Courant Research Centre

'Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis'

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 228

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

Sebastian Schneider and Martin Schlather

June 2017

Platz der Göttinger Sieben 5 · 37073 Goettingen · Germany
Phone: +49-(0)551-3921660 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

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

Sebastian Schneider
University of Goettingen, Germany
sebastian.schneider@zentr.uni-goettingen.de

Martin Schlather
University of Mannheim, Germany

May 31, 2017

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 minimizing MSE approach – uses sample characteristics to obtain balanced treatment groups. Compared to other methods, the min MSE procedure is attrition tolerant, more flexible and very fast, it can conveniently be implemented and balances different moments of the distribution of the treatment groups. Additionally, it has a clear theoretical foundation, which bases on the idea by Kasy (2016), but involves randomness, works without any parameter to be specified by the researcher and is extended to multiple treatments. The information used for treatment assignment can be multivariate and continuous and consist of arbitrary many variables. In this paper, we theoretically derive the underlying selection criteria which we then apply to various simulated treatment effect scenarios and datasets, comparing it to established approaches. Our proposed method performs superior or comparable to competing approaches such as matching in most measures of balance commonly used. We provide software to apply the min MSE approach as an ado-package for Stata.

1. Introduction

From the early days of experiments involving random treatment allocation, researchers have thought about methods to ensure the decisive fundament of any experiment – similar treatment groups in absence of treatment. When discussing Darwin's experiment comparing growth of crossed and self-fertilized plants, Fisher (1935) argues, that it is not enough to randomly assign plots to treatment or control, i.e. to crossed or self-fertilized plants. An unbiased experiment alone, he continues, does not "ensure the validity of the estimates, [...] for it might well be that some unknown circumstance, such as the incidence of different illumination at different times of the day [...], might systematically favour all the plants on one [plot] over those on the other."

Would the experiment have been repeated several times, or when a large number of plots would have been used, any difference caused by those "unknown circumstances" would diminish in expectation. The same is of course true for *known* or *observed* circumstances.

Fisher notices that an increase in *precision* of the experiment could also be achieved differently – by making the groups more similar. His suggestion to achieve both, the validity of estimates *and* an increase in precision was to change the level of randomization by allowing plants from both groups to be planted in the two available plots, thus being exposed to the circumstances of both plots.

Hence, not only from the perspective of validity, but also with respect to efficiently estimating the outcome of an experiment it is desirable to have similar or *balanced* treatment groups. A newer motivation for seeking balance across treatment groups comes from the interest in subgroup analysis. And lastly, a quantity often estimated in the recent impact evaluation literature – the conditional average treatment effect – can formally only be estimated if the so called *overlap condition* is fulfilled – a weak criterion of balance[1](Abadie and Imbens, 2006).

When relying on randomization only, balanced groups are not ensured: Following Fisher's suggestion for the specific case illustrated above, it could happen that – by pure chance – all self-fertilized and all crossed plants are still allocated to separate plots.

Therefore, it has long been recognized that group characteristics or *circumstances* should be accounted for when assigning treatment to experimental subjects and subsequently analyzing such experiments, see e.g. Cox (1957) for an early review of the possibilities to do so. Today, several strategies are widely used.

Stratification or blocking goes back to Fisher (1935). The idea is to build subgroups according to observable characteristics and randomize within those subgroups. Although this improves "balance" in comparison to purely random treatment assignment, it is impractical in several aspects: With stratification, it is only possible to balance a very limited number of variables. Furthermore, continuous variables have to be discretized arbitrarily 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.

[1]Although this is rather a condition ruling out imbalance.

Pairwise matching is often seen as the limit case of stratification, where the subgroups consist of only two individuals. The subgroups, in case of matching called *pairs*, have to be created[2] 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, King, and Nall, 2009) and an optimal matching algorithm (Greevy, Lu, Silber, and Rosenbaum, 2004; Lu, Greevy, Xu, and Beck, 2011). Matching can be realized with many, possibly continuous variables and thus eliminates some shortcomings of stratification. This however comes at the cost of analytical difficulties with the estimation of treatment effect variances (e.g. Abadie and Imbens, 2006; Imbens, 2011; Klar and Donner, 1997). Further problems arise when attrition happens, especially in small samples or when performing randomization at the cluster level: For every unit – possibly consisting of many individuals – dropping out of the experiment, its pair should also be removed, which lowers sample size and power and can be a major concern. Additionally, we are unaware of an approach to extend matching to multiple treatment arms. Also, matching can only be performed when the number of units is even. Furthermore, for the matching approach implemented in Bruhn and McKenzie (2009), the treatment assignment with a sample size of 300 units ran several days, so this approach has to be rejected if time is scarce.

Several so called rerandomization methods have evolved – probably because of the theoretical or practical limitations of the above mentioned 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 rerandomize a certain number of times and chose the best assignment, according to a specified evaluation criteria. Sometimes, also subjective judgment is 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, focuses on the mean only, completely ignoring other dimensions of the distributions of the variables to seek balance on. Irrespectively of this limitation, we are unaware of a software implementation of this approach or an extension to multiple treatment arms.

Kasy (2016) applies a decision theoretic, Bayesian approach for determining 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. The posterior MSE, i.e. the sum of bias and variance is then to be minimized across treatment assignments. When the estimator is modeled with a linear model, this leads to a decision criterion that balances not only the mean of the variables of interest, but also partial correlations.

[2]Note that this is another task as the one to be performed for matching in observational studies: Finding pairs when groups are already formed is, also from a computational aspect, by far less demanding.

Based on the introduced decision theoretic framework, Kasy (2016) argues that a deterministic assignment rule is superior to any random assignment in terms of minimizing the MSE.

Drawbacks of this method however are the limitation to only one treatment group and the number and nature of the required parameters: To apply this approach, the researcher has to specify a mean vector and a covariance matrix of the regression coefficient vector in a linear model explaining potential outcomes as function of covariates. In addition, a guess for the R^2 of that linear regression model has to be specified. Apart from that, a non-random treatment assignment rules out the possibility to perform conventional randomization inference. Lastly, to date, we are unaware of a software implementation in commonly used statistics software[3].

We simplify this method considerably, extend it to multiple treatments, provide a software implementation as ado-package for Stata and thus increase its applicability. Apart from that we interpret and implement the method as a rerandomization method, which yields the possibility of randomization inference.

