# Rerandomization to Improve Covariate Balance by Minimizing the MSE of a Treatment Effect Estimator

## A Treatment Assignment Method for One and Multiple Treatment Arms

Sebastian O. Schneider*
Max Planck Institute for Research on Collective Goods, Bonn
and
Martin Schlather
Department of Mathematics, University of Mannheim

November 30, 2021
Click here to get the newest version of this paper.

**Abstract**

We present a new approach to treatment assignment in (field) experiments for the case of one or multiple treatment groups. This approach, which we call the minimizing Mean Squared Error (min MSE) approach, uses sample characteristics to obtain balanced treatment groups. Compared to other methods, the min MSE procedure is attrition tolerant, offers greater flexibility, is very fast, it can be conveniently implemented and balances different moments of the distribution of the treatment groups. Additionally, it has a clear theoretical foundation, works without parameters being specified by the researcher and allows multiple treatments. The information used for treatment assignment can be multivariate, discrete or continuous, and may consist of any number of variables. In this paper, we derive the underlying theoretical selection criteria, which we then apply to various scenarios and datasets. Our proposed method performs better than, or comparably to, competing approaches, such as matching, in most of the commonly used measures of balance. We provide Stata, R and Python implementations of our method.

*sschneider@coll.mpg.de

# 1 Introduction

The current debate about replicability of scientific findings from experiments (Open Science Collaboration, 2015; Camerer et al., 2016) shows the importance of practices that improve the validity of experimental outcomes. One such practice is conducting randomization or treatment assignment in an appropriate way. The more similar the treatment groups, the higher is the precision of the experiment, that is: the closer is the outcome of a single experiment to the truth (Fisher, 1935). Thus, appropriate treatment assignment is directly linked to replicability. Since there is no consent on how treatment assignment should be carried out (Bruhn and McKenzie, 2009), several competing and complementary strategies to account for group characteristics in treatment assignment are widely used. Also from a theoretical perspective, a clear answer is missing; see e.g. Imbens (2011) for a brief discussion.

One method is pure randomization, which can be considered, depending on the transparency of the actual implementation, as the fairest method for treatment allocation, and is certainly the fastest. A drawback is, that imbalances appear by chance and can lead to undesired false alarms. Furthermore, it is not guaranteed, especially when the sample size is small, that all characteristics of a variable appear in all experimental groups at all and additionally with the same frequencies; this is a problem when subgroup analysis is desired to study heterogeneous treatment effects.

Stratification or blocking goes back to Fisher (1935). The idea is to build subgroups according to observable characteristics and to randomize within those subgroups. This method achieves exact balance for binary variables and improves the balance in comparison to purely random treatment assignment for other types of variables. The main advantage of stratification is to ensure the possibility of subgroup analysis while ideally increasing the efficiency of the analysis. The time needed to conduct treatment assignment using stratification depends on the actual implemen-

tation, but in simple cases, e.g. with two dichotomous variables, it takes only slightly longer than pure randomization. A disadvantage is that only a very limited number of variables can be balanced, since the number of required strata is the product of the number of parameter values and each stratum should contain several observations. Furthermore, continuous variables have to be discretized and are never really balanced with this approach. Additional problems arise when the number of participants is not divisible by the number of subgroups. Although solutions to this have been suggested, a simple implementation is no longer possible. Moreover, building the strata requires expertise on both the data and the question under investigation.

Pairwise matching is often seen as the limit case of stratification, when the subgroups consist of only two individuals. The subgroups, called pairs in the case of matching, have to be created such that the two individuals are similar, where the similarity can be measured e.g. with the so-called Mahalanobis distance of the covariate vectors of the two individuals. Two types of algorithms are commonly used: the so-called greedy algorithm (Imai et al., 2009a) and an 'optimal matching' algorithm (Greevy et al., 2004; Lu et al., 2011). Note that this is a different task to the one performed for matching in observational studies: Finding pairs when groups have already been formed is far less demanding, also from a computational aspect. Matching can be realized with many possible continuous variables and thus eliminates some of the shortcomings of stratification. Subgroup analysis, however, is not ensured in cases, where balance on a certain variable could not be achieved, which, however, should not be the case in moderate sized samples and a moderate amount of variables to balance. It is arguably considered to be fair and the design is relatively clear and easy to explain. The biggest advantage of the 'optimal matching' algorithm is that the distribution of covariates in the treatment and the control group become as similar as possible. This, however, comes at the cost of analytical difficulties when

estimating the variance and the standard error of the treatment effect (e.g. Imbens, 2011; Abadie and Imbens, 2006; Klar and Donner, 1997). Limitations arise when attrition occurs, i.e. when, for some units, the outcome remains unobserved. Imai et al. (2009a) note that an advantage of matching is that if a unit drops out, its pair can also be taken out of the experiment while the remaining sample still remains balanced. This problem becomes severe in small samples or when performing randomization at the cluster level: For every unit, possibly consisting of many individuals, dropping out of the experiment, its pair should also be removed, which lowers the sample size and power and can be of major concern. Furthermore, matching can only be performed when the number of units is even. Finally, the matching approach implemented by Bruhn and McKenzie (2009), needed several days to conduct treatment assignment with a sample size of 300 units, so this approach is inappropriate if time is a limiting factor. Yet, the software implementation of the 'optimal matching' algorithm in the *R* package *nbpMatching* (Lu et al., 2011) is considerably faster.

Rerandomization methods try to avoid the above mentioned theoretical or practical limitations. The basic idea is to choose the best assignment, according to a specified evaluation criteria. Due the complexity of the problem, a random treatment assignment is picked in a certain way, evaluateed with respect to the criteria and rerandomized either a certain number of times or until the criteria meets some prespecified condition. Sometimes, subjective judgment is also used (Bruhn and McKenzie, 2009). All of the rerandomization methods discussed here are able to consider continuous, categorial and binary variables in a theoretically unlimited number. However, we are aware of only one rerandomization approach that relies on a theoretical derivation of the statistical threshold to stop the rerandomization (Morgan and Rubin, 2012). This threshold, as well as the alternative ad-hoc thresholds, such as picking the maximum t-value minimizing treatment group assignment,

focuses only on the mean value of one or several covariates and ignores other characteristics of the distributions of the variables that might be balanced. Yet, all of those rerandomization methods aim at balancing group means, and with the exception of Morgan and Rubin (2012), fail to consider dependencies of the different variables included in treatment assignment. However, in their approach, the dependency between variables is constant accross treatment assignments. Irrespectively of these limitation, we are unaware of any software implementation.

Kasy (2016) applies a decision theoretical, Bayesian model to analyze the problem of treatment assignment. To that end, he derives the posterior mean squared error (MSE) of an estimator for the conditional average treatment effect of interest as a function of treatment assignment and argues that randomization never increases precision compared to an optimal, deterministic treatment assignment. Kasy (2016) discusses several possibilities to implement such a (binary) treatment assignment procedure and provides Matlab code for their implementation; one such possibility is his Bayesian linear model, where for an application, the researcher has to pick a mean vector and a covariance matrix for the distribution of the estimator in a model for the potential outcomes. Since, in addition, a guess for the coefficient of determination $R^2$ of such a linear regression model must be specified, the approach becomes impractical in its generality. Of course, one could use a *flat* prior, inducing nearly no prior information. In this case, however, one might also resign from using prior information, as it simplifies the objective function and consequently the method considerably.

To conclude, to date there is to our knowledge no solution available to perform randomization with multiple treatment arms and multiple possibly continuous variables. Also in the case when only one treatment is to be assigned but attrition might be of a concern, researchers might be unsatisfied with the matching approach.

# 2 The min MSE Treatment Assignment

In the spirit of Kasy (2016), our approach combines the mean squared errors of the estimators for the conditional average treatment effects within a linear model that is a function of treatment assignment. Moreover, due to developing the approach in a frequentist setting, we increase the applicability of the statistic for treatment assignment considerably: Our result works without choosing any technical parameters while still allowing for the needed flexibility. In the treatment assignment mechanism derived here, the only parameter that must be specified by the researcher is the number of treatment groups desired; other parameters, such as scaling factors for variances, can be specified, but can be left constant unless a better guess is available. The assumption of equal variances is an intuitive assumption that experienced researchers quickly can confirm or withdraw, and in the latter case, easily adjust by specifying a good guess for scaling up the variance of a treatment or an outcome.

A further advantage of the frequentist setting is that the statistic establishes an undistorted balance between treatment groups. More precisely, we show that this statistic aims at balancing the second moments of the covariate distributions, incorporates dependencies between covariates and illustrate the importance of these features. Apart from that, we interpret and implement the method as a rerandomization method, which yields the possibility of randomization inference.

Finally, our method allows multiple treatments and multiple outcomes, which both can be weighted.

## 2.1 Framework and Treatment Effect

First, we define the parameter we are ultimately interested in estimating: the conditional average treatment effect. We do so by introducing the potential-outcome

framework (Rubin, 1974, 1977). As we derive the minimizing MSE treatment assignment procedure for various treatment effects and various outcomes, we directly extend the framework to fit our needs.

