Better experimental design by hybridizing binary matching with imbalance optimization

Abba M. KRIEGER¹, David A. AZRIEL², and Adam KAPELNER³*

¹Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA, U.S.A.
²Faculty of Industrial Engineering and Management, The Technion, Technion City, Israel
³Department of Mathematics, Queens College, CUNY, Queens, NY, U.S.A.

Key words and phrases: Experimental design; matching; optimization; rerandomization.

MSC 2020: Primary 62K05; secondary 62F10.

Abstract: We present a new experimental design procedure that divides a set of experimental units into two groups in order to minimize error in estimating a treatment effect. One concern is the elimination of large covariate imbalance between the two groups before the experiment begins. Another concern is robustness of the design to misspecification in response models. We address both concerns in our proposed design: we first place subjects into pairs using optimal nonbipartite matching, making our estimator robust to complicated nonlinear response models. Our innovation is to keep the matched pairs extant, take differences of the covariate values within each matched pair, and then use the greedy switching heuristic of Krieger et al. (2019) or rerandomization on these differences. This latter step greatly reduces covariate imbalance. Furthermore, our resultant designs are shown to be nearly as random as matching, which is robust to unobserved covariates. When compared to previous designs, our approach exhibits significant improvement in the mean squared error of the treatment effect estimator when the response model is nonlinear and performs at least as well when the response model is linear. Our design procedure can be found as a method in the open source R package available on CRAN called GreedyExperimentalDesign.

Additional Supporting Information may be found in the online version of this article at the publisher’s website.
* Corresponding author: kapelner@qc.cuny.edu
1. INTRODUCTION

The setting we consider is the two-arm, nonsequential, randomized study with a continuous endpoint (e.g., a pill–placebo double-blind clinical trial assessing blood glucose improvement) that seeks inference for a treatment effect. There are 2n individuals, each assessed on p covariates and each placed into either the treatment group or the control group. These covariate measurements are assumed to be known in advance of the randomization (the setting of known measurements is sometimes called offline to distinguish it from the sequential setting, the latter not being the setting considered in this article). Under the classic randomization design, the values of the subjects’ covariates in each group are approximately the same, which is why randomization is employed when seeking causal inference (Cornfield, 1959). Thus, imbalance (of which there are many principled metrics) in the covariate values between the treatment and the control subjects should be small. Upon completion, the sample responses \(y_i\), that is, results from an unknown process we call the response model, are assessed. These sample responses are used to infer the treatment effect \(\beta\).

Our contribution is a new experimental design that provides lower error in the estimation of the treatment effect and is robust across response models that are linear and/or nonlinear. If the response model is linear in the covariates, the best experimental design is the one that optimizes the treatment assignment for minimal covariate imbalance, which we call imbalance-optimizing designs. If the response model is nonlinear, a good strategy is to optimize the treatment assignment via creating binary matches (i.e., nonbipartite pairings of subjects) with small intramatch covariate distance. If the response model is a combination of both linear and nonlinear, then both criteria (minimize covariate imbalance and matching) become important. Our hybrid design does both: first by creating optimal nonbipartite matches on covariate distance, and then optimizing intramatch assignments to provide small covariate imbalance across the whole sample. The latter is done by operating on the \(n\) intramatch covariate differences (the difference vector of the \(p\) covariates) rather than all 2n subjects’ covariate vectors. (An alternative hybrid design is to stratify and then imbalance-optimize, as in Wang, Wang & Liu, 2020.)

The optimization of intramatch covariate difference imbalance can be accomplished by many methods. One example is stratification as done in Groh & McKenzie (2016), where they first stratify and then match. Another choice is the Gram–Schmidt Walk design of Harshaw et al. (2021), who also suggested hybridizing binary matching. We investigate two means of hybridizing matching: (i) with rerandomization and (ii) with the pairwise greedy-switching heuristic of Krieger, Azriel & Kapelner (2019), and leave other hybridizations for future work. Our work demonstrates that these two hybrid designs provide formidable estimation performance in a wide range of modelling scenarios.

The rest of the article proceeds as follows. Section 2 provides our notational setup and assumptions, relevant previous literature, a detailed description of the algorithm that generates our hybrid assignments, theoretical results about our designs, and a section on how to conduct inference. Section 3 illustrates our designs’ performance in simulation. The Supplementary Material provides an expanded set of simulations. The article concludes in Section 4 with a brief discussion and future directions.

2. METHODOLOGY

2.1. Setup and Assumptions

Let \(i\) index the 1…2n experimental subjects, \(j\) index the 1…\(p\) covariates, \(x_{ij}\) denote one covariate measurement, \(x_i\) denote all covariate measurements for a subject, and \(X\) denote the
matrix of all measurements with the $x_i$'s stacked row-wise. Each subject is assigned to the treatment group or the control group denoted as $w_i \in \{-1, +1\}$, where $-1$ denotes assignment to the control group, and $+1$ denotes assignment to the treatment group. All subject assignments are $w = [w_1 \ldots w_{2n}]$, and this vector is synonymously referred to as an assignment, an allocation, or a randomization. The collection of the legal $w$ vectors and the probability distribution on the legal vectors in the experimental setting is known as an experimental design, strategy, algorithm, or procedure. We further denote the continuous uncensored responses as $y = [y_1, \ldots, y_{2n}]^T$, where each $y_i$ is assumed to be a function of $x_i, w_i$, and $\epsilon_i$, where $\epsilon_i$ is the component left unexplained. We define $\beta$ to be the expected difference between the response under T minus the response under C, the population average treatment effect (PATE). This definition requires that we view our $2n$ experimental subjects as samples from a large superpopulation of subjects; and thus we assume the population model (Rosenberger & Lachin, 2016, Chapter 6.2).

We define a design $D$ as $W_D$, a discrete, uniform random variable with support $\mathbb{W}_D \subseteq \{-1, +1\}^{2n}$ and number of vectors $|\mathbb{W}_D| \leq 2^{2n}$. We restrict the designs we consider to those that have the mirror property, that is, for any $w$, if treatment assignments and control assignments were switched, then that resulting assignment would also be in the design: $w \in \mathbb{W}_D \iff -w \in \mathbb{W}_D$. The mirror property is a common assumption in the experimental design literature (Kallus, 2018, Assumption 1b; Nordin & Schultzberg, 2020, p. 5; Kapelner et al., 2021, Assumption 2.2). We consider forced balance designs, where all allocations have the same number of treatment and control subjects (Rosenberger & Lachin, 2016, Chapter 3.3) but this is not strictly necessary for our results. Forcing balance is a minor restriction and leads to a standard design known as the balanced complete randomization design (BCRD), denoted $\mathbb{W}_{BCRD} = \{w : \sum w_i = 0\}$, which has $\binom{2n}{n}$ assignments. If a design further restricts this set, that is, $\mathbb{W}_D \subset \mathbb{W}_{BCRD}$, we call $D$ a restricted design.

