal{N}_S$ for $N = 16$ and $K_1 = 7$. Here $\lambda_1 = \lambda_2 = 2$, $\boldsymbol{\beta}_1 = (\sqrt{\rho\zeta_1}, ..., \sqrt{\rho\zeta_1})'$ and $\beta_2 = \sqrt{(1-\rho)\zeta_2}$, where ζ_1 and ζ_2 are chosen such that $\boldsymbol{\beta}_1' cov(\mathbf{X}_1)\boldsymbol{\beta}_1 = \zeta_2^2 Var(x_2) = 0.5 \times Var(\epsilon)$. We let $\rho = R_1^2/R^2$ take the values $0, 0.5$, and 1, which correspond to the binary covariates having no effect on the outcome, the binary and continuous covariates having equal effect on the outcome, and, the continuous covariate having no effect on the outcome. $\mathcal{H}$ is varied as 200, 240, and 256. Note that when $\mathcal{H} = \mathcal{N}_S = 256$, the stratified rerandomization is equal to stratification as no allocations can be excluded in the rerandomization step.

The performance under the experimental designs MR_0 and SR are considered. We denoted the Mahalanobis based rerandomization using $\mathbf{x} = (\mathbf{x}_1, x_2)$ MR_0 as we in this simulation also consider an additional one-step rerandomization design based on only the three main covariates (i.e., the interactions among the binary covariates are excluded) which we denote MR_1. MR_1 is considered to illustrate the flexibility with rerandomization as opposed to stratification; the interactions can conveniently be included or excluded based on prior beliefs of their importance. MR_1 is a reasonable design if no a priori information about the covariates relative importance is available. That is, it can be argued that it is not reasonable to include all interactions of a set of covariates solely because they are binary. Note that the interaction terms are in fact informative when $\rho > 0$ since their coefficients are non-zero, implying that this setting does not favour MR_1 by construction. For each sample, the globally best $\mathcal{H}$ allocations according to each design are chosen.

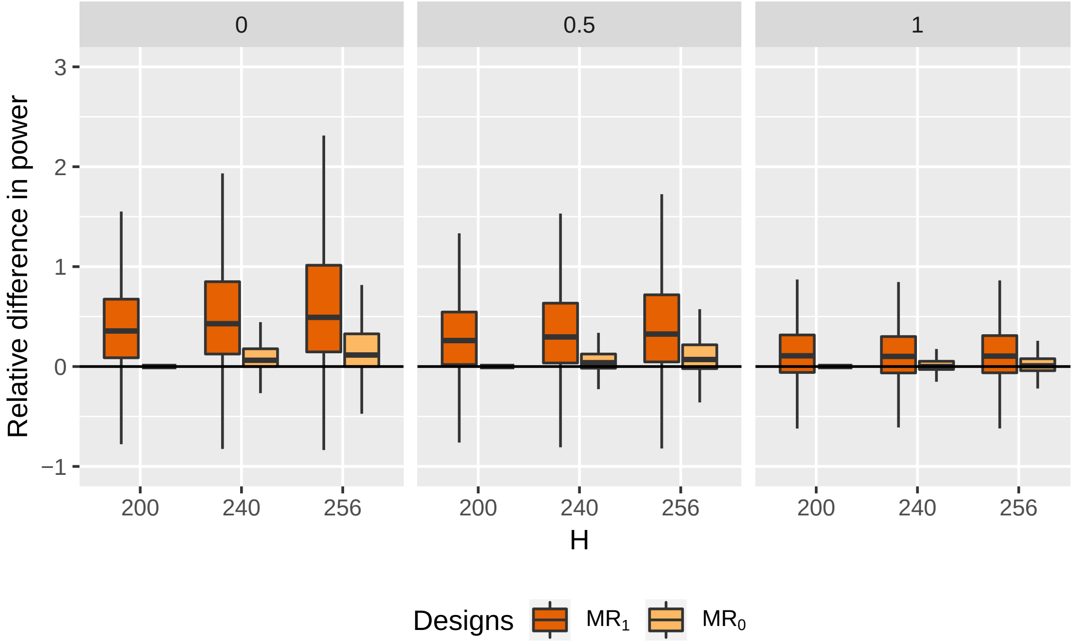

Figure 3: The distribution, over 5,000 replicates, of the relative difference in power for rerandomization as compared to stratified rerandomization with the full set of covariates and a restricted set (MR_0 and MR_1), as $\mathcal{H}$ approaches $\mathcal{N}_S = 2^8 = 256$. The panels display, from left to right, $\rho = 0$, $\rho = 0.5$, and $\rho = 1$, respectively.