Assume, we have $N$ participants, randomly selected for the experiment from the population. Individual draws of a (random) variable are indicated with a subscript $i = 1, \ldots, N$ and realizations of a random variable or vector will be denoted by the corresponding lower-case letter.

In the experiment, an individual $i$ is randomly assigned to one of $d$ experimental groups and treated with the corresponding treatment or is not treated at all if it is assigned to the control group. Let denote this random assignment with $A_i$, $A_i \in \{0, \ldots, d\}$, where 0 indicates the control group. The $m$-variate (realized) outcome for individual $i$, denoted by $Y_i = (Y_{i,1}, \ldots, Y_{i,m})^\top$, depends on $A_i$, but also on its $r$-variate covariates $X_i = (X_{1,i}, \ldots, X_{r,i})^\top$, which will be modelled by a random vector. Let $X = (X_1, \ldots, X_N)$.

For theoretical reasons we are interested also in the hypothetical outcomes if the assignment would have been different. The ensemble of realized and hypothetical outcomes are called potential outcomes and are denoted by the random variables $Y_{i,k}^{p,a}$, where $a \in \{0, \ldots, d\}$, $k \in \{0, \ldots, m\}$ and $i$ as above. Let $Y_i^{p,a} = (Y_{i,1}^{p,a}, \ldots, Y_{i,m}^{p,a})$ the $m$-variate potential outcome of individual $i$ for treatment $a$.

Let $I(a) = \{i : A_i = a\} = \{i_1, \ldots, i_{n_a}\}$, $V_{I(a)} = (V_{i_1}, \ldots, V_{i_{n_a}})^\top$ if $V$ is a vector with components $V_1, \ldots, V_N$, and $V_{I(a)} = (V_{i_1}, \ldots, V_{i_{n_a}})$ if $V$ is matrix with columns $V_1, \ldots, V_N$. We also use $I(a)$ in superscript notation with the same meaning.

The realized outcome $Y_i$ of individual $i$ can now be written by means of potential outcomes and the treatment group assignment:

$$Y_i = \sum_{a=0}^{d} \mathbb{1}_{i \in I(a)} Y_i^{p,a} = Y_i^{p,0} + \sum_{a=0}^{d} \left( Y_i^{p,a} - Y_i^{p,0} \right) \mathbb{1}_{i \in I(a)}.$$

The right-hand side of the above formula decomposes the realized outcomes for an individual in her potential outcomes. The differences $Y_i^{p,a} - Y_i^{p,0}$, which are the causal effects of the treatment $a$, would be of great interest in any study, but can never be observed. However, under certain conditions, we can estimate the population average effect of treatment $a$:

$$\tau^a = \mathbb{E}\left[Y_i^{p,a} - Y_i^{p,0}\right], \quad \text{for all } a = 1, \ldots, d \text{ and any } i,$$

which is often sufficient for most research questions.

If the main interest is to study a subpopulation (e.g. the poor), or when one is not sure whether or not the sample at hand is representative for the population, one should focus on the conditional average treatment effect (Imbens, 2004). This happens frequently in Development Economics, for instance.

**Definition 1** (Conditional Average Treatment Effect)**.** The *conditional average treatment effect* of treatment $a$, $a \in \{1, \ldots, d\}$, is defined as

$$\tau^a(X) = (\tau_1^a(X), \ldots, \tau_m^a(X)) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[Y_i^{p,a} - Y_i^{p,0} \mid X_i\right].$$

For identification of the conditional average treatment effect, further assumptions are needed and discussed, e.g. in Imbens (2004) or Abadie and Imbens (2006). The most important assumption, the *conditional independence assumption* (sometimes called unconfoundedness assumption), means that potential outcomes are independent of the group and therefore of treatment assignment, conditional on the covariates, i.e.,

$$A_i \text{ is independent of } Y_i^p \text{ conditional on } X_i = x$$

for almost every $x \in \mathbb{X}$, where $\mathbb{X}$ denotes the support of $X_i$ for any $i = 1, \ldots, N$. If this assumption holds, any potential selection bias vanishes.

The second important assumption is the so called *overlap condition*, which says that all characteristics observed in a treatment group have to be found amongst the individuals in the control group,

$$\mathbb{P}(A_i = a \,|\, X_i = x) > \eta$$

for all $a = 0, 1, \ldots, d$, almost all $x \in \mathbb{X}$ and some $\eta > 0$ (Abadie and Imbens, 2006). If the overlap condition does not hold, a comparison of the expected potential outcomes, given those covariates, is not possible. It is generally never guaranteed that this is possible, but a powerful treatment assignment procedure will make it more probable. Given these two assumptions hold, observed difference in average outcomes conditional on the observables between the treatment and the control group can be interpreted as the causal, conditional treatment effect.

## 2.2 A Mean Squared Error Based Minimization Function

The Mean Squared Error of a scalar estimator $\hat{\tau}$ conditional on $X$ is defined as

$$\mathrm{MSE}(\hat{\tau} \,|\, X) = \mathbb{E}\left[(\hat{\tau} - \tau)^2 \,|\, X\right],$$

where $\tau$ is the real-valued parameter to be estimated. The MSE can be decomposed into the variance and bias of the estimator, conditional on $X$, and thus results in a measure of efficiency for unbiased estimators, given a specific set of data $X$.

More generally, let $T$ a (random) matrix, $\hat{T}$ an estimator for $T$ and let $w = (w_1, \ldots, w_m)$ and $v = (v_1, \ldots, v_d)$ be non-negative vectors of weight that are not identically 0. Then, for the matrix of weighted estimators $\mathrm{diag}(\sqrt{v})\hat{T}\,\mathrm{diag}(\sqrt{w})$, we define the conditional weighted MSE component-wise as

$$\mathrm{MSE}(\hat{T}, v, w \,|\, X) = \mathbb{E}\left[\left\|\mathrm{diag}(\sqrt{v})(\hat{T} - T)\,\mathrm{diag}(\sqrt{w})\right\|_F^2 \,|\, X\right], \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. We assume $v$ and $w$ do not depend on $T$.

The expectation of the squared Frobenius norm of the matrix $\hat{T} - T$ with its corresponding weights is because of linearity the trace of the expected squared weighted error matrix:

$$\begin{aligned}
&\mathrm{MSE}(\hat{T}, v, w \,|\, X) \\
&\quad = \mathbb{E}\left[\mathrm{tr}\left(\mathrm{diag}(\sqrt{w})(\hat{T} - T)^\top \mathrm{diag}(v)(\hat{T} - T)\,\mathrm{diag}(\sqrt{w})\right) \,|\, X\right] \\
&\quad = \mathrm{tr}\left(\mathrm{diag}(\sqrt{w})\mathbb{E}\left[(\hat{T} - T)^\top \mathrm{diag}(v)(\hat{T} - T) \,|\, X\right] \mathrm{diag}(\sqrt{w})\right).
\end{aligned}$$

Let now $T = (\tau_k^a(X))_{a=1,\dots,d;k=1,\dots m}$ so that $w$ and $v$ weight outcomes and treatments, respectively. The objective is to minimize the weighted MSE (1) given the weights $v$ and $w$:

$$S(\hat{T}) = \mathrm{MSE}(\hat{T}, v, w \,|\, X) = \min_{\hat{T}}!.$$

As the conditional average treatment effect is a function of the covariates let $\hat{T} = f(X)$ for some function $f$. As the weights $w$ and $v$ do not depend on $\hat{T}$, $S(\hat{T})$ is a linear function of the diagonal elements of $\mathbb{E}\left[(f(X) - T)^\top (f(X) - T) \,|\, X\right]$. The minimizer of a summand of a diagonal element, $(f(X) - T)^2_{a,k}$ say, is given by the conditional expectation of $T_{a,k}$ given $X$, hence $S(\hat{T})$ is minimized by setting $f(X) = \mathbb{E}(T \,|\, X)$. With that,

$$\mathbb{E}[T \,|\, X] \in \operatorname*{argmin}_{\hat{T}} S(\hat{T}).$$

Considering the $a$-th row of the matrix $\mathbb{E}[T \,|\, X]$ and using the definition of the Conditional Average Treatment Effect, see Definition 1, yields

$$\mathbb{E}[\tau^a(X) \,|\, X] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i^{p,a} - Y_i^{p,0} \,|\, X_i\right] \,|\, X\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i^{p,a} - Y_i^{p,0} \,|\, X_i\right].$$

This, however, leaves us with the challenge of estimating $\mathbb{E}\left[Y_i^{p,a} \,|\, X_i\right]$ for all treatment groups $a = 0, 1, \dots, d$.