We employ the simple differences-in-means estimator

$$\hat{\beta} := \frac{w^T y_n}{n} = \bar{Y}_T - \bar{Y}_C,$$

which is the uniformly minimum variance unbiased estimator for $\beta$ (Lehmann & Casella, 1998, Example 4.7). The variance of $\hat{\beta}$ is difficult to express, as it is the result of a law of total variance accounted for by the three random components: the distribution of $X$, the distribution of $\epsilon$, and the random design $W_D$.

### 2.2. Previous Literature

Our design idea stems from a combination of two longstanding streams of research: first, the pure-imbalance optimizing design literature, and second, the binary-matching bias-reducing literature. For the pure-imbalance optimizing design, we first refer to the rerandomization design $R$, which dates back to Student (1938) whose theory was recently investigated by Morgan & Rubin (2012).

#### 2.2.1. Rerandomization

In the $R$ design, one begins with vectors from BCRD and discards those whose covariate imbalance is above a desired threshold. This idea goes back to Fisher, the father of randomized experiments, who was aware of imbalanced allocations and ironically warned against pure, unrestricted randomization, the procedure that he is famous for.

Here, $\mathbb{W}_{R(a)} := \{w : d_M(w) \leq a, w \in \mathbb{W}_{BCRD}\}$, where

$$d_M(w; X) := (\bar{x}_T - \bar{x}_C)^T \Sigma_X^{-1} (\bar{x}_T - \bar{x}_C),$$

DOI: 10.1002/cjs.11685 The Canadian Journal of Statistics / La revue canadienne de statistique
which is a popular metric of covariate imbalance among the treatment and control groups known as the proportional Mahalanobis distance; \( a \) is an upper-bound acceptability threshold of the covariate imbalance chosen to be sufficiently small; \( \bar{x}_T \) and \( \bar{x}_C \) are the averages across the \( p \) covariates in the treatment and control groups, respectively; and \( \hat{\Sigma}_X \) is the sample variance–covariance matrix of all \( 2n \) subjects. Since \( W_{R(a)} \subset W_{BCRD} \), rerandomization is a restricted design by construction. Asymptotically, the vector of the covariates’ sample averages is multivariate normal-distributed. By assuming this normality in the finite sample case and assuming an additive treatment effect but no assumption on the nature of the response model, the multiplicative reduction in MSE \([\hat{\beta}]\) when using the R design is \( \eta(p, a)R^2 \) where

\[
\eta(p, a) := 1 - \frac{2}{p} \frac{\gamma(p/2 + 1, a/2)}{\gamma(p/2, a/2)},
\]

where \( \gamma(\cdot, \cdot) \) denotes the lower incomplete gamma function, and \( R^2 \) is the variance explained in an OLS linear regression of the response on the covariates (Morgan & Rubin, 2012).

This reduction in treatment estimator error comes from two sources: (i) a reduction of the imbalance in the \( p \) covariates, which is under the experimenter’s control, and (ii) the degree of linearity of the covariates in the response as measured by \( R^2 \). An asymptotic result was later proved by Li et al. (2018, Corollary 2) without requiring the assumption of an additive treatment effect.

### 2.2.2. Numerical methods and greedy pair-switching

There are many ideas for achieving reduction in covariate imbalance that outperform rerandomization. First, Kallus (2018, Section 3.3) conjectured using a heuristic argument that the optimal reduction (i.e., for the best possible vector) is exponentially small in \( n \). Numerical optimization techniques such as in Bertsimas, Johnson & Kallus (2015) employ heuristics to approximate the optimal vector. Some heuristics come with theoretical guarantees, such as the greedy pair-switching design of Krieger, Azriel & Kapelner (2019). We denote this latter design G and we denote the paper by KAK19. Since our design builds on the work of KAK19, we review the design algorithm here briefly.

Design G begins with a draw from BCRD of which \( n \) subjects are allocated to T and \( n \) subjects are allocated to C. Each of the T subjects can be paired to each of the C subjects for a total of \( n \times n = n^2 \) pairs. Each of the \( n^2 \) pairs is switched in isolation, and the resulting \( w \) vector’s imbalance is recorded. The minimum imbalance switch is then performed. We then repeat this procedure of considering all pairs of (T, C), calculating all imbalances if the T and C in the pair are switched and choosing the pair that minimizes imbalance until there is no switch that reduces the imbalance (and thus the algorithm is a greedy heuristic). As each starting point (a draw from BCRD) yields a different ending point with very high probability, the algorithm is stochastic.

The resulting vectors have a provably very low imbalance, which can then be further enhanced by generating many vectors \( w \)’s and retaining only the best of those akin to rerandomization (KAK19, Corollary 1). Moreover, since there are provably few switches (KAK19, Section A.1), the degree of randomness is nearly that of BCRD. Since the algorithm is complicated with a random number of iterations, we cannot explicitly provide the final set \( W_G \) in closed form but we expect it to be very similar to \( W_{BCRD} \) for the reason just provided.

After a procedure such as KAK19, the performance of the estimator is much more sensitive to the second source of error: the strength of the covariates’ effect on the response (as gauged by \( R^2 \)). If the \( R^2 \) value is low (as in the case of a nonlinear response model and/or a high noise setting), covariate imbalance reduction will not be very fruitful. Moreover, Kallus (2018, Section 2.3.3) has proven that minimizing \( a \) in rerandomization is fruitful for minimax estimator error reduction only in the case of a linear response model. Kallus (2018) explains the reason why this advantage is limited only to linear response models: even if \( d_M(w; X) = 0 \) (i.e., the
first moments are matched perfectly), the covariate distributions in the two arms (control and treatment) could be very different. In a nonlinear response model, the estimator performance can depend on more subtle differences in the covariates’ distributions (e.g., their tail behaviour). The poor performance of G and R in the case of nonlinear response models is demonstrated in Section 3.

2.2.3. Binary matching

The literature is rich with ideas that use matching as a vehicle to improve estimation. In observational studies, the assignments are fixed a priori, and hence bipartite matching is performed. Matching in order to provide robust estimation under a nonlinear response model has an extensive body of literature in observational studies and we recommend Stuart (2010) for a broad overview. Rubin (1979) was the first to show this robustness to nonlinearity and recommended that the distance function employed in the matching procedure should be specified as the Mahalanobis distance. There are other approaches that use the idea of matching in observational studies to produce better estimators. For example, in Abadie & Imbens (2011) the counterfactual response for a treated (control) person is estimated as the average responses of nearest neighbour covariates based on the covariates among those assigned to the control (treated) group. This is ultimately used to obtain a better estimate for the treatment effect.