Bayesian modeling allows for a bigger flexibility in many cases because it relies on distributions instead of parameters. In this case, however, at least when using a linear model, we think it complicates the treatment assignment process unnecessarily: For modeling an abstract quantity – namely the potential outcome of an individual for a given treatment – the researcher has to pick a mean vector and a covariance matrix of the regression coefficient vector and further guess the R^2 of that regression. 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 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 of course means that we only get a point estimate instead of a distribution for any result. Since we are interested in the MSE and not in its distribution anyways, this even comes without limitation.

A nice side effect is that we can factor out variances of the decision criteria and thus the only parameters to specify are ratios – if we want. Otherwise 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.

Our result thus works without choosing any parameters while still allowing for the needed flexibility. In the result derived here, the only parameter that necessarily has to be chosen is the number of treatment groups desired; other parameters can be specified, but can be left constant unless a better guess is available. These default values have an intuitive interpretation and are not chosen by us, but follow from the theoretical derivation of the method as laid out in this paper.

[3]Kasy (2016) provides MATLAB code.

Another advantage of the frequentist approach compared to the Bayesian approach is the reduced computation time. This is relevant, when extending the method to various treatment groups. Also, when implementing the min MSE method in statistical software, computation time might be a concern.

The software package we provide allows assigning treatment groups with one single line of code – 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, divisible by the number of treatments, ...). Another feature is its speed: a reasonably good balance for a sample size of 100 units and 10 variables is usually achieved in less than 5 minutes.

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 and find that it is comparable to the matching methods, without being as vulnerable to attrition as they are.

Additionally, we include an optimal matching algorithm and add different measures of balance and different scenarios.

The structure of this paper is as follows: Section 2 introduces our approach to treatment assignment. Section 3 explains the design of the simulation study and explains its implementation. Section 4 concludes by reporting the results.

2. Finding an MSE Minimizing Treatment Assignment

2.1. Estimation of the Treatment Effect

First, we define the parameter we finally want to estimate: the conditional average treatment effect. We do so by introducing the potential-outcome framework (Fisher, Neyman, Rubin), 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. As usual, 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 (Treatment)**.** Assume we have n_d treatments numbered by $1, \ldots, n_d$ and let 0 denote no treatment. Let $D_1, \ldots, D_N$ be random variables with values in $\{0, 1, \ldots, n_d\}$. Then, $D = (D_1, \ldots, D_N)^\top$ is called a random *treatment group assignment*, or *treatment assignment* for short.

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

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

Treatments might affect more than one outcome, so we build our framework to consider multiple outcomes.

Definition 3 (Potential Outcome)**.** Assume we are interested in n_y outcomes numbered by $1, \dots, n_y$. 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 case of treatment t, where the superscript p indicates *potential outcomes*.

These *potential outcomes* of individual i in 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 (Realized Outcome)**.** 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, where the superscript now indicates *realized outcomes*.

The *realized outcomes* Y_i^r of individual i can now be written by means of *potential outcomes*:

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

The right-hand side of above formula decomposes the *realized outcomes* for an individual in her *potential outcomes*. The differences $Y_{i,t}^p - Y_{i,0}^p$, called causal effects of treatment, 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 lies in studying a subpopulation (e.g. the poor), or when one is not sure whether the sample at hand is representatitive for the population, one focuses on the *conditional* average treatment effect (Imbens, 2004). This happens frequently in Development Economics, for instance.

Definition 5 (Conditional Average Treatment Effect)**.** For every treatment $t \in 1, \dots, n_d$,

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

is called the *conditional average treatment effect* of treatment t.

The random matrix $T = (\tau_{t,k})_{t=1,\dots,n_d;k=1,\dots n_y}$ contains all 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 (or sometimes unconfoundedness assumption), means that potential outcomes are independent of group and therefore treatment assignment, given covariates. If the *Conditional Independence Assumption* holds, any potentially given selection bias vanishes and the observed difference in average outcomes conditioned on observables between treatment and control group has a causal, conditional treatment effect interpretation.

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 comparison of expectations of potential outcomes given those covariates is not possible. It is in general never guaranteed that this is possible, but a powerful treatment assignment procedure will make it more likely. Formally (Abadie and Imbens, 2006):

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

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

$$\eta < \Pr(D_i = 1 \,|\, X_i = x) < 1 - \eta \text{ for some } \eta > 0. \tag{Overlap}$$

2.2. 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 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 $w^d = (w_1^d, \ldots, w_{nd}^d)$ and $w^y = (w_1^y, \ldots, w_{nd}^y)$ be non-negative weights. Then, for the matrix of weighted estimators $\mathrm{diag}(\sqrt{w^d})(\hat{T} - T)\,\mathrm{diag}(\sqrt{w^y})$, we define the conditional weighted MSE component-wise as

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

where $\|\cdot\|_F$ denotes the Frobenius norm (also called Hilbert-Schmidt norm). The vectors w^d and w^y include weights on the t-th treatment (starting with $t = 0$) and k-th outcome, respectively. 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 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.

This yields the following objective function:

Assumption 2 (Objective Function)**.** The objective function S is given by the expectation of the squared Frobenius norm of the weighted estimation error matrix. Hence, we seek an estimator $\hat{T}$ minimizing this function:

$$\begin{aligned}
\eta &\in \operatorname*{argmin}_{\hat{T}} S_T(\hat{T}) \\
&= \operatorname*{argmin}_{\hat{T}} \mathbb{E}\left[|| \operatorname{diag}(\sqrt{w^d})(\hat{T}-T)\operatorname{diag}(\sqrt{w^y})||_F^2 \mid X \right] \\
&= \operatorname*{argmin}_{\hat{T}} \mathbb{E}\left[\operatorname{tr}\left(\operatorname{diag}(\sqrt{w^y})(\hat{T}-T)^\top \operatorname{diag}(w^d)(\hat{T}-T)\operatorname{diag}(\sqrt{w^y}) \right) \mid X \right] \\
&= \operatorname*{argmin}_{\hat{\tau}_{t,k}} \mathbb{E}\left[\sum_{t=1}^{n_d} w_t^d \sum_{k=1}^{n_y} w_k^y (\hat{\tau}_{t,k}-\tau_{t,k})^2 \mid X \right].
\end{aligned}$$

As the quantity of interest, the conditional average treatment effect, is a function of the covariates, it is natural to assume the same for its estimator:

Assumption 3 (Estimators are Functions of Covariates)**.** We write the estimator of the treatment effects $\hat{T}$ as (measurable) function of X, so $\hat{T} = m(X)$.

