Guiteras, Raymond P.; Levine, David I.; Polley, Thomas H.

Article

The pursuit of balance in sequential randomized trials

Development Engineering

Provided in Cooperation with:
Elsevier

Suggested Citation: Guiteras, Raymond P.; Levine, David I.; Polley, Thomas H. (2016) : The pursuit of balance in sequential randomized trials, Development Engineering, ISSN 2352-7285, Elsevier, Amsterdam, Vol. 1, pp. 12-25, https://doi.org/10.1016/j.deveng.2015.11.001

This Version is available at:
https://hdl.handle.net/10419/187782

Development Engineering 1 (2016) 12–25

Contents lists available at ScienceDirect

Development Engineering

journal homepage: www.elsevier.com/locate/deveng

The pursuit of balance in sequential randomized trials

Raymond P. Guiteras [a,*], David I. Levine [b], Thomas H. Polley [c]

[a] Department of Economics, 3114 Tydings Hall, University of Maryland, College Park, MD 20142, USA
[b] U.C. Berkeley Haas, USA
[c] Duke University, USA

ARTICLE INFO

Article history:
Received 7 August 2015
Received in revised form
6 October 2015
Accepted 7 November 2015
Available online 2 December 2015

JEL classification:
C93
O12

Keywords:
Stratification
Sequential randomization
Design of Experiments

ABSTRACT

In many randomized trials, subjects enter the sample sequentially. Because the covariates for all units are not known in advance, standard methods of stratification do not apply. We describe and assess the method of D_A-optimal sequential allocation (Atkinson, 1982) for balancing stratification covariates across treatment arms. We provide simulation evidence that the method can provide substantial improvements in precision over commonly employed alternatives. We also describe our experience implementing the method in a field trial of a clean water and handwashing intervention in Dhaka, Bangladesh, the first time the method has been used. We provide advice and software for future researchers.

1. Introduction

Randomized-controlled trials (RCTs) are an increasingly important tool for policy evaluation and estimation of economic parameters. However, they are expensive, and efficient use of limited resources (funding, inputs from implementation partners, and researchers' time) requires that they be designed carefully. In an important contribution, Bruhn and McKenzie (2009) reviewed stratification methods that were common in economics RCTs at the time, and showed that large gains in precision were available by adopting more sophisticated stratification methods from the clinical trials literature. These stratification methods require researchers to obtain stratification covariates from all subjects prior to randomization. However, this is not always feasible. In clinical trials, subjects are often allocated to treatment as they arrive. In field trials, operational constraints may prevent defining and surveying the full sample frame in advance. In such situations, subjects must be assigned *sequentially*, with the researcher only learning the value of the stratification variables for that subject's at the time of enrollment and assignment.[1]

In this paper, we propose the use of D_A-*optimal sequential allocation* (Atkinson, 1982) to improve balance and power when subjects are enrolled sequentially. The D_A-optimal method minimizes imbalance given the constraint of not knowing covariate values in advance. We describe the method and its properties, and provide an algorithm for its implementation. We conduct a set of simulations, based on Bruhn and McKenzie (2009), and show that the D_A-optimal method offers clear benefits relative to commonly used sequential alternatives. In fact, surprisingly, optimal sequential designs are comparably well-balanced to stratifications performed with full knowledge of covariates in advance. In spite of these practical advantages, the method had not, to our knowledge and according to three survey articles, ever been employed in the field.[2] We describe our experience implementing the method in a water treatment and hygiene intervention in Dhaka, Bangladesh (Guiteras et al., 2015), and offer practical advice on its implementation under field conditions. Implementation was feasible with standard software (Stata), and produced an allocation that was well-balanced both on the stratification variables chosen *ex ante* and, *ex post*, on other important variables that were not included in the stratification.

* Corresponding author.
E-mail address: guiteras@econ.umd.edu (R.P. Guiteras).

[1] Examples of sequential randomization in economics include Beaman and Magruder (2012), which randomized without stratification, and Bronchetti et al. (2013), which stratified using the block randomization method we describe in Section 5.1.2.

[2] See McEntegart (2003), Table 1 in Taves (2010), and Ciolino et al. (2011). Confirmed by personal communication with J. Cicolino, Northwestern University, January 17, 2014.

http://dx.doi.org/10.1016/j.deveng.2015.11.001

2. Theory

Our exposition follows Atkinson (2002), with some changes in notation. First, we lay out the model and notation. Second, we develop the theory for the traditional situation of a fixed population of N subjects, for whom covariates X have been collected in advance. Third, we introduce sequential designs using a simplified case where the researcher is concerned with the precision of all estimated parameters, both treatment effects and nuisance parameters (coefficients on stratification variables). Finally, we adapt the sequential design to the standard situation where only precisely estimated treatment effects are of interest.

2.1. Model and notation

Suppose the researcher is conducting an individual-level trial with J treatments, including the control treatment. We first consider a linear model with homogeneous treatment effects and i.i.d. errors. In Section 3, we discuss extensions, including heteroscedasticity, nonlinear models, and cluster designs. The model for unit i is

$$y_i = d_i'\alpha + x_i'\beta + \varepsilon_i = w_i'\theta + \varepsilon_i, \tag{1}$$

where d_i is a $J \times 1$ vector of indicator variables assigning unit i to a single treatment (i.e., exactly one element of d_i is equal to one), x_i is a $K \times 1$ vector of covariates, and ε_i is an error term. Without loss of generality, we order the treatments with the control condition first. Let $d_i(j)$ indicate assignment to the jth treatment; that is, $d_i(1) = (1\ 0 \cdots 0)'$, $d_i(2) = (0\ 1\ 0 \ldots 0)'$, etc. We are interested in estimating *contrasts* between the elements of α; that is, $\alpha_1 - \alpha_2$, $\alpha_1 - \alpha_3$, etc. The control group mean is a nuisance parameter,[3] as are the K elements of β (the coefficients on the covariates), so we have $K+1$ nuisance parameters and $J-1$ parameters of interest.[4]

2.2. Optimal designs with baseline covariates

First, consider a population of N subjects, for whom the researcher has obtained baseline covariates X prior to randomization. The population regression model is given by

$$E[Y] = D\alpha + X\beta = W\theta, \tag{2}$$

where D is the $N \times J$ matrix assigning all subjects to treatment (i.e., $D = (d_1 \cdots d_n)'$). X is the $N \times K$ matrix of covariates, and α and β are as before. Given the covariates X, our goal is to choose D to minimize the variance of our estimated treatment effect. As a simple example, with one treatment plus a control condition, $J=2$, we are interested in the *contrast* $\alpha_1 - \alpha_2$ and wish to minimize $V[\hat{\alpha}_1 - \hat{\alpha}_2]$.

A useful matrix to create contrasts is

$$L'_{(J-1)\times J} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & & -1 \end{bmatrix}.$$

Now we can create a vector of contrasts by premultiplying α by L':

$$L'\alpha = \begin{bmatrix} \alpha_1 - \alpha_2 \\ \vdots \\ \alpha_1 - \alpha_J \end{bmatrix}.$$

To annihilate the nuisance parameters, we augment L' with a $(J-1) \times K$ matrix of zeros, and define

$$A' = [L'\ \ 0].$$

The variance of $\hat{\alpha}$ is proportional to square root of the determinant of the generalized variance[5]:

$$\left|A'(W'W)^{-1}A\right| = \left|L'\left\{D'D - D'X(X'X)^{-1}X'D\right\}^{-1}L\right|. \tag{3}$$

This quantity is minimized when $D'X = 0$; that is, when the treatment assignment is *orthogonal* to the covariates, which is to say that the treatments are *balanced* across the covariates. When $D'X = 0$, the generalized variance simplifies, and the determinant is

$$\left|A'(W'W)^{-1}A\right| = \left|L'(D'D)^{-1}L\right| = J^J/N^{J-1}$$

This minimum possible value is the standard against any other treatment assignment D. Note that this value is increasing in J and decreasing in N, which matches our intuition that the variance will increase with the number of treatments and decrease with the number of observations.

The relative efficiency of a design D is the ratio of the determinant of the generalized variance to this minimal value:

$$\mathcal{E} = \left\{\frac{J^J/N^{J-1}}{A'(W'W)^{-1}A}\right\}^{1/(J-1)},$$

where $1/(J-1)$ is a scale factor. A smaller denominator $|A'(W'W)^{-1}A|$ leads to higher $\mathcal{E}$, implying a more efficient design. Note that $\mathcal{E} = 1$ for an exactly balanced design. A useful representation is the *loss*

$$\mathcal{L} = N(1 - \mathcal{E}),$$

which is expressed as the effective loss of observations relative to an optimal design. That is, a non-optimal design D with N units is as precise as an optimal design with $N(1-\mathcal{E})$ fewer units. For an exactly balanced design, $\mathcal{L} = 0$.

Although not the focus of this paper, this framework can be used for near-optimal randomization in cases where a researcher can collect baseline data prior to randomization. Specifically, create a large number S of random allocations $\{D^1, \ldots, D^s, \ldots, D^S\}$ and choose the allocation D^s with lowest associated loss.[6] Kasy (2013) considers a more general Bayesian framework, and provides a search algorithm to find an optimal allocation.[7]

2.3. Sequential D-optimality

To extend to sequential randomized trials, we first consider the simple case where all elements of $\theta = (\alpha', \beta')'$ are of interest. Our goal is to minimize the variance of $\hat{\theta}$. The variance of $\hat{\theta}$ is proportional to the inverse of the design matrix $(W'W)^{-1}$, so we want to minimize $(W'W)^{-1}$ or, equivalently, maximize $|W'W|$, which will give us a *D-optimum design*.

Suppose the first n units have been allocated, with the resulting

[3] We are not interested in α_1 *per se*, but a precise estimate $\hat{\alpha}_1$ is necessary to estimate contrasts precisely.

[4] A more familiar setup for economics readers would include an intercept term as a covariate, so $\alpha_{\sim 1}$ would have $J-1$ elements (corresponding to the $J-1$ treatment conditions excluding the control) and the augmented covariate vector $(1, x')'$ would have $K+1$ elements including the intercept. This turns out to be less convenient for some of the matrix algebra below.