Let $P_m(\text{SR}), P_m(\text{MR}_0)$ and $P_m(\text{MR}_1)$ be the the estimated power in replicate m (defined in 22) of the three designs. Figure 3 displays the distributions, as box plots, of the relative difference in power of rerandomization compared to stratified rerandomization, defined as

$$\text{RD}_m(\text{MR}_0) = \frac{P_m(\text{MR}_0) - P_m(\text{SR})}{P_m(\text{SR})} \text{ and } \text{RD}_m(\text{MR}_1) = \frac{P_m(\text{MR}_1) - P_m(\text{SR})}{P_m(\text{SR})}.$$

It is clear that in the settings where the binary covariates affect the outcome (panel 2 and 3), the power of MR_0 is exactly the same as SR for $\mathcal{H} = 200$, and larger when $\mathcal{H}$ comes close to $\mathcal{N}_S$. MR_1 has higher power both when $\rho = 0$ and, perhaps more suprisingly, when $\rho = 0.5$. When the continuous covariate has no effect on the outcome, MR_0 gives the same power as SR on average. Surprisingly, MR_1 perform slightly better than SR on average, also in this setting. Clearly, the information about $\mathbf{Y}$ sacrificed when excluding the interaction terms to lower the degrees of freedom in the rerandomization step, increases efficiency in this special case.

This small simulation study shows that the theoretical results of Section 4.1.1 translate quite well to small samples and indicates that the theoretical results are robust to violations of the normality assumptions. If the sample size is small, the stratification reduces the number of allocations to a small set. It is then important to think about the covariates relative importance before deciding on a design.

5.2 MC simulation 2

Here $N = 64$, $\beta_2 = 0$ and $\mathbf{x}_{i1}$ is a set of independent binary covariates and their interactions which imply 32 strata of size 2. λ_1 and λ_2 is chosen to obtain $var(x_2) = 0.4$ and $var(\epsilon) = 0.8$, respectively. Here $\boldsymbol{\beta}_1 = (\sqrt{\zeta_1}, ..., \sqrt{\zeta_1})'$ where ζ_1 is chosen such that $\boldsymbol{\beta}_1' cov(\mathbf{x}_1)\boldsymbol{\beta}_1 = Var(\epsilon)$. We set $\mathcal{H} = 40$, and vary $\mathcal{I} = 300, 1,000, 10,000, 20,000$, which means that $p_{a_0}^{\mathcal{I}}$ is in the range 0.133 (= 40/300) to 0.002 (= 40/20000). For CR and S, $\mathcal{H}$ allocations are randomly drawn from $\mathbf{W}$ and $\mathbf{W}^S$, respectively. For SR, $\mathcal{I}$ allocations are randomly drawn from $\mathbf{W}^S$ and the $\mathcal{H}$ allocations with the smallest Mahalanobis distance on x_2 within this set are chosen. For MR_0, $\mathcal{I}$ allocations are randomly drawn from $\mathbf{W}$ and the $\mathcal{H}$ allocations with the smallest Mahalanobis distance on $\mathbf{x}$ are chosen.

The maximum number of considered allocations in each replication, $\mathcal{I} = 20,000$, is very far from $\mathcal{N}_S = 2^{32} = 4.295 \times 10^9$. This means that there is no restriction on $\mathcal{N}^S$ in the rerandomization on x_2

in the second stage as was the case in Section 4.1.1. Instead, due to the large degrees of freedom in the Chi-square distribution in the rerandomization, and by the fact that $\beta_1 = 0$, this Monte Carlo simulation illustrates the potential problems with rerandomization in comparision to stratification and stratified rerandomization when $\mathbf{x}_1$ contains a large number of covariates. The degrees of freedom in the MR$_0$ design is 32 (31 binary, 1 continuous). This implies that, even though SR and MR$_0$ should give approximately equal designs asymptotically (see Equation 16) $p_{a_0}^{30,000} = \Pr(\chi^2(K_0) \leq a_0|30,000)$ is far from $\lim_{\mathcal{I}\to\mathcal{N}_A} p_{a_0}^{\mathcal{I}} = p_{a_0}$. Figure 4 displays the distibution (box plots) of the estimated power across replications in the FRT for the