With this assumption, we analyze the objective function of Assumption 2 a bit further: The weights do not depend on $\hat{T}$, so

$$\begin{aligned}
\eta &\in \operatorname*{argmin}_{\hat{T}} S_T(\hat{T}) \\
&= \operatorname*{argmin}_{\hat{T}} \mathbb{E}\left[\operatorname{tr}\left((\hat{T}-T)^\top (\hat{T}-T) \right) \mid X \right] \\
&= \operatorname*{argmin}_{m(X)} \operatorname{tr}\left(\mathbb{E}\left[(m(T)-T)^\top (m(T)-T) \mid X \right] \right)
\end{aligned}$$

Now,

$$\begin{aligned}
&\mathbb{E}((m(X)-T)^\top (m(X)-T) \mid X) \\
&\quad = \mathbb{E}\left[(m(X)-\mathbb{E}[T \mid X]+\mathbb{E}[T \mid X]-T)^\top (m(X)-\mathbb{E}[T \mid X]+\mathbb{E}[T \mid X]-T) \mid X \right] \\
&\quad = (m(X)-\mathbb{E}[T \mid X])^\top (m(X)-\mathbb{E}[T \mid X]) + \mathbb{E}\left[(\mathbb{E}[T \mid X]-T)^\top (\mathbb{E}[T \mid X]-T) \mid X \right]
\end{aligned}$$

Since the last summand does not involve $m(X)$, the trace 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 to estimate $\mathbb{E}\left[Y_{i,t}^p \mid X_i\right]$ for all treatment groups $t = 0, 1, \ldots, n_d$.

2.3. A linear model for Potential Outcomes

For introducing our approach to treatment assignment in this paper, we choose a linear model for the relationship between covariates and potential outcomes.

Model equation and assumptions

$$Y_{i,t}^{p,k} = X_i^\top \beta_t^{p,k} + \varepsilon_{i,t}^{p,k}, \tag{1}$$

for $i \in \{i : D_i = t\}$ and $k = 1, \ldots, n_y$, $t = 0, 1, \ldots, n_d$ 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 with $\varepsilon_{i,t}^{p,k} \mid X_i \sim \mathcal{N}(0, \sigma_{t,k}^2)$.

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: $\sigma_{t,k}^2 = s_{t,k}\sigma_{0,k}^2$ and $\sigma_{0,k}^2 = s_{0,k}\sigma_0^2$ for some $\sigma_0^2 > 0$.

Implications Let the submatrix X_t of X contain the covariate vectors of all individuals in treatment group t, that is

$$X_t := (X_{i_1}, X_{i_2}, \ldots, X_{i_{n_t}}), \qquad \{i_1, i_2, \ldots, i_{n_t}\} \in \{i : D_i = t\}.$$

Accordingly, we define the subvector of the k-th potential outcome

$$Y_{\{D_i=t\},t}^{p,k} := (Y_{i_1,t}^{p,k}, Y_{i_2,t}^{p,k}, \ldots, Y_{i_{n_t},t}^{p,k})^T$$

and the respective subvector of error terms

$$\varepsilon_{\{D_i=t\},t}^{p,k} := (\varepsilon_{i_1,t}^{p,k}, \varepsilon_{i_2,t}^{p,k}, \ldots, \varepsilon_{i_{n_t},t}^{p,k})^T,$$

where again $\{i_1, i_2, \ldots, i_{n_t}\} = \{i : D_i = t\}$. We then can write equation (1) for all t and k in matrix notation:

$$Y_{\{D_i=t\},t}^{p,k} = {X_t}^\top \beta_t^{p,k} + \varepsilon_{\{D_i=t\},t}^{p,k}. \tag{2}$$

For this linear model, it is known that the least squares estimator of $\beta_t^{p,k}$ for all $t = 0, 1, \ldots, n_d$ and $k = 1, \ldots, n_y$ is given by

$$\hat{\beta}_t^{p,k} = (X_t {X_t}^\top)^{-1} X_t Y_{\{D_i=t\},t}^{p,k}, \tag{3}$$

and further:

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

2.3.1. Derivation of the optimal treatment assignment procedure

Following from this model, we get

$$\begin{aligned}
\mathbb{E}\left[(\hat{\tau}_{t,k} - \tau_{t,k})^2 \mid X\right] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_i \left(\hat{Y}_{i,t}^{p,k} - \hat{Y}_{i,0}^{p,k}\right) - \frac{1}{N}\sum_i \left(E\left[Y_{i,t}^{p,k} \mid X_i\right] - E\left[Y_{i,0}^{p,k} \mid X_i\right]\right)\right)^2 \mid X\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_i X_i^T \left(\left(\hat{\beta}_t^{p,k} - \beta_t^{p,k}\right) - \left(\hat{\beta}_0^{p,k} - \beta_0^{p,k}\right)\right)\right)^2 \mid X\right] \\
&= \frac{1}{N^2}\sum_i X_i^T \left(\mathrm{Cov}\left(\hat{\beta}_t^{p,k} - \beta_t^{p,k} \mid X\right) + \mathrm{Cov}\left(\hat{\beta}_0^{p,k} - \beta_0^{p,k} \mid X\right)\right)\sum_i X_i \\
&= \frac{1}{N^2}\sum_i X_i^T \left(\sigma_{t,k}^2 (X_t {X_t}^\top)^{-1} + \sigma_{0,k}^2 (X_0 {X_0}^T)^{-1}\right)\sum_i X_i,
\end{aligned}$$

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

For our objective function specified in assumption 2 we get

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

where $\| \cdot \|_1 = \sum |\cdot|$ is the 1-norm of a vector and we summarize weights and scaling factors for the variance as follows: $\tilde{w}_k^y = w_k^y s_{0,k}$ and $\tilde{w}_t^d = w_t^d s_{t,k}$. Assuming the same variance for

all outcomes and treatment groups including the control group and neglecting any weights simplifies this result to

$$S_T(\hat{T}) \propto \frac{n_y}{N^2} \sum_i X_i^T \left[n_d \left(X_0 {X_0}^T \right)^{-1} + \sum_t \left(X_t {X_t}^T \right)^{-1} \right] \sum_i X_i. \tag{4}$$

To summarize, this approach allows for considerable flexibility if one has a good guess why an outcome should have a higher variance than the other in general, and more specifically that treatment t that mainly aims on outcome k might have compliance problems, so compared to the control a higher variance is to be expected.