Can binary matching play a role in experimental design? Binary matching implicitly attempts to equalize the covariate distributions in both arms. Here, the $2n$ subjects’ indices are first organized into a matched pair structure $\mathcal{M}_d$, a set whose elements are pairs (a set of two subject indices). The structure $\mathcal{M}_d$ is created by first assuming a distance function $d(x_r, x_s) : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}_{\geq 0}$ whose inputs are the two subjects’ covariates. A $2n \times 2n$ nonnegative symmetric matrix is then computed consisting of distances for all pairs of subjects. The optimal match structure $\mathcal{M}_{d}^\star$ is defined to be one of the $\binom{2n}{n}$ sets that minimizes the sum of the resulting intramatch distances, i.e., $\sum_{\{r,s\}\in\mathcal{M}_d^\star} d(x_r, x_s)$. The algorithm that produces the optimal match structure is known as optimal nonbipartite matching. This algorithm has been shown to be reducible to a polynomial-time algorithm (see Lu et al., 2011 for details and history). To then create an assignment vector $w$, the individual assignments within each of the $n$ pairs are randomly allocated to T/C or C/T via $n$ i.i.d. Bernoulli draws. We denote the design M to be the optimal binary matching design, with the distance function $\left( x_r - x_s \right)^\top \hat{\Sigma}^{-1} \left( x_r - x_s \right)$ the between-two-subjects proportional Mahalanobis distance (Stuart, 2010, Section 2.2). Its allocation space is therefore $\mathbb{W}_M = \{ w : w_r = -w_s, \{r,s\} \in \mathcal{M}_{dM}^\star \} \subset \mathbb{W}_{BCRD}$. This design M was first investigated by Greevy et al. (2004), who reported better imbalance and higher power using randomization tests.

2.3. Our Proposed Designs: MG and MR

Our contribution is relatively simple: we first compute $\mathcal{M}_{dM}^\star$ as part of an M design, and this binary pairing structure will provide robustness to nonlinear response models. Then, instead of assigning the specific pairs’ subjects via Bernoulli draws as explained above, we further restrict the pair assignments to insist on small imbalance by using R (resulting in the matching-then-rerandomization design which we term MR) or by using G (resulting in the matching-then-greedy-pair-switching design which we term MG). For both MR and MG, we first compute $\mathcal{M}_{dG}^\star$ by running the optimal nonbipartite matching algorithm implemented in the R package nbpMatching (Beck, Lu & Greevy, 2016).

Then, for MR we let $\mathbb{W}_{MR}(a) = \{ w : d_M(w; X) \leq a, w \in \mathbb{W}_M \}$, that is, we draw many different assignments from $\mathbb{W}_M$ and retain those whose Mahalanobis distances meet the threshold $a$, thereby finding small imbalances subject to the match structure.
But for MG, we cannot simply input \( w \in \mathbb{W}_M \) directly into the KAK19 algorithm. A direct input will switch pairs of subject allocations arbitrarily, thereby violating the \( M_{dM}^* \) structure (i.e., breaking the binary pairings). Our solution is straightforward with some subtle details. To ensure clarity of exposition, we write the pseudocode for how a vector \( w \) is drawn from MG in Algorithm 1.

First, we temporarily view the matched pairs as individual subjects (of which there are \( n \) assumed to be an even number) as the input into KAK19. This means we temporarily view the intramatch covariate vector differences as the subjects’ covariates, which are inputted into KAK19. Put another way, the T arm now refers to a matched pair assigned T/C, where the covariate difference vector is computed by taking the first subject’s covariates minus the second subject’s covariates. The C arm would be the opposite. We initialize the orientation of these \( n \) pairs via a draw from BCRD. The BCRD draw in this context means there are \( n \) pairs via a draw from BCRD. The BCRD draw in this context means there are \( n/2 \) T/C pairs and \( n/2 \) C/T pairs (lines 1–6, Algorithm 1).

In every iteration of KAK19, all \((n/2)^2\) pair of pairs switches are considered: from (T/C)/(C/T) to (C/T)/(T/C), or vice versa (lines 10–11). One of these will result in the largest reduction in imbalance as measured among the pair of pairs (which is algebraically the same imbalance as measured among the entire \( w \) vector for the \( 2n \) subjects). This best pair of the pair switch is retained, and the algorithm stops when an iteration cannot reduce the imbalance any further.

---

**Algorithm 1.** Drawing one assignment \( w \) from the MG design by inputting \( X, n, \) and \( p \)

1: \[
\{\{r_1, s_1\}, \ldots, \{r_n, s_n\}\} \leftarrow npb\text{Matching}(X) \quad \triangleright \text{Compute the optimal binary match set by minimizing the sum of Mahalanobis intramatch distances (denoted } M_{dM}^* \text{ in the text).}
\]

2: \( b \leftarrow \text{BCRD}(n) \) \quad \triangleright \text{Initialize a random assignment of } n/2 + 1 \text{'s and } n/2 - 1 \text{'s}

3: \( \Delta \) is initialized to be a matrix with \( n \) rows and \( p \) columns

4: for \( i \in 1, \ldots, n \) do \quad \triangleright \text{Populate covariate difference vectors based on a BCRD draw}

5: \hspace{1em} if \( b[i] = +1 \) then \( \Delta[i, \cdot] \leftarrow (x_{ri} - x_{si}) \) else \( \Delta[i, \cdot] \leftarrow (x_{si} - x_{ri}) \)

6: end for

7: \hspace{1em} \( d_0 \leftarrow d_M(b; \Delta) \) \quad \triangleright \text{Calculate the initial proportional Mahalanobis distance (Equation 2)}

8: while \( \text{TRUE} \) do \quad \triangleright \text{Note: this loop performs the KAK19 procedure on } \Delta \text{ to optimize } b

9: \hspace{1em} \( \text{flag} \leftarrow \text{TRUE} \), \( i^+_1 \leftarrow \text{NULL} \), \( i^-_1 \leftarrow \text{NULL} \)

10: \hspace{2em} for \( i +_1 \in \{i : b[i] = +1\} \) do

11: \hspace{3em} for \( i^-_1 \in \{i : b[i] = -1\} \) do

12: \hspace{4em} \( v \leftarrow \text{copy}(b), v[i +_1] \leftarrow -v[i +_1], v[i^-_1] \leftarrow -v[i^-_1], d \leftarrow d_M(v; \Delta) \)

13: \hspace{4em} if \( d < d_0 \) then \( d_0 \leftarrow d, i^+_1 \leftarrow i +_1, i^-_1 \leftarrow i^-_1, \text{flag} \leftarrow \text{FALSE} \)

14: \hspace{3em} end for

15: \hspace{2em} end for

16: if \( \text{flag} \) then break else \( b[i^+_1] \leftarrow -b[i^+_1], b[i^-_1] \leftarrow -b[i^-_1] \) \quad \triangleright \text{Greedy change}

17: end while

18: \( w \leftarrow \text{new blank array of size } 2n \)

19: for \( i_T \in \{i : b[i] = +1\} \) do \quad \triangleright \text{These two loops convert } b \text{ into an assignment } w

20: \hspace{2em} \( w[r_{i_T}] \leftarrow T, w[s_{i_T}] \leftarrow -1 \)

21: end for

22: for \( i_C \in \{i : b[i] = -1\} \) do

23: \hspace{2em} \( w[r_{i_C}] \leftarrow C, w[s_{i_C}] \leftarrow +1 \)

24: end for

25: return \( w \)
This scheme thereby preserves the structure of $M^*_d$. There is then a final step of straightforward bookkeeping that converts the $n$-length vector containing information about the (T/C)/(C/T) pairs into the full $2n$-length $w$ vector of subject treatments (lines 18–24).

