A New Approach to Treatment Assignment for One and Multiple Treatment Groups

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

November 27, 2019

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 (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 parameter being specified by the researcher and allows multiple treatments. The information used for treatment assignment can be multivariate, 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 simulated treatment effect scenarios and datasets, comparing it to established approaches. 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 and R implementation of our method.

Keywords: Optimal Experimental Design, Randomized Controlled Trial, Rerandomization, Randomization

*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. Already Fisher (1935), one of the first scholars to investigate the topic, concludes that neither randomization alone nor an unbiased experiment "ensure the validity of the estimates, [...] for it might well be that some unknown circumstance, such as [...] different illumination [...], might systematically favour all the plants on one [plot] over those on the other."

Today, in the social sciences, we allocate units to treatment groups instead of seeds to treatment and control plots as in Darwin's experiment, which Fisher (1935) analyzed. Yet, this insight is still true: valid estimates root in similar or *balanced* treatment groups in the absence of the treatment. 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.

Another practice that has gained the reputation of increasing replicability are pre-analysis plans. Researchers specify the analysis they are planning to carry out before conducting the experiment. Commonly, this involves a regression model for the conditional average treatment effect, i.e. the average treatment effect given the

characteristics of the sample.[1] Based on such a model, it might be of great use to increase the precision of the estimator for the conditional average treatment effect.

In this paper, we present a method to assign experimental units to possibly multiple treatment groups based on sample characteristics. The method unifies both of the above mentioned practices and falls in the category of rerandomization methods. It builds on a theoretically derived statistic linked to the rich literature on experimental design (e.g. Smith, 1918; Kiefer, 1959; Fedorov, 1997), which was first connected with treatment assignment by Kasy (2016).

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 the Bayesian linear model, where for 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. In addition, a guess for R^2, the coefficient of determination, of such a linear regression model must be specified. We think that even for experienced researchers, it is hard to come up with a reasonable guess on these parameters. Of course, one could use a *flat* prior, inducing nearly no prior

[1] Note that the conditional average treatment effect can formally only be estimated if the so called *overlap condition* (Abadie and Imbens, 2006) is satisfied, which can be described as a weak criterion of balance.

information. In this case, however, one can also resign from using prior information, as it simplifies the objective function and consequently the method considerably.

Therefore, we introduce the approach in a frequentist setting. This means that we only get a point estimate instead of a distribution for any result. As the method is designed to minimize the MSE (a point estimate), this comes without limitations.

In the same spirit as Kasy (2016), our statistic 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. Yet, our method allows multiple treatments and multiple outcomes, which both can be weighted. 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.

For the developed method, we provide a software implementation as an ado-package for Stata and R package via CRAN[2] and thus increase its usability in the respective areas of application.

Compared to alternative methods, this method can be applied irrespective of the number of treatment arms, the number of units in the experiment and its relation to the number of treatments and variables (even or uneven, divisibility by the number of treatments, ...). Another feature is its speed: Compared to the Bayesian approach, but also compared to competing methods of treatment assignment, a reasonably good balance for a sample size of 100 units and 10 variables is usually achieved in less than 5 minutes on a 2.3 Ghz dual CPU.

In a simulation study similar to the one by Bruhn and McKenzie (2009), we compare the performance of our min MSE method in various dimensions to competing methods when considering multivariate, including continuous, pre-treatment information and find that it is comparable to the matching methods and superior to stratification or pure randomization. In addition, the min MSE method is tolerant of attrition, i.e. of units whose outcome finally is unobserved, for example because treatment is never received although it was planned.

2 Treatment Assignment

Today, several competing and complementary strategies to account for group characteristics in treatment assignment are widely used, although there is no consent on how treatment assignment should be carried out, even among experts in field research (see e.g. the survey by Bruhn and McKenzie, 2009). Also from a theoretical perspective, a clear answer is missing; see e.g. Imbens (2011) for a brief discussion.

[2]Our R package 'minMSE' is currently under submission at CRAN.

One method is stratification or blocking and goes back to Fisher (1935). The idea is to build subgroups according to observable characteristics and to randomize within those subgroups. Although this method achieves exact balance for binary variables and improves 'balance' in comparison to purely random treatment assignment for other types of variables, it is impractical in several situations. Using stratification, it is only possible to balance a very limited number of variables. Furthermore, continuous variables have to be arbitrarily discretized and are never really balanced with this approach. Additional problems arise in implementing this method when the number of participants is not divisible by the number of subgroups.

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[3] 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., 2009) and an 'optimal matching' algorithm (Greevy et al., 2004; Lu et al., 2011). Matching can be realized with many possible continuous variables and thus eliminates some of the shortcomings of stratification. It's biggest advantage is that the distribution of covariates in the treatment and the control group become as similar as possible – given the 'optimal matching' algorithm is used. 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). Additional problems arise when attrition occurs, i.e. when for some units the outcome finally is unobserved, especially in small samples or when performing randomization at the cluster level: For

[3]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.

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. Additionally, we are unaware of an existing approach to extend matching to multiple treatment arms. 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.

Several so called rerandomization methods have evolved, probably because of the theoretical or practical limitations of the abovementioned approaches. The basic idea of rerandomization is to pick a random treatment assignment in some way, evaluate it with respect to a certain criteria and rerandomize until this criteria meets some condition to be specified or to rerandomize a certain number of times and choose the best assignment, according to a specified evaluation criteria. Sometimes, subjective judgment is also used (Bruhn and McKenzie, 2009). However, we are aware of only one rerandomization approach, the one by Morgan and Rubin (2012), that relies on a theoretical derivation of the statistical threshold to stop the rerandomization. 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, ignoring other dimensions of the distributions of the variables to be balanced. Irrespectively of this limitation, we are unaware of a software implementation of this approach or an extension to multiple treatment arms.

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. The gap of methods for these use cases is filled with our treatment assignment method.

3 The min MSE Treatment Assignment Mechanism

3.1 Framework and Treatment Effect

First, we define the parameter we are interested in estimating: the conditional average treatment effect. We do so by introducing the potential-outcome framework (Rubin, 1974, 1977), as this is the standard notation in the literature on program evaluation (Imbens, 2004). 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, each individual is randomly assigned to an experimental group and treated with the corresponding treatment or not treated at all if assigned to the control group.

Definition 1. Assume we have n_d treatments indicated by $1, \ldots, n_d$ and a control denoted by 0. Let $D_1, \ldots, D_N$ be random variables with values in $\{0, 1, \ldots, n_d\}$. Then, the vector $D = (D_1, \ldots, D_N)^\top$ is called a (random) *treatment group assignment*.

Irrespective of the treatment group assigned to, each participant has potential outcomes, observed outcomes and a vector of pretreatment information, which we call covariates.

Definition 2. Let $X = (X_{j,i})_{j=1,\dots,m;i=1,\dots,N}$ be a random matrix. Then the vector $X_i = (X_{1,i}, \dots, X_{m,i})^\top$ is called the *vector of covariates* of individual i.

Definition 3. Let $Y_i^p = (Y_{i,t}^{p,k})_{t=1,\dots,n_d;k=1,\dots,n_y}$ be a random matrix for $i = 1, \dots, N$. Then the row vector $Y_{i,t}^p = (Y_{i,t}^{p,1}, \dots, Y_{i,t}^{p,n_y})$ is called the vector of *potential outcomes* of individual i in the case of treatment t, where n_y is the number of outcomes of interest.

These *potential outcomes* of individual i in the case of treatment t exist irrespective of whether individual i was actually treated with treatment t or not. However, for every unit and outcome of interest, we only observe the *realized outcome*.

Definition 4. Let $Y^r = (Y_i^{r,k})_{i=1,\dots,N;k=1,\dots,n_y}$ be a random matrix. Then the row vector $Y_i^r = (Y_i^{r,1}, \dots, Y_i^{r,n_y})$ is called the vector of *realized outcomes* of individual i.

The realized outcomes Y_i^r of individual i can be written by means of potential outcomes and the treatment group assignment:

$$Y_i^r = \sum_{t=0}^{n_d} \mathbb{1}_{\{D_i=t\}} Y_{i,t}^p = Y_{i,0}^p + \sum_{t=0}^{n_d} \left(Y_{i,t}^p - Y_{i,0}^p\right) \mathbb{1}_{\{D_i=t\}}.$$

The right-hand side of the above formula decomposes the realized outcomes for an individual in her potential outcomes. The differences $Y_{i,t}^p - Y_{i,0}^p$, which are the causal effects of the treatment t, 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 t:

$$\tau_t = \mathbb{E}\left[Y_{i,t}^p - Y_{i,0}^p\right], \quad \text{for all } t = 1, \dots, n_d,$$

which—depending on the question—is often sufficient.

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 5 (Conditional Average Treatment Effect)**.** Let X, $Y_{i,t}^p$ and $Y_{i,0}^p$ as in Definition 2 and 3. For every treatment $t \in \{1, \ldots, n_d\}$,

$$\tau_t(X) = (\tau_{t,1}(X), \ldots, \tau_{t,n_y}(X)) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[Y_{i,t}^p - Y_{i,0}^p \,|\, X_i\right]$$

is called the *conditional average treatment effect* of treatment t. The random matrix $T = (\tau_{t,k})_{t=1,\ldots,n_d;k=1,\ldots n_y}$ contains all of the conditional treatment effects.

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 treatment assignment, conditional on covariates. If the *Conditional Independence Assumption* holds, any potential selection bias vanishes and the 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.

The second most important assumption is the so called *overlap assumption*, which basically says that all characteristics observed in a treatment group have to be found amongst the individuals in the control group, because otherwise, 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. Formally (Abadie and Imbens, 2006) we have

Assumption 1 (Conditional Independence Assumption and Overlap Condition)**.** For every $t = 0, 1, \ldots, n_d$, for almost every $x \in \mathbb{X}$, where $\mathbb{X}$ denotes the support of X_i and $i = 1, \ldots, N$, the following conditions hold:

$$D_i \text{ is independent of } Y_i^p \text{ conditional on } X_i = x; \tag{CIA}$$

$$\eta < \Pr(D_i = t \,|\, X_i = x) \text{ for some } \eta > 0. \tag{Overlap}$$

3.2 A Mean Squared Error Based Minimization Function

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

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

where τ 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 $w^d = (w_1^d, \ldots, w_{n_d}^d)$ and $w^y = (w_1^y, \ldots, w_{n_y}^y)$ be non-negative vectors that weight treatments $t \geq 1$ and outcomes, respectively. Then, for the matrix of weighted estimators $\mathrm{diag}(\sqrt{w^d})(\hat{T})\,\mathrm{diag}(\sqrt{w^y})$, we define the conditional weighted MSE component-wise as

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

where $\|\cdot\|_F$ denotes the Frobenius norm. If weights shall not be considered, w^d and w^y will be vectors with entries 1 only, so $\mathrm{diag}(w^y)$ and $\mathrm{diag}(w^d)$ will be the $n_y \times n_y$ and $n_d \times n_d$ identity matrix, respectively. We assume w^d and w^y be independent of T.

The expectation of the squared Frobenius norm of the matrix $\hat{T} - T$ with its corresponding weights is—because of linearity—simply the trace of the expected 'squared' weighted error matrix:

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

Objective Function The objective is to minimize the generalized MSE (1). Hence, we seek an estimator $\hat{T}$ minimizing this function for the given weights w^d and w^y:

$$S_T(\hat{T}) = \mathrm{MSE}(\hat{T}, w^d, w^y \mid X).$$

As the quantity of interest, the conditional average treatment effect, is a function of the covariates, it is natural to start from this point: Suppose the estimator of the treatment effects $\hat{T}$ is a function of X, so $\hat{T} = m(X)$. As the weights do not depend on $\hat{T}$, $S_T(\hat{T})$ is given by the trace of $\mathbb{E}\left[(m(X) - T)^\top (m(X) - T) \mid X\right]$, which can be written as

$$(m(X) - \mathbb{E}[T \mid X])^\top (m(X) - \mathbb{E}[T \mid X]) + \mathbb{E}\left[(\mathbb{E}[T \mid X] - T)^\top (\mathbb{E}[T \mid X] - T) \mid X\right].$$

Since the last summand does not involve $m(X)$, $S_T(\hat{T})$ is minimized by setting $m(X) = \mathbb{E}(T \mid X)$. With that,

$$\mathbb{E}[T \mid X] \in \operatorname*{argmin}_{\hat{T}} S_T(\hat{T}).$$

Considering the t-th row of the matrix $\mathbb{E}[T \mid X]$ and using the definition of the Conditional Average Treatment Effect (Definition 5) yields

$$\mathbb{E}[\tau_t \mid X] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_{i,t}^p - Y_{i,0}^p \mid X_i\right] \mid X\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_{i,t}^p - Y_{i,0}^p \mid X_i\right].$$

This, however, leaves us with the challenge of estimating $\mathbb{E}\left[Y_{i,t}^{p} \,|\, X_i\right]$ for all treatment groups $t = 0, 1, \ldots, n_d$.

3.3 A Linear Model for Potential Outcomes

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

Assumption 2 (Potential outcomes are linear functions of covariates)**.**

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

for $i = 1, \ldots, N$, $k = 1, \ldots, n_y$ and $t = 0, 1, \ldots, n_d$ with

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

We assume (Y_i^p, X_i) independent and identically distributed for all $i = 1, \ldots, N$. For the error terms, we assume $\varepsilon_{i,t}^{p,k} \,|\, X_i \sim \mathcal{N}(0, \sigma_{t,k}^2)$ for all $i = 1, \ldots, N$ and all $k = 1, \ldots, n_y$, $t = 0, 1, \ldots, n_d$. Moreover, we assume independence between $\varepsilon_{i,t}^{p,k}$ and $\varepsilon_{i,0}^{p,k}$ for $i = 1, \ldots, N, k = 1, \ldots, n_y$ and $t = 1, \ldots, n_d$. The variances are expressed in relation to a 'base' variance: $\sigma_{t,k}^2 = s_{t,k}\sigma_{0,k}^2$ for all $t = 1, \ldots, n_d$, $k = 1, \ldots, n_y$ with $s_{t,k} > 0$ and $\sigma_{0,k}^2 = s_{0,k}\sigma_0^2$ with $s_{0,k} > 0$ for all $k = 1, \ldots, n_y$ and for a $\sigma_0^2 > 0$.