## 2.3 A Linear Model for Potential Outcomes

We choose a linear model for the relationship between covariates and potential outcomes, i.e.,

$$Y_{i,k}^{p,a} = X_i^\top \beta_k^a + \varepsilon_{i,k}^a \tag{2}$$

for $i = 1, \ldots, N$, $k = 1, \ldots, m$ and $a = 0, 1, \ldots, d$ with

| | |
|---|---|
| $Y_{i,k}^{p,a}$ | a random number taking values in $\mathbb{R}$, |
| $X_i$ | a random vector of length $r$ with values in $\mathbb{R}$ and positive variance, |
| $\beta_k^a$ | the vector of deterministic parameters of dimension $r$ and |
| $\varepsilon_{i,k}^a$ | a real-valued random number. |

We assume that the $((Y_{i,k}^{p,a})_{a=1,\ldots,d;k=1,\ldots,m}; X_i)$ are independent and identically distributed for all $i = 1, \ldots, N$. For the error terms, we assume $\varepsilon_{i,k}^a \,|\, X_i \sim \mathcal{N}(0, \sigma_{ak}^2)$ for all $i = 1, \ldots, N$ and all $k = 1, \ldots, m$, $a = 0, 1, \ldots, d$. Moreover, we assume independence between $\varepsilon_{i,k}^a$ and $\varepsilon_{i,k}^0$ for $i = 1, \ldots, N, k = 1, \ldots, m$ and $a = 1, \ldots, d$. The variances are expressed in relation to a base variance: $\sigma_{ak}^2 = s_k^a \sigma_{0k}^2$ for all $a = 0, 1, \ldots, d$, $k = 1, \ldots, m$ with $s_k^a > 0$ and $\sigma_{0k}^2 = s_k^0 \sigma_0^2$ with $s_k^0 > 0$ for all $k = 1, \ldots, m$ and for some $\sigma_0^2 > 0$.

The objective function $S$ can be expressed in terms of the weights and the submatrix $X_{I(a)}$ of covariate vectors of all individuals in treatment group $a$. Let $\bar{X} = N^{-1} \sum_i X_i$ and denote by $C^-$ the Moore-Penrose inverse of a matrix $C$ and by id the identity.

**Theorem 1.** *Under Assumption 2, the minimization criterion* (1) *equals*

$$\begin{aligned} &\bar{X}^\top \left[ \|\tilde{w}\|_1 \, \|v\|_1 \left( X_{I(0)} X_{I(0)}^\top \right)^- + \sum_k \tilde{w}_k \sum_{a>0} \tilde{v}_k^a \left( X_{I(a)} X_{I(a)}^\top \right)^- \right] \bar{X} \\ &+ \sum_{a>0} v_a \sum_k w_k (\bar{X}^\top ((\mathrm{id} - H_a)\beta_k^a - (\mathrm{id} - H_0)\beta_k^0))^2, \end{aligned} \tag{3}$$

*where* $\|\cdot\|_1$ *is the* $l_1$ *norm of a vector,* $\tilde{w}_k = w_k s_k^0$ *and* $\tilde{v}_k^a = v_a s_k^a$ *for* $k = 1, \ldots, m$, $a = 1, \ldots, d$, *and* $H_a = (X_{I(a)} X_{I(a)}^\top)^- X_{I(a)} X_{I(a)}^\top$.

The proof of all theorems and propositions can be found in Appendix A.

Since, in general, the $\beta_k^a$ are unknown and since the first summand of (3) tends to zero as $N \to \infty$ for any reasonable choice of assignment, see Lemma 1, the condition

$$H_a = \mathrm{id}, \qquad \text{for all } a \geq 0 \text{ and } N \text{ large enough} \tag{4}$$

must hold. Of course, knowledge about the $\beta_k^\alpha$ may lead to $H_a \neq \mathrm{id}$, even for large $N$.

**Corollary 1.** *Let the condition of Theorem 1 and* $N$ *large enough so that* (4) *holds. Assume that variances for all outcomes and treatment groups are the same, including the control group (i.e.* $s_k^a = 1$ *for all* $a = 0, 1, \ldots, d$ *and all* $k = 1, \ldots, m$*) and that all weights are* 1. *Then, minimizing* (1) *through choice of* $A$ *is equivalent to minimizing*

$$\bar{X}^\top \left[ d \left( X_{I(0)} X_{I(0)}^\top \right)^{-1} + \sum_{a>0} \left( X_{I(a)} X_{I(a)}^\top \right)^{-1} \right] \bar{X}. \tag{5}$$

**Proposition 1.** *Let the conditions of Corollary 1 Assume that the* $X_{ij}$ *have compact support (sollte auch schwaecher gehen, ist aber ausreichend fuer Praxis). Then, for* $N$ *large enough, the optimization problem* (13) *is invariant under a transformation of the vector* $(X_{j,1}, \ldots, X_{j,N}) \mapsto (cX_{j,1}, \ldots, cX_{j,N})$ *for any* $c \neq 0$ *and for any* $j = 1, \ldots, r$.

## 2.4 Balanced Treatment Groups

**Proposition 2.** *Assume that* $\sum_j |\mathbb{E}X_{j1}| > 0$, *and* $M_{\alpha,j} = \mathbb{E}|X_{j1}|^{2+\alpha} \in (0, \infty)$ *for all* $j$ *and some* $\alpha > 0$. *Assume that* $X_{j,1}$ *and* $X_{l,1}$ *are uncorrelated for* $j, l = 1, \ldots, r$,

*$k \neq j$. Furthermore, assume that all covariates have the same mean, i.e. $\mathbb{E}X_{j,1} = c$ for some $c \neq 0$. Then, for $N \to \infty$, a solution to the minimization problem according to Corollary 1 is obtained, if*

$$d\frac{\sum_{b\geq 0}\sqrt{c_b}}{\sqrt{c_0}}\frac{s_0}{1+s_0} + \sum_{a>0}\frac{\sum_{b\geq 0}\sqrt{c_b}}{\sqrt{c_a}}\frac{s_a}{1+s_a} \tag{6}$$

*where $s_a = \sum_{j=1}^{r}\hat{\eta}_j^2/\hat{v}_{j,I(a)}$ and $\hat{\eta}_j^2$ is the sample mean, and $\hat{v}_{j,I(a)}$ are the sample variance of the $\{X_{j,i} : i \in I(a)\}$.*

*is minimized.*

If $d = 1$ and under the conditions of Proposition 2 the value $\sum_{j,a}\left(\sum_{i\in I(a)} X_{j,i}^2\right)^-$ must be minimized. That is, for each $i$, the square deviation from 0 must be balanced for all covariates across groups. In the simple case of equally sized groups, this is equivalent to balancing the variance of each of the covariates. This makes the min MSE procedure a unique method in the sense that balance incorporates not just the mean, but a higher moment of the distribution of covariates. It is exactly this property that makes the groups comparable in the sense that the different subgroups– if any–are to be found in all experimental groups.

Without the restricting assumptions of Proposition 2 we get a partial answer.

**Proposition 3.** *Under the conditions of Corollary 1, the diagonal elements of the matrix $\left(X_{I(0)}X_{I(0)}^\top\right)^- + \left(X_{I(a)}X_{I(a)}^\top\right)^-$ in (13) are given by*

$$\mathrm{Var}(\hat{\beta}_{k,j}^a - \hat{\beta}_{k,j}^0 \,|\, X) = c_k \sum_{b\in\{0,a\}} \frac{1}{(1-R_{b,k,j}^2)\sum_{i=1}^{n_b}(X_{I(b),j,i} - \overline{X_{I(b),j}})^2}$$

*for any $a \geq 1$, $j \in \{1,\ldots,r\}$, $k = \{1,\ldots,m\}$, and some $c_k > 0$. The value of the $j$-th covariate of individual $i$ in treatment group $a$ is denoted by $X_{I(a)j,i}$ and $\overline{X_{I(a)j}}$ denotes the mean. $R_{a,k,j}^2$ is the coefficient of determination of a regression between*

*the variable $X_{I(a),j}$ as response and all $X_{I(a),p}$ for $p = 1, \ldots, r$, $p \neq j$ as explanatory variables, and for all $a = 0, 1, \ldots, d$. The number of individuals in treatment group $a$ and in the control group is denoted by $n_a$ and $n_0$, respectively.*

The sum on the right-hand side of (7) is decreased for every $a \geq 1$ and every covariate $j = 1, \ldots, r$, if the linear dependencies between covariates within groups are decreased. Thus by having this criterion in the objective formula, we reward a grouping that avoids multicollinearity and punish a high level of similarity amongst the combination of covariates in a group. This characteristic is also inherent in matching: Two very similar subjects should not be allocated to the same group. This characteristic alsp distinguishes the min MSE procedure from other rerandomization methods, as it considers the complete composition of covariate values in a group instead of considering all covariates independently. Note, however, that this grouping might not minimize off-diagonal entries of the sum of the covariance matrices.

The second part influencing (7) is the within-group variance of variable $x_j$ around its mean. The higher its value, the lower the variance of $\hat{\beta}$. An overall decrease of the latter variance can only be achieved if an increase of the variance of $x_j$ in one group does not lower the variance in the other groups to the same extent or more.