One can consider alternative implementations. First, instead of drawing $b$ from BCRD (line 2), we can draw it from the Bernoulli trial, allowing for more randomness upon initialization. Then the iteration in lines 10 and 11 could be an iteration over pairs of matched sets independent of the allocation of $b$ (i.e., for $i+1 \in \{1, \ldots, n-1\}$ do: for $i-1 \in \{i+1, \ldots, n\}$ do then continue with the rest of the algorithm). This allows for a fuller exploration of the allocation space as we are examining $\binom{n}{2}$ pairs instead of $(n/2)^2$ pairs. We expect this method to perform nominally better than MG but we did not opt for this procedure as we could neither leverage the software already written that performs G nor could we leverage the theory about G found in KAK19.

Another implementation could be as follows. After M, each iteration of G can perform one switch of one pair (instead of a switch of a pair of pairs). Here, we have fewer switch possibilities: $n$ instead of $\binom{n}{2}$, and we conjecture this procedure would not perform as well as our proposed MG or the alternative above.

2.4. Theoretical Discussion of MG and MR

2.4.1. Introduction

The ultimate aim is to create designs that estimate the treatment effect in the most efficient way. This is affected by the relationship between observed covariates and the response. One consideration, as described above, is covariate imbalance as measured by the Mahalanobis distance between the mean covariate values in the treatment and control. This imbalance does not necessarily ensure that the distributions of the covariate values between the two arms are similar. Particularly, this affects the performance if the covariates are related to the response in a nonlinear way. This consideration is what motivates matching. Finally, the performance of the estimator of the treatment effect might depend on unobserved covariates as well. This consideration favours designs that are more random, that is, the space of possible assignment vectors is as large as possible. A discussion of the effect that matching followed by greedy pair-switching or rerandomization has on imbalance and on randomness can be found in the respective two subsections. The general point is that it is in essence the combination of the two ideas: matching and rerandomization or matching and greedy pair-switching. It is not mathematically tractable to derive specific results, and so this section covers a discussion that is illustrated by the simulations of the next section.

2.4.2. Imbalance

As the covariates are random and i.i.d. under the superpopulation model, the imbalance as measured by the Mahalanobis distance (squared units) is $O_p(n^{-1})$ for BCRD. In KAK19, it was shown that the imbalance using G is $O_p(n^{-2+4/p})$. R has an imbalance of $O_p(n^{-1})$ with a reduced multiplicative constant over BCRD that is governed by the choice of $a$. Since in our proposal each of G and R is applied after matching, the unit of measurement is now the difference of the covariate values in each of the $n$ pairs after assignments are made in each pair using the flip of a fair coin. Since these differences are typically smaller than the original observations, this leads to an even further reduction in imbalance. We are not aware of any theoretical results for the imbalance after optimal nonbipartite matching when $p > 1$ and the theoretical results when $p = 1$ are also limited. The theory is complicated in part due to (i) the differences no longer being independent and (ii) possibly having different distributions depending on the underlying distribution of the covariate.
In the univariate case, consider the order statistics $X_{(1)}, \ldots, X_{(2n)}$. Optimal matching creates $\mathcal{M} = \{2n, 2n - 1, \ldots, 2, 1\}$, that is, pairing the first largest covariate with the second largest covariate value, then pairing the third largest with the fourth largest, and so on. We will consider the case where $X_1, \ldots, X_{2n}$ are i.i.d. U(0, 1). It is well known that the i-th order statistic has a beta distribution with mean $i/(2n + 1)$. Hence, the difference of each pair is of order $n^{-2}$. If we multiply each covariate value by this factor, they are then commensurate in size with the original data. Define the spacings $\Delta_i = X_{(2i)} - X_{(2i-1)}$ for $i = 1, 2, \ldots, n$. By DasGupta (2011, Theorem 6.6c), $[\Delta_1, \ldots, \Delta_n]$ has the same distribution as $\frac{1}{n^2} [E_1, \ldots, E_n]$, where $E_1, \ldots, E_n, \ldots E_{2n+1}$ are i.i.d. Exp(1) and $S := \sum_{i=1}^{2n+1} E_i = O_p(n)$. Since the denominator $S$ is common to all entries of the vector, applying G to $[\Delta_1, \ldots, \Delta_n]$ is the same as (in terms of distribution) as applying G to $[E_1, \ldots, E_n]$ and then dividing by $S$. Since G produces an imbalance of $O_p(n^{-6})$, applying G on $[E_1, \ldots, E_n]$ likewise produces $O_p(n^{-6})$. Thus when G is applied to $[\Delta_1, \ldots, \Delta_n]$, we will get an imbalance of $O_p(n^{-6})$ and when we divide by $S = O_p(n)$ and use the squared units of Mahalanobis distance, the resulting rate of imbalance is $O_p(n^{-8})$. This argument suggests that MG will have lower balance than both M and G and that MR will have lower balance than both M and R.

To demonstrate these rates, we conduct a brief simulation before the main MSE simulations of Section 3. We vary $n \in \{32, 100, 1000\}$ and $p \in \{1, 2, 5, 10\}$, and for each $n \times p$ setting we draw 50 sets of covariate values all drawn i.i.d. from a $U(-\sqrt{3}, +\sqrt{3})$ random variable (this parallels our MSE simulations in Section 3). For each $n \times p$ covariate set, we draw 500 assignments from each of our six designs $D = BCRD, G, R, M, MG, MR$ for each $n \times p$ setting, where R and MR retain only the best 1% of assignments. For each assignment we calculate $\log_{10}(d_M(w;X))$, the log proportional Mahalanobis distance of Equation (2).

Figure 1 plots the average log proportional Mahalanobis distances versus log sample size in all $D \times p \times$ covariate set settings.

These log–log plots show linear relationships demonstrating empirical evidence that the rates can be expressed in the form $d_M = cn^{-r}$ with $c$ and $r$ being functions of the design $D$ and the number of covariates $p$ (the effect of the specific idiosyncratic values within $X$ appears minimal, as all lines within each design setting are nearly equal).

![Figure 1: Average log proportional Mahalanobis distances by log sample size for all six designs faceted by the number of covariates.](image-url)
Since the relationships are now assumed to be linear, we can estimate the rate $r$ within $d_M = \Theta_p(n^{-r})$, by running a log $d_M$ versus log $n$ regression. We tabulate the estimates of $r$ (i.e., the slope coefficients in these log–log linear regressions) in Table 1.

Table 1 illustrates that the rate of BCRD is 1 regardless of $p$, which is expected as the variance of $\bar{X}_T - \bar{X}_C$ is $\Theta_p(n^{-1})$ with $p$ only affecting the constant $c$. The rate of R is also $\Theta_p(n^{-1})$, and this is also expected as rerandomization herein was defined as a fixed subset of the BCRD vectors regardless of the sample size (i.e., the best 1% of the assignments). The lower imbalance of the R design is due to its significantly lower constant $c$ when compared to BCRD (i.e., the difference in y-intercepts in Figure 1). For all other designs, the rates decline with the introduction of more covariates.