Let the submatrix X_t of X contain the covariate vectors of all individuals in treatment group t, that is $X_t := (X_{i_1}, X_{i_2}, \ldots, X_{i_{n_t}})$ for $\{i_1, i_2, \ldots, i_{n_t}\} = \{i : D_i = t\}$. Then, the objective function can be expressed conveniently in terms of covariates, treatment group assignment and possibly weights, as the following theorem shows.

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

$$\frac{1}{N^2}\sum_i X_i^\top \left[\parallel \tilde{w}^y \parallel_1 \parallel w^d \parallel_1 \left(X_0 X_0{}^\top\right)^{-1} + \sum_k \left\{ \tilde{w}_k^y \left(\sum_{t>0} \tilde{w}_{t,k}^d \left(X_t X_t{}^\top\right)^{-1} \right) \right\} \right] \sum_i X_i,$$

where $\|\cdot\|_1$ *is the* l_1 *norm of a vector,* $\tilde{w}_k^y = w_k^y s_{0,k}$ *and* $\tilde{w}_t^d = w_{t,k}^d s_{t,k}$ *for* $k = 1, \ldots, n_y$ *and* $t = 1, \ldots, n_y$.

The proof of Theorem 1 is in Appendix A.

Corollary 1. *Under the condition of Theorem 1 and assuming the same variance for all outcomes and treatment groups, including the control group (i.e.* $s_{t,k} = 1$ *for all* $t = 0, 1, \ldots, n_d$ *and all* $k = 1, \ldots, n_y$*) and neglecting any weights (i.e. assuming* $w_t^d = w_k^y = (1, \ldots, 1)$*), minimizing* (1) *through choice of* D *is equivalent to minimizing*

$$\frac{1}{N}\sum_i X_i^\top \left[n_d \left(X_0 X_0{}^\top\right)^{-1} + \sum_{t>0} \left(X_t X_t{}^\top\right)^{-1} \right] \frac{1}{N} \sum_i X_i. \tag{3}$$

Contrary to the result by Kasy (2016), our approach is more applicable as the researcher is relieved from guessing any (absolute) values. This is because we do not assume a prior distribution for $\beta_t^{p,k}$, as such there is no need to specify its parameters: a mean or—more difficult—a covariance matrix for this parameter vector for every combination of k and t in case of an assumed normal prior distribution. Furthermore, there is no need of specifying the R^2 for the model of each potential outcome in order to express the model's variance. Instead, one can simply specify the relative scaling factors.

4 Characteristics of the min MSE Treatment Assignment Method

4.1 The Relation to the Classical Approach of Optimal Experimental Design

In what follows, we establish the link to the literature on experimental design. In this literature, starting with Smith (1918), the goal has also been the 'minimization' of the variance of an estimator θ within an e.g. linear regression model $y = x^\top \theta + \varepsilon$ such as (3) that determines how observations x (or a function thereof) affect the outcome or response y (Fedorov, 1997).[4] The task in this literature usually is (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

$$\zeta_L = \begin{Bmatrix} x_1 & x_2 & \cdots & x_N \\ p_1 & p_2 & \cdots & p_N, \end{Bmatrix} \tag{4}$$

where $\sum_i p_i = 1$ and $p_i L \in \mathbb{N} \cup \{0\}$ where $L = \sum_i l_i$. As such, ζ_L is a discrete probability measure defined on the design space X. The optimization problem may formally be written as $\operatorname{argmin}_{\zeta_L} \operatorname{Var}[\hat{\theta}(\zeta_L)] = \operatorname{argmin}_{\zeta_L} M^{-1}(\zeta_L)$, where M is called *information matrix*. Due to $\operatorname{Var}[\hat{\theta}]$ being a matrix, this has usually to be replaced with

$$\operatorname{argmin}_{\zeta_L} \Phi[M^{-1}(\zeta_L)],$$

for a suitable function Φ called *optimality criterion*. Several such optimality criteria have evolved (Kiefer, 1959; Fedorov, 1997).

[4] Note that the MSE corresponds to the variance in case of an unbiased estimator.

In case a linear combination of estimators $c^\top\theta$ for a p-vector c is of interest, c-optimality is the criterion that must be considered (Fedorov, 1997). This criterion, which aims at minimizing $\text{Var}[c^\top\theta]$, i.e. $\text{argmin}_{\zeta_L}\, \Phi[M^{-1}(\zeta_L)] = \text{argmin}_{\zeta_L}\, c^\top M^{-1}(\zeta_L)c$, is a special case of the min MSE Treatment Assignment procedure, as the following proposition shows.

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

The proof of Proposition 1 is in Appendix A. This proposition shows how our result builds on an old and rich literature that has been dedicated to increasing the precision of experimental outcomes for decades. From the simple case of one treatment, one outcome, equal model variances and no weights, where the optimization criteria coincide, we extend the min MSE Treatment Assignment in a suitable way to allow for the needed flexibility in practice.

4.2 Scale Invariance

Proposition 2. *Under the conditions of Corollary 1,* (3) *is constant 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 $j = 1, \ldots, m$.*

The proof of Proposition 2 is in Appendix A.

Proposition 2 states that the min MSE Treatment Assignment Procedure is scale invariant in the sense that the corresponding minimization criterion (3) is unaffected by changing the scale of a covariate. This feature is desirable, since it renders rescaling of the data unnecessary, but at the same time leaves the freedem to do so.

4.3 Balanced Treatment Groups

Proposition 3. *Assume $X_{k,i}$ is orthogonal to $X_{j,i}$ for $k, j = 1, \ldots, m$, $k \neq j$ with respect to the inner product $< \cdot, \cdot >_2$ of L^2, i.e. $\mathbb{E}[X_{k,i} X_{j,i}] = 0$. Furthermore, assume all covariates have the same mean, i.e.$\mathbb{E}[X_i] = c(1, \ldots, 1)^\top$ for any $c \neq 0$. Then, for $N \to \infty$, a solution to the minimization problem according to Corollary 1 is obtained, if*

$$\sum_j \left[n_d \left(\sum_{\{i:D_i=0\}} X_{j,i}^2 \right)^{-1} + \sum_{t>0} \left(\sum_{\{i:D_i=t\}} X_{j,i}^2 \right)^{-1} \right] \tag{5}$$

is minimized.

The proof of Proposition 3 is in Appendix A.

Proposition 3 states that—provided the covariates are orthogonal and have equal mean—the MSE for $n_d = 1$ is decreased, if the sum of squared observations of the covariates are increased in both groups and among all covariates. That is, the absolute deviation from 0 is 'balanced' for all covariates across groups. In the simple case of one treatment and one control group, equally sized, with one covariate considered for treatment assignment, this is equivalent to balancing the second moment of the distribution of the covariate of interest. 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.

Proposition 4. *Under the conditions of Corollary 1, now allowing for arbitrary mean values of the covariates and arbitrary relationships between covariates, the diagonal elements of the matrix* $\left(X_0 X_0{}^\top\right)^{-1} + \left(X_t X_t{}^\top\right)^{-1}$ *in* (3) *are given by*

$$\mathrm{Var}(\hat{\beta}_{t,j} - \hat{\beta}_{0,j} \,|\, X) \propto 1/[(1 - R_j^{t\,2}) \sum_{i=1}^{n_t} (X_{j,i}^t - \bar{X}_j^t)^2] + 1/[(1 - R_j^{0\,2}) \sum_{i=1}^{n_0} (X_{j,i}^0 - \bar{X}_j^0)^2] \tag{6}$$

for every $t \geq 1$ *and every* $j = 1, \ldots, m$*, where* $\propto$ *denotes equality up to a multiplicative constant and the value of the* j*-th covariate of individual* i *in treatment group* t *is denoted by* $X_{j,i}^t$ *and* $\bar{X}$ *denotes the mean.* $R_j^{t\,2}$ *is the coefficient of determination of a regression between the variable* X_j^t *as response and all* X_p^t *for* $p = 1, \ldots, m$*,* $p \neq j$ *as explanatory variables; for all* $t = 0, 1, \ldots, n_d$*. The number of individuals in treatment group* t *and the control group is denoted by* n_t *and* n_0 *respectively.*

The proof of Proposition 4 is in Appendix A.

The sum on the right-hand side of (6) is decreased for every $t \geq 1$ and every covariate $j = 1, \ldots, m$, if the linear dependencies between covariates within groups are decreased. This introduces a certain orthogonality criterion, with $R_j^t = 0$ when all other covariates in the group are (partially) uncorrelated with the covariate X_j. When covariates are perfectly collinear, we would have $R_j^t = 1$ (however, in this case, the covariance matrix of the estimator of the parameter vector does not even exist, as $X_t X_t^\top$ is not invertible). 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. Note, however, that this grouping might not minimize off-diagonal entries of the sum of the covariance matrices.

Would the off-diagonal entries be minimized in the same way than the diagonal entries according to (6), then, when balancing the covariates age and household income for example, a family's twin children living in the same household should not

be placed in the same group. Consider the extreme case with two twin pairs and two groups: If twin pairs are in the same group, within each group, both variables are perfectly predictable by the other. The more combinations of covariates can be found in a group, the smaller $R_j^{t\,2}$. However, a reduction in $R_j^{1\,2}$ might lead to an increase in $R_j^{0\,2}$. Therefore, (6) is decreased for every $t \geq 1$ and every covariate $j = 1, \ldots, m$, if for one group, more combinations of covariates can be achieved without limiting the combinations of covariates to the same amount expressed in $R_j^{1\,2}$ and $R_j^{0\,2}$.

The second part influencing (6) is the within-group variation of variable j around its mean. The higher the variation, the lower the variance. Again, an increase in overall variance can only be achieved if an increase in a variable's variation in one group does not lower that variable's variation in the other to the same extent or more.

Especially the first part is interesting, as it shows a characteristic that is also inherent in matching: Two very similar subjects should not be allocated to the same group. This characteristic additionally 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.

5 Comparison of the min MSE Treatment Assignment Method with Alternatives

5.1 Pair-Wise Matching

Consider 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 sum of variances' approach.*

The proof of Theorem 2 is in Appendix A.

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.

5.2 An (Alternative) Linear Model for Potential Outcomes

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 t, a linear, additive model is specified as follows:

$$Y_t = \begin{bmatrix} Z_t & X_t \end{bmatrix} \begin{bmatrix} \tau_t \\ \beta^t \end{bmatrix} + \varepsilon,$$

where the subscript t of Y_t, Z_t and X_t indicates that row entries are from individuals of $\{i : D_i = t \vee D_i = 0\}$. Y_t contains the potential outcomes for the control group or treatment group t, depending on Z_t, which is the treatment status, with $Z_{i,t} = \mathbb{1}_{\{D_i=t\}}$ for those in treatment group t and $Z_{i,t} = -\mathbb{1}_{\{D_i=0\}}$ for the control group. X_t contains the covariate vectors $X_i^\top$ of individuals in treatment group t and in the control group. In this model, $2\tau_t$ is the estimate for the conditional average treatment effect.

Under the Gauss-Markov assumptions (additive errors, that are uncorrelated conditional on X_t with constant variance σ^2), the MSE of the estimated treatment effect is proportional to $1/Z_t^\top P_t Z_t$ with $P_t = I - X_t \left(X_t^\top X_t\right)^{-1} X_t^\top$ and is minimized for $X_t^\top Z_t = 0$ (Greevy et al., 2004). Thus, with the assumption of constant variances

across treatments and without imposing weights, an alternative objective function for minimization would be

$$S_T^*(\hat{T}) \propto \sum_t 1/Z_t^\top P_t Z_t. \tag{7}$$

In this model, 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. Contrarily to a simple difference in means estimator for the average treatment effect, here covariates are controlled for, which induces a criterion of balance for covariates. As this criterion is minimized by $X_t^\top Z_t = 0$, 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.

Thus, comparing this result with the results derived in Section 4 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.

5.3 Morgan and Rubin (2012)

The approach by Morgan and Rubin (2012) considers the Mahalanobis distance between the vector of covariate means of the control group and the vector of covariate means of the treatment group. When group averages are equal, the distance is minimal. For the derived statistic, a threshold to stop re-randomization is derived. This approach is closely related to the omnibus test for multivariate covariate balance by Hansen and Bowers (2008); in fact, for the case without strata or matching pairs, the statistic is the same.

With respect to the notion of balance, the statistic shares its properties with the objective function (7) of the just discussed alternative linear model as considered in

Greevy et al. (2004). Balance in these approaches equals balancing group means,[5] thus balance is limited to the first moment of the distribution of covariates.

6 Simulation Study

In order to investigate the performance of the treatment assignment procedure described in Section 3, we perform a simulation study similar to the one by Bruhn and McKenzie (2009), henceforth cited as BK09, to compare our new mechanism to established ones.

BK09 compare five treatment assignment methods (purely random assignment, pairwise greedy matching, stratification and two rerandomization schemes) in terms of 'balance' of relevant observable and "unobservable" variables when creating one treatment and one control group. To rule out the possibility that results depend on the characteristics of a specific dataset or sample size, they consider several data sets.

We extend this study by adding the scenario of multiple treatment arms and a scenario where attrition occurs randomly. In terms of treatment assignment mechanisms, we also include an 'optimal matching' approach (Lu et al. (2011) as introduced by Greevy et al. (2004)) and our new min MSE procedure, as introduced in Section 3 of this paper.

6.1 Study Design

6.1.1 Data

We use the same data as BK09 for reasons of comparability. It consists of four panel datasets, with different data from different contexts.

[5] For (7) this holds when the treatment and control groups are of equal size.

The first dataset contains data on microenterprises in Sri Lanka and is from an actual randomized experiment by De Mel et al. (2008). The outcome variable of interest is firms' profits, and data on firm and owner characteristics at the time of the baseline study is available. It is either used for treatment assignment or treated as "unobservable" and studied after treatment assignment to assess the effect of the different methods on "unobservables".

The second dataset consists of a subsample of the Mexican employment survey (ENE), where we used the same subsamples as BK09. In this dataset, the outcome of interest is the income of household heads that were employed and between age of 20 and 65 when the baseline survey was conducted in 2002. In addition to this, the dataset includes additional characteristics on the household and its head, which again are used either for treatment assignment or as "unobservables".