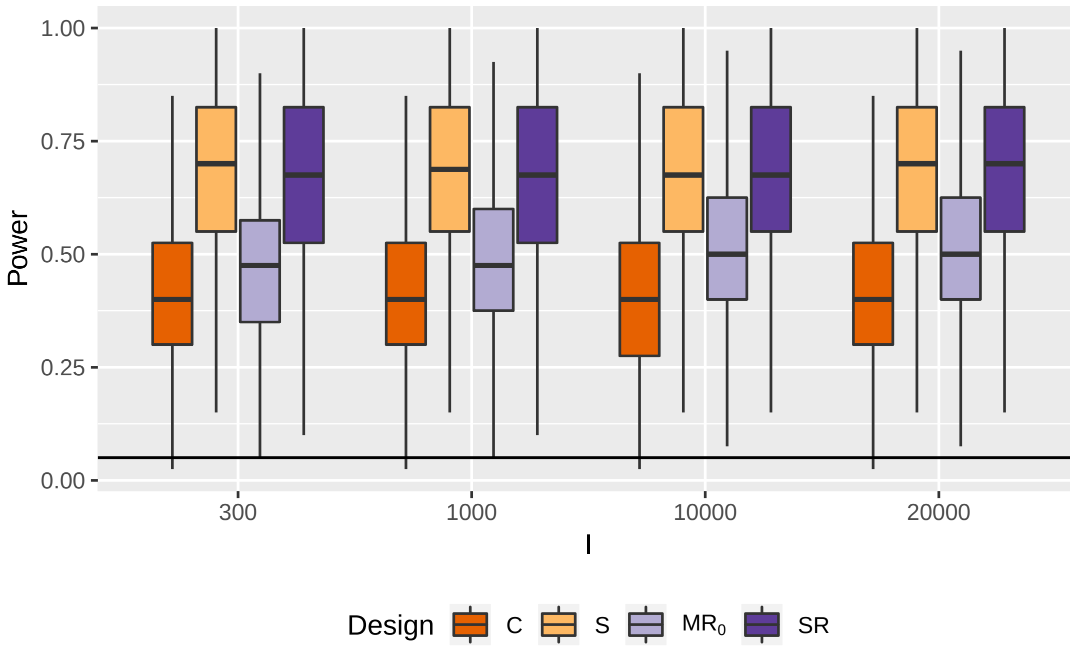

Figure 4: The distribution, over 5,000 replicates, of the power of complete randomization (C), stratification (S), rerandomization (MR$_0$) and stratified rerandomization (SR) as a functions of number of considered allocations, $\mathcal{I}$.

four designs; $P_m(\text{C}), P_m(\text{S}), P_m(\text{SR})$ and $P_m(\text{MR}_0)$. As expected, stratification and stratified randomization achieves the full efficiency gain already with $\mathcal{I} = 300$, and the stratified randomization does not improve by rerandomizing on x_2 but is not distorted either. The rerandomization design do improves slowly with $\mathcal{I}$. However, as expected, the improvement is hardly visible across the span of $\mathcal{I}$ presented here. Table 2 displays the empirical variance of SATE for each design over $\mathcal{I}$, averaged over the replications. As expected,

Table 2: Empricial variance of SATE estimator averaged over replications for complete randomization (C), rerandomization (MR_0), stratification (S), and stratifed rerandomization (SR) as a functions of number of considered allocations, $\mathcal{I}$.

$\mathcal{I}$	C	S	MR_0	SR
300	0.099	0.050	0.086	0.050
1000	0.099	0.050	0.081	0.050
10,000	0.098	0.050	0.074	0.050
20,000	0.099	0.049	0.074	0.050

the only design for which the variance decreases with $\mathcal{I}$ is MR_0. That is, for S, and SR the full maximum variance reduction of the SATE estimator, in this case $100 \times R^2 = 50\%$, is as expected achieved for all $\mathcal{I}$, whereas MR_0 only achieve 25% for $\mathcal{I} = 20,000$. This is in line with Figure 2, from which we expect that with $K_0 = 32$ and $\mathcal{I} = 20,000$ we should have on average 60% variance reduction in $\mathbf{x}$, i.e., $\nu_0 = 0.6$, implying $PRIV_1 = 20\%$, i.e. $PRIV_1 = 100 \times R^2(1 - \nu_0) = 100 \times 0.5 \times 0.4$.

5.3 MC simulation 3

Here $\beta_1 = \sqrt{\rho 2}$ and $\beta_2 = \sqrt{(1-\rho)2}$, where ρ is varied as 1/3, 1/2 and 2/3. This means that the binary covariate is half, equally and twice as as important in explaining the outcome as the continuous. Furthermore, we let $\lambda_1 = 2$, $\lambda_2 = \sqrt{2}$, $N \equiv 28$, and $\mathcal{H} \equiv 40$. We vary $\mathcal{I}$ by letting $\mathcal{I}$ =60, 100, 500 and 800, which means that $p_{a_0}^{\mathcal{I}}$ varies in the range 0.67 (= 40/60) to 0.05 (= 40/800). The sampling of the $\mathcal{I}$ allocations is performed as in section (5.2).