[5] Recall that $W = [D\ X]$, so $W'W = \left[\begin{smallmatrix} D' \\ X' \end{smallmatrix}\right][D\ X] = \left[\begin{smallmatrix} D'D & D'X \\ X'D & X'X \end{smallmatrix}\right]$. Then use results on inverses of partitioned matrices and use the zero block of the matrix A to zero out several terms.

[6] To conduct randomization inference, rather than choose the allocation with minimum $\mathcal{L}$, the researcher can instead specify an acceptable maximum $\overline{\mathcal{L}}$, retain $R+1$ draws with loss less than $\overline{\mathcal{L}}$, select one of these $R+1$ at random, and retain the remaining R for randomization inference. Code is available from the authors on request.

[7] This optimal allocation is unique if any element of x is continuous, and may be unique (in finite samples) even for discrete x with a large number of treatments and covariate cells. See also Bertsimas et al. (2015).

design given by $W_n = [D_n\ X_n]$. Suppose that unit $n+1$ arrives, with covariate vector x_{n+1}. By the Generalized Equivalence Theorem of Kiefer and Wolfowitz (1960), maximizing $|W'W|$ is equivalent to minimizing the maximum variance of the predicted response, where this maximum is taken over the space of w. That is, given the current allocation W_n, the variance of $\hat{y}$ at a point $w = (d', x')'$ is

$$V\left[\hat{y}\left(w\right)\right] \propto s(w, W_n) = w'\left(W_n'W_n\right)^{-1}w, \tag{4}$$

where $s(w, W_n)$ is the *standardized variance* at w given the allocation W_n. Because minimizing $V[\hat{\theta}]$ is equivalent to minimizing the maximum variance of the predicted mean, we can restate our objective as minimizing

$$\sup_w V\left(\hat{y}(w)\right).$$

That is, given the existing design W_n and the covariate values x_{n+1}, the optimal assignment for unit $n+1$ is

$$d^*_{n+1}(x_{n+1}, W_n) = \arg\min_d\left\{\sup_w V(\hat{y}(w))\right\}.$$

In other words, we want to allocate this unit to the treatment where the variance is greatest. To accomplish this, for unit $n+1$, write the set of possible values for w_{n+1} as $w_{n+1}(1) = (d(1)', x_{n+1}')', \ldots w_{n+1}(J) = (d(J)', x_{n+1}')'$, where $d(j)$ denotes allocation to treatment j. For each $w_{n+1}(j)$, we calculate

$$s_j = s(w_{n+1}(j), W_n) = w_{n+1}(j)'\left(W_n'W_n\right)^{-1}w_{n+1}(j). \tag{5}$$

The best allocation for unit $n+1$ is the $d(j)$ with the largest value of s_j. In this simplest case, we mechanically assign person $n+1$ to this treatment. The intuition is that the unit is being assigned to the treatment where it is most needed, which is where the variance is highest. In Section 3.7, we discuss non-deterministic or "biased coin" assignment.

2.4. Sequential D_A-optimality

The D-optimal procedure in the previous subsection minimizes $V[\hat{\theta}]$. That is, it maximizes precision for estimates of both treatment effects and the coefficients on baseline characteristics. However, in most cases our goal is to maximize precision for the estimated treatment effects. That is, our objective is to minimize $V[\hat{\alpha}]$, and we are not *per se* interested in minimizing $V[\hat{\beta}]$. Atkinson calls this problem D_A-optimality. The intuition and the basic procedures are the same, but the formula for the standardized variance $s_A(w_{n+1}(j), W_n)$ is slightly more complicated:

$$s_A(w_{n+1}(j), W_n) = w_{n+1}(j)'\left(W_n'W_n\right)^{-1}A\left\{A'\left(W_n'W_n\right)^{-1}A\right\}^{-1}A'\left(W_n'W_n\right)^{-1}w_{n+1}(j), \tag{6}$$

where $A' = [L'\ 0]$, as above.

The assignment algorithm follows the logic of Section 2.3. Suppose that n units have been allocated, and the current matrix of assignments and covariates is $W_n = [D_n\ X_n]$. When unit $n+1$ arrives with covariates x_{n+1}, check the value of $s_j = s_A(w_{n+1}(j), W_n)$ for each possible assignment j. The optimal allocation of unit $n+1$ is where s_j is greatest.

2.5. Algorithm

The procedure described above cannot be used for the first units, because $W'W$ is singular as long as the number of observations, n, is less than the number of incidental parameters, J. These first units could be assigned randomly, or the $W'W$ matrix could be made invertible by adding a small amount of random noise to the diagonal.

Having allocated n units, allocate unit $n+1$ as follows:

1. Subject $n+1$ arrives with x_{n+1}
2. For each treatment j, calculate $s_A(w_{n+1}(j), W_n)$ using (6).
3. Assign treatment to the study arm where $s_A(w_{n+1}(j), W_n)$ is greatest.
4. Update W_n to W_{n+1}.

3. Extensions

3.1. Unequal allocations

The basic exposition assumes that the researcher places equal weight on the precision of each element of α (that is, on the coefficient of each treatment effect). However, the researcher may wish to weight these unequally – if, for example, the cost of treatments vary, or if external constraints require unequal numbers of treated units. To create unequal allocations, inflate or deflate the corresponding values of $s_j = s_A(w_{n+1}(j), W_n)$. For example, to overweight treatment j, premultiply s_j by an appropriate weight m_j. These weights $\{m_j\}$ can be calculated analytically in some simple cases, or the researcher can conduct simulations to tune the weights. See Section 6 for an example.

3.2. Heteroscedasticity

If the variance of the outcome of interest is a function of treatment, equal allocations may be inefficient because coefficients corresponding to treatments that increase variance will be less precisely estimated. If the researcher has a strong prior that the variance of the outcome of interest is likely to be greater under certain treatment conditions, she can allocate more units to that treatment following the weighting strategy described in the previous subsection.

3.3. Unequal penalties for imbalance

The researcher may wish to emphasize balance in one covariate or set of covariates over others. Suppose one set of K_1 covariates x_1 have very strong predictive power for the outcome variable, while the remaining K_2 covariates x_2 have some predictive power but less than x_1. The researcher wants to balance against both covariates, but imbalance in x_1 will cause greater efficiency loss than imbalance in x_2. To assign a greater penalty to imbalance in x_1, note from Eq. (3) that efficiency is maximized when $D'X$ is zero, and decreases as $D'X(X'X)^{-1}X'D$ becomes larger. Partition X into $X = [X_1\ X_2]$. Now

$$\begin{aligned} D'X\left(X'X\right)^{-1}X'D &= D'[X_1\ X_2]\left(X'X\right)^{-1}\left[X_1\ X_2\right]'D \\ &= [D'X_1\ D'X_2]\left(X'X\right)^{-1}\left[D'X_1\ D'X_2\right]', \end{aligned} \tag{7}$$

so to penalize imbalance in x_1 simply multiply $D'X_1$ – but not $D'X_2$ – by a scalar weight $m > 1$. The same logic applies in the sequential algorithm: in the standardized variance formula (6), replace $D_n'X_{1n}$ with $mD_n'X_{1n}$.[8] Again, simulations can help choose the appropriate weights.

[8] Note that, as with (3), the middle term $A'(W_n'W_n)^{-1}A$ of (6) reduces to $L'\{D_n'D_n - D_n'X_n(X_n'X_n)^{-1}X_n'D_n\}^{-1}L$, with the loss coming from $D_n'X_n \neq 0$. Partitioning as above, the second term in the inverse becomes $[D_n'X_{1n}\ D_n'X_{2n}](X_n'X_n)^{-1}[D_n'X_{1n}\ D_n'X_{2n}]'$, so we increase the penalty for imbalance in x_1 by replacing $D_n'X_{1n}$ with $mD_n'X_{1n}$.

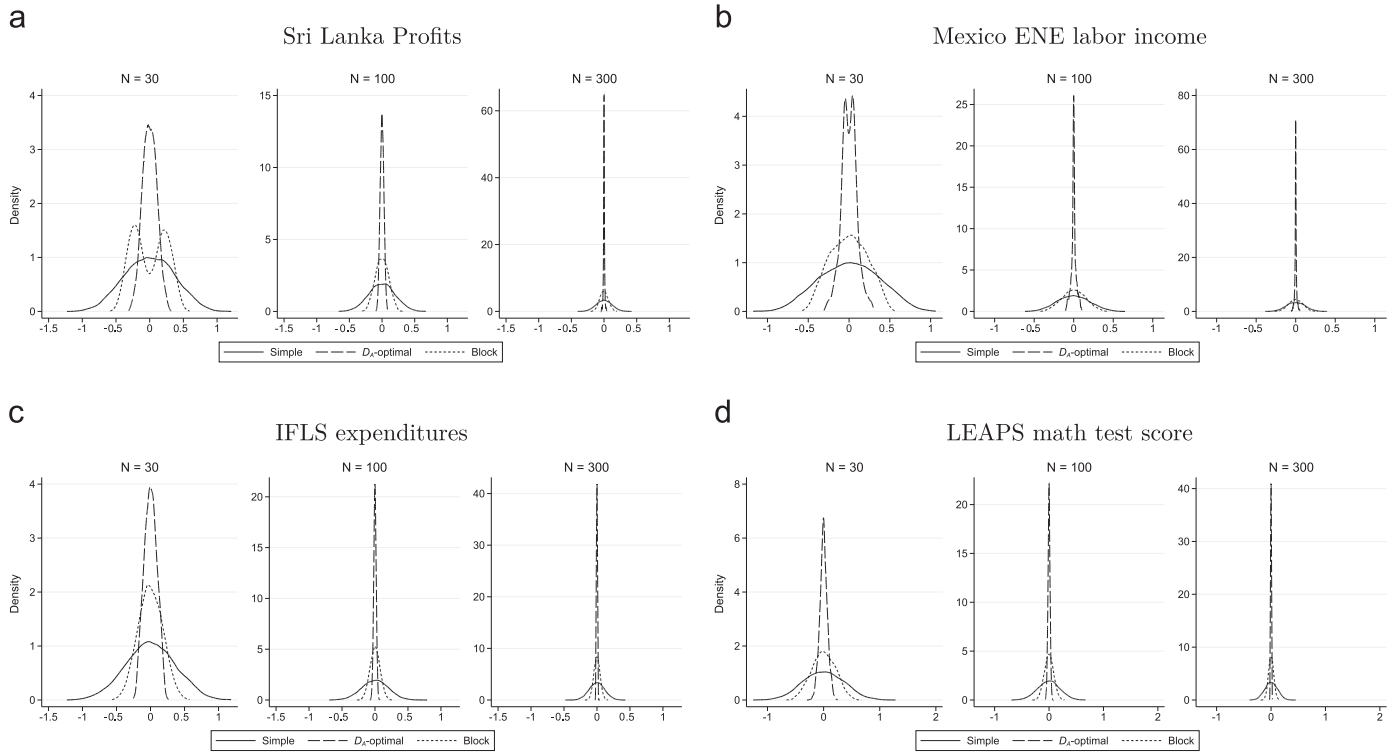

Fig. 1. Imbalance at baseline. *Notes*: (a–d) show the distribution of differences in means between the treatment and control groups at follow-up for each dataset at all three sample sizes: 30, 100 and 300 observations. Distributions are kernel density plots, using the Epanechnikov kernel, based on 10,000 bootstrap iterations. In each iteration, the difference in means is divided by the standard deviation of the outcome variable. No mean normalization is done; means are close to zero as the result of the balancing methods used.