The third dataset is comprised of subsamples with two waves (IFLS 2 and IFLS 3) from the Indonesian Family Live Survey (IFLS):[6] The year 1997 (IFLS 2) is used as the baseline and the data from 2000 (IFLS 3) is treated as the follow-up. We only use data on household expenditure from this survey.[7]

The fourth dataset is from the Learning and Educational Achievement project in Pakistan, which is also used by Andrabi et al. (2015). It contains child and household data, and the outcome variables of interest are math test scores and z-scores of children between the ages of 8 to 12 at the baseline.

It is noteworthy that the subsamples of 30 and 100 observations sometimes differed considerably: The share of the variation in the follow-up variable explained by the

[6]See http://www.rand.org/labor/FLS/IFLS.html.

[7]BK09 also use data on the schooling of children from this dataset. Given that they do not report results for this dataset in all graphs and tables, we limited ourselves to the inclusion of household expenditure.

group used for treatment assignment in the dataset on firms' profit for the small subsample is around 6 percent, whereas it amounts to 18 percent for the subsample of 100 observations. A larger difference is observed in the dataset from Mexico[8] and in the data on height z-scores, in the smaller subsample, an even higher share of variation in z-scores could be explained than in the larger sample (64% and 51%, respectively). For the remaining datasets, however, no meaningful differences are found.

Nevertheless, this observation gives rise to a method for drawing comparative samples of a 'universe'. The treatment assignment procedure derived in Section 3 can also be used in this setting.[9]

6.1.2 Treatment Assignment Mechanisms

A number of treatment assignment mechanisms is common (BK09).

In this study, we additionally consider the minimal MSE procedure as introduced in this paper, and the matching method called 'optimal matching' as implemented in the *R* package *nbpMatching* (Lu et al., 2011), going back to the work of Greevy et al. (2004). We give a short overview over the assignment mechanisms applied in this study. In Section 8, we discuss the weaknesses and strengths of the different treatment assignment mechanisms in comparison with our min MSE method.

Pure randomization refers to the realization of a single random draw, performed by using anything from a coin to a random number generator on a computer. Stratification, also known as blocking, is attributed to Fisher (1935). The idea is to build subgroups according to observable characteristics (covariates) and randomize within those subgroups.

[8] Roughly 7 and 32 percent (for the subsamples of 30 and 100 observations, respectively) of the variation in household expenditure is explained by "observable" variables.

[9] A Stata software package for this purpose can be obtained from the author.

Pairwise matching is in a certain sense the limit case of stratification, with only two units per strata, which are then randomly assigned to the treatment or the control group. BK09 apply a 'greedy algorithm' laid out in Imai et al. (2009), an approach popular in the literature on matching observational data at least since Rubin (1973). This implementation of the greedy algorithm computes pairwise Mahalanobis distances[10] between two units for the whole sample and pairs the two with the smallest distance; those then are taken out of the sample of units to be matched and the procedure is repeated. Overall distance is not necessarily minimized by this approach, because it is not 'forward-looking' (Rosenbaum, 1989; Greevy et al., 2004). The approach called 'optimal matching', as introduced by Rosenbaum (1989) for observational and Greevy et al. (2004) for experimental studies aims at achieving this goal.[11] We use both implementations in our study.

Finally, different rerandomization methods have evolved and are widely used (BK09). A rerandomization approach is basically any method that performs a somehow random treatment assignment, and repeats randomization until a certain condition is reached. This condition might either be a certain number of iterations or a statistical threshold or even subjective judgment. In the first case, the "best" assignment is chosen; in the second, usually the first to reach the statistical threshold is kept. In this sense, the "best" assignment can be determined in various ways. A representative of the first group is, for example, the min max t-stat method, in which 1000 random assignments are made, and the one chosen is the one in which

[10] The use of Mahalanobis distance in matching for observational data has been discussed e.g. in Cochran and Rubin (1973).

[11] Forming pairs is referred to as non-bipartite matching in the optimization but also the matching literature. This procedure (non-bipartite matching) is different than finding matches in already existing groups (bipartite matching) and is considerably more difficult (Lu et al., 2011). Therefore, not all results for matching, and more importantly, software implementations, can be applied in this setting.

the maximal t-statistic on any variable to consider is the smallest. A variant of the second group is the 'big stick' method in BK09, in which a new treatment assignment is drawn if any difference in means between treatment and control group is significantly different from zero. A more sophisticated approach is the one by Morgan and Rubin (2012, 2015), where the Mahalanobis distance between the group means of the covariates is considered and a theoretical threshold is derived. Since the ad-hoc approaches lack a theoretical foundation and as for the last method, a software implementation is missing, we do not consider these approaches in this study.

Min MSE Method The Min MSE Method can be considered as a rerandomization method, in which a certain number of iterations is drawn. Unlike the rerandomization approaches considered in BK09, our implementation of the min MSE approach improves on previous draws, as we use the stochastic simulated annealing optimization algorithm (Kirkpatrick et al., 1983): Given a treatment assignment, a new one is obtained by randomly exchanging the treatment status for a certain number of units. The new assignment is then evaluated according to the formula derived in Section 3 and either kept or withdrawn. Thus, the min MSE method maximizes the balance in a more efficient way with respect to time than the other discussed approaches of rerandomization.

Furthermore, apart from the approach by Morgan and Rubin (2012), we are not aware of a rerandomization criterion that has a theoretical foundation and is not an ad-hoc measure. For their approach—to the best of our knowledge—there is no extension to multiple treatment groups[12] and the criterion is only based on the mean differences of the treatment groups. Additionally, we were unable to find any software implementation of their approach.

[12] Although they name a possible way of extending their criteria in this sense.

6.1.3 Variables for Balancing

We used the same variables for treatment assignment as BK09. They include the baseline outcome of an outcome of interest, and add six other variables that may affect the outcome of interest, with the exception of stratification, where only subsamples are used. This means, however, that the results of the study regarding stratification can only be conditionally compared to the other results, since stratification is tested with a lower number of variables and thus has a higher likelihood to achieve balance on those, and in particular, the baseline outcome variable. For reasons of comparability, we stick to this approach despite its shortcomings. For the exact reasoning for the choice of variables to be balanced, we refer to Bruhn and McKenzie (2009).

We use the same variables with the newly added treatment assignment mechanisms as we did for greedy matching: the baseline outcome and six additional variables.

6.1.4 Attrition

Researchers might be concerned about the consequences of attrition when a sophisticated method of treatment assignment has been applied. In case obtaining the outcome from a unit fails when having used a matching approach, it is common to also exclude its pair from the analysis (Imai et al., 2009). While mostly perceived as a disadvantage for the diminished sample size—especially when performing cluster randomization—Imai et al. (2009) consider this practice an advantage, as they argue the remaining sample is still balanced.

We investigate this claim by randomly removing $1, 3, 5$ and 7 units after the treatment groups have been assigned with a sample size of $N = 30$. While we exclude the pair of a randomly removed unit from the treatment groups assigned

by the matching approaches, for the other treatment assignment methods, such a possibility is missing and we leave the groups unaltered after simulated attrition. Subsequently, we investigate how attrition has affected balance.

6.2 Comparing Treatment Assignment Mechanisms

6.2.1 Pre-Treatment Balance

We investigate balance using the measures of pre-treatment balance on baseline variables as BK09 for the cases of one treatment arm. In the main text, we will report suitably aggregated results over all variables used for treatment assignment, as our interest lies rather in overall performance than in performance on an arbitrarily selected variable. Results for the latter case are printed in Appendix B

For reasons of comparability, we also assess balance in follow-up outcomes. However, we think that balance on follow-up outcomes is rather important when assessing the general value of covariate based treatment assignment mechanisms in panel studies, which is beyond the scope of this study. We therefore report those results in Appendix B.

For the cases of multiple treatment arms, we extend the measures used by BK09 in a suitable way.

Balance in a single variable, one treatment arm To assess balance in a single variable for the case of one treatment and one control group, BK09 compare the difference in means for one draw, expressed in the variable's standard deviation. Of all draws, they then graphically compare the distribution of the differences and report the average, and the 95% quantile of the distribution of (absolute) differences in the group means. Additionally, they perform a t-test to assess whether or not estimates for differences are "significantly" different from zero, and report the share

of draws in which this was the case. We assess balance using these measures and report results in Appendix B for comparison reasons.

Balance in a group of variables, one treatment arm First, standardized differences in means are calculated for every single variable of the group of variables for one draw. Then, for the average (absolute) difference and for the share of estimates significantly different from zero as calculated for a single variable, overall averages are built. For assessing the 95% quantile of differences, first the 95% quantile of every single variable in the group is determined. Then, the maximum among the variables in the group is reported as the 95% quantile. To detect extreme imbalances, we also compute the maximum difference of group means of a group of variables and evaluate the distribution of these maximal differences across 10,000 iterations graphically.

Balance in a group of variables, several treatment arms When aggregating the balance of several treatment arms, taking the average difference of the means between the several treatment groups and the control groups before taking the average over all variables of interest in all performed draws is one option. Another possibility is to take the overall average over the largest difference in the means between the treatment and control groups in one variable and in one draw. We think both are relevant and perform both.

7 Results

All results are based on 10,000 simulations, unless otherwise stated. The sample size, which was used for the tables and graphs, is indicated in the respective caption. For sample sizes, where results are not reported in the text, we provide the respective tables and graphs in Appendix B.

7.1 Scenario 1: One Treatment Group

We first present the results for the scenario considered in BK09: Units have to be assigned to either one treatment or the control group. In the main text, we focus on aggregate measures over all variables considered for treatment assignment, since no single variable has received a higher focus or a higher weight. In Appendix B, we present results for one single variable, as BK09 report these results.

Table 1 shows the average differences between the group means, average absolute differences between group means and the 95% quantile of the differences in group means among the 10,000 iterations performed. A lower measure indicates a better balance in group means. The main message of Table 1(a) is that, on average, all differences equalize and we observe balance (note that results are reported in 1000 standard deviations and thus are zero to the third or fourth digit in the upper panel of Table 1(a)).

A measure that is informative with respect to imbalance in a single random draw is the absolute difference in group means. The averaged absolute difference is reported in Table 1(b) for every treatment assignment mechanism and every dataset. On average and for the variables of every single dataset, the min MSE method performs better than all competing methods. Compared to the second best method, 'optimal matching', it reduces the average absolute difference in the balance of group means by nearly 30%.

Finally, Table 1(c) shows the average over all 95% quantiles of the absolute differences in a single variable considered for treatment assignment. Averaging all of the datasets, the min MSE method again performs superior to competing methods.

In summary, the min MSE method not only performs better than to competing methods on average, but also reduces extreme differences in the mean values between the treatment and control groups the most as compared to a single random draw.

This last finding is supported by the results in Figure 1, which shows the distribution of the largest differences between the group means of any variable for a single draw. In all graphs, we see that the mass of differences close to zero is largest for the min MSE approach, which also always yields a favourable mass in the tails as compared to competing methods.

7.2 Scenario 2: Multiple Treatment Arms

The second scenario we consider is an experiment, in which multiple (variants of) interventions are tested. Units shall be assigned to the control or one of the treatment groups while keeping all groups comparable.

For this scenario, we were unable to find a software implementation of a competing method, so we compare the min MSE procedure to a single random draw.

The findings are graphically presented in Figure 2. We first compute the maximum and the mean difference between the treatment group means and the control group mean of one variable for a single draw. We aggregate the measure over the variables of one draw and over all iterations by averaging over the mean or maximal difference. The first aggregate measure is shown by the dashed lines, the latter aggregate measure by the solid lines. Both lines, the solid and the dashed one, start at the same point, as for only one group, the maximum and the mean difference in group means is the same, as there is only one group difference to consider.

When applying the min MSE procedure, the maximum difference—typically the one a researcher is worried about—is always increasing at a lower rate with an additional treatment group to assign than when drawing completely random. For 9 treatment groups, which means 10 groups of 10 units, the average maximal difference (in SD) is—for all datasets—between .4 and .6, whereas when randomly drawing, the average maximal group difference is mostly around .75 or .8 SD. In one case (house-

Table 1: Comparison of Treatment Assignment Methods Regarding Balance in a Group of Baseline Variables (N=30)

(a) Average difference in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Indonesia	0.025	0.444	1.134	-1.010	1.729	3.101
Pakistan (height scores)	-0.875	0.785	-0.872	1.035	-0.489	-0.062
Pakistan (test scores)	1.944	-1.579	0.949	-0.653	-1.903	-0.138
Mexico	2.191	1.996	-0.477	0.481	-0.021	-0.310
Sri Lanka	0.624	2.111	-0.748	-0.328	-1.594	-1.667

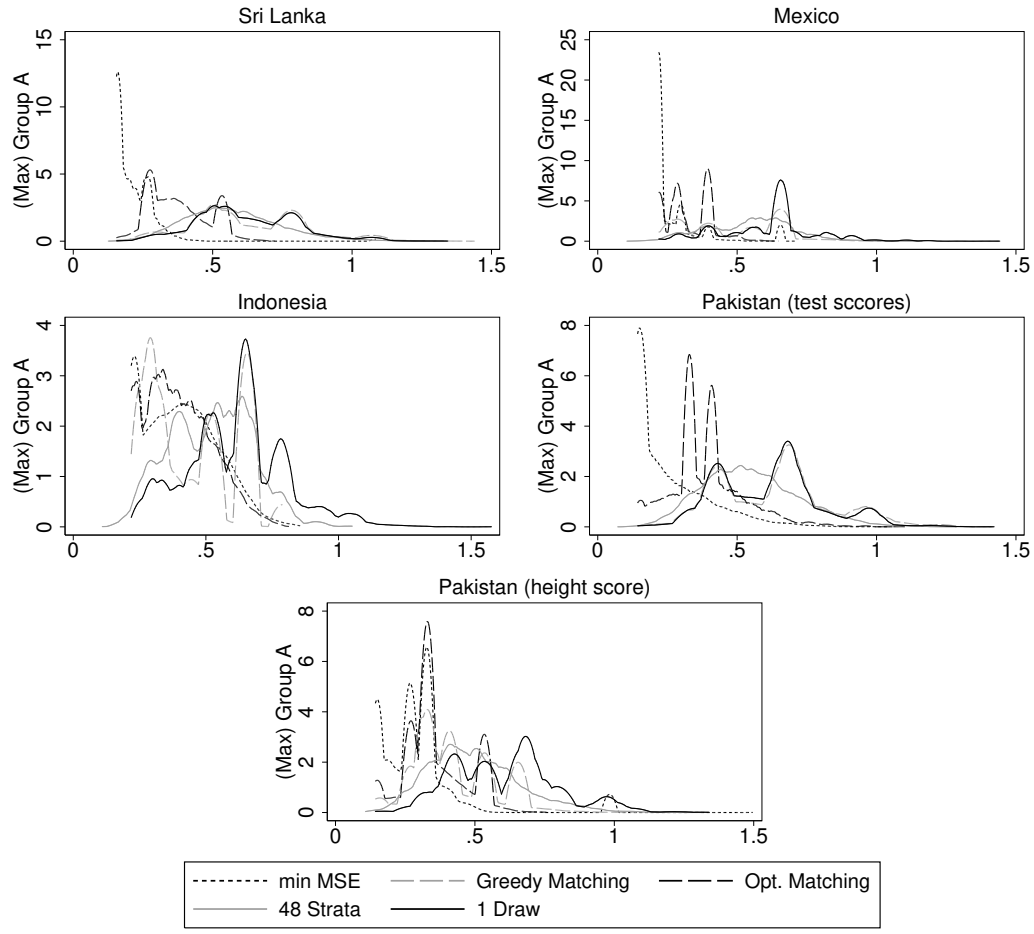
(b) Average abs. difference in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Indonesia	295.9	198.1	174.0	126.2	233.5	243.8
Pakistan (height scores)	291.0	180.1	172.3	134.6	235.4	230.5
Pakistan (test scores)	293.1	287.9	180.3	103.0	244.7	257.3
Mexico	299.1	223.8	174.1	147.8	258.5	262.1
Sri Lanka	292.4	253.9	170.9	93.4	248.8	267.0
Total	294.3	228.8	174.3	121.0	244.2	252.1