Figure 5 displays the distribution (box plots) of $P_m(\mathrm{C}), P_m(\mathrm{S}), P_m(\mathrm{SR})$ and $P_m(\mathrm{MR}_0)$ across $\mathcal{I}$. As expected, the stratified design does not improve by increasing $\mathcal{I}$. All gain in efficiency from stratification is immediate since only allocations allowed under stratification are allowed. The stratified rerandomization is always better or equally good as rerandomization. This is expected as in the stratified rerandomization only allocations allowed under stratification on x_1 are allowed, and therefore it immediately starts balancing on x_2. When $\mathcal{I}$ becomes larger there is no difference between the stratified rerandomization and the rerandomi-

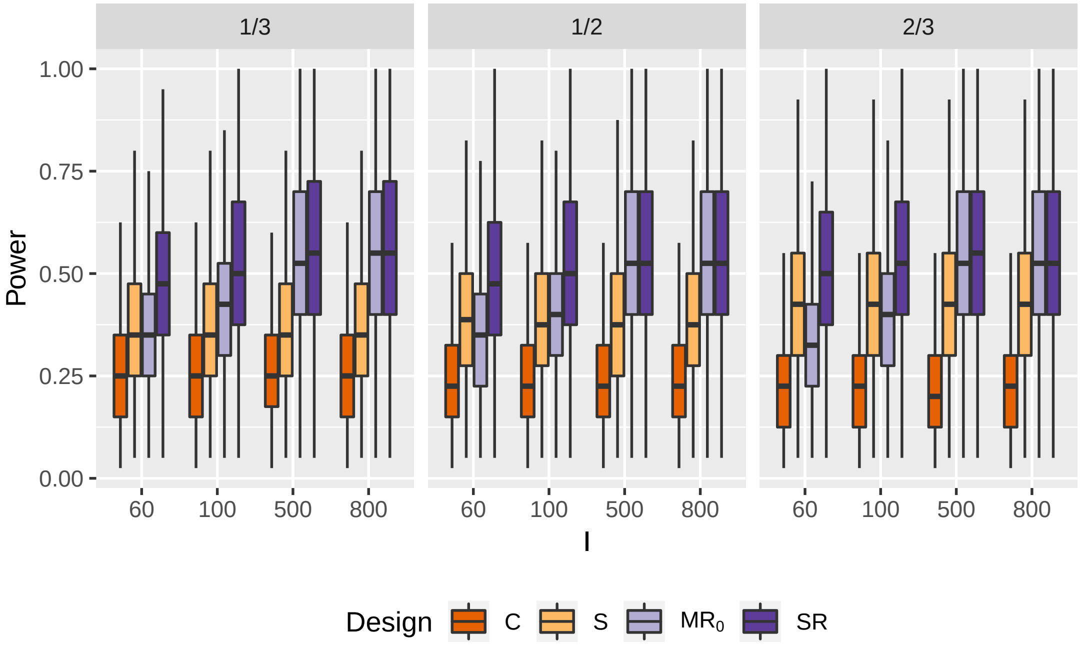

Figure 5: The distribution, over 5,000 replicates, of the power of complete randomization (C), stratification (S), rerandomization (MR_0) and stratified rerandomization (SR) as a function of the number of considered allocations $\mathcal{I}$. The panels display, from left to right, $\rho = 1/3$, $\rho = 1/2$, and $\rho = 2/3$, respectively.

zation as is expected from the theoretical results derived above. The small $\mathcal{I}$ needed to obtain good power improvements in SR and MR_0 is because the number of covariates in the rerandomization is only 1 and 2, respectively.

5.4 Summary of the Monte Carlo study

The Monte Carlo simulations show that the theoretical findings apply in finite sample settings and indicate that they are robust to violations of the normality assumption of the covariate means and error term (needed to derive PRIV_T). If the sample size is small, stratified designs can be suboptimal. If the sample size is moderately large, around 30, rerandomization without stratification run into problems when the number of strata is large. In such situations, stratified rerandomization is a good strategy for reducing the number of degrees of freedom in the Chi-square distribution of the Mahalanobis distance. This enables the Mahalanobis-based rerandomization design to benefit from informative continuous covariates. If the total number of covariates is small, say less than 5, rerandomization and stratified rerandomization gives very similar efficiency gains in comparison to complete randomization also for small N and number of considered allocations $\mathcal{I}'s$ that are manageable for ordinary computers.

6 Empirical example

In this section we make use of an electricity consumption data set, explored in Lundgren and Schultzberg (2019). Within the field of electricity-use research, there is an increasing focus on how to change users' electricity consumption to reduce peak load as well as total load to enable further integration of weather-dependent electricity production as well as to cope with an increase in demand. Several types of interventions have been proposed and evaluated, such as financial incentives in the form of dynamic price signals (Öhrlund et al., 2019; Faruqui et al., 2017) and non-financial incentives in the form of information campaigns or energy feedback (Darby, 2006; Karlin et al., 2015). To help plan future experiments, Lundgren and Schultzberg (2019) present a exploratory study aimed at evaluating the prospects of interventions targeting the attitudes towards electricity use, and savings in particular. Due to the large natural variation in householders' electricity consumption, it is crucial to employ rigorous experimental designs to achieve acceptable power, which makes this data set suitable for illustrating the designs discussed in this paper.