Contrary to the result by Kasy (2016), our approach does not require guessing any (absolute) values: Since we do not assume a prior distribution for $\beta_t^{p,k}$, there is no need to specify its parameters: a mean and – more difficult – a covariance matrix for this parameter vector for every combination of k and t. Further, there is no need of specifying the R^2 for the model of each potential outcome in order to express the model variance; instead, one can simply specify relative scaling factors.

2.4. An (alternative) linear model for Potential Outcomes

If we assume the treatment effect to be constant across individuals, there is no need to specify different models for the potential outcomes of the control group and the treatment group of interest: they only differ by a constant. We can still control for covariates, however, instead of taking a simple difference in means estimator. Thus, for every treatment t we specify a linear, additive model as follows:

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

where the subscript t of Y_t, Z_t and W_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. W contains the covariate vectors $X_i^\top$ of individuals in treatment group t and in the control group.

In this model, τ_t is half the estimated treatment effect.

Under Gauss-Markov assumptions (additive errors, that are uncorrelated conditional on W_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 $P_t = I - W_t \left(W_t^\top W_t \right)^{-1} W_t^\top$ and minimized for $W_t^\top Z_t = 0$ (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

$$S_T^*(\hat{T}) \propto \sum_t 1/Z_t^\top P_t Z_t. \tag{5}$$

This is the criteria considered by Greevy et al. (2004) to compare the efficiency of treatment assignment.

3. Simulation Study

3.1. Study Design: Treatment Assignment Mechanisms and Data

In order to investigate the performance of the treatment assignment procedure described in section 2, we ran a simulation study to compare the new mechanism to established ones. We did so using a similar approach as Bruhn and McKenzie (2009), BK from hereon.

BK compare five treatment assignment methods (purely random assignment, pairwise greedy matching, stratification and two rerandomization schemes) when "building" one treatment and one control group in terms of "balance" of relevant observable and "unobservable" variables using different datasets and different sample sizes.

We extend this study by adding the scenario of multiple treatment arms.

In terms of treatment assignment mechanisms, we also include an optimal matching approach and of course the min MSE procedure as introduced in section 2 of this paper.

To rule out that results in this study are limited to only a specific dataset or sample size, we consider several data sets, following BK.

Thus, for a variety of datasets, variables and sample sizes, we can credibly compare the treatment assignment method introduced in this paper to different other methods, and also add insights about the optimal matching approach.

3.1.1. Data

The data used for the simulation results of this paper is the same as the data used by BK for reasons of comparability. It consists of four panel datasets, with different data from different contexts.

The first dataset contains data about microenterprises in Sri Lanka and is from an actual randomized experiment by De Mel, McKenzie, and Woodruff (2008). The outcome variable of interest is firms' profits, and data about firm and owner characteristics at 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 took the same subsamples as BK. In this dataset, the outcome of interest for us is income of household heads, that were employed and between age 20 and 65 when the baseline was conducted in 2002. In addition to this, the dataset includes further characteristics about the household head and the household, which are again used either for treatment assignment or as "unobservables".

The third dataset uses subsamples of two waves (IFLS 2 and IFLS 3) of the Indonesian Family Live Survey (IFLS)[4]: The year 1997 (IFLS 2) is used as baseline and the data from

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

2000 (IFLS 3) is treated as follow-up. We only use data on household expenditure from this survey[5].

The fourth dataset is from the Learning and Educational Achievement project in Pakistan, used e.g. in Andrabi, Das, and Khwaja (2015). It contains child and household data, and the outcome variables of interest are math test scores and z-scores of children aged 8 to 12 at baseline.

It is interesting to note that the subsamples of 30 and 100 observations sometimes differed considerably in terms of the share of variation of the follow-up variable explained by the group used for treatment assignment: In the data on firms' profit, for the small subsample, around 6 percent of variation could be explained by the corresponding baseline variables; for the subsample of 100 observations it was 18 percent.

A similar pattern could be observed for the Mexican data: roughly 7 and 32 percent (for the subsamples of 30 and 100 observations, respectively) of variation was explained, depending on the subsample used.

In the data on household expenditure (Indonesian Family Live Survey), this share was roughly the same in both subsamples (31/28 %), which was also the case for the math test score in the data from Pakistan (46/48 %).

In the other data from the Learning and Educational Achievement project in Pakistan, the variation in the follow-up variable on height z-scores, in the small subsample an even higher share could be explained than in the subsample covering 100 observations (64/51 %).

This observation gives rise to a method for drawing comparative samples of a "universe". The treatment assignment procedure explained in section 2 can be used in this setting as well[6].

3.1.2. Treatment Assignment Mechanisms

In order to allocate subjects in an experiment to the different experimental or treatment groups before treatment is conducted, different methods can be used. According to the survey conducted by BK, there is no consent about how treatment assignment should be done, even among experts in field research. Also from a theoretical perspective, a clear answer is missing; see e.g. Imbens (2011) for a brief discussion.

To the methods of treatment assignment studied in the work by BK, we added the minimal MSE procedure as introduced in this paper, and an optimal matching method as implemented in the (still active and maintained) *R* package *nbpMatching* (Lu et al., 2011), going back to the work of Greevy et al. (2004).

[5] BK also use data on schooling of children from this dataset. Given that they do not report results for this dataset in all graphs and tables and given that we the dataset for another outcome variable, we limited ourselves to the other data.

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

We give a short overview over the assignment mechanisms applied in this study and discuss weaknesses and strengths:

Pure Randomization The simplest form of randomization possibly is flipping a coin and allocating a subject to either the treatment or the control group, depending on whether the upper side of the coin shows heads or tails. This or any other form of randomization (rolling a dice, ...) can be carried out very quickly without any preparation or computation even in the most remote field research location. Depending on the transparency of the actual implementation, it can be considered as the fairest method for treatment allocation.