(c) 95% quantile of the difference in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Indonesia	788.3	655.5	579.9	643.4	742.6	702.3
Pakistan (height scores)	802.8	655.5	535.2	393.8	783.3	722.9
Pakistan (test scores)	715.2	905.4	628.4	533.8	729.3	715.2
Mexico	701.3	677.1	445.0	348.2	727.7	702.3
Sri Lanka	802.8	863.6	535.2	311.8	792.3	744.7
Total	762.1	751.4	544.7	446.2	755.1	717.5

Note: Statistics based on 10,000 iterations. Details on the study and the computation of each measures are explained in Section 6.2. For every dataset, several variables were considered for treatment assignment. The results in this table report aggregate measures of differences in treatment group means for the group of considered variables. Differences are weighted by standard deviation. Lower values indicate better balance with respect to equality of group



Note: Distributions of the (maximal) differences in treatment group means among the group of variables to consider for treatment assignment are based on 10,000 treatment assignments. Differences in group means are expressed in standard deviations. A high mass around a difference of 0 indicates a good balance with respect to equality of group means.

Figure 1: Distributions of the Maximal Differences in Group Means (N=30)

hold expenditure in Indonesia), this average *maximum* difference for 6 treatments (thus 7 groups) when using the min MSE procedure was as high as the average *mean* difference when relying on a single random draw.

In all datasets, the min MSE procedure was able to lower the average maximum difference across group means compared to drawing randomly by between .1 SD (height z-score in Pakistan and labor income in Mexico) and up to .3-.4 SD (math test score in Pakistan and household expenditure in Indonesia).

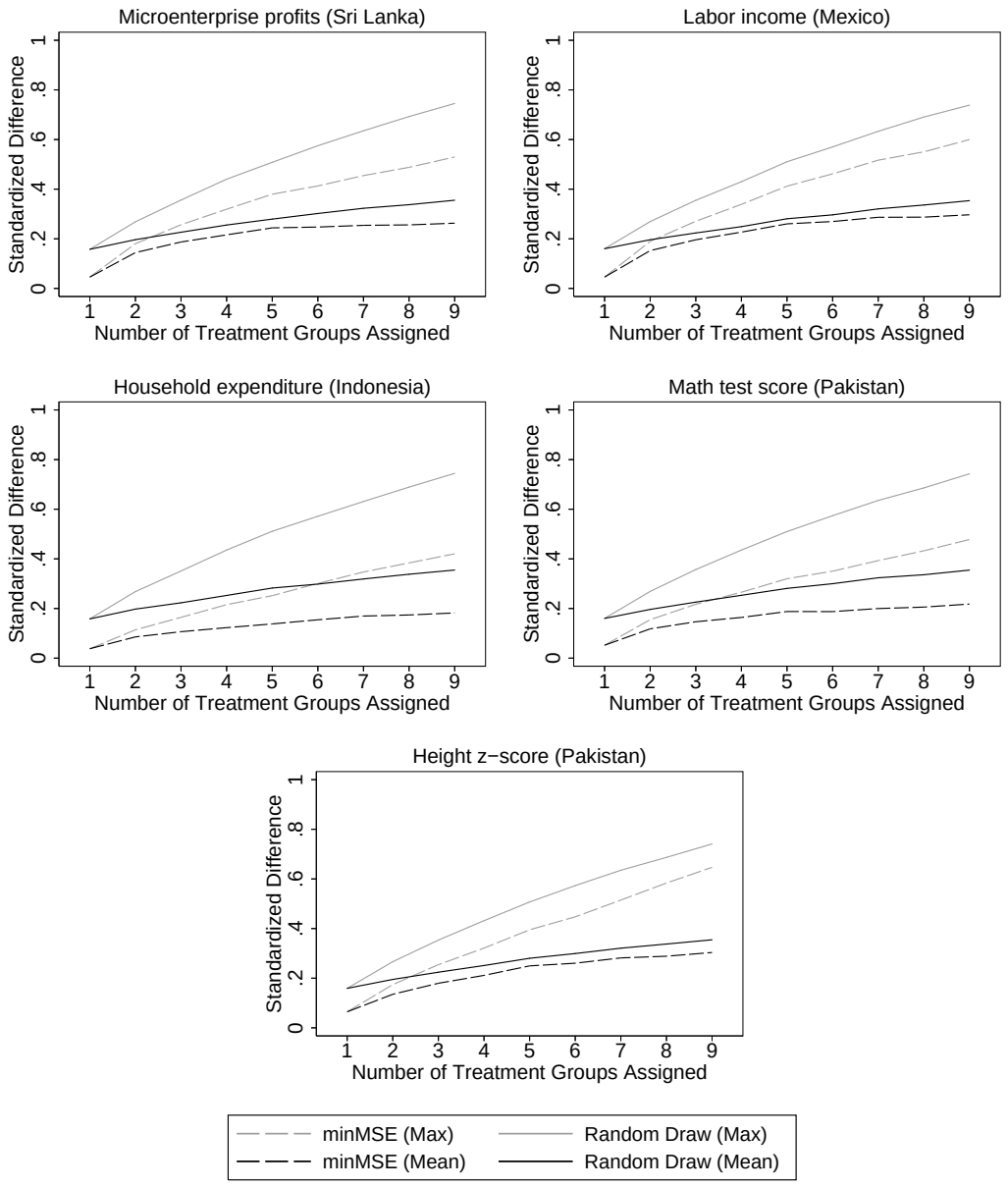
It is also worth noting that with the min MSE procedure, we can assign between 2 to 5 more treatments compared to randomly drawing with the same maximum difference in group means to be expected.

7.3 Scenario 3: Attrition

The third considered scenario corresponds to the first scenario, where units have to be assigned to either the treatment or the control group. After treatment assignment, however, some units fail to provide the outcome of interest, i.e. the study suffers from attrition. We randomly remove $1, 3, 5$ and 7 units after the treatment groups have been assigned with a sample size of $N = 30$. We "correct" for attrition in case matching approaches were used for treatment assignment, see Section 6.1.4 for details.

Table 2 reports the results of the attrition scenario. Table 2(a) shows the absolute difference between group means, averaged over variables considered for treatment allocation, datasets and iterations. We report the average absolute difference instead of the average difference, since average differences are close to zero for all mechanisms, and imbalances obtained for the different attrition levels could average out.

The first observation is that attrition always strictly worsens balance for all treatment assignment mechanisms. Except for the case when 7 units (roughly 25% of the



Note: Multiple Treatment Assignment. Evolution of the (mean and maximal) difference in group means between the treatment groups and the control group among the group of variables to consider for treatment assignment. X-axis: Number of treatment groups to assign. For each dataset, the difference in group means in all variables used for treatment assignment are computed. The line labeled 'max' shows the average over all iterations and variables of the maximum of these differences amongst the treatment groups. The line labeled 'mean' shows the differences amongst the group differences by building the average. Distributions are based on 10,000 treatment assignments. Differences in group means are expressed in standard deviations. Lower mean and maximal differences in group means between the treatment groups and the control group indicate a better balance with respect to equality of group means.

Figure 2: Multiple Treatment Assignment: Evolution of the Differences Between Treatment Groups and Control Group in the Group of the Baseline Variables for an

sample) drop out, balance achieved with the min MSE mechanism is best for all levels of attrition, as indicated by the lowest average absolute difference. Only the 'optimal matching' algorithm achieves a better balance in that case, and the difference between both mechanisms in that case is only marginal, compared to the balance of the other mechanisms. However, the sample assigned with the min MSE mechanism then still consists of 23 units, whereas for the matching approaches, it diminishes to 16 (see Section 6.1.4). On average, however, it is the min MSE mechanism, that performs best when attrition happens with respect to the absolute difference between group means. Compared to the min MSE mechanism and 'optimal matching', stratification and greedy matching cannot considerably improve balance compared to a single random draw.

The second panel of Table 2 shows the worst case scenario across 10,000 iterations. For every method, every level of attrition and every variable in every dataset, we compute the 95% quantile of the absolute difference in group means of all iterations. This measure is then averaged across all variables and datasets and reported for every method and level of attrition. When assessing balance according to this measure, again the balance strictly worsens when attrition occurs. On average, it is again the min MSE approach yielding the most favorable results, and with exception of the 'optimal matching' approach, other mechanisms only provide a limited improvement on a single random draw.

In Appendix B, we report results for individual datasets and when assessing balance using t-tests.[13]

[13]Testing for equality of group means in a single variable of a random draw is often mistaken as a test for successful randomization. BK09 report the share of treatment assignments for every mechanism, where a test for equal group means yields a p-value below .1. In Table 6(a) in Appendix B, we report corresponding results as these tests are frequently conducted. However, we think that it is misleading to assess the probability that a statistic excesses a certain value by pure chance

Summing up, we cannot confirm the claim by Imai et al. (2009) that the practice of removing the matched pair of a unit dropping out an experiment is actually beneficial. Only when using the 'optimal matching' approach, is balance in cases of attrition comparable to the min MSE approach. On average, however, the min MSE approach outperforms all competing mechanisms in our study in cases of attrition, while mantaining the maximal possible sample size in contrast to the matching approaches.

8 Discussion

8.1 Treatment Assignment Mechanism

In what follows, we discuss the strengthes and weaknesses of the min MSE approach as compared to alternative mechanisms.

Pure Randomization Depending on the transparency of the actual implementation, randomization can be considered to be the fairest method for treatment allocation and it certainly is the fastest.

When comparing the means of randomly allocated groups across 20 variables with a conventional t-test and a significance level of 5%, we have to expect that for one variable, the hypothesis of no difference will be rejected. Pure randomization does not yield any device for controlling undesired imbalances that may happen by chance.

when actually *knowing* that it is pure chance that drives the differences and thus the statistic. Therefore, we forgo presenting these results in the main text. Figure 8 in Appendix B shows the results of Panel (b) of Table 2 for the individual datasets.

Table 2: Comparison of Treatment Assignment Methods Regarding Balance in a Group of Baseline Variables in Cases of Attrition (N=30)

(a) Average abs. differences in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
No attrition	294.3	228.8	174.3	121.0	244.1	252.1
1 (2) unit removed	300.7	235.9	179.9	139.1	252.0	258.5
3 (6) units removed	312.3	263.8	187.2	163.3	265.6	271.3
5 (10) units removed	328.3	277.0	198.4	186.9	280.5	285.7
7 (14) units removed	352.7	294.3	205.3	211.7	297.8	302.9
Total	317.6	259.9	189.0	164.4	268.0	274.1

(b) 95% quantile of the differences in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
No attrition	705.8	520.7	356.3	231.2	575.4	600.2
1 (2) unit removed	723.6	544.4	372.0	289.1	593.3	614.5
3 (6) units removed	765.6	601.7	381.8	377.1	634.9	650.8
5 (10) units removed	781.1	636.2	395.6	437.5	674.7	687.6
7 (14) units removed	844.8	676.1	408.9	503.6	714.9	730.0
Total	764.2	595.8	382.9	367.7	638.6	656.6

Note: Statistics based on 10,000 iterations. Details on the study and the computation of each measures are explained in Section 6.2. For every dataset, several variables were considered for treatment assignment. After treatment assignment, units were randomly removed from the study to simulate attrition. The results in this table report aggregate measures of differences in treatment group means for the group of considered variables. Differences are weighted by standard deviation. Lower values indicate better balance with respect to equality of group means.

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 The main advantage of stratification is to ensure the possibility of subgroup analysis while ideally increasing the efficiency of the analysis. The idea is to build subgroups according to observable characteristics (covariates) and randomize within those subgroups. This design is probably still considered relatively fair.

One problem of this approach is that continuous variables have to be discretized arbitrarily, and that stratification is only possible for a limited number of variables: Consider a sample of 50 units, where subgroup analysis for age, income and gender is desired. If three categories for age and income are desired, 18 strata have to be created, where at most 3 persons are in one strata. Stratification on another variable is thus not feasible with a comparable sample size. This example points to another drawback: Difficulties arise in implementation if sample size is not divisible by the number of strata. 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.

The time needed to conduct treatment assignment using stratification depends on the actual implementation, but in simple cases, e.g. with two dichotomous variables, it takes only slightly longer than pure randomization.

Matching Pairwise matching resolves the problems of stratification. Theoretically, pairwise matching may be performed with an unlimited number of variables considered for treatment allocation. Moreover, the possibility to balance continuous

variables is an advantage (e.g. Greevy et al., 2004). It is arguably considered to be fair and the design is relatively clear and easy to explain.

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.

Implementation of pairwise matching may take considerable time[14] when relying on the 'greedy algorithm' used by BK09. Yet, the software implementation of the 'optimal matching' algorithm in the *R* package *nbpMatching* (Lu et al., 2011) is considerably faster.