Electricity consumption data were collected at the monthly level for several years for 510 households of which we used the 102 with no missing data. We extracted the two last time periods, November and December for 2017 where the November measurements are being used in the experimental design. The electricity consumption in December 2017, Y_{Dec}, is the outcome. As there were no experiments in December, this electricity consumption is what is observed under the Fisher sharp null. This fact enables us to study the relative performance of the different experimental designs under the Fisher alternative and calculate the power for hypothetical treatment effects as described in Section in 5, Equations (21) and (22).

The November data contains the electricity consumption, Y_{Nov}, and the number of residents in each household, *Residents*. Out of the 102 households, there were 43 with 1 resident, 30 with 2 residents, 11 with 3 residents, 16 with 4 residents, 1 with 5 residents, and 1 with 6 residents.

The following six designs where considered: (i) Stratification (S) on *Residents*[4], (ii) Stratification (S_1) on the quantiles of Y_{Nov} (Y_{Nov}^Q) and *Residents*,[5] (iii) Stratified rerandomization (SR) where we stratify on *Residents* and then using Mahalanobis-based rerandomization on Y_{Nov}, (iv) Rerandomization (MR_0) on Y_{Nov} and the incidence matrix implied by *Residents*, (v) Rerandomizing (MR_1) on Y_{Nov} and *Residents* and (vi), finally, Complete randomization (C) was performed as a benchmark.

In this special case, where the outcome under no treatment is observed, the power under hypothetical treatment effects can be studied. In a real experiment, however, only covariates are observed. This means that in order to compare the efficiency across designs, only the expected variance reduction in $\mathbf{x}$ can be used. Since, R^2 unknown but fixed, this gives a valid relative comparison as long as $R^2 > 0$. Table 3 displays the expected PRIV in the covariates under the different designs. It is clear that if only a small set of all allocations is considered, there are large differences in the variance reduction in the covariates across the designs. For $\mathcal{I} = 10^5$ ($p_{a_1}^{\mathcal{I}} = 0.15$) the variance reduction is 100% and 98.8% for Residents and Y_{Nov}, respectively. As the same variance reduction is obtained for all covariates by definition with the Mahalanobis distance measure, the corresponding variance reduction for MR_0 and MR_1 is 73.45% and 92.09%, respectively. When the number of allocations $\mathcal{I} = 10^8$ (i.e. $p_{a_1}^{\mathcal{I}} = p_{a_0}^{\mathcal{I}} = 0.00015$) we get almost 100% variance reductions on all covariates for these three designs.

Returning to simulating the power under hypothetical effects, we let

$$Y_i(0) = Y_{i,Dec}$$

$$Y_i(1) = Y_i(0) - \tau,$$

where τ is varied as 0, 10, 20, and 30 (kWh), which correspond almost exactly to 0, 0.1, 0.2, and 0.3 standard deviation of $\mathbf{Y}_{Dec}$ ($s_{\mathbf{Y}_{Dec}} = 102.2$). The negative sign of the effect is motivated by the intervention aiming at decreasing consumption.

We follow the procedure in Section in 5 but let $\mathcal{H} = 15,000$ ($= 2/r$) which implies that $r = 0.00013$. For the Mahalanobis distances rerandomization we randomize among the $\mathcal{H}$ allocations with smallest Mahalanobis distances in a random set of sizes $\mathcal{I} = 10^5, 10^6, 10^7$ and 10^8. This means that $p_{a_0}^{\mathcal{I}}$ and $p_{a_1}^{\mathcal{I}}$ vary in the range

[4] Since only one household each has 5 and 6 residents,respectively, only 5 strata were created merging 5 and 6 for all designs using stratification. The strata are not all evenly sized which means that Corollary 4.1 does not apply.

[5] This design implies a maximum of 16 strata, in this case a few strata had zero units resulting in 13 strata.

Table 3: Expected variance reduction in the covariates under stratified rerandomization (SR) and Mahalanobis-based rerandomization (MR$_0$ and MR$_1$).