The rapid rate of G agrees with what is expected by theory for all $p$. The rate of M is also expected by theory to be $r = 3$ for $p = 1$. The reason is that spacings (differences of successive order statistics) are $\Theta_p(n^{-1})$ for uniform random variables. Since Mahalanobis distance is measured in squared units of spacing, this suggests a rate of $r = 2$. We then take the average over the matched sets to produce an additional order, resulting in $r = 3$. We are unaware of any theory that predicts M’s specific rates for $p > 1$. As before, the rate for MR is the same as that of M but its constant is significantly lower for the same reason that the rate for R is the same as that of BCRD with a lower constant (see Figure 1).

The best performing design is MG. This rate is expected to be the sum of $r = 3$ from M and $r = 5$ from G to result in $r = 8$ (although G has an observed value of $r = 6$, we already accounted for the additional order when averaging in M). Thus, M and G are operating independently to decrease imbalance.

Asymptotics are always meant to be suggestive and may not reflect the reality in small sample sizes. However, it is evident from the slopes of each design in Figure 1 that the asymptotic rates are good approximations even for samples as small as 100.

2.4.3. Degree of randomness

Since the MR and MG designs are more restrictive, the gain in imbalance potentially comes at a price. This leads to a discussion of the degree of randomness in MG and MR. We reiterate that in experimental designs, it is important for the assignments to be highly random, as this randomness provides insurance against a large variance in the treatment effect estimator due to unobserved covariates (Kapelner et al., 2021). M, G, and R are less random than BCRD because they are forms of restricted randomization. In the case of G, KAK19 (Theorem 2 and Proposition 1) shows that G’s assignments are asymptotically as random as those of BCRD according to three degree-of-randomness metrics: (i) the pairwise entropy metric of assignments, (ii) the standard error metric of probability of pairwise assignments, and (iii) the accidental bias metric of Efron (1971, Section 5) defined as the maximum eigenvalue of the variance–covariance

| $D \downarrow, p \to$ | 1     | 2     | 5     | 10    |
|-------------------|-------|-------|-------|-------|
| BCRD              | 0.999 | 0.992 | 1.002 | 1.009 |
| R                 | 1.016 | 1.017 | 1.023 | 1.043 |
| G                 | 6.011 | 4.054 | 2.889 | 2.515 |
| M                 | 2.969 | 2.069 | 1.510 | 1.324 |
| MR                | 2.968 | 2.070 | 1.502 | 1.317 |
| MG                | 8.184 | 5.111 | 3.384 | 2.873 |

DOI: 10.1002/cjs.11685

The Canadian Journal of Statistics / La revue canadienne de statistique
matrix of $W$, the random variable producing assignments in any strategy (see Section 2.1). The reason for $G$’s high degree of randomness is that the number of pairwise switches made in $G$ is small, being only $O_p(\sqrt{n})$. Thus, since $G$ starts with BCRD, and its assignments are not changed significantly, the resulting $\mathbb{W}_G$ is very large. In the case of $R$, we directly limit the number of designs based on the cutoff $a$. As $a$ decreases, $\mathbb{W}_R$ shrinks relative to $\mathbb{W}_{BCRD}$, making $R$ less random. Since $a$ can never be too small due to computational limitations, it follows that randomness is not substantially deteriorated in $R$. Thus, $MG$ is as random as $M$ asymptotically and $MR$ in practice is likely as random as $M$ asymptotically as well.

We apply $G$ after $M$ and $R$ after $M$. Therefore, the degree of randomness of $MG$ and $MR$ is essentially the same as the degree of randomness of $M$. It is easy to verify that the maximum eigenvalue of the variance–covariance of $W$ for $M$ is essentially twice that of BCRD. Since this is a constant effect, rather than an effect that is an order in $n$, the loss in randomness using $M$ will yield a trivial diminution in MSE due to the relationship between unobserved covariates and the response. Hence the gain that $MR$ and $MG$ provide on the observed covariates is almost certain to outweigh the risk one takes due to covariates that are not measured.

2.5. Inference

We briefly describe here how frequentist hypothesis tests are conducted and confidence intervals can be generated under our designs. We first note that traditional Gaussian distribution-based inference is justified by the standard assumption in the superpopulation sampling model only if the design is unrestricted. In our case of a restricted design, “the omission of [assignments] would involve the danger of vitiating normal-law tests” (Anscombe, 1948, p. 188). For example, when employing $R$, the improvement in “balance in the covariates will typically create more precise estimated treatment effects, making traditional Gaussian distribution-based forms of analysis [of $\hat{\beta}$] statistically too conservative” (Morgan & Rubin, 2012, p. 1267). Li, Ding & Rubin (2018) elucidated $\hat{\beta}$’s non-Gaussian asymptotic distribution under $R$ and Wang, Wang & Liu (2020) elucidated $\hat{\beta}$’s non-Gaussian asymptotic distribution under $R$ after stratification. It is possible that our procedure can be shown to fit the criterion proved therein but we leave this to future work.

An inferential procedure that is always valid and exact even for low-sample-size scenarios is the randomization test. After citing Fisher, Tukey, and others, Morgan & Rubin (2012, Section 2.2) recommends this procedure, as it “can incorporate whatever rerandomization procedure was used, will preserve the significance level of the test, and works for any estimator” and can assess the validity of any sharp null hypothesis conditional on the covariate values (Fisher, 1935). To do so, we assume in this section that the only source of the randomness in the response is the treatment assignments $w$. This assumption on the source of randomness is termed the randomization model (Rosenberger & Lachin, 2016, Chapter 6.3), the Fisher model, or the Neyman model, whereby “the $2n$ subjects are the population of interest” (Lin, 2013, p. 297). Even in the randomization-based Neyman model, the resulting discrete estimator can often be approximated by a normal distribution using a finite-population central limit theorem. For a history of this theory and its necessary conditions, see Li & Ding (2017). However, in our restricted designs, these theorems do not apply.

We now describe how to perform the randomization test for the sharp null hypothesis (i.e., all responses will be identical under both treatment groups, $H_0 : Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, n$) under any experimental design (including our MG and MR). We first draw one assignment $w$ from the design $W$, and run the experiment to collect the response $y$. Then, we compute the differences-in-means estimate $\hat{\beta}$. We repeat this process to obtain $R$ replicates, by redrawing a new assignment vector $w_r$ from the design $W$ and recomputing a differences-in-means estimate $\hat{\beta}_r := w_r y/n$, a computation that reuses the fixed experimental responses.
The draws of \( w_r \) may be quite elaborate. For instance, in MG, we begin with a random draw from \( M^* \), where each pair is flipped \((T/C)/(C/T)\) at random and the covariate vector differences are computed. We then run the KAK19 algorithm on the pairs of pairs and iterate until no improvement is possible. Any resulting vector from this procedure is equally likely to have been used in the experiment.