However, the biggest disadvantage is probably attrition, but also with perfect compliance, analysis is a major concern. Regarding the analysis, Abadie and Imbens (p. 236, 2006) note that matching estimators for the average treatment effect "include a conditional bias term whose stochastic order increases with the number of continuous matching variables". They show that the simple matching estimator is not $N^{1/2}$ efficient and propose an alternative. Imai et al. (2009) claim that the variance can be consistently estimated, but they refer to the variance not conditional on covariates (Imbens, 2011). Imbens (p. 17, 2011) writes that "this variance is larger than the conditional one if treatment effects vary by covariates. In stratified randomized experiments we typically estimate the variance conditional on the strata shares, so the natural extension of that to paired randomized experiments is to also condition on covariates." In contrast to this, BK09 estimate the variance conditional on pair dummies, but not conditional on covariates. It thus seems that even among experts, it is unclear how to correctly assess the variance of estimates.

With respect to attrition, Imai et al. (2009) note that an advantage of matching is that if a unit drops out, its pair can also be taken out of the experiment while

[14] BK09 note that in the 300 observation sample, the algorithm takes several days to run and that ample time is needed to perform matching techniques.

the remaining sample still remains balanced. In the simulation study, we have seen that even with attrition, the matching techniques perform, on average, worse than the min MSE approach, which by design is unable to "correct" for attrition in this way. It thus might be perceived an overall disadvantage that for every unit dropping out of the experiment, its pair also has to be discarded, as this leads to a lower sample size, and consequently, lower power—irrespective of the exact nature of the treatment effect.

Rerandomization Methods Bruhn and McKenzie (p. 210, 2009) name a key advantage of ad-hoc rerandomization methods: "[They] may offer a way of obtaining approximate balance on a set of relevant variables in a situation of multiple treatment groups of unequal sizes." As pointed out before, for the approach by Morgan and Rubin (2012), an extension allowing the assignment of multiple treatment groups is theoretically possible, but missing to date.

Furthermore, implementation time may differ considerably—depending on the approach. Drawing a thousand treatment assignments may take some time, and in small samples with many variables, the 'big stick' method that aims at finding a treatment assignment where no difference in group means exceeds a certain t-value might need even more time.

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.

Nevertheless, all of the rerandomization methods discussed here are able to consider continuous, categorial and binary variables in a theoretically unlimited number.

The approach by Kasy (2016) In a Bayesian setting, Kasy (2016) analyzes the task of treatment assignment from a decision theoretical perspective, where the mean squared error of an estimator is to be minimized. Kasy (2016) argues that randomization never increases precision. In the technical appendix of his paper, he discusses several modelling aspects of conditional expectations of potential outcomes, which are the basis of his analysis. One of the discussed models for potential outcomes is the bayesian linear model, which gives rise to a treatment assignment mechanism using the framework of his paper.

While Kasy (2016) provides software implementation in Matlab, an extension to treatment assignment for multiple groups is neither discussed nor implemented. Moreover, the Bayesian setting requires the choice of parameter values that are hard to guess without analyzing the pre-treatment version of the outcome of interest such as the covariance matrix of the estimator of the parameter vector in the linear model or the coefficient of determination of this model. Yet, in practice, a pre-treatment version of the outcome of interest may be unavailable. The choices, however, are consequential, as they distort the balance of treatment groups, which might be desired if the interest is limited to the precision of estimation of the specified treatment effect. In case the researcher is interested in comparable treatment groups for the sake of credible research, or in treatment effects for a specific subgroup, the balance in treatment groups might be an equally important goal.

However, the approach is—as the other rerandomization methods—able to consider any type of variable and any number of covariates desired. Moreover, dependencies between variables are taken into consideration and dispersion of variables within groups is encouraged—if not impeded by the researcher through an unwise choice of parameter values.

The min MSE Treatment Assignment Mechanism The min MSE Treatment Assignment Mechanism retains the advantages of the linear model as discussed by Kasy (2016). In particular, it is able to perform treatment assigment using possibly various continuous, categorical and binary variables. In that aspect, it is as powerful as all other rerandomization schemes and the matching approach.

Contrarily to the Bayesian model applied in Kasy (2016), we rely on a standard frequentist model. This relieves the researcher from the obligation to choose values for abstract parameters, for which arguably a good guess is sometimes impossible. Moreover, this setup aims at maximizing the dispersion of variables within treatment groups—without any possibly unwanted distortion. Thus, if possible, subgroup analysis is ensured as more characteristics of the distribution of variables in treatment groups than just their mean, namely a generalized second moment, is taken into consideration.

Yet, also the min MSE approach considers dependencies amongst variables in the different treatment groups and still allows for considerable flexibility if one has reasons to believe that an outcome will have a higher variance than the other in general; and more specifically, that treatment t, which mainly affects outcome k, might have compliance problems, resulting in an expected higher variance than for the control group.

By the sum of its characteristics, of all treatment assignment mechanisms we are aware of, the min MSE approach comes closest to the ability of the 'optimal matching' approach to achieve approximate equality of the distributions of the covariates in the treatment and control groups—while being able to assign multiple treatment groups. Moreover, the min MSE treatment assignment mechanism is attrition tolerant in the sense that the balance stays favorable and the sample size is reduced only by the units dropping out.

Finally, software implementation for Stata is available. Thus, implementation takes only a few minutes and may easily be performed in the field.

Appendix

A Proofs

Proof of Theorem 1. Similar to the definition of X_t, define the subvector of the k-th potential outcome $Y^{p,k}_{\{i:D_i=t\},t} := (Y^{p,k}_{i_1,t}, Y^{p,k}_{i_2,t}, \ldots, Y^{p,k}_{i_{n_t},t})^\top$ and the respective subvector of error terms $\varepsilon^{p,k}_{\{i:D_i=t\},t} := (\varepsilon^{p,k}_{i_1,t}, \varepsilon^{p,k}_{i_2,t}, \ldots, \varepsilon^{p,k}_{i_{n_t},t})^\top$, where again $\{i_1, i_2, \ldots, i_{n_t}\} = \{i : D_i = t\}$. That is, using observed information, (2) for all t and k in matrix notation writes $Y^{p,k}_{\{i:D_i=t\},t} = X_t^\top \beta^{p,k}_t + \varepsilon^{p,k}_{\{i:D_i=t\},t}$. For this linear model, it is well-known that for all $t = 0, 1, \ldots, n_d$ and $k = 1, \ldots, n_y$,

$$\hat{\beta}^{p,k}_t - \beta^{p,k}_t \sim \mathcal{N}(0, \sigma^2_{t,k}(X_t {X_t}^\top)^{-1}).$$

Using this result, the squared error of the estimator of the treatment effect for treatment t and outcome k becomes

$$\begin{aligned}
&\mathbb{E}\left[(\hat{\tau}_{t,k} - \tau_{t,k})^2 \mid X\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_i \left(\hat{Y}^{p,k}_{i,t} - \hat{Y}^{p,k}_{i,0}\right) - \frac{1}{N}\sum_i \left(E\left[Y^{p,k}_{i,t} \mid X_i\right] - E\left[Y^{p,k}_{i,0} \mid X_i\right]\right)\right)^2 \mid X\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_i X_i^\top \left(\left(\hat{\beta}^{p,k}_t - \beta^{p,k}_t\right) - \left(\hat{\beta}^{p,k}_0 - \beta^{p,k}_0\right)\right)\right)^2 \mid X\right] \\
&= \frac{1}{N^2}\sum_i X_i^\top \left(\mathrm{Cov}\left(\hat{\beta}^{p,k}_t - \beta^{p,k}_t \mid X\right) + \mathrm{Cov}\left(\hat{\beta}^{p,k}_0 - \beta^{p,k}_0 \mid X\right)\right)\sum_i X_i \\
&= \frac{1}{N^2}\sum_i X_i^\top \left(\sigma^2_{t,k}(X_t {X_t}^\top)^{-1} + \sigma^2_{0,k}(X_0 {X_0}^\top)^{-1}\right)\sum_i X_i,
\end{aligned}$$

where we used independence of the error terms $\varepsilon^{p,k}_{i,t}$ and $\varepsilon^{p,k}_{i,0}$.

Now denote the l_1 norm of a vector with $\|\cdot\|_1 = \sum|\cdot|$ and summarize weights and scaling factors for the variance as $\tilde{w}^y_k = w^y_k s_{0,k}$ and $\tilde{w}^d_t = w^d_t s_{t,k}$. Then, applying

the just derived result to the objective function, the generalized MSE (1), completes the proof:

$$
\begin{aligned}
S_T(\hat{T}) = \frac{\sigma_0^2}{N^2} \sum_i X_i^\top & \\
& \left[\sum_k \left\{ w_k^y s_{0,k} \left(\sum_t w_t^d s_{t,k} \left(X_t {X_t}^\top \right)^{-1} + \parallel w^d \parallel_1 \left(X_0 {X_0}^\top \right)^{-1} \right) \right\} \right] \\
& \sum_i X_i \\
\propto \frac{1}{N^2} \sum_i X_i^\top & \\
& \left[\parallel \tilde{w}^y \parallel_1 \parallel w^d \parallel_1 \left(X_0 {X_0}^\top \right)^{-1} + \sum_k \left\{ \tilde{w}_k^y \left(\sum_t \tilde{w}_t^d \left(X_t {X_t}^\top \right)^{-1} \right) \right\} \right] \\
& \sum_i X_i,
\end{aligned}
$$

where $\propto$ denotes equality up to multiplicative constants. □

Proof of Proposition 1. Assume $n_d = 1$ and $n_y = 1$. 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):

$$
\begin{pmatrix} Y^p_{\{i:D_i=0\}} \\ Y^p_{\{i:D_i=1\}} \end{pmatrix} = \begin{pmatrix} X_0^\top & 0 \\ 0 & X_1^\top \end{pmatrix} \theta, \tag{8}
$$

where $\theta = \begin{pmatrix} \beta_0^p \\ \beta_1^p \end{pmatrix}$ and $Y^p_{\{i:D_i=t\}}, X_t, \beta_t^p$ for $t = 0, 1$ are vectors and matrices as defined in the proof of Theorem 1. Recall that—with only one treatment group—we are interested in minimizing $\mathbb{E}[(\hat{\tau} - \tau)^2 \,|\, X] = \mathrm{Var}[\hat{\tau} \,|\, X]$. Now

$$
\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} X_i^\top (\hat{\beta}_1^p - \hat{\beta}_0^p) = c^\top \hat{\theta}
$$

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

$$\begin{aligned}\mathrm{Var}[c^\top\hat{\theta}\,|\,X] &= c^\top M(\zeta)^{-1}c &(9)\\ &= \sigma^2\frac{1}{N}\sum_{i=1}^N X_i^\top\left((X_0X_0^\top)^{-1}+(X_1X_1^\top)^{-1}\right)\frac{1}{N}\sum_{i=1}^N X_i, &(10)\end{aligned}$$

with $M(\zeta) = \sigma^{-2}\left(\begin{pmatrix}X_0 & 0\\ 0 & X_1\end{pmatrix}\begin{pmatrix}X_0 & 0\\ 0 & X_1\end{pmatrix}^\top\right)$, where, in our case,

$$\zeta = \begin{Bmatrix} z_{1,0} & z_{1,1} & z_{2,0} & \cdots & z_{N,n_d}\\ p_{1,0} & p_{1,1} & p_{2,0} & \cdots & p_{N,n_d}\end{Bmatrix} \tag{11}$$

with $\sum_{i,j} p_{i,j} = 1$, $p_{i,j}\in\{1/N, 0\}$ and $\forall i:\ \sum_{j=0}^{n_d} p_{i,j} = 1/N$ and $z_{i,0}^\top = X_i^\top, 0, \ldots, 0$, $z_{i,1}^\top = 0, \ldots, 0, X_i^\top$. □

Proof of Proposition 2. According to Cramer's rule, the inverse of the $p\times p$ matrix $\boldsymbol{A}$ is given by

$$(A^{-1})_{ij} = \frac{(-1)^{i-j}\det(\boldsymbol{A_{ji}})}{\det(\boldsymbol{A})}, \tag{12}$$

where $\boldsymbol{A_{ji}}$ is the $(p-1)\times(p-1)$ matrix resulting from deleting row j and column i.

Assume $n_d = 1$ and consider the first summand of (3) with a realization of the sample, i.e., $\bar{\boldsymbol{x}}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\bar{\boldsymbol{x}}$, where, for notational simplicity, we omit group indicators and $\bar{\boldsymbol{x}}$ is the vector containing the mean values of the covariates. Suppose the pth covariate vector $\boldsymbol{x_p}^\top$ is multiplied with some scalar $c\neq 0$ and denote the product with $\boldsymbol{w_p}^\top$. The covariate matrix changes accordingly from $\boldsymbol{X}$ to $\boldsymbol{W} = \mathrm{diag}(1,1,\ldots,c)\boldsymbol{X}$. Thus, $\det(\boldsymbol{W}\boldsymbol{W}^\top) = c^2\det(\boldsymbol{X}\boldsymbol{X}^\top)$.

Now consider the denominator of (12) for the possible combinations of i and j and denote $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{X}^\top$, $\boldsymbol{N} = \boldsymbol{W}\boldsymbol{W}^\top$, and $\boldsymbol{M_{ij}}$, $\boldsymbol{N_{ij}}$ the matrices resulting from deleting row i and column j of the matrices $\boldsymbol{M}$ and $\boldsymbol{N}$, respectively. Since

$\det(\boldsymbol{N_{ji}}) = c^2 \det(\boldsymbol{M_{ji}})$ for $i \neq p$ and $j \neq p$, $(M^{-1})_{ij} = (N^{-1})_{ij}$ in those cases. If either $j = p$ or $i = p$, we have $\det(\boldsymbol{N_{ji}}) = c \det(\boldsymbol{M_{ji}})$, thus $(N^{-1})_{ij} = 1/c(M^{-1})_{ij}$. Finally, for $i = p$ and $j = p$, as $\boldsymbol{N_{pp}} = \boldsymbol{M_{pp}}$, we have $(N^{-1})_{ij} = 1/c^2(M^{-1})_{ij}$.