	Expected PRIV			
Covariate/$p_{a_1} = p_{a_0}$	15%	1.5%	0.15%	0.015%
Stratified rerandomization (SR)				
Residents (factor)	100	100	100	100
Y_{Nov}^{Q}	98.81	99.99	100	100
Rerandomization (MR$_0$)				
Ressidents (factor) and Y_{Nov}	73.45	90.75	96.48	98.62
Rerandomization (MR$_1$)				
Residents and Y_{Nov}	92.09	99.25	99.92	99.99

0.15 to 0.00015. For the C, S and S$_1$ designs, 15,000 allocations were randomly drawn from $\mathbf{W}$. For the SR design the randomization is conducted in the $\mathcal{H}$ allocations with the smallest Mahalanobis in the random set from $\mathbf{W}^S$ of sizes $\mathcal{I}$ =$10^5, 10^6, 10^7$ and 10^8.[6]

Figure 6 displays the power of the five designs and complete randomization for increasing number numbers of considered allocations, $\mathcal{I}$. For $I = 10^5$ ($p_{a_0}^{\mathcal{I}} = p_{a_1}^{\mathcal{I}} = 0.15$) there are large differences between the different rerandomization designs and complete randomization. Among the designs using rerandomization, the price of the degrees of freedom in the rerandomization designs, discussed in Section 4.1.2, is clearly seen as MR$_0$ has the lowest power, MR$_1$ is in the middle, and SR has the highest power. When $\mathcal{I} = 10^8$ ($p_{a_0}^{\mathcal{I}} = 0.00015$) these differences are negligible. Stratification on Y_{Nov}^{Q} and *Residents* gives substantial power improvements as compared to CR. For $\mathcal{I} = 10^5$, S$_1$ has higher power than MR$_0$, however, as S$_1$ does not improve with $\mathcal{I}$, this does not hold when $\mathcal{I}$ increases. The difference between SR and S$_1$ clearly illustrates the (unnecessary) information loss associated with discretizing Y_{Nov}. S also gives higher power than C, but as expected, the importance of balancing the pre-treatment outcome is far more rewarding than perfect balance on number

[6] Note that an alternative is to conduct a Monte Carlo approximations from the set $\mathbf{W}$. The number of Monte Carlo draws needs to be large in order for the FRT to have the right level. In a single analysis this is not a problem, however, in a Monte Carlo simulation this procedure would be very time consuming.

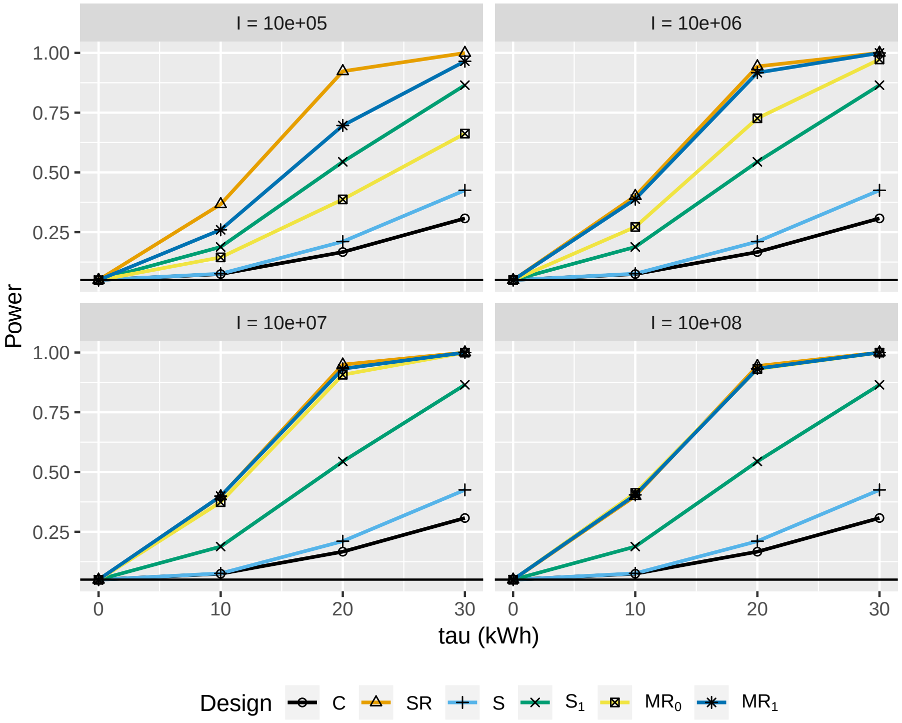

Figure 6: Power of FRT's under hypothetical homogeneous treatment effects for the five designs (Stratification (S) on *Residents*, Stratification (S$_1$) on Y^Q_{Nov} and *Residents*, Stratified rerandomization (SR) where we stratify on *Residents* and then using Mahalanobis-based rerandomization on Y_{Nov}, Mahalanobis-based rerandomization on Y_{Nov} and the incidence matrix implied by *Residents* (MR$_0$), Mahalanobis-based rerandomizing on Y_{Nov} and *Residents* (MR$_1$) and complete randomization (C).

of residents.