3.4. Subgroups and interactions

As described, the algorithm maximizes marginal balance. That is, it minimizes the variance of the estimates of the treatment effects overall, not for any particular value of the covariates. To optimize estimates for subgroups, augment the vector containing the covariates of interest, α, with the relevant interaction coefficients. Similarly, if the interaction of two treatments is of interest, augment α and d appropriately. See Section 6 for an example.

3.5. Nonlinear models

The algorithm is motivated by linear regression, so the allocation it produces may not be optimal for nonlinear designs. In particular, for linear regression, the optimal design does not depend on the value of the unknown parameters, but it does for nonlinear or generalized linear models (Atkinson and Haines, 1996). Therefore, the optimal design, for a non-sequential or sequential trial, in a nonlinear model requires that the researcher specify her prior belief distribution about the parameter of interest. While the intuition is similar (maximizing the log of the determinant of the information matrix), the calculation can be quite difficult. However, it is difficult to imagine a scenario where balancing to minimize the variance of OLS estimates would severely worsen the precision of nonlinear estimators. Therefore, we speculate that such concerns are of second order, and that, while the method may not produce the optimal design for a nonlinear model, it is likely to produce a good approximation. In highly specialized situations, there may be some efficiency gains to more specialized solutions.[9]

3.6. Time trends

Because the algorithm seeks to maintain balance at each point in the sequence, it is robust to trends or fluctuations in potential outcomes that occur as sample enrollment proceeds. For example, neither a geographic pattern to enrollment nor a change in recruitment methods would cause bias, even if these were correlated with potential outcomes (e.g., moving from richer to poorer neighborhoods, or making a greater effort to recruit poor subjects).

3.7. Biased coin methods

The D_A-optimal method will produce unbiased estimates as long as each unit's exact place in the sequence is uncorrelated with potential outcomes. This assumption could be violated if, for example, an intake nurse in a clinical trial knows the algorithm and current allocation. The nurse could then manipulate the order in which subjects are processed to ensure that a particular subject receives a particular treatment.

To reduce the possibility of gaming, a "biased coin" version of the sequential allocation algorithm allocates a subject probabilistically, putting highest probability on the study arm that would reduce the variance of estimated treatment effect, $s_A(w_{n+1}(j), W_n)$, the most. Following the logic of Efron (1971) and of Atkinson (1982) suggests this formula for the probability for allocation to study arm:

$$\pi(j) = \frac{s_A\left(w_{n+1}(j), W_n\right)}{\sum_{j'} s_A\left(w_{n+1}(j'), W_n\right)}.$$

[9] For an example, see Zocchi and Atkinson (1999), who derive the optimal design to study the relationship between a discrete treatment (dose of gamma radiation) and a multinomial, ordered outcome (whether housefly pupae die before opening, die during emergence, or survive past emergence).

Table 1
How do the different methods compare in terms of baseline balance?

	(Sample size of 100)				
	Simple random	Block (2 variables)	Block (4 variables)	D_A optimal (2 variables)	D_A optimal (4 variables)
Panel A. Average difference in BASELINE between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	−0.003	0.001	−0.001	0.000	0.000
Housing expenditure (Indonesia)	0.000	0.000	−0.001	0.000	0.000
Labor income (Mexico)	−0.002	−0.002	0.000	0.000	0.000
Height *z*-score (Pakistan)	0.000	−0.002	0.001	0.000	0.000
Math test score (Pakistan)	−0.005	0.001	0.000	0.000	0.000
Baseline unobservables (Sri Lanka)	0.000	0.000	0.000	0.000	0.000
Baseline unobservables (Mexico)	0.000	−0.001	0.000	0.001	0.000
Panel B. Ninety-fifth percentile of difference in BASELINE between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	0.390	0.194	0.244	0.055	0.064
Housing expenditure (Indonesia)	0.391	0.148	0.193	0.038	0.058
Labor income (Mexico)	0.389	0.283	0.305	0.074	0.078
Height *z*-score (Pakistan)	0.389	0.160	0.203	0.038	0.055
Math test score (Pakistan)	0.396	0.164	0.237	0.039	0.055
Baseline unobservables (Sri Lanka)	0.434	0.417	0.414	0.434	0.427
Baseline unobservables (Mexico)	0.457	0.448	0.439	0.448	0.429
Panel C. Proportion of p-values <0.1 for testing difference in BASELINE means					
Microenterprise profits (Sri Lanka)	0.101	0.000	0.005	0.000	0.000
Housing expenditure (Indonesia)	0.100	0.000	0.001	0.000	0.000
Labor income (Mexico)	0.103	0.017	0.031	0.000	0.000
Height *z*-score (Pakistan)	0.098	0.000	0.001	0.000	0.000
Math test score (Pakistan)	0.106	0.000	0.006	0.000	0.000
Baseline unobservables (Sri Lanka)	0.101	0.096	0.094	0.096	0.092
Baseline unobservables (Mexico)	0.108	0.094	0.095	0.102	0.096

Notes: Statistics are based on 10,000 simulations of each method.

Randomization may also be useful for expositional purposes, to explain to subjects that the process is fair and to regulators or other consumers of the research who are used to hearing about "randomized," rather than "optimally allocated," trials.

3.8. Cluster designs

The development above is based on individual-level treatments, but can be adapted for cluster designs. The first application is to interventions where the treatment is assigned at a cluster level. In this case, the D_A-optimal method applies directly, using cluster-level covariates.

A second application is when treatments are assigned to individuals, but individuals belong to clusters and for logistical reasons these clusters are enrolled sequentially. For example, consider a study of the demand for water filters among N households in V villages.[10] The researcher wishes to vary a sales treatment at the household level. However, villages are enrolled – and stratification covariates collected – sequentially, so at the time of assigning treatments in village v, the researcher only knows covariates for households in villages $1, \ldots, v$, and the history of assignments in villages $1, \ldots, v-1$. This is not a purely sequential allocation, since the researcher knows the values of the covariates for all households in village v, and can assign treatments simultaneously within village v. However, the researcher does not know the value of covariates for households in future villages $v+1, \ldots V$. The tools of sequential allocation can be usefully applied in this situation. Suppose that the matrix of assignments and covariates through village $v-1$ is $W_{v-1} = [D_{v-1}\, X_{v-1}]$. The researcher obtains covariate data for households in village v, resulting in the covariate matrix X_v for all v villages. The researcher then creates a large number S of random treatment allocations for village v, resulting in a set of assignment matrices $\{D_v^1, \ldots, D_v^s, \ldots, D_v^S\}$ and corresponding design matrices $\{W_v^1, \ldots, W_v^s, \ldots, W_v^S\}$. For each s, we can calculate the associated determinant $A'(W_v^{s'} W_v^s)^{-1}A$, where, as in Section 2.2, the matrix A allows us to focus on the parameters of interest. Since this determinant is proportional to the expected variance of $\hat{\alpha}$, we select the allocation that minimizes this determinant.

4. Inference

Confidence intervals can be constructed from the usual regression-based methods, and the standard covariance matrices can also be used for t-tests of hypotheses. Shao et al. (2010) prove that controlling for balancing variables will yield tests of the correct size. As emphasized by Bruhn and McKenzie (2009), researchers should commit ex-ante to controlling for the balancing variables, since this increases power on average, but retaining the option to analyze without controlling for the balancing variables gives the researcher a degree of freedom that can distort the size of a test.

Randomization inference can be conducted by following the "reasoned basis for inference" logic of Fisher (1935): because the design assumption is that the precise order of arrival of units is arbitrary, one can construct counterfactual distributions by reshuffling this order in which subjects arrive (Simon, 1979). If the study incorporates a biased coin (as in Section 3.7), then one can instead re-randomize the biased coin flips, holding the order of arrival fixed. Shao and Yu (2013) also propose a covariate-augmented bootstrap method and show that it provides valid tests for generalized linear models. Bugni et al. (2015) generalize the results of Shao et al. (2010) by showing that the traditional t-test provides asymptotically conservative inference, and that asymptotically non-conservative inference can be achieved by using tests based on covariate adaptive permutations or based on regressions with strata-specific dummies. The use of these asymptotically non-

[10] This example is inspired by Berry et al. (2015), who, regrettably, were not aware of the D_A-optimal method at the time of implementation, and used a complicated version of the block method described in Section 5.1.2.