Then, $\bar{\boldsymbol{w}}^\top \boldsymbol{N^{-1}} \bar{\boldsymbol{w}} = (1/n)^2 \sum \sum \bar{x}_i \bar{x}_j (M^{-1})_{ij} = \bar{\boldsymbol{x}}^\top \boldsymbol{M^{-1}} \bar{\boldsymbol{x}}$ applies to both summands of (3), also for $n_d > 1$, which completes the proof. □

Proof of Proposition 3. For $N \to \infty$, $\sigma_{t,k}^2 \left(X_t {X_t}^\top\right)^{-1}$, the sample covariance matrix of $\hat{\beta}_t^{p,k}$, converges to the population covariance matrix for any $t = 0, 1, \ldots, n_d$ and any $k = 1, \ldots, n_y$. The elements of the inverse of the population covariance matrix are given by $\sigma_{t,k}^{-2} \mathbb{E}[X_{g,i} X_{j,i}]$ for all $g, j = 1, \ldots, m$. As $\mathbb{E}[X_{g,i} X_{j,i}] = 0$ for $g \neq j$, this is a diagonal matrix and so is its inverse, the population covariance matrix. Hence, $\left(X_t {X_t}^\top\right)^{-1}$ for $t = 0, 1, \ldots, n_t$ in (3) converges to a diagonal matrix with elements $\mathbb{E}[X_{j,i}^2]^{-1}$, $j = 1, \ldots, m$ for $N \to \infty$. Since $\frac{1}{N} \sum_i X_i$ is independent of treatment assignment and equals $c(1, \ldots, 1)^\top$ for some $c \neq 0$, for $N \to \infty$, (3) is minimized if

$$\sum_j \left[n_d \mathbb{E}[X_{j,i}^2]^{-1} + \sum_{t>0} \mathbb{E}[X_{j,i}^2]^{-1} \right]$$

is minimized. Noting that (5) converges to this sum as $N \to \infty$ completes the proof. □

Proof of Proposition 4. 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

$$\mathrm{Var}(\hat{\beta}_j \mid X) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^n (X_{j,i} - \bar{X}_j)^2},$$

for $j = 1, \ldots, m$, with notations as in Proposition 4 (see e.g. Wooldridge, 2013).

As in the proof of Theorem 1, the covariance matrix of $\hat{\beta}_t^{p,k} - \hat{\beta}_0^{p,k}$ for any $t = 1, \ldots, n_d$ and any $k = 1, \ldots, n_y$ is given by

$$\left(X_t {X_t}^\top\right)^{-1} + \left(X_0 {X_0}^\top\right)^{-1},$$

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

Proof of Theorem 2. We use $m_{i,j} = m_{j,i} = 1$ to indicate individual i is matched to individual j; otherwise $m_{i,j} = 0$. Every individual is matched exactly once, so $\sum_i \sum_j m_{i,j} = N$, assuming the sample consists of N individuals. Usually, in that case, the goal is to minimize $\sum_i \sum_j m_{i,j}(y_i - y_j)^2$ 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

$$\text{argmin}_{(m_{i,j})} \sum_i \sum_j m_{i,j}(y_i^2 + y_j^2 - 2y_i y_j) = \text{argmax}_{(m_{i,j})} \sum_i \sum_j m_{i,j}(y_i y_j).$$

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

$$(N/2)^{-1} \sum_{\{i:D_i=0\}} y_i^2 - \bar{y}_0^2 + (N/2)^{-1} \sum_{\{j:D_j=1\}} y_j^2 - \bar{y}_1^2 = (N/2)^{-1} = \sum_i y_i^2 - \bar{y}_0^2 - \bar{y}_1^2. \tag{13}$$

Since $\bar{y}_t^2 = (N/2)^{-2}(\sum_{\{i:D_i=t\}} y_i^2 + \sum_{\{i:D_i=t\}} \sum_{\{j:D_j=t, j\neq i\}} y_i y_j)$, for $t = 0, 1$, (13) can be rewritten as

$$((N/2)^{-1} - (N/2)^{-2}) \sum_i y_i^2 - (N/2)^{-2}(\sum_{\{i:D_i=0\}} \sum_{\{j:D_j=0, j\neq i\}} y_i y_j + \sum_{\{i:D_i=1\}} \sum_{\{j:D_j=1, j\neq i\}} y_i y_j).$$

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

$$\sum_{\{i:D_i=0\}} \sum_{\{j:D_j=0, j\neq i\}} y_i y_j + \sum_{\{i:D_i=1\}} \sum_{\{j:D_j=1, j\neq i\}} y_i y_j \tag{14}$$

$$= \sum_i \sum_j y_i y_j - \sum_i y_i^2 - 2 \sum_{\{i:D_i=0\}} \sum_{\{j:D_j=1\}} y_i y_j \tag{15}$$

$$= \sum_i \sum_j y_i y_j - \sum_i y_i^2 - 2 \sum_{\{i:D_i=0\}} \sum_{\{j:D_j=1\}} m_{i,j} y_i y_j - 2 \sum_{\{i:D_i=0\}} \sum_{\{j:D_j=1\}} (1 - m_{i,j}) y_i y_j, \tag{16}$$

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 assignments after matching $\{D : D_i = |D_j - 1| \text{ for } \hat{m}_{i,j} = 1, \hat{m} \in \text{argmax}_{(m_{i,j})} \sum_i \sum_j m_{i,j} x_i x_j\}$ is still maximized over m. □

B Additional Results

Table 3: Comparison of Treatment Assignment Methods Regarding Balance in a Group of Baseline Variables (N=100)

(a) Average difference in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Indonesia	-0.056	-0.634	0.044	-0.151	0.111	0.892
Pakistan (height scores)	1.508	-0.667	-0.098	-0.105	0.578	0.263
Pakistan (test scores)	0.383	2.001	-0.330	-0.334	0.315	-0.112
Mexico	-0.383	-0.492	-0.097	0.074	0.565	-1.546
Sri Lanka	-0.633	0.632	0.708	-0.294	0.219	0.106

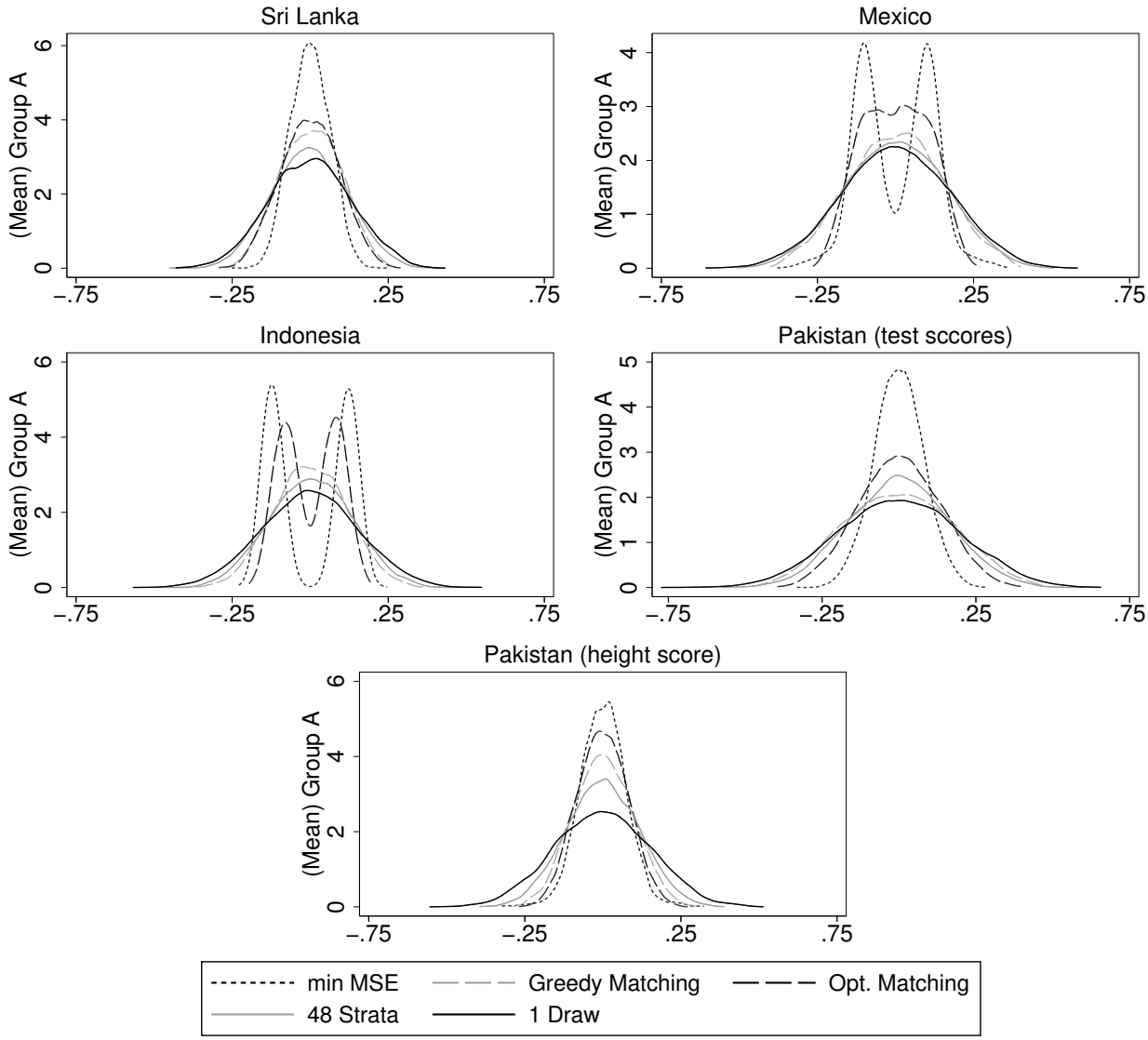
(b) Average abs. difference in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Indonesia	159.2	95.6	62.6	37.8	121.5	121.8
Pakistan (height scores)	159.7	85.7	52.3	65.4	116.8	95.2
Pakistan (test scores)	160.1	137.4	67.2	51.4	123.5	122.0
Mexico	159.2	118.0	56.5	44.3	134.6	124.4
Sri Lanka	159.1	128.5	75.7	46.1	126.2	120.6
Total	159.4	113.0	62.9	49.0	124.5	116.8

(c) 95% quantile of the difference in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Indonesia	409.8	263.9	199.7	260.9	390.3	397.5
Pakistan (height scores)	426.6	350.3	193.6	219.8	403.4	398.6
Pakistan (test scores)	394.7	438.6	218.4	276.4	394.7	407.8
Mexico	445.2	372.8	196.0	140.2	399.9	372.8
Sri Lanka	399.3	453.7	319.4	131.3	401.2	380.0
Total	415.1	375.8	225.4	205.7	397.9	391.3

Note: Statistics based on 10,000 iterations. Details on the study and the computation of each measures are explained in section 6.2. For every dataset, several variables were considered for treatment assignment. The results in this table report aggregate measures of differences in treatment group means for the group of considered variables. Lower values indicate a better balance with respect to equality of group means.



Note: Distributions of the (mean) differences in the group means of the respective group of variables are based on 10,000 treatment assignments, where the difference in one draw is the average of the distances in group means of all variables included in the assignment of treatments. Group means are expressed in standard deviations. A high mass around a difference of 0 indicates a good balance with respect to equality of the group means.

Figure 3: Distributions of the Mean Differences in Group Means (N=30)

We present the results for the scenario considered in BK09, when focusing on one of the up to seven considered variables for treatment assignment. For two datasets, groups of variables—referred to as "unobservables"—are available, but not considered during treatment assignment. Balance on these variables is reported 'group-wise' (see Section 6.2.1).

Figures 4 and 5 show the distribution of the differences in group means for the indicated variables, which are the baseline and follow-up variable of interest, for the five datasets considered with sample sizes 30 and 100. Tables 4 and 5 consist of three panels: The upper panel shows the average difference in group means, the middle panel shows the 95% quantile of this difference and the lower panel shows the proportion of draws, in which the p-value of a t-test of the differences in group means was lower than 0.1.

For the single random draw method as well as the stratification methods, results are identical to those in BK09. Differences in the pairwise greedy matching approach are probably due to the order in which we run the scripts. However, the essential part of the do-file for performing the greedy matching is the same as the one provided by BK09.

The newly introduced methods—the optimal matching approach and the min MSE procedure—perform comparable to the others and the conclusion here is the same as in BK09, namely that "on average all methods lead to balance".

In terms of average balance in baseline variables, we conclude that the min MSE procedure outperforms the other methods: For four of the five baseline variables and all unobservables, an average difference of zero to the third digit was achieved.

With respect to the whole distribution, as shown in Figures 4 and 5, the min MSE procedure shows the most favorable distribution with the highest mass at 0 and thinnest tails in half of the cases considered. Stratification seems to be superior in one dataset, where household expenditure is studied, whereas pairwise greedy matching dominates the competing mechanisms in achieving balance with the height z-score data. These findings are numerically underlined not only by the group means as discussed above, but also by the 95% quantile of the differences in group means as shown in the middle panel of Table 5, although in this panel, no mechanism clearly shows more favorable figures than another.

Consistent with the findings of BK09, we also note that with increasing sample size, balance improves. This can be seen in Figures 4 and 5, where the distributions of group means for the bigger sample sizes are mostly nearly half as wide as the distributions on the left.

With respect to the balance observed in follow-up variables, we find, and consistent with BK09, no major differences; especially for the cases in which baseline vari-

ables explained little of the variation in follow-up outcomes (Microenterprise profits in Sri Lanka and Indonesian expenditure figures). For the bigger sample size, there is hardly any difference between the covariate based treatment assignment mechanisms.

Summing up, in most comparisons, either one of the matching methods or the min MSE procedure dominates the competing mechanisms. All methods achieve balance, on average, and all decrease extreme imbalances considerably in comparison with a single random draw. However, we think results that consider the overall balance of all variables considered in treatment assignment are more informative. These results are discussed in the main text, see Section 7.1, Tables 1 and 3 and Figures 1 and 3.