When evaluating an experiment with random treatment assignment, issues of self selection into one of the groups can be ruled out[7] and analysis is very simple.

In order to draw valid conclusions from an experiment, groups ideally have to be identical so that any difference between treatment and control groups after treatment can be attributed to the intervention only. However, when comparing means of randomly allocated groups across 20 variables with a conventional t-test and significance level of 5%, we have to expect that for one variable, the hypothesis of no difference will be rejected.

Further, it is not guaranteed, especially when 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 Stratification (sometimes also referred to as blocking) is attributed to Fisher (1935). Its main advantage is to ensure the possibility of subgroup analysis, while ideally increasing efficiency of analysis. The idea is to build subgroups according to observable characteristics (covariates) and randomize within those subgroups.

The problem however 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 we build three categories for age and income, we end up with 18 strata, where maximally 3 persons are in one group. Stratification on another variable thus is not feasible with a comparable sample size.

Also, sometimes building the strata requires expertise with the data and the question to study.

The example above already 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 not possible anymore.

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.

Regarding the nature and the number of variables to stratify on (or more generally to consider for treatment allocation), a clear recommendation is not possible. The variables

[7] The sample to be split, however, could still differ considerably from the population.

that one would like to use for subgroup analysis should be included. With respect to statistical efficiency, the use of including a variable is the bigger, the higher its correlation with the outcome to be studied will be.

Since stratification is a quite subjective matter, we use the code from BK and the same stratification strategy.

Pairwise Matching 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. Forming pairs however is considerably more difficult and is referred to as non-bipartite matching in the optimization but also the matching literature[8].

The advantages of pairwise matching are that the number of variables to consider for treatment allocation is not limited and that it is also possible to balance continuous variables. It is perceived as 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.

The biggest disadvantage is probably the analysis, but also attrition is a major concern. Regarding the analysis, Abadie and Imbens (2006) note that the simple matching estimator for the average treatment effect "includes 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 estimated consistently, but they refer to the variance not conditional on covariates. Imbens (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, BK estimate the variance conditional on pair dummies, but not conditional on covariates.

With respect to attrition, Imai et al. (2009) note that it is an advantage of matching, that for a unit that drops out, its pair can also be taken out of the experiment while the remaining sample still remains balanced. However, in a small sample, it is a considerable 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 a lower power.

One disadvantage depends on the implementation of the pairwise matching. BK note that for the 300 observations samples, the "optimal greedy matching" algorithm took several days to run. Thus quickly testing the effect of adding one variable to the group of variables on the balance of the other variables is just impossible in a field setting.

[8]Note, that forming groups that match (non-bipartite matching) is a different task than finding matches in already existing groups (bipartite matching). Therefore, not all results for matching and more importantly, software implementations, can be applied in this setting.

BK apply a "optimal greedy algorithm" laid out in Imai et al. (2009), who are strong advocates of pairwise matching. The greedy algorithm computes pairwise distances using the Mahalanobis distance between two pairs 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. This is a "naive" implementation, as overall distance is not necessarily minimized by this approach.

The better solution would be an optimal matching algorithm as introduced by Greevy et al. (2004), however, BK note that it would be more computationally intense. Since the study by BK was published, a software implementation of the optimal matching algorithm was released in the *R* package *nbpMatching*; see Lu et al. (2011). With that implementation, the "optimal matching" algorithm is by far faster than the "optimal greedy algorithm" used in BK.

We use both implementations in our study.

Rerandomization Methods A rerandomization method is basically any method that performs a somehow random treatment assignment, and repeats this step 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 normally 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 e.g. the min max t-stat method, in which 1000 random assignments are made, and the one is chosen in which the maximal t-statistic on any variable to consider is minimal. A variant of the second group is the "big stick" method in BK, 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 (e.g. 2012, 2015), where the Mahalanobis distance is involved.

On the advantages of these methods, BK write: "[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." Thus, in contrast to the methods discussed before, this is the most flexible approach. Whereas pairwise matching is limited to only two groups and stratification is limited to only a very limited number of variables to balance on, rerandomization methods are – depending on the implementation – able to provide a solution for both, and also for cases of unequal sizes of subgroups or treatment groups. As with pairwise matching, subgroup analysis is not ensured and only possible, if enough units of a subgroup are assigned to treatment and control group. If the relevant variable is included in the treatment mechanism and the sample and the number of variables to consider are of moderate size, one can be confident that subgroup analysis is possible.

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 approaches mentioned above, subsequent draws are not independent (although conditional random) from

each other: Given a treatment assignment, a new one is obtained by randomly exchanging treatment status of a certain number of units. The new assignment is then evaluated according to the formula derived in section 2 and either kept or withdrawn. Thus, the Min MSE method maximizes the balance in a more efficient way than the other discussed approaches of rerandomization.

Further, 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[9] and the criterion is based only on the mean differences of the treatment groups. Additionally, we could not find any software implementation of their approach.

3.2. Comparing Treatment Assignment Mechanisms

3.2.1. Variables for Balancing

For every treatment assignment method, we used the same variables to balance on as BK. They always balance on the baseline outcome of an outcome of interest, and add six other variables that may affect the outcome of interest, with exception of stratification, where only subsamples are used.

This means however, that the results of the study regarding stratification can be compared to the other results only vaguely, since stratification is tested with a lower number of variables, thus having higher chances to achieve balance on those and in particular, the baseline outcome variable.

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

For the exact reasoning of the choice of variables to be balanced, we refer to Bruhn and McKenzie (2009).

3.2.2. Pre-Treatment Balance

We apply the same measures of pre-treatment balance on baseline variables as BK for the cases of one treatment arm. For the sake of completeness and transparency, we also do so for some measures performed on follow-up outcomes. However, we think that balance on follow-up outcomes is rather relevant when assessing the general value of covariate based treatment assignment mechanisms in panel studies, which is not the focus of this study. We therefore do not study balance on follow-up outcomes in detail.

For the cases of multiple treatment arms, we extend the measures used by BK 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, BK compare the difference in means

[9] Although they vaguely name a possible way of extending their criteria in this sense, without formally laying out this argument.

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 group means. Additionally, they perform a t-test to assess whether estimates for differences are significantly different from zero, and report the share of draws in which this was the case.