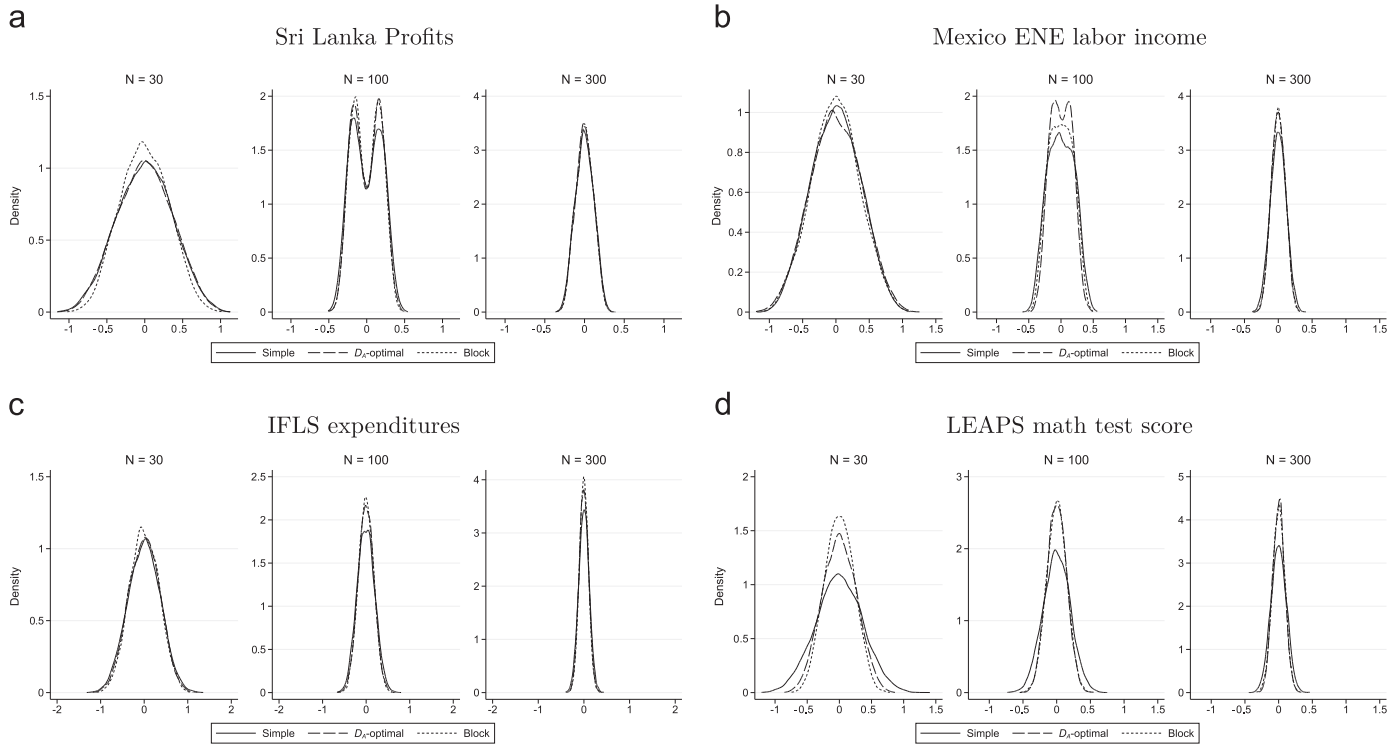

Fig. 2. Imbalance at follow-up. *Notes*: (a–d) show the distribution of differences in means between the treatment and control groups at follow-up for each dataset at all three sample sizes: 30, 100 and 300 observations. Distributions are kernel density plots, using the Epanechnikov kernel, based on 10,000 bootstrap iterations. In each iteration, the difference in means is divided by the standard deviation of the outcome variable. No mean normalization is done; means are close to zero as the result of the balancing methods used.

Table 2
How do the different methods compare in terms of balance on future outcomes?

	(*Sample size of 30*)				
	Simple random	Block (2 variables)	Block (4 variables)	D_A optimal (2 variables)	D_A optimal (4 variables)
Panel A. Average difference in FOLLOW-UP between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	−0.001	0.004	−0.001	−0.001	−0.001
Child schooling (Indonesia)	−0.003	0.001	0.000	0.008	−0.002
Housing expenditure (Indonesia)	0.000	−0.004	0.000	−0.002	0.005
Labor income (Mexico)	0.001	0.006	0.000	0.000	−0.001
Height *z*-score (Pakistan)	−0.004	0.005	−0.005	−0.002	0.001
Math test score (Pakistan)	−0.003	0.000	0.001	−0.001	−0.004
Panel B. Ninety-fifth percentile of difference in FOLLOW-UP between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	0.713	0.610	0.717	0.700	0.700
Child schooling (Indonesia)	0.556	0.745	0.699	0.652	0.652
Housing expenditure (Indonesia)	0.722	0.655	0.673	0.687	0.652
Labor income (Mexico)	0.703	0.713	0.696	0.731	0.754
Height *z*-score (Pakistan)	0.712	0.619	0.663	0.497	0.504
Math test score (Pakistan)	0.716	0.441	0.641	0.515	0.526
Panel C. Proportion of p-values <0.1 for testing difference in FOLLOW-UP means with inference as if pure randomization was used (e.g., no adjustment for strata or balancing variables)					
Microenterprise profits (Sri Lanka)	0.103	0.054	0.105	0.101	0.105
Child schooling (Indonesia)	0.050	0.110	0.088	0.084	0.059
Housing expenditure (Indonesia)	0.105	0.074	0.080	0.087	0.071
Labor income (Mexico)	0.101	0.103	0.092	0.117	0.128
Height *z*-score (Pakistan)	0.102	0.056	0.073	0.015	0.016
Math test score (Pakistan)	0.098	0.005	0.064	0.019	0.023
Panel D. Proportion of p-values <0.1 for testing difference in FOLLOW-UP means with inference which takes account of randomization method (i.e., controls for stratum, or balancing variables)					
Microenterprise profits (Sri Lanka)	0.103	0.084	0.085	0.111	0.109
Child schooling (Indonesia)	0.095	0.120	0.087	0.087	0.101
Housing expenditure (Indonesia)	0.104	0.099	0.126	0.102	0.100
Labor income (Mexico)	0.101	0.105	0.082	0.115	0.122
Height *z*-score (Pakistan)	0.105	0.093	0.148	0.101	0.095
Math test score (Pakistan)	0.095	0.095	0.105	0.114	0.102

Notes: The coefficients in panels A and B are for specifications without controls for balancing variables or stratum dummies. Statistics are based on 10,000 simulations of each method.

Table 3
How do the different methods compare in terms of balance on future outcomes?

	(*Sample size of* 300)				
	Simple random	Block (2 variables)	Block (4 variables)	D_A optimal (2 variables)	D_A optimal (4 variables)
Panel A. Average difference in FOLLOW-UP between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	0.001	0.000	0.003	−0.002	0.000
Child schooling (Indonesia)	0.000	0.002	0.001	−0.002	−0.001
Housing expenditure (Indonesia)	0.001	0.000	0.001	−0.001	0.001
Labor income (Mexico)	0.000	0.000	−0.001	−0.001	−0.002
Height *z*-score (Pakistan)	0.001	0.001	0.000	−0.001	−0.001
Math test score (Pakistan)	0.001	0.000	0.001	0.000	0.000
Panel B. Ninety-fifth percentile of difference in FOLLOW-UP between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	0.219	0.211	0.211	0.210	0.213
Child schooling (Indonesia)	0.227	0.219	0.212	0.227	0.216
Housing expenditure (Indonesia)	0.226	0.195	0.190	0.199	0.193
Labor income (Mexico)	0.226	0.196	0.195	0.198	0.195
Height *z*-score (Pakistan)	0.227	0.188	0.193	0.187	0.188
Math test score (Pakistan)	0.224	0.182	0.183	0.175	0.174
Panel C. Proportion of p-values <0.1 *for testing difference in FOLLOW-UP means with inference as if pure randomization was used (e.g., no adjustment for strata or balancing variables)*					
Microenterprise profits (Sri Lanka)	0.095	0.086	0.086	0.085	0.091
Child schooling (Indonesia)	0.115	0.087	0.083	0.111	0.088
Housing expenditure (Indonesia)	0.099	0.059	0.050	0.060	0.054
Labor income (Mexico)	0.101	0.058	0.058	0.062	0.058
Height *z*-score (Pakistan)	0.099	0.048	0.054	0.046	0.047
Math test score (Pakistan)	0.097	0.042	0.043	0.033	0.034
Panel D. Proportion of p-values <0.1 *for testing difference in FOLLOW-UP means with inference which takes account of randomization method (i.e., controls for stratum, or balancing variables)*					
Microenterprise profits (Sri Lanka)	0.095	0.109	0.139	0.103	0.112
Child schooling (Indonesia)	0.103	0.101	0.107	0.113	0.098
Housing expenditure (Indonesia)	0.098	0.100	0.100	0.099	0.096
Labor income (Mexico)	0.099	0.100	0.094	0.119	0.117
Height *z*-score (Pakistan)	0.097	0.098	0.100	0.103	0.101
Math test score (Pakistan)	0.102	0.100	0.096	0.097	0.096

Notes: The coefficients in panels A and B are for specifications without controls for balancing variables or stratum dummies. Statistics are based on 10,000 simulations of each method.

conservative methods often results in power advantages relative to the *t*-test.

5. Simulation

Using the four panel datasets and simulation structure of Bruhn and McKenzie (2009), we compare the D_A-optimal method to two sequential methods, simple randomization and block randomization (permutations within block), which are commonly used in clinical and other trials (McEntegart, 2003). We describe these methods in greater detail in Section 5.1.2. For each dataset and each allocation method, we simulate 10,000 randomizations. In each iteration, the enrollment order is randomized, treatments are assigned according to the given method, a simulated output variable is created by adding a true treatment effect (possibly zero) to those assigned treatment, and we obtain an estimated treatment effect with associated standard errors and *p*-values.

5.1. Methods

5.1.1. Data

We use the four panel datasets used in Bruhn and McKenzie (2009), who provide further detail on these data. The first dataset covers microenterprises in Sri Lanka and contains information on firms' profits, assets and other firm and owner characteristics. The second dataset is a subsample from the Mexican employment survey (ENE) containing data on heads of household between 20 and 65 years of age who were first interviewed in the second quarter of 2002. The third dataset comes from the 1997 and 2000 waves of the Indonesian Family Life Survey (IFLS) and contains child schooling outcomes as well as household level data such as weekly expenditure. The final dataset comes from the Learning and Educational Achievement Project (LEAPS) in Pakistan and includes math and height *z*-scores as well as other covariates for children aged 8–12 at baseline. From each dataset, we draw subsamples of 30, 100 and 300 observations to allow a comparison of the methods over small, medium and large samples.

5.1.2. Allocation methods

In the simulations, we use three allocation methods: a simple, unstratified randomization; stratified permuted block randomization; and the D_A-optimal method.