If the sharp null were to be true, the experimental estimate \( \hat{\beta} \) would be indistinguishable from the replicates \( \hat{\beta}_r \), and thus \( \hat{\beta}_1, \ldots, \hat{\beta}_R \) constitute an empirical null distribution. If testing at level \( \alpha \), the \( \alpha/2 \) and \( (1 - \alpha/2) \) quantiles of the set define the null’s retainment region and the \( P \)-value can be assessed by computing the proportion of replicates that are more extreme than the experimental estimate \( \hat{\beta} \).

This test is exact only for \( R = |\mathbb{W}| \), which is computationally infeasible as the number of assignments is exponentially large. However, the \( P \)-value can be approximated to an arbitrary precision \( 2/R \) as long as the replicate estimates are unique, which is expected in the case of a continuous response (Johansson & Schultzberg, 2020, p. 801). Since it is computationally fast to draw an assignment from our designs, it is not difficult to let \( R \) be large enough to obtain any reasonable resolution in practice. Note that the \( R \) assignments must be drawn from \( W \) and not merely be permutations of the original experimental \( w \) as “one should analyze as one designs” (Rosenberger & Lachin, 2016, Chapter 6.4); the randomization test is not a type of permutation test (Hemerik & Goeman, 2021).

As the above testing procedure is valid for any sharp null (i.e., any scenario where the responses under both treatment groups can be explicitly specified), we can use this procedure to compute confidence intervals via the duality of testing and interval construction (Rosenbaum, 2002, Section 2.6.2) by inverting the null

\[
H_0^\beta : Y(1) = Y(0) + \beta \quad \text{for} \quad \beta \in \mathbb{R}; \quad \text{for all} \quad i = 1, \ldots, n
\]

to produce a valid \( 1 - \alpha \) confidence interval for an additive treatment effect.

Parenthetically, note that the unbiasedness of the \( \hat{\beta} \) result cited in Section 2.1 assumed the population model. However, because of the mirror assumption, \( \hat{\beta} \) is still unbiased in the Neyman model. This is an intuitive result, as there is symmetry among each of the \( \pm w \) pairs in the design (see Theorem S1 in the Supplementary Material).

We demonstrate our titular claim in the next section via simulations that estimate the MSE of \( \hat{\beta} \) under our designs. It is typically the case that low squared error implies tighter confidence intervals and higher power.

3. SIMULATIONS OF ESTIMATOR PERFORMANCE

3.1. Setup

To demonstrate the conjecture, we consider \( 2n = 100, p = 2, 5, 10 \) and eight varied response models that take the form \( Y = \beta w + f(X_1, X_2, w) + \epsilon \). We independently draw \( x_1, x_2 \) from \( U(-\sqrt{3}, +\sqrt{3}) \) so that their variance is equal to 1. We repeat this procedure 50 times. For each of these draws from the covariate distribution, we take 50 nested draws of the error terms \( \epsilon \) i.i.d. from \( N(0, \sigma^2) \), where \( \sigma = 0.5 \). For each of these draws, we run a total of 500 nested experiments by drawing one \( w \) vector from the design; computing \( y \) using the fixed covariates, epsilon, \( \beta \), and the drawn \( w \); computing the value of \( \hat{\beta} \) of Equation (1); and finally computing the squared error \( (\beta - \hat{\beta})^2 \).

The eight response models \( f \) are tabulated in Table 2. Since the response model incorporates only the first two covariates, the settings where \( p = 5 \) and \( p = 10 \) simulate three and eight covariates, respectively, which are unrelated to the response. Experimental designs are not privy to this lack of relationship (as they see only \( X \) and not \( f \)) and thus these settings allow

DOI: 10.1002/cjs.11685
Table 2: Eight response models simulated to investigate the differential performance of the experimental designs.

| Name                                | Code | \( f(x_1, x_2, w) \) |
|-------------------------------------|------|-----------------------|
| Zero effects                        | Z    | 0                     |
| Purely linear                       | L    | 3x_1 + 3x_2           |
| Linear with a small nonlinearity    | LsNL | 3x_1 + 3x_2 + x_1^2   |
| Linear and nonlinear                | LNL  | 3x_1 + 3x_2 + x_1^2 + x_2^2 + x_1x_2 |
| Purely nonlinear                    | NL   | x_1^2 + x_2^2 + x_1x_2 |
| Purely linear with heterogeneous treatment effects | HL   | 3x_1 + 3x_2 + x_1w + x_2w |
| Linear and nonlinear with heterogeneous treatment effects | HLNL | 3x_1 + 3x_2 + x_1^2 + x_2^2 + x_1x_2 + x_1w + x_2w |
| Purely nonlinear with heterogeneous treatment effects | HNL  | x_1^2 + x_2^2 + x_1x_2 + x_1w + x_2w |

us to gauge commensurate degradation in performance as the number of nonsense covariates increases. It is known that matching via the Mahalanobis distance metric degrades in performance as dimensionality increases (Branson & Shao, 2021, Section 6) and this type of matching is integral to our proposed designs. The three models HL, HLNL, and HNL simulate heterogeneous treatment effects; specifically, we have first-order interactions of \( w \) with both \( x_1 \) and \( x_2 \). Since the covariates are centred at zero, \( \beta \) is still the PATE. However, estimating \( \hat{\beta} \) in these situations is more difficult, as the dependence of the observed effect on the covariates induces much higher variance. It is also not clear if our matching plus imbalance-optimization designs are a priori advantageous in this case of heterogeneity.

We consider the six designs of the previous section all of which feature an equiprobable draw from a set of mirrored allocations: balanced and completely randomized (BCRD), rerandomization of Morgan & Rubin (2012), where only the top 1% of BCRD vectors are retained (R); the greedy pair-switching of KAK19 (G); optimal nonbipartite matching (M); and our two new hybrid approaches of optimal nonbipartite matching followed by greedy pair-switching on the within-pair covariate differences (MG), and optimal nonbipartite matching followed by rerandomization on the within-pair covariate differences, where, once again, only the top 1% of pairings are retained (MR).

Thus, there are three values of \( p \), six design settings, and eight response models. For each of these 144 settings, we simulate 50 sets of error values \( \times 50 \) sets of covariate values \( \times 500 \) allocations from the design for a total of \( 144 \times 1,250,000 = 180,000,000 \) total calculations of \( \hat{\beta} \).

All simulations were performed with GreedyExperimentalDesign, an \( \text{R} \) package available on CRAN whose core is implemented in Java for speed.