Balance in a group of variables, one treatment arm First, the measures above are calculated for every single variable in a group. Then, for the average difference and for the share of estimates significantly different from zero, 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 95% quantile.

Balance in a group of variables, several treatment arms When aggregating the balance of several treatment arms, taking the average difference of 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 maximal difference in means between the treatment groups and the control group in one variable in one draw. We think both are relevant and perform both.

4. 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 A.

Scenario 1: One Treatment Group We first present the results for the scenario considered in BK: Units have to be assigned to either the treatment or the control group. Although up to seven variables are selected for treatment assignment, BK only report the balance for the main variable of interest. Further, 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 3.2.2).

Figures 1 and 2 show the distribution of 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. Table 1 consists 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 BK. 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 BK.

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 BK, 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 five baseline variables and all unobservables, an average difference of zero to the third digit was achieved.

With respect to the whole distributions as shown in figures 1 and 2, in half of the cases considered, the min MSE procedure shows the most favorable distribution with the highest mass at 0 and thin tails. 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 differences in group means as shown in the middle panel of table 1, although in this panel, no mechanism clearly shows more favorable figures than another.

Consistently with the findings by BK, we also note that with increasing sample size, balance improves. This can be seen in figures 1 and 2, 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, consistently with BK, no major differences; especially for the cases in which baseline variables 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.

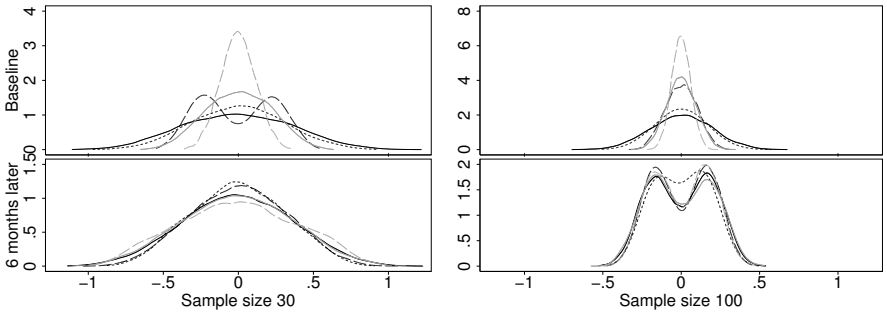
Summing up, there is no treatment assignment mechanism performing best in every dataset and every measurement. Also, with the exception of a single random draw, none performs inferior than all other mechanisms throughout. In most comparisons, either one of the matching methods or the min MSE procedure dominates the competing mechanisms. All methods achieve balance on average and all decrease extreme imbalances considerably in comparison with a single random draw.

Scenario 2: Multiple Treatment Arms The second scenario we consider is an experiment, in which multiple (variants of) interventions are tested. In such experiments, there is one control group, which receives a placebo treatment or no treatment at all and several treatment groups. Units shall be assigned to the control or one of the treatment groups while keeping all groups comparable.

For this scenario, we haven't found any software implementation of a competing method, so we compare the min MSE procedure to a single random draw.

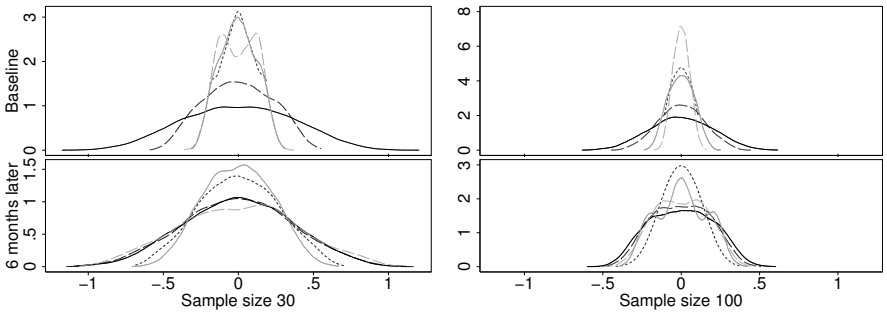
The findings are presented most conveniently graphically; this is done in figure 3.

Panel A. Sri Lanka profits



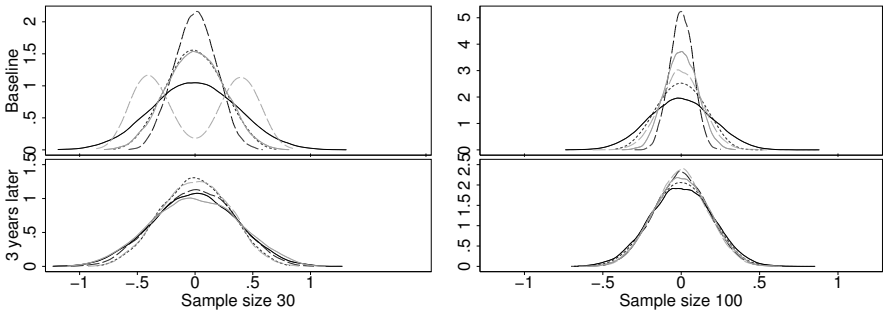
Difference in average profits (weighted by standard deviation)

Panel B. Mexico ENE labor income



Difference in average income (weighted by standard deviation)

Panel C. IFLS expenditure data



Difference in average *hh* expenditure *p cap* (weighted by standard deviation)

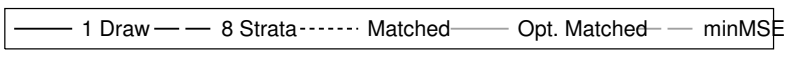
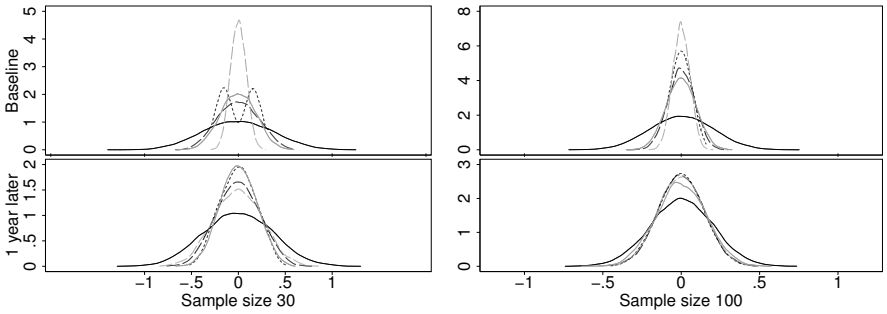