## 2.5 The Relation to the Classical Approach of Optimal Experimental Design

In what follows, we establish the link to the literature on experimental design. Starting with Smith (1918), the goal has also been the minimization of the variance of an estimator $\theta$ under a certain criterion. The task in optimal design is usually a version of the following: For a set $X$ of $N$ measurement points called the design region, choose the multiplicity $l_i$ of measurements at the measurement points $X_i$ such

that the precision of the outcome to be estimated is maximized. This choice is called *design* and it can be represented as a collection of variables

$$\zeta = \begin{Bmatrix} X_1 & X_2 & \cdots & X_N \\ p_1 & p_2 & \cdots & p_N, \end{Bmatrix} \tag{7}$$

where $\sum_i p_i = 1$, $p_i L \in \mathbb{N} \cup \{0\}$ and $L = \sum_i l_i$. As such, $\zeta$ is a discrete probability measure defined on the design space $X$. Then, $\zeta$ is chosen such that $\Phi(\operatorname{Var} \hat{\theta}(\zeta))$ is minimized for a suitable optimality criterion $\Phi$ (Kiefer, 1959; Fedorov and Hackl, 1997). In case a linear combination $c^\top \theta$ of the estimator of interest for some $c \in \mathbb{R}^r$, c-optimality should be considered (Fedorov and Hackl, 1997) as criterion, i.e., $\Phi(\operatorname{Var} \hat{\theta}(\zeta)) = c^\top \operatorname{Var} \hat{\theta}(\zeta) c$.

**Proposition 4.** *Assume $d = m = 1$. Then, under the conditions of Corollary 1, the min MSE minimization criterion coincides with the c-optimality criterion.*

## 2.6 Comparison to Alternatives Methods

### 2.6.1 Pair-Wise Matching

Consider a treatment assignment for a treatment and a control group, where for every individual $i$, one covariate $x_i$ is observed and the treatment should be assigned such that this covariate is balanced across the treatment and control groups.

**Theorem 2.** *Pair-wise matching before treatment assignment is a max-min approach for the sum of the variances.*

In essence, this theorem shows that also matching aims at balancing a higher moment of the covariate distribution than the mean, as does the min MSE approach. Basically, it ensures that the most similar observations are assigned to different groups.

### 2.6.2 First Moment Optimizers

The criteria considered by Greevy et al. (2004) to compare the efficiency of treatment assignment could also be used as an optimization criterion for treatment assignment.

For every treatment $a$, a linear, additive model is specified as follows:

$$Y_{I(a)} = \begin{bmatrix} Z_{I(a)} & X_{I(a)} \end{bmatrix} \begin{bmatrix} \tau^a \\ \beta^a \end{bmatrix} + \varepsilon,$$

where $Z_{I(a)}$ gives the treatment status, i.e., ${Z_{I(a)}}_i = \mathbb{1}_{i \in I(a)}$ for those in treatment group $a$ and ${Z_{I(0)}}_i = -\mathbb{1}_{i \in I(0)}$ for the control group. In particular, the treatment effect is assumed to be constant across individuals, so potential outcomes of the control group and the treatment group of interest are assumed to differ only by a constant.

Under the Gauss-Markov assumptions, i.e. additive errors that are uncorrelated conditional on $X_{I(a)}$ with constant variance $\sigma^2$, the MSE of the estimated treatment effect is proportional to $({Z_{I(a)}}^\top P_a Z_{I(a)})^{-1}$ with $P_a = \mathrm{id} - X_{I(a)} \left( X_{I(a)}^\top X_{I(a)} \right)^{-1} X_{I(a)}^\top$ and is minimized for $X_{I(a)}^\top Z_{I(a)} = 0$ (Greevy et al., 2004). Thus, with the assumption of constant variances across treatments and constant weights, an alternative objective function for minimization would be

$$S^*(\hat{T}) \propto \sum_a ({Z_{I(a)}}^\top P_a Z_{I(a)})^{-1}. \tag{8}$$

In contrast to simple differences of mean estimators for the average treatment effect, here covariates are controlled for. The induces criterion of balance for covariates is minimized by $X_{I(a)}^\top Z_{I(a)} = 0$, i.e., it is enough to have equal mean values of a covariate to minimize this criterion (given equal group sizes), independent of the distribution in the respective groups. Also the approach by Morgan and Rubin (2012) leads to a solution with similar properties. By way of contrast, Section 2.3

shows that if there is reason to assume that any of the treatment effects might differ across individuals and be a function of the covariates, it is necessary to focus on more distributional characteristics of the covariates than their means.

# 3 Application in Two Case Studies

We illustrate the applicability and the benefits of our method with two case studies. Two research teams applied the minMSE method for treatment assignment in two scenarios in which the method is particularly suited: Cluster-randomized settings and settings with several treatment arms. In the first setting, we focus on the impact of attrition on pre-treatment covariate balance and on the ability to detect significant treatment effects. In the second setting, we show how balance varies as, for a constant sample, the number of treatment groups increases. Riener et al. (2021) experimentally investigate how balance affects precision in a similar setting, using the minMSE method. For the standard case of one control and one treatment group, we have also run simulations on five different datasets in the spirit of Bruhn and McKenzie (2009). These results are not reported here in the interest of brevity, but are available from the authors. They show that our method performs superior to all alternatives but the optimal matching approach, to which it performs comparable.

## 3.1 Study 1: Cluster Randomized Health Intervention

The goal of the intervention study in Indonesia was to introduce and assess the effectiveness of the World Health Organisation's Safe Childbirth Checklist in Aceh Province in Indonesia by use of a Randomized Controlled Trial (Diba et al., 2020). Outcomes of interest were the maternal and neonatal mortality rate and the stillbirth rate. The study shows that neonatal mortality rate and the stillbirth could both be

significantly lowered ($\alpha = 10\%$) at the birth level in hospitals applying the Safe Childbirth Checklist.

Treatment assignment was conducted at the health facility level (using a phase-in design) for ethical and practical reasons. Stratification was discarded, as more than a handful of variables were identified that were likely to influence the outcome and should hence be balanced across treatment groups, such as the cluster size as expressed in yearly delieveries, the location (district and rural/urban) or the capabilities of a facility (e.g. blood transfusion), and moreover, some of them were continuous, such as the yearly number of delieveries of a hospital.

Pair-wise matching is a common solution in that case. However, the 32 health facilities differed considerably in the number of yearly deliveries: In the year prior to the study, this number ranged from 5 to 3220 with a mean of about 360 and a median of 85. The number of delieveries in the largest hospital amounted to roughly 30% of all delieveries in the year prior to the intervention, and, together with its most likely paired facility, even to about 40% of delieveries. Thus, attrition was of major concern, and in particular that one of the large units could have dropped out of the study between treatment assignment and final measurements. Pair-wise matching would have caused another large unit to be taken out of the study in such a situation. For this reason, the research team opted for the minMSE method as an attrition tolerant option for treatment assignment.

We illustrate the impact of attrition on balance and the power to detect significant treatment effects by contrasting pair-wise matching and the minMSE method. For the pair-wise matching approach, we picked the 'greedy' approach by Imai et al. (2009a) in the R package `experiment` as well as the optimal matching algorithm going back to Greevy et al. (2004), and used the implementation in the R package of Lu et al. (2011). For the minMSE approach we used our own Stata ado-package

(Schneider, 2021), with 300 iterations for actual treatment implementation, and our own R package `minMSE` (Schneider and Baldini, 2021) for the simulation study.

**Attrition and Balance of Experimental Groups Before Treatment** To study the effect of attrition on balance, we produce 1000 treatment assignment vectors either purely at random, with the minMSE method or with the two pair-wise matching approaches. For treatment assignment, we consider a set of seven categorical or continous pre-treatment variables containing information on the cluster size, the location, the type (private or public) and the capabilities of a facility. For each treatment assignment vector and each of these variables, we assess balance by the difference in a variable's mean values in the two assigned treatment groups, expressed in standard deviations.

Dropouts are simulated: For one and two facilities, this is done exhaustively by dropping every facility and every possible combination of two facilities once. For more than two facilities, 10000 combinations of facilities are randomly sampled. For every dropout or combination of dropout, we consider the 95% quantile of group differences across the seven variables and the 1000 treatment assignments. For every number of dropout facilities, these are averaged over the dropouts. Importantly, when treatment assignment is conducted with a pair-wise matching approach, we also remove the pair of a dropout from the sample, according to common praxis.