In simple randomization, each subject is randomly assigned to treatment as she arrives, with each treatment given the desired probability. In our simulations, there are only two arms (treatment and control), and each is given probability 0.5.

In stratified permuted block randomization ("block randomization" for brevity in what follows), the researcher creates a separate randomization list for each stratum (unique combination of balancing covariates). For example, when balancing on 3 binary variables, there are 8 ($=2^3$) strata.[11] Since the number of subjects who will fall in each stratum is unknown ex ante, a list should

[11] When using the Block method, the researcher must discretize continuous variables, e.g. above and below the median. However, the median (or other sensible cut point) may not be known in advance. One advantage of the D_A-optimal method is that it allows for continuous covariates. In our application, we included the number of households in each compound as a balancing variable. Although we did not know the distribution of this variable in advance, the D_A-optimal method produced an allocation that was well-balanced both on the mean of the continuous variable and on the proportion above/below the median.

Table 4
How do the different methods compare in terms of power in detecting a given treatment effect?

	(*Sample size of* 30)		
	Simple random	Block (2 variables)	D_A optimal (2 variables)
Panel A. Proportion of p-values <0.10 *when no adjustment is made for method of randomization*			
Microenterprise profits (Sri Lanka)	0.144	0.105	0.145
Child schooling (Indonesia)	0.124	0.149	0.133
Housing expenditure (Indonesia)	0.387	0.375	0.378
Labor income (Mexico)	0.187	0.178	0.196
Height *z*-score (Pakistan)	0.174	0.132	0.075
Math test score (Pakistan)	0.155	0.050	0.081
Panel B. Proportion of p-values <0.10 *when adjustment is made for randomization method*			
Microenterprise profits (Sri Lanka)	0.144	0.142	0.154
Child schooling (Indonesia)	0.122	0.134	0.135
Housing expenditure (Indonesia)	0.396	0.415	0.407
Labor income (Mexico)	0.176	0.171	0.191
Height *z*-score (Pakistan)	0.248	0.199	0.257
Math test score (Pakistan)	0.217	0.315	0.234

Notes: Statistics are based on 10,000 simulations of each method. Simulated treatment effects are as follows: Microenterprise profits – an Rs. 1000 (LKR) increase in profits (about 25% of average baseline profits); child schooling – one in three randomly selected children in the treatment group who would have dropped out do not; household expenditure – an increase of 0.4 in ln household expenditure per capita, which corresponds to about one-half of a standard deviation or moving a household from the twenty-fifth to the fiftieth percentile; labor income – a MEX $920 increase in income (about 20% of average baseline income); height *z*-score – an increase of one quarter of a standard deviation in the *z*-score, where the *z*-score is defined as standard deviations from mean US height for age; math test score – an increase of one quarter of a standard deviation in the test score.

Table 5
How do the different methods compare in terms of power in detecting a given treatment effect?

	(*Sample size of* 300)		
	Simple random	Block (4 variables)	D_A optimal (4 variables)
Panel A. Proportion of p-values <0.10 *when no adjustment is made for method of randomization*			
Microenterprise profits (Sri Lanka)	0.286	0.283	0.279
Child schooling (Indonesia)	0.609	0.573	0.593
Housing expenditure (Indonesia)	0.998	1.000	1.000
Labor income (Mexico)	0.488	0.482	0.481
Height *z*-score (Pakistan)	0.729	0.756	0.759
Math test score (Pakistan)	0.622	0.656	0.653
Panel B. Proportion of p-values <0.10 *when adjustment is made for randomization method*			
Microenterprise profits (Sri Lanka)	0.303	0.346	0.305
Child schooling (Indonesia)	0.602	0.599	0.607
Housing expenditure (Indonesia)	1.000	1.000	1.000
Labor income (Mexico)	0.589	0.541	0.581
Height *z*-score (Pakistan)	0.862	0.849	0.853
Math test score (Pakistan)	0.807	0.783	0.807

Notes: Statistics are based on 10,000 simulations of each method. Simulated treatment effects are as follows: Microenterprise profits – an Rs. 1000 (LKR) increase in profits (about 25% of average baseline profits); child schooling – one in three randomly selected children in the treatment group who would have dropped out do not; household expenditure – an increase of 0.4 in ln household expenditure per capita, which corresponds to about one-half of a standard deviation or moving a household from the twenty-fifth to the fiftieth percentile; labor income – a MEX $920 increase in income (about 20% of average baseline income); height *z*-score – an increase of one quarter of a standard deviation in the *z*-score, where the *z*-score is defined as standard deviations from mean US height for age; math test score – an increase of one quarter of a standard deviation in the test score.

maximize overall balance at each point in the sequence while keeping predictability of each allocation low. The block randomization method uses lists that are sequences of blocks of random but balanced permutations of the treatments. The size of the blocks can be as low as the number of treatments – 2 if there is one treatment and a control (e.g. (T,C), (C,T), (C,T), (T,C) …) – or higher to reduce predictability (e.g. 2 treatments, block size 4: (T,C, C,T), (C,T,C,T), (T,C,T,C), (C,T,T,C), …). The size of the blocks may also be varied randomly to minimize predictability. In our simulations, we abstract from concerns about allocation manipulation and set the block size equal to the number of treatments, which achieves the greatest balance and therefore provides the most conservative test of the D_A-optimal method.

Our third method is the D_A-optimal method, for which we follow the algorithm laid out in Section 2.5. For the first J units, we added random noise to the diagonal of the matrix. We did not use the biased coin variant.

For the block and D_A-optimal methods, we use the same balancing variables as Bruhn and McKenzie (2009). The outcome at baseline is always included. Other variables are chosen based on their likelihood of being good predictors of the follow-up outcome.

5.2. Simulation results

5.2.1. Balance at baseline

We first examine balance on the baseline outcome variable. After each randomization we regress the baseline outcome on the assigned treatment. Since the treatment is fictitious, the distribution of these coefficients gives an idea of the potential for imbalance under each method. Fig. 1 presents the distributions for the three methods over a number of datasets at all three sample sizes. Table 1 gives the average coefficient, the 95th percentile of the distribution, and the proportion of iterations with a p-value less than 0.1 for each dataset and method, for the sample size 100.[12]

Panel A of Table 1 shows that all methods, even simple randomization, achieve balance on average at baseline since all have means close to zero. However, Panel B shows the methods that balance over the baseline outcome do significantly better than simple randomization at avoiding extremes and that the D_A-optimal method clearly dominates the Block method in this respect. The ninety-fifth percentile of the distribution drops from around 0.4 standard deviations for simple randomization to around 0.2 for the Block method and to below 0.08 for the D_A-optimal method. This point is illustrated by the kernel density plots in Fig. 1, where the distribution of coefficients for D_A-optimal appears as a sharp spike around zero compared to the more rounded bell curves for Block and simple randomization.

We also see that including additional balancing variables in either method reduces balance on existing balancing variables, although only slightly.

Table 1 also shows how the methods may affect balance on

[12] Results for $N=30$ and $N=300$ are provided in Tables A1 and A2 in the Appendix. Results do not differ substantively across all three sample sizes.

"unobservables." Of course, it is not possible in practice to assess balance on true unobservables, but in a simulation we can mimic unobservables with observed variables we expect to be correlated with the outcome of interest, but which we intentionally exclude from the set of stratification variables. Standard stratification methods should not, on average, lead to more imbalance on unobservables than unstratified randomization (Aickin, 2000). Here, we obtain a similar result for the sequential balancing methods. All result in balance on unobservables on average (zero average coefficients) and all have roughly the same or lesser likelihood of extreme imbalance, with 95th percentiles of the coefficient distributions at or below that of simple randomization (Sri Lanka: 0.434 for simple and 0.414–0.434 for balancing methods; Mexico: 0.457 for simple and 0.429–0.448 for balancing methods).

5.2.2. Balance at follow-up

We now address balance on the outcome of interest at follow-up. Just as for the baseline case, we have regressed the follow-up outcome on the fictitious treatment variable in each iteration. Fig. 2 shows the distribution of coefficients. Tables 2 and 3 present the same statistics as Table 1 but for the follow-up case, for sample sizes 30 and 300 and with the addition of Panel D.[13] In all cases, the p-values are based on regressions that take into account the method of randomization: for Block randomization, stratum dummies are added to the regression, for D_A-optimal, the balancing covariates are included as controls.

Panel A of Tables 2 and 3 shows that balance on the follow-up outcome is achieved on average by all methods. We expect to see the likelihood of extreme imbalance reduced only when the balancing covariates used are good predictors of the outcome. Bruhn and McKenzie note that covariates in the LEAPS datasets (specifically, lagged values for both outcomes, math and height z-scores) have high predictive power, with 43% or more of the variation in the outcome explained by the balancing covariates. The Sri Lankan microenterprise data and the IFLS schooling data fall at the other end of the spectrum with 17% or less explained. The results in Panel B of Table 2 confirm this hypothesis. For both the LEAPS outcomes, we see substantial improvements in balance when using block randomization or D_A-optimal methods compared to simple randomization. However, for outcomes where balancing covariates have low predictive power, there is no detectable improvement. This is not surprising: given that these baseline covariates have low predictive power for endline outcomes, we would not expect any balancing method to make much difference.

Additionally, as with Bruhn and McKenzie (2009), we see that the benefit of using covariates to balance attenuates as the sample size increases (compare Panel B of Table 2 to Panel B of Table 3). We also confirm that test statistics are incorrect when not controlling for the method of randomization (see Panel C vs. Panel D in Tables 2 and 3).

5.2.3. Power to detect a given treatment effect

Lastly, we turn to perhaps the most important question: How do the methods compare in terms of power for detecting a given treatment effect? To answer this question, we add a constant treatment effect to the output of all subjects assigned to treatment,[14] and then regress the modified outcome on treatment. Tables 4 and 5 present the proportion of estimated treatment effects that are significant at the 0.10 level, with and without proper controls for the method of randomization.[15] As with Bruhn and McKenzie (2009), we see an increase in power in nearly all cases when proper controls for the method of randomization are included.

In Table 4, Panel B we see the D_A-optimal method improves power for all six datasets. The size of the improvements is modest with the given balancing variables, but improvement is consistent. The Block method on the other hand decreases power as often as it increases it. Table 5 shows that the differences in power are small across methods as N grows large. Tables A4–A7 in the Appendix present similar results for various combinations of sample size and number of balancing variables.