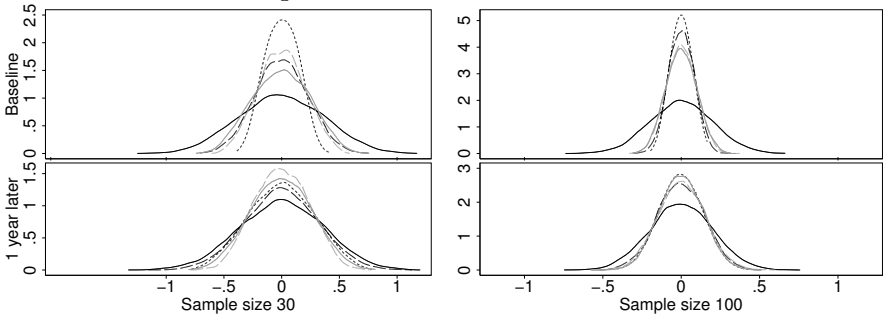
Figure 1: Distribution of the differences in group means between the treatment and control group in the baseline variable and the follow-up variable.

Panel D. LEAPS math test score data



Difference in average in math test score (weighted by standard deviation)

Panel E. LEAPS height score data



Difference in average z-score (weighted by standard deviation)

Figure 2: Distribution of the differences in group means between the treatment and control group in the baseline variable and the follow-up variable.

Table 1: Comparison of Treatment Assignment Methods Regarding balance in baseline outcomes, Sample Size 100

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

	Single random draw	Pairwise greedy matching	Optimal matching	Min MSE procedure	Stratified on two variables	Stratified on four variables
Microenterprise profits (Sri Lanka)	0.001	0.001	0.001	-0.001	-0.000	-0.001
Household expenditure (Indonesia)	-0.002	-0.002	-0.000	0.000	0.001	-0.001
Labor income (Mexico)	-0.000	-0.000	-0.001	-0.000	0.000	-0.000
Height z-score (Pakistan)	0.001	0.000	0.000	0.000	0.001	0.000
Math test score (Pakistan)	0.003	-0.000	-0.001	-0.000	-0.000	-0.001
Baseline unobservables (Sri Lanka)	-0.000	-0.001	0.001	-0.000	0.000	0.000
Baseline unobservables (Mexico)	0.000	-0.000	-0.000	-0.000	0.000	-0.000

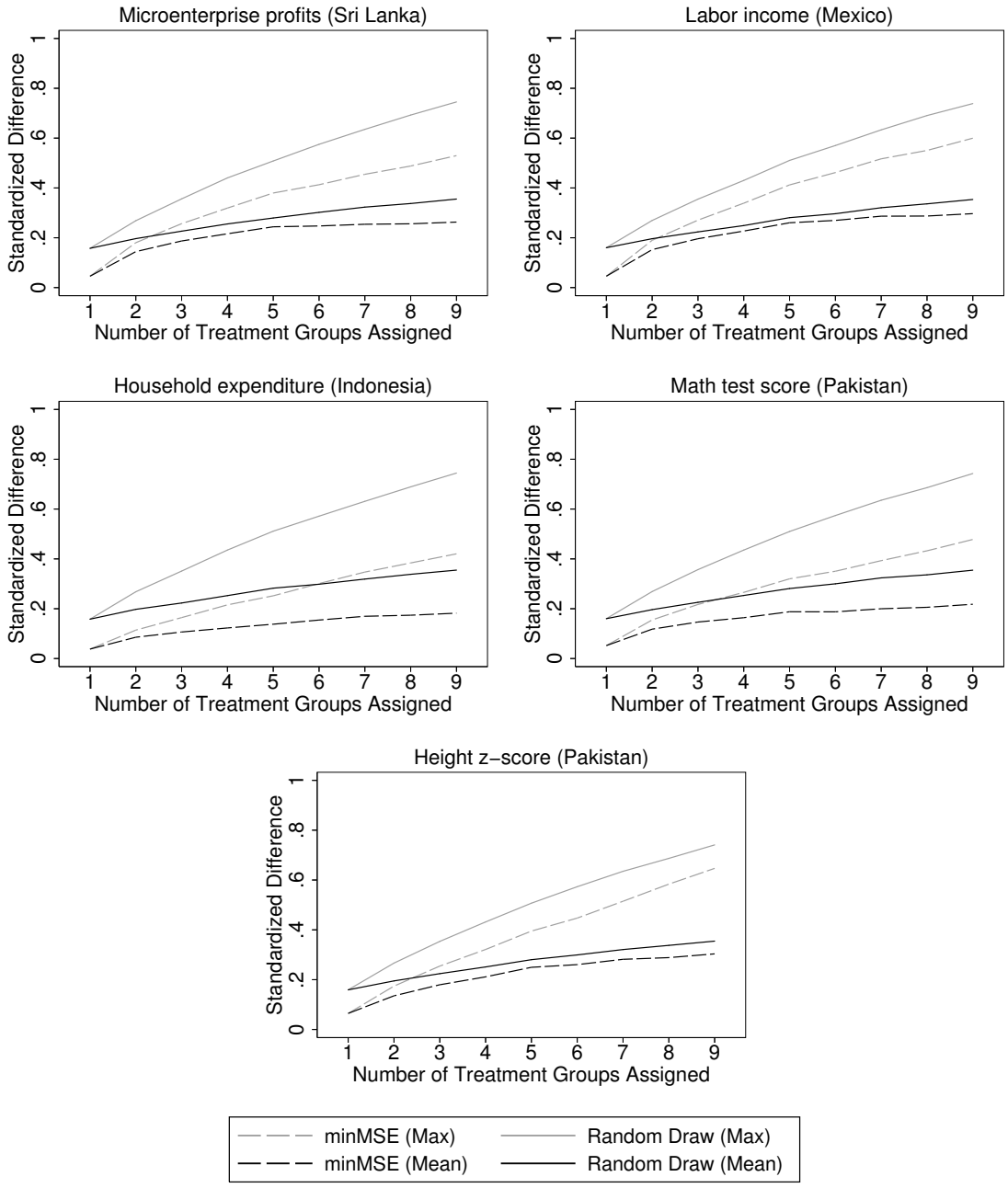
(b) 95% quantile of the difference in baseline group means between treatment and control group in SD