The results confirm the hypothesis that taking out the pair of a dropped out facility allow a relatively stable degree of balance for the matching approaches: The 95% quantile of differences in group means is about .4 and .56 for the optimal and the greedy matching approach, respectively, without attrition and it increases only very moderate to an average of about .5 and .66, thus increases, on average, only by .02 standard deviations per dropped facility. The increase in imbalance with the minMSE approach is about twice as large. However, as it starts with a 95% quantile

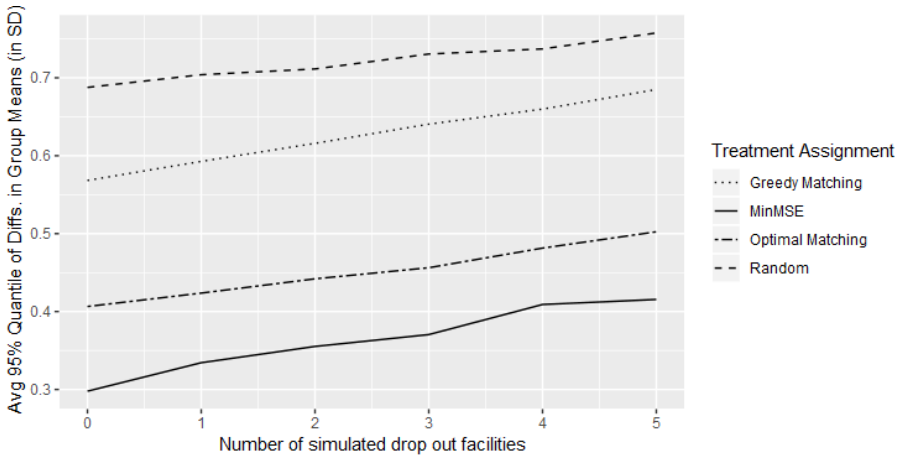


Figure 1: The effect of attrition on imbalance (differences in group means)

of differences in group means of only about .13 SD and increases to .33, the increase is not a threat to balance, as even with five facilites dropping out, balance is still better than with the matching approaches without any attrition; see Figure 1.

**Attrition and the Ability to Detect Treatment Effects** Actual treatment assignment was performed by the minMSE method. However, when building pairs of health facilties with the optimal matching approach, the two facilities in each pair were – by the minMSE approach – always assigned to different treatment groups, except for two of the pairs. In other words, for a subsample excluding these two pairs, the actual treatment assignment could have been a realisation of both, treatment assignment using optimal pair-wise matching or the minMSE method. To be able to use actual treatment effects instead of simulated or predicted ones, we focus on this subsample in this case study. We do so to abstract from issues of pre-treatment balance that otherwise might differ in realizations of the two methods, and to avoid predicting potential outcomes based on rather few data and/or strong assumptions.

For estimation of treatment effects in matched-pair and unmatched cluster-randomized experiments, we use the same estimators as in Imai et al. (2009a):

$$\hat{\theta} = \frac{2}{N} \sum_{i=1}^{N} Y_{i,1}^{p} D_i - (1 - D_i) Y_{i,0}^{p} \tag{9}$$

is used as an unbiased estimator for the sample average treatment effect for the unmatched design, i.e. when assigning treatment with the minMSE method.For the matched-pair design, we denote the number of units in the $i$th cluster of the $k$th pair with $n_{ik}$, where $i = 1, 2$ and $k = 1, \ldots, m$. $Z_k$ is the treatment status for the first cluster of the $k$th pair, i.e. for $Z_k = 1$ we have $D_1 k = 1$ and $D_2 k = 0$ and define

$$\begin{aligned} \hat{\kappa} = \frac{1}{\sum_{k=1}^{m} w_k} \sum_{k=1}^{m} w_k \Bigg\{ Z_k \left( \frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} - \frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} \right) \\ + (1 - Z_k) \left( \frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} - \frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} \right) \Bigg\}, \end{aligned} \tag{10}$$

where, following Imai et al. (2009a), for $w_k$ we used either harmonic means of the cluster sizes within a pair, $w_k = n_{i1k} n_{i2k} / (n_{i1k} + n_{i2k})$ or just $n_{i1k} + n_{i2k}$, i.e., their total size. We focus on these design-based, model-free estimators for their simplicity, and to abstract from model misspecifications, particularly so with our small sample. See Freedman (2008), Lin (2013) and Imai et al. (2009b) for discussion of the otherwise possible complications in our setting.

To keep the method of inference comparable for both methods of treatment assignment, to refrain from distributional assumptions that may or may not hold in small samples, and to abstract from complications and controversies when estimating the variance and the standard error of the treatment effect resulting from a matching design (see, e.g. Imbens, 2011; Abadie and Imbens, 2006; Klar and Donner, 1997; Imai et al., 2009a, on this topic), we compute p-values using randomization inference. We follow the recommendations by Morgan and Rubin (2012), and create alternative treatment assignment vectors that could have arisen from the treatment assignment

method. For the minMSE method, these are all treatment assignments that result from 300 iterations of the algorithm using the same covariates that were used for the original treatment assignment. Similarly, for the pair-wise matching approach, we randomize within pairs as established from the non-stochastic matching algorithm (Greevy et al., 2004; Lu et al., 2011). Then, the p-value is the proportion of alternative treatment assignment vectors with corresponding estimated treatment effect as extreme or more extreme than the actually observed treatment effect resulting from the actual treatment assignment.

Attrition actually happened during the intervention and one hospital had to close before the end of the study. Naturally, however, we cannot assess the effect of this hospital dropping out of the study on the ability to detect a significant treatment effect. Therefore, we simulate attrition, and compare the significance of the estimators (9) and (10) with randomization inference and one to five health facilities taken out of the sample. For the matching-pair cluster randomized design estimator (10), we also remove the paired health facility from the sample. Just as before, we simulated attrition exhaustively for the first two hospitals to drop out with every possible combination of two hospitals. For simulating attrition with three to five facilities, we sampled 10,000 combinations of health facilities, take them, and, if applicable, their pair out of the sample, estimate the treatment effect and assess its significance.

**Attrition and the Ability to Detect Treatment Effects** Without attrition, both estimators perform comparable in the sense that a significant treatment effect can be detected in both outcome variables of interest that we considered: the neonatal death rate, and the stillbirth rate. The minMSE method and the associated estimator demonstrate a higher ability to detect significant treatment effects at all levels of attrition compared to the matching approach and the associated estimators. While the difference is moderate for one hospital dropping out of the sample (significance in

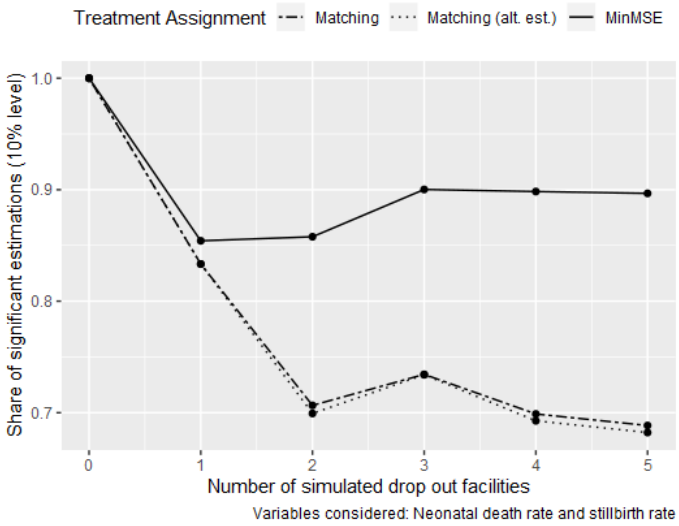


Figure 2: The effect of attrition on the ability to detect treatment effects

one variable in one case of the 26 possibilities of dropping a first facility), it becomes larger in case two hospitals are missing in the endline: In 167 comparisons, or 13% of all performed estimations when iteratively dropping all possible combinations of two hospitals, no significant difference between treatment and control in the considered variables can be detected after using the matching approach, while this is still possible if using the minMSE approach. For higher levels of attrition, the results are similar, as apparent from Figure 2. It should be noted, of course, that these results are partly due to the small sample size and the rather low baseline variance. Yet, exactly in these settings, treatment assignment matters, and might make a difference.

## 3.2 Study 2: Experiment with Multiple Treatments

The minMSE method has also been used in a social science experiment where participants need to be assigned to one of three treatments. The main goal of the study is to understand the underpinnings of effort provision among pupils (Bašić et al., 2021).

The primary outcome is the effort level as measured by the number of successfully solved tasks in a counting exercise. The authors test the effect of each treatment in improving this primary outcome. In this case study we illustrate the advantages of the minMSE method in this setting, using data from pilot sessions (N = 102; data collection for the main study is still ongoing).

Treatment is assigned at the individual level. Stratification was discarded as assignment method, as several – some of them continuous – variables were identified that were likely to influence the outcome. Pair-wise matching is limited to the situation of only one binary treatment. Therefore, and for the lack of any software implementation of a (formal) rereandomization method, we compare the minMSE method with purely random treatment assignment in this setting, again using our own R package (Schneider and Baldini, 2021).

To study (im)balance of pre-treatment variables, we produce 1000 treatment assignment vectors either purely at random or with the minMSE method, assigning individuals to two to ten experimental groups. For treatment assignment, we consider a set of eight either categorical or continous pre-treatment variables containing individual information on the pre-treatment effort level provided, the IQ, patience, risk tolerance, age and gender. For each treatment assignment vector and each of these variables, we assess balance by the average difference in a variable's mean value between the control group and any of the treatment groups, expressed in standard deviations. For every number of treatment groups to assign, we finally consider the 95% quantile of average differences between control and treatment groups across the eight variables and the 1000 treatment assignments.