Table 4: Comparison of Treatment Assignment Methods Regarding Balance in the Baseline Outcome (N=30)

(a) Average difference in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Microenterprise profits (Sri Lanka)	-4.214	-0.636	4.004	-1.302	-5.657	0.239
Household expenditure (Indonesia)	-2.094	-2.908	2.218	-2.794	0.644	2.739
Labor income (Mexico)	2.627	1.375	-1.346	-1.062	-0.774	-0.195
Height z-score(Pakistan)	-2.502	0.670	-3.486	0.034	-0.506	-0.228
Math test score (Pakistan)	-1.741	-0.632	-2.306	-0.301	-1.464	-1.571
Baseline unobservables (Sri Lanka)	-0.641	-0.331	0.837	0.896	-1.192	1.144
Baseline unobservables (Mexico)	-0.112	-0.130	-0.066	-0.390	-0.114	-0.673

(b) 95% quantile of the difference in baseline group means between the treatment and the control group in 1000 SD

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Microenterprise profits (Sri Lanka)	705.6	598.0	415.9	227.6	415.9	538.0
Household expenditure (Indonesia)	716.2	458.1	478.0	643.4	346.9	500.9
Labor income (Mexico)	690.8	176.5	223.7	228.1	409.1	581.8
Height z-score(Pakistan)	710.1	257.9	467.3	393.8	444.8	445.6
Math test score (Pakistan)	712.8	256.8	362.6	161.2	408.8	586.3
Baseline unobservables (Sri Lanka)	802.8	879.4	879.4	889.3	824.4	804.6
Baseline unobservables (Mexico)	834.3	834.3	879.4	834.3	771.3	774.9

(c) Proportion p-values <0.1 when testing the difference in baseline group means

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Microenterprise profits (Sri Lanka)	0.097	0.049	0.001	0.000	0.000	0.021

Table 5: Comparison of Treatment Assignment Methods Regarding Balance in the Baseline Outcome (N=100)

(a) Average difference in baseline group means between the treatment and the control group in 1000 SD

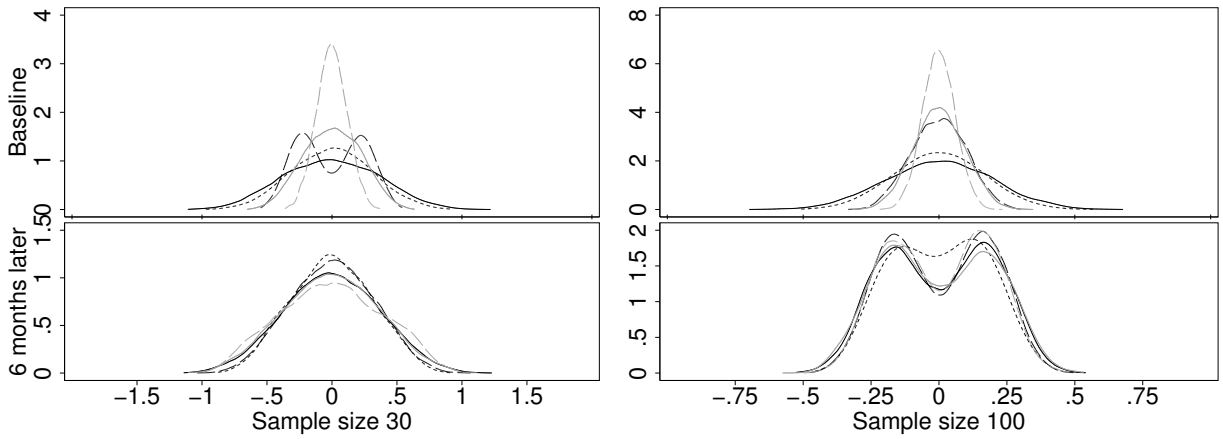
	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Microenterprise profits (Sri Lanka)	1.388	0.590	0.655	-0.638	-0.053	-0.777
Household expenditure (Indonesia)	-2.223	-1.665	-0.153	0.409	0.810	-0.679
Labor income (Mexico)	-0.428	-0.493	-1.051	-0.002	0.024	-0.295
Height z-score(Pakistan)	1.336	0.025	0.413	0.287	0.832	0.117
Math test score (Pakistan)	2.946	-0.260	-1.419	-0.288	-0.116	-0.555
Baseline unobservables (Sri Lanka)	-0.205	-0.593	1.102	-0.447	0.305	0.031
Baseline unobservables (Mexico)	0.135	-0.087	-0.113	-0.062	0.250	-0.218

(b) 95% quantile of the difference in baseline group means between the treatment and the control group in 1000 SD

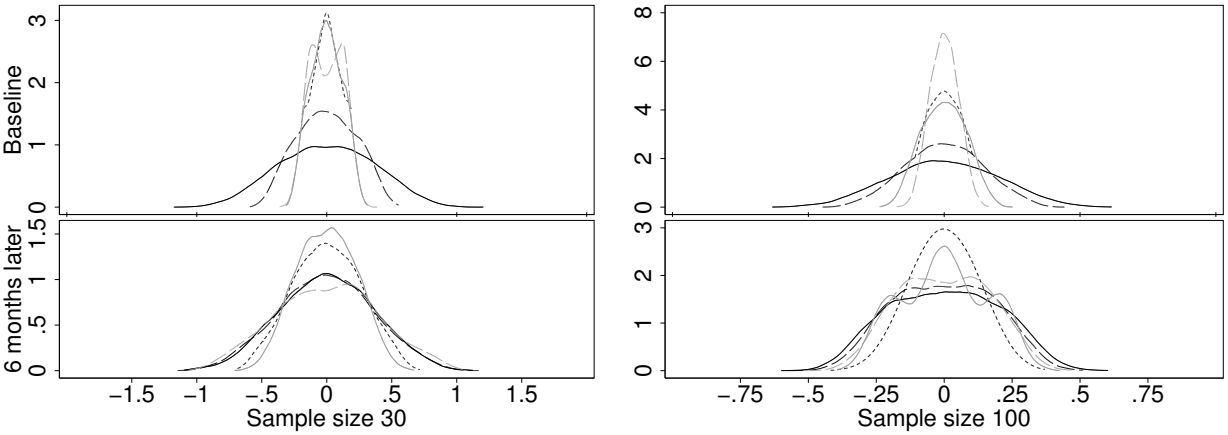
	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Microenterprise profits (Sri Lanka)	386.3	314.6	183.0	119.3	195.5	240.7
Household expenditure (Indonesia)	390.2	263.9	198.5	260.9	145.0	191.0
Labor income (Mexico)	383.9	99.5	153.8	100.1	280.2	304.0
Height z-score(Pakistan)	394.9	102.7	189.9	185.2	160.1	206.0
Math test score (Pakistan)	392.2	74.5	184.5	106.5	163.6	237.3
Baseline unobservables (Sri Lanka)	434.2	434.2	434.2	434.2	417.0	414.2
Baseline unobservables (Mexico)	456.5	456.5	456.5	456.5	447.6	439.0

(c) Proportion p-values <0.1 when testing the difference in baseline group means

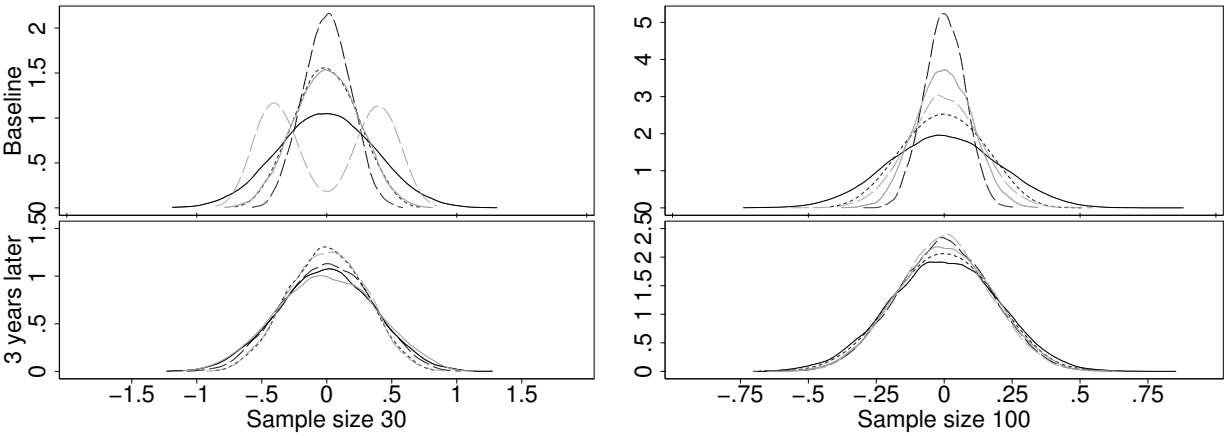
	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
Microenterprise profits (Sri Lanka)	0.097	0.039	0.000	0.000	0.000	0.005



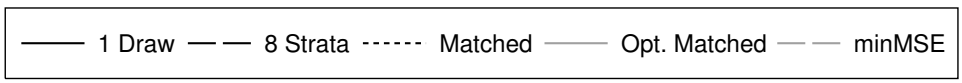
(a) Panel A. Sri Lanka. Differences in average profits (weighted by standard deviation)



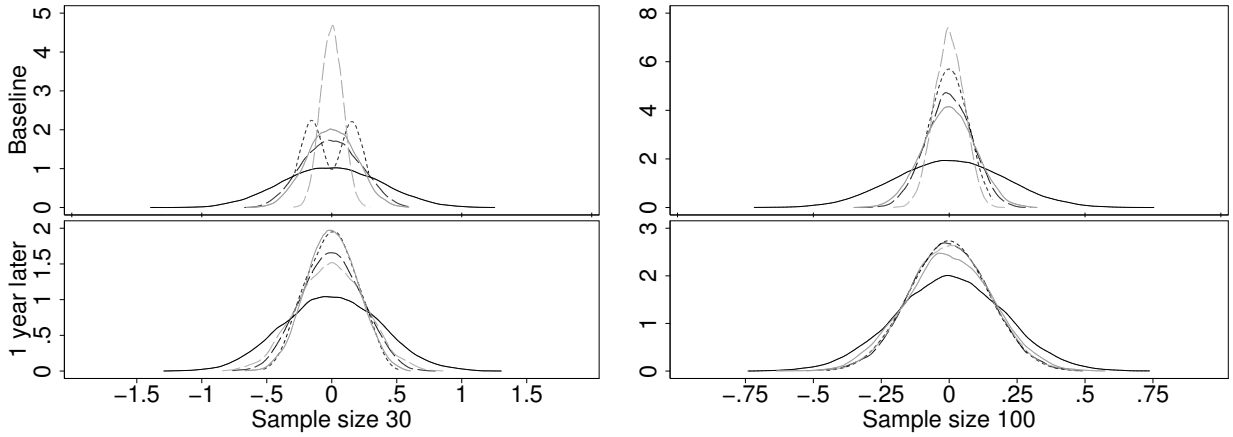
(b) Panel B. Mexico ENE. Differences in average income (weighted by standard deviation)



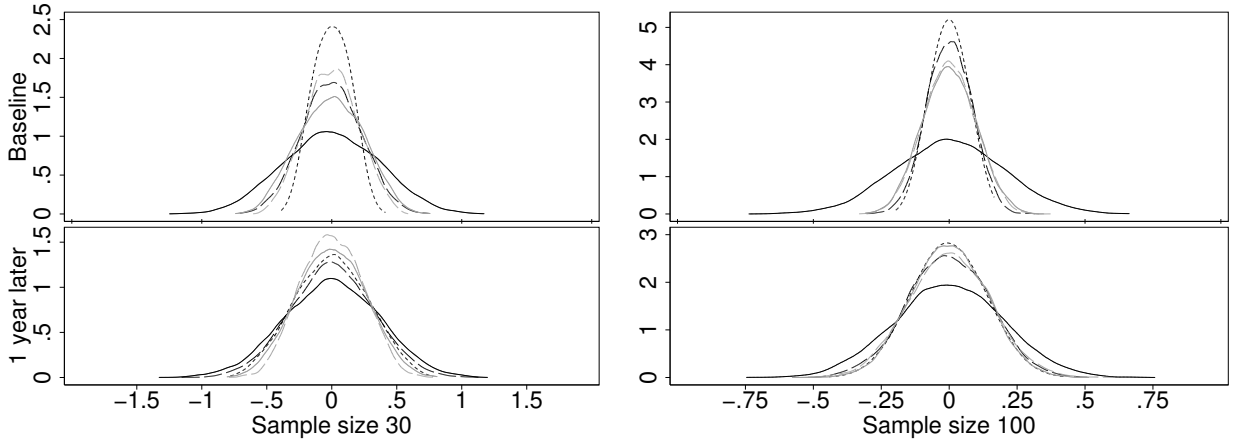
(c) Panel C. IFLS. Differences in average *hh* expenditure *p cap* (weighted by standard deviation)



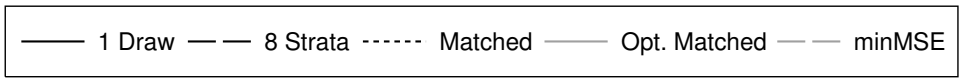
Note: Panel A, B and C show the distributions of the differences in treatment group means of the indicated variable for 10,000 treatment assignments, expressed in standard deviations. The higher the mass around a difference of 0, the better the treatment assignment with respect to balancing group means. Baseline variable is the outcome of interest; one of up to six variables used for treatment assignment. Follow-up variable is the same variable, measured six months after the baseline variable was collected; note that it is not included in treatment assignment.