6. Experience from the field

We implemented the D_A-optimal method in a randomized trial of safe water and handwashing interventions among 435 compounds in slums of Dhaka, Bangladesh. To our knowledge, this study was the first to apply the D_A-optimal method. Details of the study and results are provided in Guiteras et al. (2015).

All participants received behavior change communication and a free trial of a "chlorine dispenser," a compound-level device for treating water with chlorine (Evidence Action, 2014). At the end of the free trial, we measured the compound's willingness to pay for a one-year subscription for use and maintenance of the chlorine dispenser. We had 8 study arms in a 2-by-2-by-2 interaction of

- Behavior change message was a standard health message vs. messages based on disgust and shame.
- Included handwashing messages and soapy water bottle vs. not included.
- Measured collective willingness to pay for the compound vs. individual household willingness to pay.

Because we were especially interested in the effect of the disgust and shame treatment on handwashing in the handwashing arm, we allocated 2/3 of compounds to handwashing. The other treatments were allocated equally. By trial and error, we found that we could achieve this allocation by the using weight $m = \sqrt{5}$ on the handwashing arms in the weighting procedure described in Section 3.1. Our balancing variables were the number of households in the compound and an indicator for the presence of gas burners connected to the municipal supply.

To implement the allocation, the enumerator collected baseline data on compound size and gas status from each eligible compound and transmitted these data to the field office, either by SMS or a phone call. The field supervisor then entered the covariates into a Stata program that assigned the compound to a treatment cell using the D_A-optimal method.

Manipulation was not likely in our context, because the enumerators could not plausibly have anticipated which assignment any given compound would receive. First, the enumerators collected several baseline variables, but did not know which would be used for stratification. Second, they did not have access to the full list of compounds, covariates and assignments. Third, they were unfamiliar with the D_A-optimal algorithm. If manipulation is a concern, the biased-coin variant (Section 3.7) may be appealing.

The resulting sample was well-balanced on both the balancing variables and other plausibly important covariates that were not explicitly included in the randomization. See Tables A1–A4 of Guiteras et al. (2015) for detailed results. Compounds were well-balanced both on the mean of the (continuous) number of households in the compound and on an indicator for whether the

[13] In the main text, we display only the $N=30$ and $N=300$ cases; the $N=100$ case is presented in Table A3 of the Appendix.

[14] See the notes to Tables 4 and 5 for the values of these imposed treatment effects.

[15] In the main text, we display only the $N=30$ and $N=300$ cases; the $N=100$ case is presented in Table A4 of the Appendix.

(*footnote continued*)

compound had the median (8) or fewer number of households, even though we did not know in advance what the median number would be.

These results were obtained in spite of choosing one balancing variable poorly. In our piloting, we observed that water treatment practices varied importantly by whether the compound had a connection to the municipal gas supply. A gas connection also appeared to be a useful proxy for better overall socio-economic status. However, gas coverage in our study area turned out to be much higher than in the pilot area (even though the pilot area was nearby and similar in many other respects), and in fact was nearly universal (> 95%). Despite this unhelpful balancing variable, the algorithm produced a sample that was well-balanced on the other balancing variable and on our main SES variable, household monthly income. We view the robustness of the method as encouraging.

7. Discussion and conclusion

Given the theoretical optimality of the D_A-optimal method, and the evidence in its favor from our simulations and other simulation studies (Atkinson, 2002; Senn et al., 2010), it is somewhat puzzling that this free lunch has previously gone unclaimed, both in economics and more broadly. We consider this situation in two ways: first, by comparing D_A-optimal allocation to the two most popular covariate-adaptive[16] allocation methods, block randomization and minimization (McEntegart, 2003; Taves, 2010; Pond et al., 2010); second, by addressing concerns that have inhibited all covariate-adaptive methods.

Block randomization, described in Section 5.1.2, is intuitively appealing, simple to implement, and allows for randomization inference through re-randomization within blocks. However, it can only be used with discrete (or discretized) covariates and as the number of blocks (the number of cells of all interacted balancing covariates) grows relative to the overall sample size, a "remainder problem" can arise where marginal balance (overall balance of the number of units per treatment) must be traded off against balance within block.[17] Minimization (Taves, 1974; Pocock and Simon, 1975), in which each unit is assigned to minimize the sum of absolute (or squared) imbalances by balancing variable (see McEntegart, 2003 for an extended discussion and examples), is simple to implement and can be used when there are many balancing covariates. As with block randomization, it requires discrete or discretized balancing variables. Relative to block randomization and minimization, we suspect that the primary drawback of D_A-optimal allocation is its complexity. McEntegart (2003) notes that "the method is difficult to explain to nonstatisticians and it is probably for this reason that it has never been used in practice." Furthermore, the algorithm is not trivial to program and implementation requires rapid communication between the field and a centralized database. However, mobile computing power and communication technology continue to improve, which will reduce these barriers.

All covariate-adaptive methods have faced academic and regulatory resistance, chiefly on two grounds: predictability and inference.[18] Predictability occurs when the implementer can anticipate, either with certainty or high probability, the treatment to which the next unit with a given set of covariates will be allocated. The implementer may be tempted to manipulate the order in which units are processed in a way that is correlated with potential outcomes. This manipulation is clearly a concern in minimization, because the algorithm's simplicity can make it easily predictable. Block randomization is also subject to predictability if the implementer knows the number of treatments and the length of the random block. As the trial moves towards the end of each successive block, the implementer knows which treatments remain and therefore has at least some information on which treatment the next unit may receive. Both minimization and block randomization can be adapted to reduce predictability by adding a probabilistic element similar to the "biased coin" approach we describe in Section 3.7, at the cost of some increase in operational complexity and loss of balance. D_A-optimal allocation is also subject to predictability, although in this case its relative complexity may be a virtue because it makes prediction difficult. As mentioned in Section 6, predictability can be made virtually impossible without resorting to a biased coin by including "placebo" covariates in the intake form and not telling implementers which covariates are operative.

The second major criticism of covariate-adaptive methods is that balancing can complicate inference. In our view, this concern is overstated because many simulation studies have shown that simply controlling for balancing covariates is sufficient for conservative inference in most cases. In addition, several permutation or rerandomization-like methods provide appropriate inference (see references in Section 4). Proponents of covariate-adaptive methods appear to be winning this battle, in that these methods are becoming increasingly common in at least some forms of clinical trials (Pond et al., 2010).

To conclude, the D_A-optimal method is a useful tool for improving balance in situations when *ex ante* stratification is not feasible. The benefits are most pronounced when sample sizes are modest relative to the number of balancing covariates and the number of parameters of interest (including subgroup analysis). Our field experience demonstrates that implementation is feasible and yields a well-balanced sample. In our simulations, inference using the usual regression-based methods performed well, provided we controlled for stratification variables. However, we expect that there are gains to both validity and power from using permutation tests such as those proposed by Bugni et al. (2015). Our implementation required a qualified field supervisor to run the treatment assignment. Future trials could improve on our implementation by coding the algorithm into tablets or smartphones. With such an app, enumerators in the field would then be able to allocate treatments optimally in real time.

Acknowledgments

We thank James Berry, Federico Bugni, Jody Cicolino, John Klopfer and Jordan Norris for helpful comments, and Stephen P. Luby, Kaniz Khatun-e-Jannat, and Leanne Unicomb for their collaboration on the field research that led to this paper. Stata code for implementation is provided at http://papers.ccpr.ucla.edu/papers/PWP-MPRC-2015-012/PWP-MPRC-2015-012_supplement.7z, and we encourage interested researchers to contact us with questions. All errors are our own. We gratefully acknowledge financial support for related research from the Bill and Melinda Gates Foundation, the Clausen Center for International Business

[16] *Covariate-adaptive* methods adjust allocation probabilities based on the history of covariates and assignments. *Response-adaptive* methods adjust based on the history of covariates, assignments and outcomes of previously treated units.

[17] Bruhn and McKenzie (2011) describe a solution that can be applied when stratification is conducted *ex ante*, and provide sample code for their particular application. See Quistorff (2005) for more generally applicable code.

[18] For example, the European Medicines Agency's Committee for Proprietary Medicinal Products once wrote that "techniques of dynamic allocation such as minimization … remain highly controversial…. Dynamic allocation is strongly discouraged" (Committee for Proprietary Medicinal Products, 2004). See Buyse and

(*footnote continued*)
McEntegart (2004) and Senn (2004) for a lively discussion.

and Policy at U.C. Berkeley, and the International Initiative for Impact Evaluation, Inc. (3ie) through the Global Development Network (GDN). The views expressed in this paper are those of the authors alone and may not reflect the views of the funders. The funders had no role in study design or execution, analysis, writing of the paper, or decision to submit for publication. The authors declare no conflicts of interest.

Appendix

See Tables A1–A7.

Table A1
How do the different methods compare in terms of baseline balance?

	(*Sample size of 30*)				
	Simple random	Block (2 variables)	Block (4 variables)	D_A optimal (2 variables)	D_A optimal (4 variables)
Panel A. Average difference in BASELINE between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	−0.003	0.003	0.002	0.000	0.000
Housing expenditure (Indonesia)	0.001	0.001	−0.001	−0.001	0.001
Labor income (Mexico)	0.005	0.003	0.003	−0.001	0.000
Height *z*-score (Pakistan)	−0.003	0.003	0.000	0.000	0.001
Math test score (Pakistan)	−0.003	0.000	−0.001	0.000	0.001
Baseline unobservables (Sri Lanka)	0.001	0.000	0.000	0.000	0.001
Baseline unobservables (Mexico)	0.001	0.000	0.000	−0.001	0.000
Panel B. Ninety-fifth percentile of difference in BASELINE between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	0.701	0.412	0.536	0.205	0.242
Housing expenditure (Indonesia)	0.714	0.351	0.495	0.176	0.210
Labor income (Mexico)	0.697	0.413	0.576	0.196	0.259
Height *z*-score (Pakistan)	0.719	0.440	0.448	0.173	0.204
Math test score (Pakistan)	0.700	0.408	0.578	0.138	0.203
Baseline unobservables (Sri Lanka)	0.803	0.824	0.805	0.803	0.803
Baseline unobservables (Mexico)	0.834	0.771	0.790	0.775	0.771
Panel C. Proportion of p-values <0.1 for testing difference in BASELINE means					
Microenterprise profits (Sri Lanka)	0.106	0.000	0.021	0.000	0.000
Housing expenditure (Indonesia)	0.105	0.000	0.013	0.000	0.000
Labor income (Mexico)	0.101	0.000	0.036	0.000	0.000
Height *z*-score (Pakistan)	0.100	0.006	0.006	0.000	0.000
Math test score (Pakistan)	0.094	0.001	0.038	0.000	0.000
Baseline unobservables (Sri Lanka)	0.090	0.094	0.095	0.089	0.089
Baseline unobservables (Mexico)	0.088	0.078	0.080	0.082	0.077

Notes: Statistics are based on 10,000 simulations of each method.