Results show that when assigning treatment groups purely at random, for every number of treatment groups to assign, imbalance is by between 41 to 120%, or, on average 56% higher than when conducting treatment assignment with the minMSE

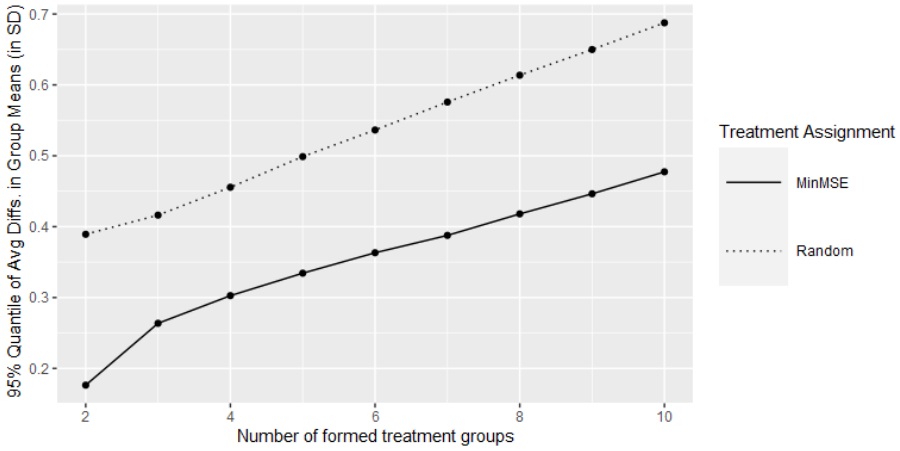


Figure 3: Imbalance in pre-treatment information with multiple treatment groups

method. This means that, keeping the size of the sample constant, the minMSE method can assign five more groups than purely random treatment assignment with about the same level of imbalance; see Figure 3. As the increase in imbalance associated with an additional treatment group when assigning more than two groups is clearly lower with the minMSE method than when assigning treatment groups purely at random, we can expect this result to hold beyond the results presented here, i.e. also for six groups (random assignment) vs. eleven (minMSE method), etc.

# Appendix

## A Proofs

*Proof of Theorem 1.* Let $Y^{p,a}_{I(a),k}$ be the subvector of the observed $k$-th component under treatment $a$ and $\varepsilon^a_{I(a),k}$ be the respective subvector of error terms. That is, equality (2) writes in matrix notation $Y^{p,a}_{I(a),k} = X^\top_{I(a)}\beta^a_k + \varepsilon^a_{I(a),k}$ for any $a$ and $k$. It is well-known that the best unbiased estimator $\hat{\beta}^a_k$ for $\beta^a_k$ has the distribution

$$\hat{\beta}^a_k \sim \mathcal{N}(H_a\beta^a_k, \sigma^2_{ak}(X_{I(a)}X^\top_{I(a)})^{-1}).$$

and $\hat{Y}^{p,a}_{I(a),k}$ is best chosen as $\hat{Y}^{p,a}_{I(a),k} = X^\top_{I(a)}\hat{\beta}^a_k$. Hence,

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\tau}^a_{t,k} - \tau^a_k\right)^2 \mid X\right] - (\bar{X}^\top((\mathrm{id} - H_a)\beta^a_k - (\mathrm{id} - H_0)\beta^0_k))^2 \\
&\quad = \mathbb{E}\left[\left(\frac{1}{N}\sum_i \left(\hat{Y}^{p,a}_{i,k} - \hat{Y}^{p,0}_{i,k}\right) - \frac{1}{N}\sum_i \left(E\left[Y^{p,a}_{i,k} \mid X_i\right] - E\left[Y^{p,0}_{i,k} \mid X_i\right]\right)\right)^2 \mid X\right] \\
&\quad = \frac{1}{N^2}\mathbb{E}\left[\left(\sum_i X_i^\top\left(\left(\hat{\beta}^a_k - \beta^a_k\right) - \left(\hat{\beta}^0_k - \beta^0_k\right)\right)\right)^2 \mid X\right] \\
&\quad = \frac{1}{N^2}\sum_i X_i^\top\left(\mathrm{Cov}\left(\hat{\beta}^a_k - \beta^a_k \mid X\right) + \mathrm{Cov}\left(\hat{\beta}^0_k - \beta^0_k \mid X\right)\right)\sum_i X_i \\
&\quad = \bar{X}^\top\left(\sigma^2_{ak}(X_{I(a)}X^\top_{I(a)})^- + \sigma^2_{0k}(X_{I(0)}X^\top_{I(0)})^-\right)\bar{X},
\end{aligned}$$

where we used that the error terms $\varepsilon^a_{i,k}$, $a > 0$ are i.i.d. and $I(a)$ depends only on $X$.

Now denote the $l_1$ norm of a vector by $\|\cdot\|_1$ and summarize weights and scaling factors for the variance as $\tilde{w}_k = w_k s_k^0$ and $\tilde{v}_k^a = v_k^a s_k^a$. Then, applying the just derived result to the objective function, the generalized MSE (1), completes the proof:

$$
\begin{aligned}
S(\hat{T}) &= \sigma_0^2 \bar{X}^\top \left[ \sum_k \left\{ w_k s_k^0 \left( \sum_a v_k^a s_k^a \left( X_{I(a)} X_{I(a)}^\top \right)^- + \|v\|_1 \left( X_{I(0)} X_{I(0)}^\top \right)^- \right) \right\} \right] \bar{X} \\
&\quad + \sum_{a>0} v_a \sum_k w_k (\bar{X}^\top ((\mathrm{id} - H_a)\beta_k^a - (\mathrm{id} - H_0)\beta_k^0))^2 \\
&= \sigma_0^2 \bar{X}^\top \left[ \|\tilde{w}\|_1 \|v\|_1 \left( X_{I(0)} X_{I(0)}^\top \right)^- + \sum_k \left\{ \tilde{w}_k \sum_a \tilde{v}_k^a \left( X_{I(a)} X_{I(a)}^\top \right)^- \right\} \right] \bar{X} \\
&\quad + \sum_{a>0} v_a \sum_k w_k (\bar{X}^\top ((\mathrm{id} - H_a)\beta_k^a - (\mathrm{id} - H_0)\beta_k^0))^2
\end{aligned}
$$

□

**Lemma 1.** *Assume that* $\sum_j |\mathbb{E}X_{j1}| > 0$, *and* $M_{\alpha,j} = \mathbb{E}|X_{j1}|^{2+\alpha} \in (0,\infty)$ *for some* $\alpha > 0$ *and all* $j$. *Assume further that* $X_{ji}$ *and* $X_{li}$ *are uncorrelated for all* $j \neq l$ . *Let* (4) *hold. Let* $c_0 = \|\tilde{w}\|_1 \|v\|_1$ *and* $c_a = \sum_k \tilde{w}_k \tilde{v}_k^a$ *for* $a > 0$. *Let* $A$ *be any assignment that is in the limit at least as good with respect to the minimization problem as any assigment that is independent of* $X$, *given* (4) *holds. Denote by* $|I(a)|$ *the number of elements in* $I(a)$. *Then,* $N^{-1}|I(a)| \to \sqrt{c_a}/\sum_{b\geq 0}\sqrt{c_b}$ *for all* $a \geq 0$ *as* $N \to \infty$ *(and consequently,* $\frac{1}{|I(a)|}X_{I(a)}X_{I(a)}^\top \to (\mathbb{E}X_{j1}X_{l,1})_{j,l=1,\ldots,r}$*).*

*Proof.* Let $f(z_0,\ldots z_d) = \sum_a c_a/z_a$ for $z_a \geq 0$. Under the constraint that $\sum z_a \leq 1$ the function $f$ takes its infimum at $z^0 = (z_0^0,\ldots,z_d^0) = (\sqrt{c_0},\ldots,\sqrt{c_d})/\sum_a \sqrt{c_a}$. In particular, $z_a^0 \in (0,\infty)$ for all $a$ and $f$ is continuous in an environment around $z^0$. Let $\eta = (\mathbb{E}X_{11},\ldots,\mathbb{E}X_{r1})$, $M = (\mathbb{E}X_{j1}X_{l1})_{j,l=1,\ldots,r}$ and $\xi = \eta^\top M^{-1}\eta$. By assumption the covariance matrix of $(X_{11},\ldots,X_{r1})$ is a strictly positive definite matrix and hence $M$ and $M^-$ are strictly positive definite matrices and $\xi \in (0,\infty)$.