3.2. Results

We tabulate the average squared errors over the draws from the covariate error distribution for all design and response model settings in Table 3 for \( p = 2 \) and in Table 4 for \( p = 10 \) (an equivalently presented tabulation for \( p = 5 \) can be found in Table S1 of the Supplementary Material). Within these tabulations, there is one sub-table per response model and the six designs are ordered from the lowest squared error to the highest squared error within response model. We include Tukey–Kramer comparisons for MSE equivalence with the best design in each sub-table. Performances that are statistically equal to the best design are coloured...
Table 3: Estimated MSE of the estimator and the three salience metrics tabulated for all designs and all response models for $p = 2$.

| Design | Avg. sqd. error | % better than worst design | Mult. of ideal | Design | Avg. sqd. error | % better than worst design | Mult. of ideal |
|--------|-----------------|-----------------------------|----------------|--------|-----------------|-----------------------------|----------------|
| **Model: Z** | | | | **Model: L** | | | |
| R | 0.010 | 1.041 | 0.859 | 1.000 | G | 0.010 | 1.042 | 98.612 | 1.000 |
| G | 0.010 | 1.042 | 0.698 | 1.002 | MG | 0.010 | 1.050 | 98.602 | 1.007 |
| BCRD | 0.010 | 1.044 | 0.535 | 1.003 | MR | 0.011 | 1.056 | 98.593 | 1.013 |
| M | 0.010 | 1.049 | 0.069 | 1.008 | R | 0.014 | 1.424 | 98.103 | 1.366 |
| MR | 0.010 | 1.049 | 0.046 | 1.008 | M | 0.023 | 2.318 | 96.913 | 2.223 |
| MG | 0.010 | 1.050 | 1.009 | BCRD | 0.751 | 75.080 | 72.020 |
| **Model: LsNL** | | | | **Model: LNL** | | | |
| MG | 0.013 | 1.343 | 98.275 | 1.000 | MG | 0.018 | 1.801 | 97.870 | 1.000 |
| MR | 0.014 | 1.352 | 98.263 | 1.007 | MR | 0.018 | 1.820 | 97.848 | 1.010 |
| M | 0.026 | 2.626 | 96.627 | 1.955 | M | 0.031 | 3.075 | 96.364 | 1.707 |
| G | 0.041 | 4.105 | 94.728 | 3.056 | G | 0.115 | 11.548 | 86.343 | 6.411 |
| R | 0.046 | 4.604 | 94.088 | 3.427 | R | 0.124 | 12.434 | 85.296 | 6.903 |
| BCRD | 0.779 | 77.873 | 57.970 | BCRD | 0.846 | 84.560 | 46.947 |
| **Model: NL** | | | | **Model: HL** | | | |
| M | 0.018 | 1.787 | 85.292 | 1.000 | G | 0.027 | 2.657 | 97.431 | 1.000 |
| MG | 0.018 | 1.801 | 85.175 | 1.008 | MG | 0.027 | 2.664 | 97.424 | 1.003 |
| MR | 0.018 | 1.813 | 85.077 | 1.015 | MR | 0.027 | 2.672 | 97.417 | 1.006 |
| G | 0.115 | 11.548 | 4.942 | 6.463 | R | 0.032 | 3.180 | 96.925 | 1.197 |
| R | 0.121 | 12.061 | 0.721 | 6.750 | M | 0.044 | 4.395 | 95.750 | 1.654 |
| BCRD | 0.121 | 12.148 | 6.799 | BCRD | 1.034 | 103.417 | 38.923 |
| **Model: HLNL** | | | | **Model: HNL** | | | |
| MG | 0.034 | 3.416 | 96.967 | 1.000 | MG | 0.034 | 3.416 | 78.018 | 1.000 |
| MR | 0.034 | 3.434 | 96.951 | 1.005 | MR | 0.034 | 3.425 | 77.962 | 1.003 |
| M | 0.052 | 5.158 | 95.420 | 1.510 | M | 0.034 | 3.444 | 77.836 | 1.008 |
| G | 0.132 | 13.162 | 88.313 | 3.853 | G | 0.132 | 13.162 | 15.306 | 3.853 |
| R | 0.142 | 14.198 | 87.393 | 4.156 | R | 0.137 | 13.697 | 11.865 | 4.009 |
| BCRD | 1.126 | 112.625 | 32.967 | BCRD | 0.155 | 15.541 | 4.549 |

Note: Shading of cells indicates performance that is not statistically significantly different from the best performing design.

DOI: 10.1002/cjs.11685

The Canadian Journal of Statistics / La revue canadienne de statistique
in gray. Besides displaying MSE estimates, we display three other performance metrics for comparison purposes: (i) **Multiple of ideal error.** The ideal MSE is the MSE of BCRD in model Z \(2\sigma^2/n = 0.01\). (ii) **Percentage better than worst design.** The MSE percentage reduction of the five best performing designs compared to the worst performing design (which is almost invariably BCRD). (iii) **Multiple of best design.** The squared error multiple of the five worst performing design’s MSE compared to the best performing design’s MSE. Figure S1 in the Supplementary Material provides the density plots for squared error and marks the MSE estimates for all combinations of model, design, and the number of covariates. Even though the squared error is displayed on a log scale, this plot is useful only in discerning large differences in performances (as many design differences are impossible to discern as the vertical lines seem to directly coincide) but it is useful for seeing the large variances. We first examine the performance for the Z model. Here, performance is statistically equal for all designs; thus, there is no penalty for using the MG design when the covariates do not matter (this result is expected due to Kapelner et al., 2021, Eq. 7).

The results of the simulations for all nontrivial response models are definitive: MG, MR, and G provide massive performance gains frequently between 85% and 99% over BCRD. MG finishes as the top performer (or statistically equal to the top performer) for all response models and for all \(p = 2, 5, 10\). Furthermore, there is no significant penalty to using the MG design when the covariates do not matter (Z) or when the covariates are linear (L and HL). When the covariates are linear, it is proven that the imbalance-optimization design is always the best minimax-optimal design (Kallus, 2018, Section 2.3.3). Since the M step in the MG design further reduces imbalance (Section 2.4.2), MG provides a performance edge over pure imbalance-optimizing designs. Our simulations are not sensitive enough to detect this slight edge, so we observe that MG and G are statistically equal. When the response model is completely nonlinear (NL, HNL), binary matching is the proven best minimax design (Kallus, 2018, Section 2.3.2). The G step in MG and the R step in MR do not seem to provide any cost atop M, as the allocation space of MG and MR are subsets of M’s assignments and there are exponentially many such assignments. Thus we observe that M, MG, and MR are all statistically equal in NL for all \(p\). HNL seems to slightly favour MG and MR but only for \(p = 10\). As we do not have a theoretical reason to anticipate an advantage over M, we believe this difference to be a Type I error (we are accounting for multiplicity within each panel using Tukey HSD but we are not accounting for multiplicity across tables). When the response model is a hybrid of both linear and nonlinear components, both the matching and the imbalance optimization properties of designs are hypothesized to be advantageous. This is exactly what we observe: MG performs the best in LNL and HLNL for \(p = 5, 10\) and statistically equal to the best for \(p = 2\). The mostly linear LsNL results are qualitatively similar.

In Section 2.5, we discussed inference for our designs assuming the Neyman model (where only \(w\) is random). To check if our findings hold under fixed covariates and errors, we repeated these simulations under the Neyman model assumption and the results are largely the same (see the parallel Tables S2–S4, and Figure S2 in the Supplementary Material).