Table A2
How do the different methods compare in terms of baseline balance?

	(*Sample size of* 300)				
	Simple random	Block (2 variables)	Block (4 variables)	D_A optimal (2 variables)	D_A optimal (4 variables)
Panel A. Average difference in BASELINE between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	0.000	0.001	0.001	0.000	0.000
Housing expenditure (Indonesia)	−0.001	0.001	−0.001	0.000	0.000
Labor income (Mexico)	0.000	0.000	0.000	0.000	0.000
Height *z*-score (Pakistan)	0.001	0.000	0.000	0.000	0.000
Math test score (Pakistan)	0.001	−0.001	0.000	0.000	0.000
Baseline unobservables (Sri Lanka)	0.000	0.000	0.001	0.000	0.001
Baseline unobservables (Mexico)	0.000	0.000	0.000	0.001	0.000
Panel B. Ninety-fifth percentile of difference in BASELINE between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	0.227	0.122	0.130	0.055	0.023
Housing expenditure (Indonesia)	0.230	0.093	0.099	0.038	0.020
Labor income (Mexico)	0.222	0.163	0.166	0.074	0.029
Height *z*-score (Pakistan)	0.227	0.084	0.104	0.038	0.018
Math test score (Pakistan)	0.226	0.093	0.107	0.039	0.019
Baseline unobservables (Sri Lanka)	0.241	0.235	0.234	0.434	0.235
Baseline unobservables (Mexico)	0.259	0.259	0.257	0.448	0.259
Panel C. Proportion of p-values <0.1 for testing difference in BASELINE means					
Microenterprise profits (Sri Lanka)	0.102	0.001	0.002	0.000	0.000
Housing expenditure (Indonesia)	0.104	0.000	0.000	0.000	0.000
Labor income (Mexico)	0.096	0.021	0.024	0.000	0.000
Height *z*-score (Pakistan)	0.104	0.000	0.000	0.000	0.000
Math test score (Pakistan)	0.101	0.000	0.000	0.000	0.000
Baseline unobservables (Sri Lanka)	0.101	0.097	0.093	0.096	0.095
Baseline unobservables (Mexico)	0.092	0.088	0.083	0.102	0.087

Notes: Statistics are based on 10,000 simulations of each method.

Table A3
How do the different methods compare in terms of balance on future outcomes?

	(*Sample size of* 100)				
	Simple random	Block (2 variables)	Block (4 variables)	D_A optimal (2 variables)	D_A optimal (4 variables)
Panel A. Average difference in FOLLOW-UP between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	−0.003	0.003	0.000	0.004	−0.001
Child schooling (Indonesia)	0.000	−0.002	0.000	−0.002	−0.001
Housing expenditure (Indonesia)	−0.003	0.001	0.001	0.002	0.003
Labor income (Mexico)	0.000	0.002	−0.001	0.001	0.000
Height *z*-score (Pakistan)	0.001	−0.001	0.002	0.000	0.002
Math test score (Pakistan)	0.000	−0.002	0.002	0.001	0.000
Panel B. Ninety-fifth percentile of difference in FOLLOW-UP between treatment and control means (in SD)					
Microenterprise profits (Sri Lanka)	0.346	0.320	0.340	0.323	0.322
Child schooling (Indonesia)	0.427	0.399	0.369	0.369	0.341
Housing expenditure (Indonesia)	0.389	0.337	0.343	0.353	0.335
Labor income (Mexico)	0.369	0.335	0.345	0.304	0.311
Height *z*-score (Pakistan)	0.390	0.289	0.313	0.273	0.276
Math test score (Pakistan)	0.393	0.287	0.308	0.284	0.293
Panel C. Proportion of p-values <0.1 for testing difference in FOLLOW-UP means with inference as if pure randomization was used (e.g., no adjustment for strata or balancing variables)					
Microenterprise profits (Sri Lanka)	0.069	0.040	0.063	0.042	0.042
Child schooling (Indonesia)	0.132	0.106	0.077	0.119	0.090
Housing expenditure (Indonesia)	0.101	0.056	0.061	0.066	0.053
Labor income (Mexico)	0.090	0.057	0.066	0.029	0.034
Height *z*-score (Pakistan)	0.096	0.027	0.040	0.019	0.022
Math test score (Pakistan)	0.103	0.024	0.035	0.022	0.027
Panel D. Proportion of p-values <0.1 for testing difference in FOLLOW-UP means with inference which takes account of randomization method (i.e., controls for stratum, or balancing variables)					
Microenterprise profits (Sri Lanka)	0.077	0.089	0.117	0.070	0.065
Child schooling (Indonesia)	0.100	0.095	0.094	0.120	0.098
Housing expenditure (Indonesia)	0.104	0.100	0.096	0.100	0.096
Labor income (Mexico)	0.091	0.086	0.099	0.095	0.097
Height *z*-score (Pakistan)	0.100	0.096	0.096	0.099	0.098
Math test score (Pakistan)	0.103	0.102	0.091	0.096	0.105

Notes: The coefficients in panels A and B are for specifications without controls for balancing variables or stratum dummies. Statistics are based on 10,000 simulations of each method.

Table A4
How do the different methods compare in terms of power in detecting a given treatment effect?

	(*Sample size of* 100)		
	Simple random	Block (2 variables)	D_A optimal (2 variables)
Panel A. Proportion of p-values <0.10 when no adjustment is made for method of randomization			
Microenterprise profits (Sri Lanka)	0.145	0.123	0.132
Child schooling (Indonesia)	0.260	0.298	0.262
Housing expenditure (Indonesia)	0.805	0.847	0.844
Labor income (Mexico)	0.207	0.183	0.144
Height *z*-score (Pakistan)	0.340	0.283	0.272
Math test score (Pakistan)	0.307	0.236	0.241
Panel B. Proportion of p-values <0.10 when adjustment is made for randomization method			
Microenterprise profits (Sri Lanka)	0.149	0.174	0.169
Child schooling (Indonesia)	0.278	0.274	0.266
Housing expenditure (Indonesia)	0.870	0.905	0.887
Labor income (Mexico)	0.248	0.216	0.259
Height *z*-score (Pakistan)	0.540	0.485	0.547
Math test score (Pakistan)	0.462	0.458	0.461

Notes: Statistics are based on 10,000 simulations of each method. Simulated treatment effects are as follows: Microenterprise profits – an Rs. 1000 (LKR) increase in profits (about 25% of average baseline profits); child schooling – one in three randomly selected children in the treatment group who would have dropped out do not; household expenditure – an increase of 0.4 in ln household expenditure per capita, which corresponds to about one-half of a standard deviation or moving a household from the twenty-fifth to the fiftieth percentile; labor income – a MEX$920 increase in income (about 20 percent of average baseline income); height *z*-score – an increase of one quarter of a standard deviation in the *z*-score, where the *z*-score is defined as standard deviations from mean US height for age; math test score – an increase of one quarter of a standard deviation in the test score.

Table A5
How do the different methods compare in terms of power in detecting a given treatment effect?

	(*Sample size of* 300)		
	Simple random	Block (2 variables)	D_A Optimal (2 variables)
Panel A. Proportion of p-values <0.10 *when no adjustment is made for method of randomization*			
Microenterprise profits (Sri Lanka)	0.286	0.276	0.271
Child schooling (Indonesia)	0.609	0.576	0.591
Housing expenditure (Indonesia)	0.998	1.000	1.000
Labor income (Mexico)	0.488	0.485	0.482
Height *z*-score (Pakistan)	0.729	0.771	0.765
Math test score (Pakistan)	0.622	0.654	0.658
Panel B. Proportion of p-values <0.10 *when adjustment is made for randomization method*			
Microenterprise profits (Sri Lanka)	0.302	0.305	0.296
Child schooling (Indonesia)	0.580	0.589	0.593
Housing expenditure (Indonesia)	1.000	1.000	1.000
Labor income (Mexico)	0.594	0.556	0.586
Height *z*-score (Pakistan)	0.866	0.854	0.857
Math test score (Pakistan)	0.811	0.794	0.813

Notes: Statistics are based on 10,000 simulations of each method. Simulated treatment effects are as follows: Microenterprise profits – an Rs. 1000 (LKR) increase in profits (about 25% of average baseline profits); child schooling – one in three randomly selected children in the treatment group who would have dropped out do not; household expenditure – an increase of 0.4 in ln household expenditure per capita, which corresponds to about one-half of a standard deviation or moving a household from the twenty-fifth to the fiftieth percentile; labor income – a MEX$920 increase in income (about 20% of average baseline income); height *z*-score – an increase of one quarter of a standard deviation in the *z*-score, where the *z*-score is defined as standard deviations from mean US height for age; math test score – an increase of one quarter of a standard deviation in the test score.

Table A6
How do the different methods compare in terms of power in detecting a given treatment effect?