Let $A$ be an assignment independent of $X$, for which $N^{-1}|I(a)| \to \sqrt{c_a}/\sum_{b\geq 0}\sqrt{c_b}$ for all $a \in \{0,\ldots,d\}$ as $N \to \infty$. As $|I(a)| \to \infty$, we have

$$\frac{1}{|I(a)|}X_{I(a)}X_{I(a)}^\top \to (\mathbb{E}X_{j1}X_{l1})_{j,l=1,\ldots,r}$$

and $\bar{X} \to (\mathbb{E}X_{11},\ldots,\mathbb{E}X_{r1})$. Then

$$\begin{aligned}\lim_{N\to\infty} NS(T_0) &= \lim_{N\to\infty} \sigma_0^2 N\bar{X}^\top \left[\|\tilde{w}\|_1\|v\|_1 \left(X_{I(0)}X_{I(0)}^\top\right)^- + \sum_k \left\{\tilde{w}_k \sum_{a>0} \tilde{v}_k^a \left(X_{I(a)}X_{I(a)}^\top\right)^-\right\}\right]\bar{X}, \\ &= \sigma_0^2\xi f(z^0) \in (0,\infty). \end{aligned} \tag{11}$$

We stick to the pseudoinverse to see clearer where condition (4) comes in. For ease of notation we will consider below only the whole sequence, but all considerations can also be applied to subsequences. Assume now that we have $A$ that is in the limit at least as good as the assignment above, i.e., (11) yields an upper bound for the limit behaviour of $NS(T)$. Then $N\bar{X}^\top \left(X_{I(a)}X_{I(a)}^\top\right)^- \bar{X}$ is bounded for all $a \geq 0$.

Let

$$B_N := \frac{1}{N}(X_{I(a)}X_{I(a)}^\top)_{jj} = \frac{|I(a)|}{N}\frac{(X_{I(a)}X_{I(a)}^\top)_{jj}}{|I(a)|}$$

and let $F_j$ be the distribution function of $X_{j1}$. Then for $u \geq 1$,

$$\int_{\{y:|y|\geq u\}}^{\infty} |x|^2 F_j(dx) \leq u^{-\alpha}\int_{\{y:|y|\geq u\}}^{\infty} |x|^{2+\alpha} F_j(dx).$$

Hence for all $u \geq 0$

$$\int_{\{y:|y|\geq u\}}^{\infty} |x|^2 F_j(dx) \leq \min\{M_{0,j}, u^{-\alpha}M_{\alpha,j}\}.$$

Assume that $A^*$ assigns to the group $a \geq 0$ the largest $|I^*(a)|$ values of $|X_{j1}|, \ldots, |X_{jN}|$ with $|I^*(a)| \geq |I(a)|$ and $|I^*(a)| \to \infty$. Let $F_j^{\leftarrow}$ the pseudoinverse of $F_j$. Then

$$\begin{aligned}\lim_{N\to\infty} B_N &\leq \lim_{N\to\infty}\frac{(X_{I^*(a)}X_{I^*(a)}^\top)_{jj}}{N} = \lim_{N\to\infty}\frac{|I^*(a)|}{N}\lim_{N\to\infty}\frac{(X_{I^*(a)}X_{I^*(a)}^\top)_{jj}}{|I^*(a)|} \\ &\leq \min\left\{M_{0,j}, \left(\lim_{N\to\infty} F_j^{\leftarrow}(1-|I^*(a)|/N)\right)^{-\alpha} M_{\alpha,j}\right\}\lim_{N\to\infty}\frac{|I^*(a)|}{N}. \quad (12)\end{aligned}$$

Hence, the eigenvalues of $N^{-1}X_{I(a)}X_{I(a)}^\top$ are bounded in the limit. Hence the eigenvalues of the (pseudo-)inverse are bounded away from zero in the limit, exept they had been zero already in the original matrix. Hence, $N\bar{X}^\top\left(X_{I(a)}X_{I(a)}^\top\right)^-\bar{X}$ can converge to 0 if and only if $\eta$ is in the kernel of $\lim_{N\to\infty} I^{-1}(a)X_{I(a)}X_{I(a)}^\top$, since $\eta \neq 0$. Thus, we have a contradiction to (4). If $N\bar{X}^\top\left(X_{I(a)}X_{I(a)}^\top\right)^-\bar{X}$ does not converge to zero, then at least for one $j$ the value of $B_N$ must be greater than 0 in the limit. Replacing $I^*$ by $I$ in (12) we get that $I(a) = O(N)$ and the same argument in (11), now for the assignment $A$, shows the assertion of the Lemma.

□

*Proof of Proposition 1.* Instead of $C = \mathrm{diag}(1, \ldots, 1, c, 1, \ldots, 1)$ we consider an arbitrary $r \times r$ matrix $C$ that is invertible. Let $W = CX$ and $I$ a non-empty subset $\{1, \ldots, N\}$. Then

$$\bar{W}^\top(W_I W_I^\top)^{-1}\bar{W} = \bar{X}^\top C^\top(CX_I X_I^\top C^\top)^{-1}C\bar{X} = \bar{X}^\top(X_I X_I^\top)^{-1}\bar{X}.$$

□

**Lemma 2.** *Let $\eta \in \mathbb{R}^r$ and $v_1, \ldots, v_r > 0$. Then*

$$\eta^\top (\eta\eta^\top + \mathrm{diag}(v_1, \ldots, v_r))^{-1}\eta = \frac{s}{1+s} \qquad \textit{with } s = \sum_{j=1}^{r} \eta_j^2 / v_j$$

*Proof.* If $\eta = 0$ then the assertion is correct. Otherwise, assume without loss of generality that $\eta_r \neq 0$. Let $x = \eta^\top (\eta\eta^\top + \mathrm{diag}(v_1, \ldots, v_r))^{-1}\eta$. Then

$$x = \eta^\top u \quad \text{ with } (\eta\eta^\top + \mathrm{diag}(v_1, \ldots, v_r))u = \eta.$$

Then $u$ solves

$$\begin{pmatrix} v_1 & 0 & \ldots & 0 & -v_r\eta_1/\eta_r \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \ldots & 0 & v_{r-1} & -v_r\eta_{r-1}/\eta_r \\ \eta_r\eta_1 & \eta_r\eta_2 & \ldots & \eta_r\eta_{r-1} & v_r + \eta_r^2 \end{pmatrix} u = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ \eta_r \end{pmatrix}$$

Hence the $j$th component of $u$ equals $u_j = \eta_j v_j^{-1}(1 + \sum_{l=1}^{r} v_l^{-1}\eta_l^2)^{-1}$. □

*Proof of Proposition 2.* Lemma 1 shows that

$$\frac{|I(a)|}{N}\frac{1}{|I(a)|}X_{I(a)}X_{I(a)}^\top \to \sqrt{c_a}/\sum_{b\geq 0}\sqrt{c_b}(\mathbb{E}X_{j,1}X_{l,1})_{j,l=1,\ldots,r} \qquad (N\to\infty) \qquad \text{for all } a \geq 0$$

Let $\eta = \mathbb{E}(X_{1,1}, \ldots, X_{r,1})$ and $v_j = \mathrm{Var}(X_{j,1})$. Then Lemma 2 yields that

$$N\bar{X}^\top \left[ d\left(X_{I(0)}X_{I(0)}^\top\right)^{-1} + \sum_{a>0}\left(X_{I(a)}X_{I(a)}^\top\right)^{-1} \right] \bar{X} \to dc_0^{-1/2}\sum_{b\geq 0}\sqrt{c_b}\frac{s}{1+s} + \sum_{a>0}c_a^{-1/2}\sum_{b\geq 0}\sqrt{c_b}\frac{s}{1+s} \tag{13}$$

where $s = \sum_{j=1}^{r} \eta_j^2 / v_j$. A finite approximation of the right hand side of (13) is

$$d \frac{\sum_{b \geq 0} \sqrt{c_b}}{\sqrt{c_0}} \frac{s_0}{1 + s_0} + \sum_{a > 0} \frac{\sum_{b \geq 0} \sqrt{c_b}}{\sqrt{c_a}} \frac{s_a}{1 + s_a} \tag{14}$$

where $s_a = \sum_{j=1}^{r} \hat{\eta}_j^2 / \hat{v}_{j,I(a)}$ and $\hat{\eta}_j^2$ is the sample mean, and $\hat{v}_{j,I(a)}$ are the sample variance of the $\{X_{j,i} : i \in I(a)\}$. □

*Proof of Proposition 3.* It is known that the diagonal elements of the covariance matrix of the estimator for the parameter vector in linear regression models such as (2) are given by

$$\mathrm{Var}(\hat{\beta}_{k,j}^a \,|\, X) = \frac{\sigma_{ak}^2}{(1 - R_{a,k,j}^2) \sum_{i=1}^{n} (X_{I(a)j,i} - \overline{X_{I(a)j}})^2},$$

for $j = 1, \ldots, r$, with notations as in Proposition 3 (see e.g. Wooldridge, 2014).