7 Discussion

This paper clarifies the properties of and the relationship between stratification and rerandomization designs. The comparison is facilitated by noting that stratification is equivalent with rerandomization in some settings which enables using the theoretical results for rerandomization in Morgan and Rubin (2015) to compare the designs. Stratification, or blocked randomization, is the most common design used to improve balance in experiments. However, as the cost for computations has fallen, the computationally intensive rerandomization design, originally suggested by R A. Fisher in the early twentieth century, is today a viable alternative, or complement, to stratification. Most scholars in experimental design, including R A. Fisher (via Cochran and Rubin), and more recently Morgan and Rubin (2012, 2015), recommend rerandomization as a complement to stratification, a recommendation that was summarized in a lecture by D.B Rubin as 'Block on what you can and rerandomize on what you cannot'. Others seem to view rerandomization as an alterntaive rather than a complement, and do not recommend rerandomization (Athey and Imbens, 2016).

The contribution of this paper is to investigate the properties and limitations of *stratification*, *rerandomization*, and *combinations of both stratification and rerandomization* with the aim of clarifying their pros and cons in practice. The comparison is focused on the efficiency in terms of the variance of sample average treatment effect (SATE) estimator, and power in exact inference to the units in the sample.

The main conclusion is that there are three aspects to consider when choosing between these designs: (i) the number of available binary and continuous covariates, (ii) the relative importance of all observed covariates, and (iii) the number of allocations remaining after a possible stratification. If the number of allocations remaining after a stratification is large, say 10,000 or larger, it is a good idea to use stratified rerandomization, i.e., stratifying on the binary covariates and then rerandomizing on the remaining covariates. This holds regardless of the importance of binary covariates as starting with a stratification step increases the efficiency in following rerandomization steps. If the number of allocations remaining after a stratification is small, it is possible that rerandomization on all covariates at once, or only the continuous covariates, is more efficient. If the sample size is moderately large, say 20 or larger, and the total number of covariates is small, say 5 or less, the difference in efficiency gain of rerandomization compared to stratified rerandomization is negligible.

The focus in this paper is limited to the inference to the units of the sample. This is mainly done for clarity, as we believe that some of the intuition may become masked by the mathematics surrounding inference to a superpopulation. However, all variance reduction formulas apply also under random sampling from a population. Since, as opposed to Fisher's exact inference, the efficiency of Neyman's inference, commonly used to draw inference to PATE, is directly related to the variance of the estimator, the main conclusions of the power simulations should therefore also apply for 'Neyman inference' to the units of a superpopoulation.

Since large efficiency gains can be made from rerandomizing on important continuous covariates, we argue that stratified rerandomization should always be the first-hand-choice. Stratification can always be encompassed by rerandomization and, correctly used, gives equivalent or better efficiency than stratification. Therefore, rerandomization, especially *stratified rerandomization* or *rerandomization in tiers* in general, should be a good starting point for any experimental design.

References

Peter M. Aronow, Donald P. Green, and Donald K.K. Lee. Sharp bounds on the variance in randomized experiments. *Annals of Statistics*, 42(3):850–871, 2014. ISSN 00905364. doi: 10.1214/13-AOS1200.

Susan Athey and Guido Imbens. The State of Applied Econometrics - Causality and Policy Evaluation. 31 (2):3–32, 2016. ISSN 0895-3309. doi: 10.1257/jep.31.2.3. URL `http://arxiv.org/abs/1607.00699`.

Sarah Darby. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 486, 2006.

Ahmad Faruqui, Sanem Sergici, and Cody Warner. Arcturus 2.0: A meta-analysis of time-varying rates for electricity. *The Electricity Journal*, 30(10):64–72, 2017. ISSN 10406190. doi: 10.1016/j.tej.2017.11.003. URL `http://linkinghub.elsevier.com/retrieve/pii/S1040619017302750`.

Ronald A Fisher. *The design of experiments.* Oliver and Boyd, 1935.

G W Imbens and D B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015. ISBN 9780521885881.

Per Johansson and Mårten Schultzberg. Experimental design using longitudinal data. 2018.

Beth Karlin, Joanne F. Zinger, and Rebecca Ford. The effects of feedback on energy conservation: A meta-analysis. *Psychological Bulletin*, 141(6):1205–1227, 2015. ISSN 00332909. doi: 10.1037/a0039650.

Xinran Li, Peng Ding, and Donald B Rubin. Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157 LP – 9162, sep 2018.

Berndt Lundgren and Mårten Schultzberg. Does energy-effective behavior matter for energy conservation? 2019.

Kari Lock Morgan and Donald B. Rubin. Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282, 2012. ISSN 00905364. doi: 10.1214/12-AOS1008.

Kari Lock Morgan and Donald B. Rubin. Rerandomization to Balance Tiers of Covariates. *Journal of the American Statistical Association*, 110(512):1412–1421, 2015. ISSN 1537274X. doi: 10.1080/01621459.2015.1079528.

Jerzy Neyman. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Translated in Statistical Science(1990)*, 5(4):465–472, 1923. ISSN 08834237. URL `http://www.jstor.org/stable/2245382`.