	Single random draw	Pairwise greedy matching	Optimal matching	Min MSE procedure	Stratified on two variables	Stratified on four variables
Microenterprise profits (Sri Lanka)	0.386	0.315	0.183	0.119	0.195	0.241
Household expenditure (Indonesia)	0.390	0.264	0.199	0.261	0.145	0.191
Labor income (Mexico)	0.384	0.100	0.154	0.100	0.280	0.304
Height z-score (Pakistan)	0.395	0.103	0.190	0.185	0.160	0.206
Math test score (Pakistan)	0.392	0.074	0.185	0.107	0.164	0.237
Baseline unobservables (Sri Lanka)	0.434	0.434	0.434	0.434	0.417	0.414
Baseline unobservables (Mexico)	0.457	0.457	0.457	0.457	0.448	0.439

(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 four variables
Microenterprise profits (Sri Lanka)	0.097	0.039	0.000	0.000	0.000	0.005
Household expenditure (Indonesia)	0.102	0.010	0.001	0.013	0.000	0.000
Labor income (Mexico)	0.100	0.000	0.000	0.000	0.015	0.029
Height z-score (Pakistan)	0.100	0.000	0.000	0.000	0.000	0.001
Math test score (Pakistan)	0.100	0.000	0.000	0.000	0.000	0.006
Baseline unobservables (Sri Lanka)	0.101	0.083	0.097	0.091	0.096	0.095
Baseline unobservables (Mexico)	0.108	0.104	0.091	0.089	0.095	0.093

Note: Statistics base on 10,000 iterations. Details on the study and the computation of each measures are explained in section 3.2.



Note: Differences are weighted by standard deviation. In each dataset, the difference of group means in all variables used for treatment assignment are compared. The line labeled with 'max' shows the average over all iterations and variables of the maximum of these differences amongst the treatment groups. The line labeled with 'mean' shows the differences amongst the group differences by building the average.

Figure 3: Differences in group means between the treatment groups and the control group in the group of baseline variables for an increasing number of treatments to assign; sample size: 1000.

For the graphs in figure 3, we first computed the maximum and the mean difference between the treatment group means and the control group mean of one variable for a single draw. We aggregated the measure over the variables of one draw and over all iterations by averaging over the mean or maximal maximal difference. The first case is shown by the dashed lines, the latter one by the solid lines. Both lines, the solid one 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.

We see that 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 a growing number of treatments 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 drawing randomly, the average maximal group difference is mostly around .75 or .8 SD. In one case (household 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.

The min MSE procedure was in all datasets 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 noteworthy, that with the minMSE procedure, we can assign between 2 to 5 more treatments compared to drawing randomly with the same maximum difference in group means to be expected.

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). 6. delivering education: a pragmatic framework for improving education in low-income countries. *Handbook of International Development and Education 75*, 85.

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.

Cox, D. R. (1957, June). The Use of a Concomitant Variable in Selecting an Experimental Design. *Biometrika 44*(1/2), 150.

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

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

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

Imai, K., G. King, and C. Nall (2009, February). The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science 24*(1), 29–53.

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, Working Paper, Harvard University.

Kasy, M. (2016, July). Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead: Table 1. *Political Analysis 24*(3), 324–338.

Klar, N. and A. Donner (1997, 8). 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.

Appendix

A. Further results

The comparison of treatment assignment methods regarding balance in baseline outcomes for a sample size of 30 is reported in table 3.

Table 3: Comparison of Treatment Assignment Methods Regarding balance in baseline outcomes, Sample Size 30

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

	Single random draw	Pairwise greedy matching	Optimal matching	Min MSE procedure	Stratified on two variables	Stratified on four variables
Microenterprise profits (Sri Lanka)	-0.004	-0.001	0.004	-0.001	-0.006	0.000
Household expenditure (Indonesia)	-0.002	-0.003	0.002	-0.003	0.001	0.003
Labor income (Mexico)	0.003	0.001	-0.001	-0.001	-0.001	-0.000
Height z-score (Pakistan)	-0.003	0.001	-0.003	0.000	-0.001	-0.000
Math test score (Pakistan)	-0.002	-0.001	-0.002	-0.000	-0.001	-0.002
Baseline unobservables (Sri Lanka)	-0.001	-0.000	0.001	0.001	-0.001	0.001
Baseline unobservables (Mexico)	-0.000	-0.000	-0.000	-0.000	-0.000	-0.001

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

	Single random draw	Pairwise greedy matching	Optimal matching	Min MSE procedure	Stratified on two variables	Stratified on four variables
Microenterprise profits (Sri Lanka)	0.706	0.598	0.416	0.228	0.416	0.538
Household expenditure (Indonesia)	0.716	0.458	0.478	0.643	0.347	0.501
Labor income (Mexico)	0.691	0.177	0.224	0.228	0.409	0.582
Height z-score (Pakistan)	0.710	0.258	0.467	0.394	0.445	0.446
Math test score (Pakistan)	0.713	0.257	0.363	0.161	0.409	0.586
Baseline unobservables (Sri Lanka)	0.803	0.879	0.879	0.889	0.824	0.805
Baseline unobservables (Mexico)	0.834	0.834	0.879	0.834	0.771	0.775

(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 four variables
Microenterprise profits (Sri Lanka)	0.097	0.049	0.001	0.000	0.000	0.021
Household expenditure (Indonesia)	0.100	0.005	0.011	0.089	0.000	0.012
Labor income (Mexico)	0.099	0.000	0.000	0.000	0.000	0.037
Height z-score (Pakistan)	0.102	0.000	0.007	0.000	0.005	0.005
Math test score (Pakistan)	0.103	0.000	0.000	0.000	0.001	0.041
Baseline unobservables (Sri Lanka)	0.090	0.088	0.068	0.079	0.094	0.094
Baseline unobservables (Mexico)	0.088	0.077	0.083	0.079	0.077	0.079

Note: Statistics base on 10,000 iterations. Details on the study and the computation of each measures are explained in section 3.2.