As in the proof of Theorem 1, the covariance matrix of $\hat{\beta}_k^a - \hat{\beta}_k^0$ for any $a = 1, \ldots, d$ and any $k = 1, \ldots, m$ is given by

$$\left(X_{I(a)} X_{I(a)}^\top\right)^{-1} + \left(X_{I(0)} X_{I(0)}^\top\right)^{-1},$$

the claim follows, noting that we assume equal variances in Corollary 1. □

*Proof of Proposition 4.* This criterion, which aims at minimizing $\mathrm{Var}\, c^\top \theta$, i.e. $\arg\min_\zeta \Phi(C_\zeta^{-1}) = \arg\min_\zeta c^\top C_\zeta^{-1} c$, $C = XX^\top$ is the so-called *information matrix*. In case of our linear model 2 $C_\zeta^{-1}$ of $\hat{\theta}(\zeta)$ is minimized. Since $C_\zeta^{-1}$ is a matrix, the optimization problem

$$\arg\min_\zeta \Phi(C_\zeta^{-1}),$$

is considered is a special case of the min MSE Treatment Assignment procedure, as the following proposition shows.

Assume $d = 1$ and $m = 1$. We start by noting that we can write the equations that we need to solve in order to estimate the linear model (2) in one single equation system (without loss of generality assumed to be in block-diagonal form):

$$\begin{pmatrix} Y^{p,0}_{I(0)} \\ Y^{p,1}_{I(1)} \end{pmatrix} = \begin{pmatrix} X^\top_{I(0)} & 0 \\ 0 & X^\top_{I(1)} \end{pmatrix} \theta, \tag{15}$$

where $\theta = \begin{pmatrix} \beta^0 \\ \beta^1 \end{pmatrix}$ and $Y^{pa}_{I(a)}, X_{I(a)}, \beta^a$ for $a \in \{0, 1\}$ are vectors and matrices as defined in the proof of Theorem 1. Recall that—with only one treatment group—we are interested in minimizing $\mathbb{E}[(\hat{\tau}^a - \tau)^2 \,|\, X] = \mathrm{Var}[\hat{\tau}^a \,|\, X]$. Now

$$\hat{\tau}^a = \frac{1}{N} \sum_{i=1}^{N} X_i^\top (\hat{\beta}^1_p - \hat{\beta}^0_p) = c^\top \hat{\theta}$$

for $c = (\frac{1}{N}\sum_{i=1}^N X_i^\top,\ -\frac{1}{N}\sum_{i=1}^N X_i^\top)^\top$. Finally,

$$\mathrm{Var}[c^\top \hat{\theta} \,|\, X] = c^\top C_\zeta^{-1} c \tag{16}$$

$$= \sigma^2 \frac{1}{N} \sum_{i=1}^{N} X_i^\top \left( (X_{I(0)} X^\top_{I(0)})^{-1} + (X_1 X_1^\top)^{-1} \right) \frac{1}{N} \sum_{i=1}^{N} X_i, \tag{17}$$

with $C_\zeta = \sigma^{-2} \left( \begin{pmatrix} X_{I(0)} & 0 \\ 0 & X_1 \end{pmatrix} \begin{pmatrix} X_{I(0)} & 0 \\ 0 & X_1 \end{pmatrix}^\top \right)$, where, in our case,

$$\zeta = \begin{Bmatrix} z_{1,0} & z_{1,1} & z_{2,0} & \cdots & z_{N,1} \\ p_{1,0} & p_{1,1} & p_{2,0} & \cdots & p_{N,1} \end{Bmatrix} \tag{18}$$

with $\sum_{i,j} p_{i,j} = 1$, $p_{i,j} \in \{1/N, 0\}$, $p_{i,0} + p_{i,1} = 1/N$, $z_{i,0}^\top = (X_i^\top, 0, \ldots, 0)$ and $z_{i,1}^\top = (0, \ldots, 0, X_i^\top)$ for all $i$. □

*Proof of Theorem 2.* We use $m_{i,j} = m_{j,i} = 1$ to indicate individual $i$ is matched to individual $j$, and 0 otherwise. Every individual is matched exactly once, so

$\sum_{i,j} m_{i,j} = N$. Usually, the goal is to minimize $\sum_{i,j} m_{i,j}(y_i - y_j)^2$ through the choice of $m_{i,j}$, although sometimes the absolute difference is also used (Rubin, 1973). For being a special case of the squared Mahalanobis distance, we prefer the squared euclidean distance. The set of solutions to this optimization problem is given by

$$\underset{(m_{i,j})_{i<j}}{\arg\min} \sum_{i,j} m_{i,j}(y_i^2 + y_j^2 - 2y_i y_j) = \underset{(m_{i,j})_{i<j}}{\arg\max} \sum_{i,j} m_{i,j} y_i y_j.$$

We now show that elements of this set maximize the minimal sum of the variances of the groups to be created. This sum of variances is given by

$$\frac{2}{N} \sum_{i \in I(0)} y_i^2 - \bar{y}_0^2 + \frac{2}{N} \sum_{j \in I(1)} y_j^2 - \bar{y}_1^2 = \frac{2}{N} = \sum_i y_i^2 - \bar{y}_0^2 - \bar{y}_1^2. \tag{19}$$

Since

$$\overline{y_{I(a)}}^2 = \frac{4}{N^2} \left( \sum_{i \in I(a)} y_i^2 + \sum_{i \in I(a)} \sum_{j \in I(a), j \neq i} y_i y_j \right)$$

for $a = 0, 1$, (19) can be rewritten as

$$\left( \frac{2}{N} - \frac{4}{N^2} \right) \sum_i y_i^2 - \frac{4}{N^2} \left( \sum_{i \in I(0)} \sum_{j \in I(0), j \neq i} y_i y_j + \sum_{i \in I(1)} \sum_{j \in I(1), j \neq i} y_i y_j \right).$$

The first summand is independent of group or treatment assignment. We rewrite the elements of the subtrahend as

$$\sum_{i \in I(0)} \sum_{j \in I(0), j \neq i\}} y_i y_j + \sum_{i \in I(1)} \sum_{j \in I(1), j \neq i} y_i y_j \tag{20}$$

$$= \sum_i \sum_j y_i y_j - \sum_i y_i^2 - 2 \sum_{i \in I(0)} \sum_{j \in I(1)} y_i y_j \tag{21}$$

$$= \sum_i \sum_j y_i y_j - \sum_i y_i^2 - 2 \sum_{i \in I(0)} \sum_{j \in I(1)} m_{i,j} y_i y_j - 2 \sum_{i \in I(0)} \sum_{j \in I(1)} (1 - m_{i,j}) y_i y_j, \tag{22}$$

where we have split the cross product between group observations into those that are matched and those that are unmatched. The first two parts are again independent of

group or treatment assignment, and so is the third for a fixed $m$. Thus, by matching we have maximized the sum of group variances across feasible treatment assignments. In other words, the sum of group variances resulting from the worst treatment assignment in this aspect from the set of possible treatment group assignments after matching $\{A : A_i = |A_j - 1| \text{ for } \hat{m}_{i,j} = 1, \hat{m} \in \arg\max_{(m_{i,j})} \sum_i \sum_j m_{i,j} x_i x_j\}$ is still maximized over $m$. □

# References


Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Bašić, Z., S. Bortolotti, D. Salicath, S. O. Schneider, S. Schmidt, and M. Sutter (2021). Heterogeneity in effort provision: Evidence from a lab-in-the-field experiment. AEA RCT Registry, Trial No. AEARCTR-0008360. Mimeo.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433–1436.

Diba, F., Ichsan, Muhsin, Marthoenis, K. Richert, L. Kaplan, S. Susanti, M. Andalas, H. Sofyan, Samadi, and S. Vollmer (2020). Impact of the WHO Safe Childbirth Checklist on quality of care and birth outcomes: a cluster-randomized controlled trial in Aceh, Indonesia. AEA RCT Registry, Trial No. AEARCTR-0003548. Mimeo.

Fedorov, V. and P. Hackl (1997). *Model-Oriented Design of Experiments*. New York: Springer.

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

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

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

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

Imai, K., G. King, and C. Nall (2009b). Rejoinder: Matched pairs and the future of cluster-randomized experiments. *Statistical Science 24*(1), 65–72.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics 86*(1), 4–29.

Imbens, G. W. (2011). Experimental design for unit and cluster randomid trials. Technical report, Harvard University.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis 24*(3), 324–338.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological) 21*(2), 272–319.

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

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

Lu, B., R. Greevy, X. Xu, and C. Beck (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician 65*(1), 21–30.

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

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science 349*(6251), aac4716.

Riener, G., S. O. Schneider, and V. Wagner (2021). Addressing validity and generalizability concerns in field experiments. Discussion Paper 2020/16, Max Planck Institute for Research on Collective Goods.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics 29*(1), 159–183.

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

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics 2*(1), 1–26.

Schneider, S. O. (2021). MINMSE: Stata module to create balanced groups for treatment in experiments with one or several treatment arms. Statistical Software Components S458939, Boston College Department of Economics.

Schneider, S. O. and G. Baldini (2021). *minMSE: Implementation of the minMSE treatment assignment method for one or multiple treatment groups.* R package version 0.1.1.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika 12*, 1–85.

Wooldridge, J. M. (2014). *Introductory Econometrics: A Modern Approach* (5 ed.). Mason: Cengage Learning.