	(*Sample size of* 30)		
	Simple random	Block (4 variables)	D_A optimal (4 variables)
Panel A. Proportion of p-values <0.10 *when no adjustment is made for method of randomization*			
Microenterprise profits (Sri Lanka)	0.144	0.146	0.148
Child schooling (Indonesia)	0.124	0.112	0.110
Housing expenditure (Indonesia)	0.387	0.376	0.375
Labor income (Mexico)	0.187	0.184	0.200
Height *z*-score (Pakistan)	0.174	0.143	0.084
Math test score (Pakistan)	0.155	0.134	0.081
Panel B. Proportion of p-values <0.10 *when adjustment is made for randomization method*			
Microenterprise profits (Sri Lanka)	0.135	0.102	0.156
Child schooling (Indonesia)	0.109	0.068	0.117
Housing expenditure (Indonesia)	0.402	0.367	0.433
Labor income (Mexico)	0.161	0.139	0.188
Height *z*-score (Pakistan)	0.233	0.165	0.255
Math test score (Pakistan)	0.206	0.192	0.218

Notes: Statistics are based on 10,000 simulations of each method. Simulated treatment effects are as follows: Microenterprise profits – an Rs. 1000 (LKR) increase in profits (about 25% of average baseline profits); child schooling – one in three randomly selected children in the treatment group who would have dropped out do not; household expenditure – an increase of 0.4 in ln household expenditure per capita, which corresponds to about one-half of a standard deviation or moving a household from the twenty-fifth to the fiftieth percentile; labor income – a MEX$920 increase in income (about 20% of average baseline income); height *z*-score – an increase of one quarter of a standard deviation in the *z*-score, where the *z*-score is defined as standard deviations from mean US height for age; math test score – an increase of one quarter of a standard deviation in the test score.

Table A7
How do the different methods compare in terms of power in detecting a given treatment effect?

	(*Sample size of* 100)		
	Simple random	Block (4 variables)	D_A optimal (4 variables)
Panel A. Proportion of p-values <0.10 *when no adjustment is made for method of randomization*			
Microenterprise profits (Sri Lanka)	0.145	0.142	0.128
Child schooling (Indonesia)	0.260	0.277	0.266
Housing expenditure (Indonesia)	0.805	0.852	0.857
Labor income (Mexico)	0.207	0.186	0.142
Height *z*-score (Pakistan)	0.340	0.299	0.277
Math test score (Pakistan)	0.307	0.255	0.241
Panel B. Proportion of p-values <0.10 *when adjustment is made for randomization method*			
Microenterprise profits (Sri Lanka)	0.147	0.254	0.162
Child schooling (Indonesia)	0.287	0.251	0.324
Housing expenditure (Indonesia)	0.890	0.882	0.910
Labor income (Mexico)	0.242	0.200	0.248
Height *z*-score (Pakistan)	0.530	0.463	0.541
Math test score (Pakistan)	0.453	0.399	0.456

Notes: Statistics are based on 10,000 simulations of each method. Simulated treatment effects are as follows: Microenterprise profits – an Rs. 1000 (LKR) increase in profits (about 25% of average baseline profits); child schooling – one in three randomly selected children in the treatment group who would have dropped out do not; household expenditure – an increase of 0.4 in ln household expenditure per capita, which corresponds to about one-half of a standard deviation or moving a household from the twenty-fifth to the fiftieth percentile; labor income – a MEX$920 increase in income (about 20% of average baseline income); height *z*-score – an increase of one quarter of a standard deviation in the *z*-score, where the *z*-score is defined as standard deviations from mean US height for age; math test score – an increase of one quarter of a standard deviation in the test score.

References

Aickin, Mikel, 2000. Randomization, balance, and the validity and efficiency of design-adaptive allocation methods. J. Stat. Plan. Infer. 94 (1), 97–119. http://dx.doi.org/10.1016/S0378-3758(00)00228-7.

Atkinson, Anthony C., 1982. Optimum biased coin designs for sequential clinical trials with prognostic factors. Biometrika 69 (1), 61–67 ⟨http://www.jstor.org/stable/2335853⟩.

Atkinson, Anthony C., 2002. The comparison of designs for sequential clinical trials with covariate information. J. R. Stat. Soc.: Ser. A 165 (2), 349–373. http://dx.doi.org/10.1111/1467-985X.00564.

Atkinson, Anthony C., Haines, Linda M., 1996. Designs for nonlinear and generalized linear models. In: Ghosh, Subir, Rao, C.R. (Eds.), Design and Analysis of Experiments, Handbook of Statistics vol. 13. Elsevier, pp. 437–475. http://dx.doi.org/10.1016/S0169-7161(96)13016-9 (Chapter 14) ISSN 0169-7161, ISBN 9780444820617.

Beaman, Lori, Magruder, Jeremy, 2012. Who gets the job referral? Evidence from a social networks experiment. Am. Econ. Rev. 102 (7), 3574–3593. http://dx.doi.org/10.1257/aer.102.7.3574.

Berry, James, Fischer, Gregory M., Guiteras, Raymond P., 2015. Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana. BREAD Working Paper No. 459. ⟨http://ibread.org/bread/working/459⟩.

Bertsimas, Dimitris, Johnson, Mac, Kallus, Nathan, 2015. The power of optimization over randomization in designing experiments involving small samples. Oper. Res. 63 (4), 868–876. http://dx.doi.org/10.1287/opre.2015.1361.

Bronchetti, Erin Todd, Dee, Thomas, Huffman, David, Magenheim, Ellen, 2013. When a Nudge isn't enough: defaults and saving among low-income tax filers. Natl. Tax J. 66 (3), 609–634. http://dx.doi.org/10.3386/w16887.

Bruhn, Miriam, McKenzie, David, 2009. In pursuit of balance: randomization in practice in development field experiments. Am. Econ. J.: Appl. Econ. 1 (4), 200–232. http://dx.doi.org/10.1257/app.1.4.200.

Bruhn, Miriam, McKenzie David, 2011. Tools of the Trade: Doing Stratified Randomization with Uneven Numbers in Some Strata. ⟨http://blogs.worldbank.org/impactevaluations/tools-of-the-tradedoing-stratified-randomization-with-uneven-numbers-in-some-strata⟩.

Bugni, Federico, Canay Ivan, Shaikh, Azeem, 2015. Inference Under Covariate-Adaptive Randomization. Cemmap Working Paper CWP45/15. ⟨http://www.cemmap.ac.uk/publication/id/7936⟩.

Buyse, Marc, McEntegart, Damian, 2004. Achieving balance in clinical trials: an unbalanced view from EU regulator. Appl. Clin. Trials 13 (5), 36–40 ⟨http://www.appliedclinicaltrialsonline.com/achieving-balance-clinical-trialsunbalanced-view-eu-regulators⟩.

Ciolino, Jody, Zhao, Wenle, Martin, Renee', Palesch, Yuko, 2011. Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization. Contemp. Clin. Trials 32 (2), 250–259. http://dx.doi.org/10.1016/j.cct.2010.11.005.

Committee for Proprietary Medicinal Products, 2004. Committee for Proprietary Medicinal Products (CPMP) Points to Consider on Adjustment for Baseline Covariates. Stat. Med. 23(5), 701–709. http:dx.doi.org/10.1002/sim.1647.

Efron, Bradley, 1971. Forcing a sequential experiment to be balanced. Biometrika 58 (3), 403–417 ⟨http://www.jstor.org/stable/2334377⟩.

Evidence Action, 2014. Chlorine Dispensers Work for Safer Water: A Review of the Evidence. ⟨http://www.evidenceaction.org/blog-full/chlorineevidence⟩.

Fisher, Ronald A., 1935. The logic of inductive inference. J. R. Stat. Soc., 39–82. http://dx.doi.org/10.2307/2342435.

Guiteras, Raymond P., Levine, David I., Luby, Stephen P., Polley, Thomas H., Khatun-e-Jannat, Kaniz, Unicomb, Leanne, 2015. Disgust, shame and soapy water: tests of novel interventions to promote safe water and hygiene. J. Assoc. Environ. Resour. Econ., Forthcoming. ⟨http://papers.ccpr.ucla.edu/abstract.php?preprint=1092⟩.

Kasy, Maximilian, 2013. Why Experimenters Should Not Randomize, and What They Should Do Instead. Working Paper. ⟨http://scholar.harvard.edu/kasy/publications/whyexperimenters-should-not-randomize-and-what-they-should-do-instead⟩.

Kiefer, Jack, Wolfowitz, Jacob, 1960. The equivalence of two extremum problems. Can. J. Math. 12 (5), 363–365 ⟨http://cms.math.ca/cjm/a145179⟩.

McEntegart, Damian J., 2003. The pursuit of balance using stratified and dynamic randomization techniques: an overview. Drug Inf. J. 37 (3), 293–308 ⟨http://dij.sagepub.com/content/37/3/293.short⟩.

Pocock, Stuart J., Simon, Richard, 1975. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 31 (1), 103–115. http://dx.doi.org/10.2307/2529712.

Pond, Gregory R., Tang, Patricia A., Welch, Stephen A., Chen, Eric X., 2010. Trends in the application of dynamic allocation methods in multi-arm cancer clinical trials. Clin. Trials 7 (3), 227–234. http://dx.doi.org/10.1177/1740774510368301.

Quistorff, Brian, 2015. Stratified Randomization. ⟨http://bquistorff.blogspot.com/2015/09/stratified-randomization.html⟩.

Senn, Stephen, 2004. Unbalanced claims for balance. Appl. Clin. Trials. ⟨http://www.appliedclinicaltrialsonline.com/letters-editor-22⟩.

Senn, Stephen, Anisimov, Vladimir V., Fedorov, Valerii V., 2010. Comparisons of minimization and Atkinson's algorithm. Stat. Med. 29 (7–8), 721–730. http://dx.doi.org/10.1002/sim.3763.

Shao, Jun, Yu, Xinxin, 2013. Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. Biometrics 69 (4), 960–969. http://dx.doi.org/10.1111/biom.12062.

Shao, Jun, Yu, Xinxin, Zhong, Bob, 2010. A theory for testing hypotheses under covariate-adaptive randomization. Biometrika 97 (2), 347–360. http://dx.doi.org/10.1093/biomet/asq014.

Simon, Richard, 1979. Restricted randomization designs in clinical trials. Biometrics 35 (2), 503–512. http://dx.doi.org/10.2307/2530354.

Taves, D.R., 1974. Minimization: a new method of assigning patients to treatment and control groups. Clin. Pharmacol. Ther. 15 (5), 443–453.

Taves, Donald R., 2010. The use of minimization in clinical trials. Contemp. Clin. Trials 31 (2), 180–184. http://dx.doi.org/10.1016/j.cct.2009.12.005.

Zocchi, Silvio S., Atkinson, Anthony C., 1999. Optimum experimental designs for multinomial logistic models. Biometrics 55 (2), 437–444. http://dx.doi.org/10.1111/j.0006-341X.1999.00437.x.