Jerzy Neyman. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472, 1990. ISSN 08834237. URL `http://www.jstor.org/stable/2245382`.

I. Öhrlund, M. Schultzberg, and C. Bartusch. Identifying and estimating the effects of a mandatory billing demand charge. *Applied Energy*, 237, 2019. ISSN 03062619. doi: 10.1016/j.apenergy.2019.01.028.

Donald B. Rubin. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591, 1980. ISSN 01621459. doi: 10.2307/2287653.

Anne Penfold Street and Deborah J Street. *Combinatorics of Experimental Design*. Oxford science publications. Clarendon Press, 1987. ISBN 9780198532569. URL `https://books.google.se/books?id=bhDbIjpL0eAC`.

A Example 1

To give intuition for Theorem 4.1, consider two binary covariates x_1 and x_2. Table 4 gives the columns of $\mathbf{x} = (x_1, x_2, x_1x_2)$ together with the unit id numbers and the implied strata.

Table 4: Example 1

Unit	x_1	x_2	x_1x_2	Stratum
1	1	1	1	1
2	1	1	1	1
3	1	0	0	2
4	1	0	0	2
5	0	1	0	3
6	0	1	0	3
7	0	0	0	4
8	0	0	0	4

In the example of two binary covariates, the incidence matrix implied by $\mathbf{x}$ has columns x_1x_1, $x_1(1-x_2)$, $x_2(1-x_1)$, and $(1-x_1)(1-x_2)$. That is

$$\mathbf{d} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ \end{pmatrix}, \text{ and } \mathbf{q} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \end{pmatrix}. \tag{23}$$

Clearly, the mean-difference vector, $\overline{\mathbf{Q}}_1^j - \overline{\mathbf{Q}}_0^j$, will only be the zero vector if $\mathbf{W}^j$ implies that: one of units 1 and 2 are treated, one of units 3 and 4 are treated and one of units 5 and 6 are treated. This implies that only one of units 7 and 8 are treated (to fulfill $N_1 = N_0 = N/2$).

It is slightly less straight forward to see why $M(\mathbf{x}, \mathbf{W}^j|\mathbf{W}) = 0$ only for allocations j in $\mathbf{W}^S$. Consider $\mathbf{x}$ in Table 4. Following a similar reasoning as above, starting with x_1x_2, it is clear that one of the units 1-2 must be treated. Moving to x_2, since we already have one unit from units $1-2$, one of the units $5-6$ has to be treated. Finally, in x_1, one of the units $3-4$ has to be treated. This directly implies that to obtain 4 units in total (to obtain $N_1 = N_0$), one unit from units $7-8$ must be treated. In summary, for the treatment group to fulfill the criterion of equal number of ones in the treatment groups in all covariates, it has to contain: one unit from units $1-2$ (stratum 1), one unit from units $3-4$ (stratum 2), one unit from units $5-6$ (stratum 3), and 2 units from units $7-8$ (stratum 4), i.e, it has to be an allocation allowed under stratified randomization design.

For completeness, the equivalence of rerandomizing on $\mathbf{q}$ and $\mathbf{x}$ is illustrated using the reasoning in the proof of Corollary 4.1. The Mahalanobis distance is affinely invariant which means that the Mahalobis distance based on $\overline{\mathbf{x}}$ is the same as the Mahalanobis distance for any $\overline{\mathbf{z}}$

$$\overline{\mathbf{z}} = \mathbf{a} + \mathbf{B}\overline{\mathbf{x}}$$

where $\mathbf{a}$ is any given $K \times 1$ vector and $\mathbf{B}$ a $K \times K$ pd matrix.

Returning to the example, $\mathbf{x}_{1i} = (x_{1i}, x_{2i})$ then $\mathbf{x}_i = (x_{1i}, x_{2i}, x_{1i}x_{2i})$ and $\mathbf{d}_i = (x_{1i}x_2, x_{1i}(1-x_{2i}), x_{2i}(1-x_{1i}), (1-x_{1i})(1-x_{2i}))'$ and $\mathbf{q}_i = (x_{1i}x_2, x_{1i}(1-x_{2i}), x_{2i}(1-x_{1i}))'$. Since $n_S^1 = n_S^2 = n_S^3 = n_S^4$, it follows that $\pi^1 = \pi^2 = \pi^3 = 0.25$. In addition we have that $p_1 = 0.5$, $p_2 = 0.5$, and $p_3 = p_1 \times p_3$ such that

$$\overline{\mathbf{q}} = \mathbf{0} + \mathbf{C}\overline{\mathbf{x}}$$

$$\Leftrightarrow$$

$$\begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.5 \\ 0.5 \\ 0.25 \end{pmatrix}.$$

Clearly, since $\mathbf{C}$ is full rank, $\overline{\mathbf{q}}$ is a affine transformation of $\overline{\mathbf{x}}$.