(a) Panel D. LEAPS. Differences in average in math test score (weighted by standard deviation)

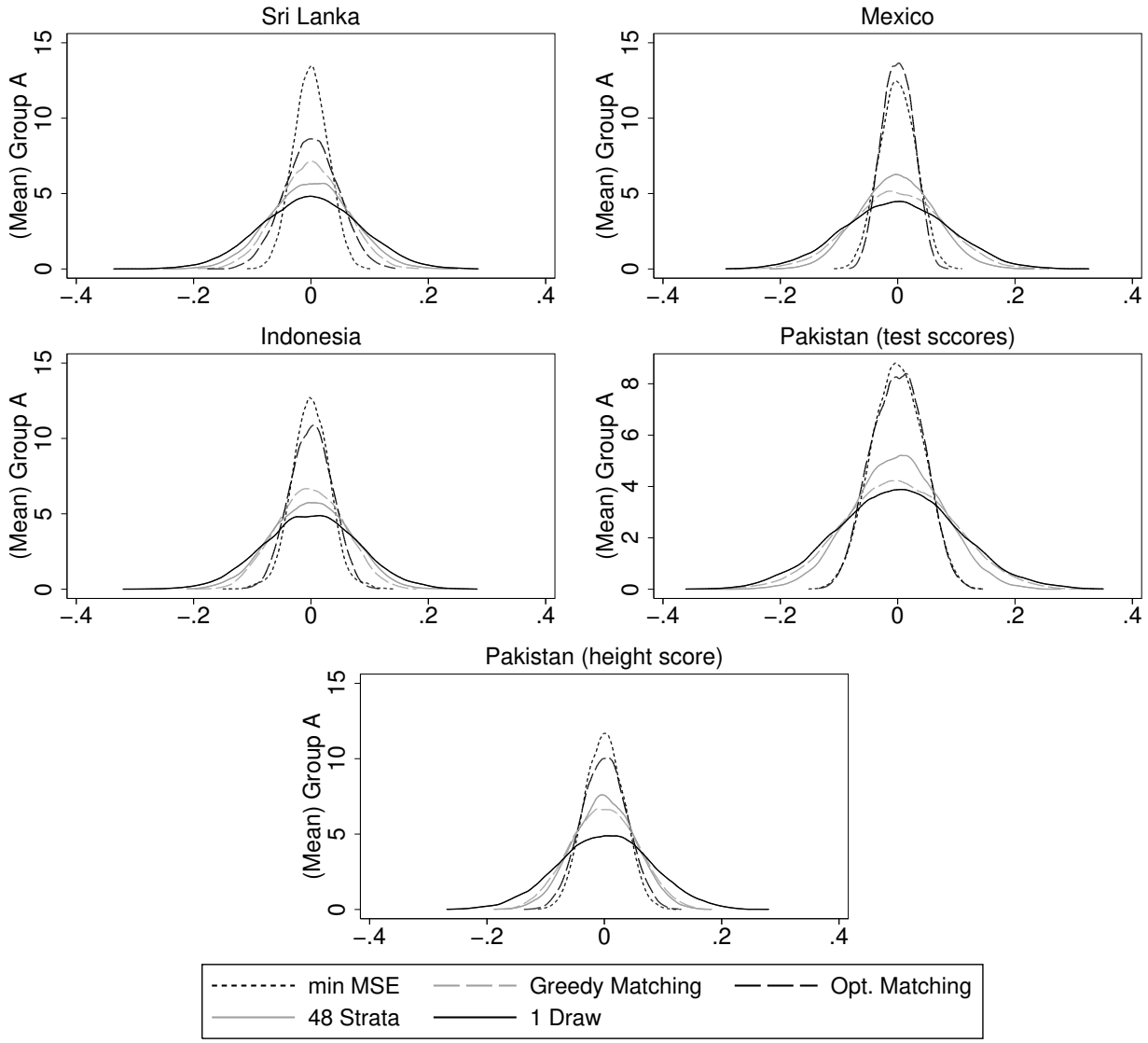


(b) Panel E. LEAPS. Differences in average z-score (weighted by standard deviation)



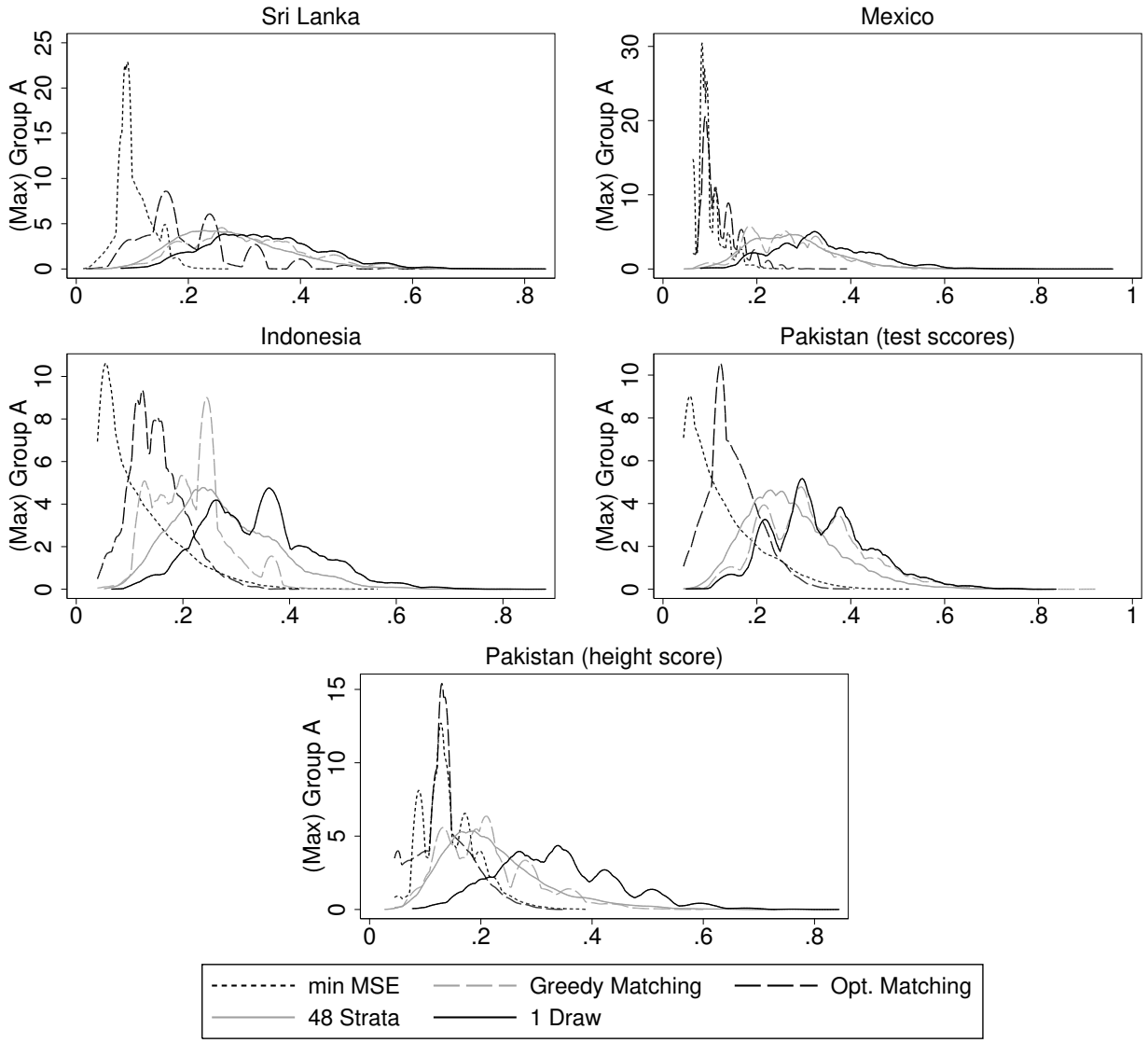
Note: Panel D and E show the distributions of the differences in treatment group means in standard deviations of the indicated variable for 10,000 treatment assignments. The higher the mass around a difference of 0, the better the treatment assignment with respect to balancing group means. Baseline variable is the outcome of interest; one of up to six variables used for treatment assignment. Follow-up variable is the same variable, measured six months after the baseline variable was collected; note that it is not included in treatment assignment.

Figure 5: Distributions of the Differences in Group Means Between the Treatment and the Control Group in the Baseline Variable and the Follow-up Variable (N=30)



Note: Distributions of the (mean) differences in treatment group means among the group of variables to consider for treatment assignment are based on 10,000 treatment assignments. Differences in group means are expressed in standard deviations. A high mass around a difference of 0 indicates a good balance with respect to equality of group means.

Figure 6: Distributions of the Mean Differences in Group Means (N=100)



Note: Distributions of the (maximal) differences in treatment group means among the group of variables to consider for treatment assignment are based on 10,000 treatment assignments. Differences in group means are expressed in standard deviations. A high mass around a difference of 0 indicates a good balance with respect to equality of group means.

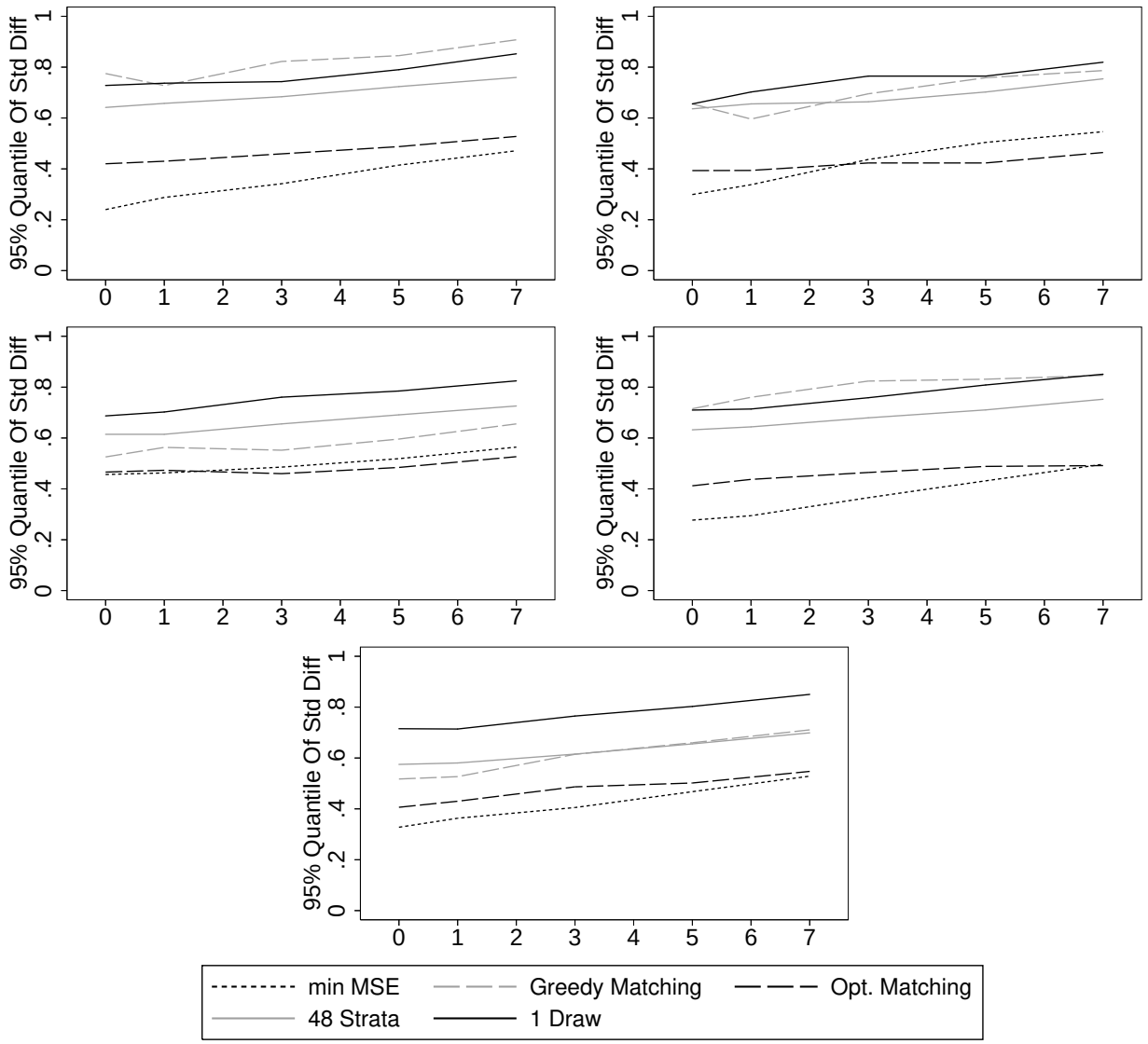
Figure 7: Distributions of the Maximal Difference in Group Means (N=100)

Table 6: Comparison of Treatment Assignment Methods Regarding Balance in Baseline Variables in Case of Attrition (Assessing Balance Using T-tests, N=30)

(a) Proportion p-values <0.1 when testing the difference in baseline group means

	Single random draw	Pairwise greedy matching	'Optimal matching'	Min MSE procedure	Stratified on two variables	Stratified on eight variables
No attrition	0.1131	0.0652	0.0045	0.0058	0.0678	0.0562
1 (2) unit removed	0.1123	0.0597	0.0047	0.0061	0.0665	0.0545
3 (6) units removed	0.0909	0.0642	0.0035	0.0069	0.0640	0.0571
5 (10) units removed	0.0926	0.0671	0.0048	0.0089	0.0635	0.0580
7 (14) units removed	0.0918	0.0686	0.0051	0.0131	0.0644	0.0606
Total	0.1001	0.0650	0.0045	0.0082	0.0652	0.0573

Note: Statistics based on 10,000 iterations. The smaller the proportion of p-values $< .1$, the better the balance with respect to similar group means. Details on the study and the computation of each measure are explained in section 6.2.



Note: Attrition. Evolution of the largest 95% quantile of differences in group means between treatment groups among the group of variables to consider for treatment assignment. X-axis: Number of units that are randomly removed from the study. Distributions are based on 10,000 treatment assignments. Differences in group means are expressed in standard deviations. A lower (maximal) 95% quantile of differences in group means indicates a better balance with respect to equality of group means.

Figure 8: Simulated Attrition: Evolution of the Largest 95% Quantile of the Differences in Group Means Among The Group of Baseline Variables, N=30

References

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

Andrabi, T., J. Das, and A. I. Khwaja (2015). Delivering education: A pragmatic framework for improving education in low-income countries. In P. Dixon, S. Humble, and C. Counihan (Eds.), *Handbook of International Development and Education*, Volume 75, Chapter 6, pp. 85–130. Cheltenham, UK: Edward Elgar Publishing.

Bruhn, M. and D. McKenzie (2009, September). 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*.

Cochran, W. G. and D. B. Rubin (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A 35*(4), 417–446.

De Mel, S., D. McKenzie, and C. Woodruff (2008). Returns to capital in microenterprises: Evidence from a field experiment. *The Quarterly Journal of Economics 123*(4), 1329–1372.

Fedorov, V. V. (1997). *Model-oriented design of experiments.* New York [u.a.]: Springer. Literaturverz. S. 111 - 112.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, Scotland: Oliver & Boyd.

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

Hansen, B. B. and J. Bowers (2008, May). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science 23*(2), 219–236.

Imbens, G. W. (2004, February). 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, July). 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.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). Optimization by simulated annealing. *Science 220*(4598), 671–680.

Klar, N. and A. Donner (1997, 08). The merits of matching in community intervention trials: A cautionary tale. *Statistics in Medicine 16*(15), 1753–1764.

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

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

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

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

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

Rubin, D. B. (1973, March). 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.

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*(1/2), 1–85.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5 ed.). Mason, OH: South-Western, Cengage Learning.