4. CONCLUDING REMARKS

We proposed a randomized experimental design that is a hybrid between two well-studied strategies: those that optimize covariate imbalance, and those that use binary matching. The former is known to be minimax-optimal when the response model is linear in the covariates, and the latter is known to be minimax-optimal when the response model is continuous with limits on its derivative with respect to all \(p\) covariates. Additionally, we provided evidence to show that the fusing of both strategies fortuitously enhances the covariate imbalance while retaining a degree of randomization close to that of the classic completely random designs.
Table 4: Estimated MSE of the estimator and the three salience metrics tabulated for all designs and all response models for $p = 10$.

| Design | Avg. sqd. error | Multiple of ideal | % better than worst design | Mult. of best design | Design | Avg. sqd. error | Multiple of ideal | % better than worst design | Mult. of best design |
|--------|-----------------|------------------|-----------------------------|----------------------|--------|-----------------|------------------|-----------------------------|----------------------|
| Model: Z |                 |                  |                             |                      | Model: L |                 |                  |                             |                      |
| R      | 0.010           | 1.040            | 0.440                       | 1.000                | G      | 0.014           | 1.383            | 98.160                      | 1.000                |
| BCRD   | 0.010           | 1.041            | 0.295                       | 1.001                | MG     | 0.018           | 1.753            | 97.669                      | 1.267                |
| M      | 0.010           | 1.042            | 0.268                       | 1.002                | MR     | 0.064           | 6.353            | 91.549                      | 4.592                |
| G      | 0.010           | 1.042            | 0.251                       | 1.002                | R      | 0.173           | 17.306           | 76.978                      | 12.510               |
| MR     | 0.010           | 1.042            | 0.222                       | 1.002                | M      | 0.256           | 25.615           | 65.925                      | 18.516               |
| MG     | 0.010           | 1.044            | 1.004                       | BCRD                 | 0.752  | 75.174          | 54.339                       |                      |
| Model: LsNL |            |                  |                             |                      | Model: LNL |               |                  |                             |                      |
| MG     | 0.043           | 4.349            | 94.421                      | 1.000                | MG     | 0.096           | 9.636            | 88.621                      | 1.000                |
| G      | 0.046           | 4.563            | 94.145                      | 1.049                | G      | 0.123           | 12.267           | 85.513                      | 1.273                |
| MR     | 0.089           | 8.865            | 88.627                      | 2.038                | MR     | 0.142           | 14.159           | 83.278                      | 1.469                |
| R      | 0.204           | 20.404           | 73.823                      | 4.692                | R      | 0.280           | 27.993           | 66.940                      | 2.905                |
| M      | 0.278           | 27.772           | 64.370                      | 6.386                | M      | 0.331           | 33.122           | 60.883                      | 3.438                |
| BCRD   | 0.779           | 77.947           | 17.923                      | BCRD                 | 0.847  | 84.675          | 8.788                        |                      |
| Model: NL |               |                  |                             |                      | Model: HL |               |                  |                             |                      |
| M      | 0.088           | 8.848            | 27.001                      | 1.000                | G      | 0.031           | 3.123            | 96.984                      | 1.000                |
| MR     | 0.089           | 8.910            | 26.485                      | 1.000                | MG     | 0.036           | 3.628            | 96.495                      | 1.162                |
| MG     | 0.089           | 8.938            | 26.258                      | 1.010                | MR     | 0.099           | 9.879            | 90.457                      | 3.164                |
| G      | 0.119           | 11.934           | 1.534                       | 1.349                | R      | 0.248           | 24.805           | 76.039                      | 7.944                |
| R      | 0.120           | 12.041           | 0.656                       | 1.361                | M      | 0.361           | 36.083           | 65.145                      | 11.555               |
| BCRD   | 0.121           | 12.120           | 1.370                       | BCRD                 | 1.035  | 103.522         | 33.152                       |                      |
| Model: HLNL |            |                  |                             |                      | Model: HNL |               |                  |                             |                      |
| MG     | 0.115           | 11.510           | 89.792                      | 1.000                | MG     | 0.106           | 10.576           | 31.886                      | 1.000                |
| G      | 0.140           | 14.002           | 87.582                      | 1.217                | MR     | 0.107           | 10.663           | 31.325                      | 1.008                |
| MR     | 0.177           | 17.675           | 84.324                      | 1.536                | M      | 0.111           | 11.098           | 28.523                      | 1.049                |
| R      | 0.354           | 35.434           | 68.574                      | 3.079                | G      | 0.136           | 13.554           | 12.706                      | 1.282                |
| M      | 0.435           | 43.538           | 61.387                      | 3.783                | R      | 0.141           | 14.053           | 9.497                       | 1.329                |
| BCRD   | 1.128           | 112.755          | 9.797                       | BCRD                 | 0.155  | 15.527          | 1.468                        |                      |

Note: Shading of cells indicates performance that is not statistically significantly different from the best performing design.
Because we fuse both designs together, sacrificing very little performance of either, our designs provides very low MSE performance when estimating a treatment effect in the purely linear case (because we optimize covariate imbalance), the purely nonlinear case (as our assignments are optimal nonbipartite matched pairs), response models that feature both linear and nonlinear components, and response models that feature nonadditive heterogeneous treatment effects. Thus we expect our design to be very powerful when employed in general real-world settings such as clinical trials and internet-based experimentation.

There are many extensions to this work. First, there are different types of matching such as ratio matching (Stuart, 2010, Section 3.1.2), matching in more than two arms (Stuart, 2010, Section 6.1.4), matching with discrete covariates (where the Mahalanobis distance function is not appropriate as discussed in Gu & Rosenbaum, 1993), and matching with some covariates being more important than others (e.g., Morgan & Rubin, 2015; Kapelner & Krieger, 2021). As some matches may not be acceptable, one can caliper match; this would result in a subset of the subjects being matched and a subset being unmatched. In this case, the G algorithm would have to be redesigned but the R procedure would be fairly straightforward. One can also combine G and R together using the recent work of Zhu & Liu (2021) and create an MGR design.

Furthermore, one may wish to use the OLS estimator instead of the classical difference-in-averages estimator (Eq. 1). The OLS estimator improves the asymptotic MSE by an order of magnitude in sample size when the response is linear in the provided covariates (Kapelner et al., 2021, Eq. 14). However, imbalance-optimizing designs (such as R and G) still provide finite-sample improvement even when using the OLS estimator in the case of a linear response model. However, the covariate adjustment in OLS cannot aid with the nonlinear component of the response model and thus we expect our hybrid designs will perform very well in conjunction with the OLS estimator on nonlinear models (as seen in Kapelner & Krieger, 2021, Table 1). Moreover, the OLS estimator is actually preferred in matching designs (Rubin, 1979). Future work can also adapt this design to the sequential setting, where individuals in the experiment arrive one by one and must be assigned quickly to an arm. We can apply the on-the-fly matching procedure in Kapelner & Krieger (2014) with the rerandomization approach in Zhou et al. (2018) in the same vein as our hybrid procedure.

4.1. Replication
All figures and tables can be reproduced by running the R code found at https://github.com/kapelner/GreedyExperimentalDesign/tree/master/hybrid_paper.

ACKNOWLEDGEMENTS
We thank the three referees for helpful comments and suggestions. This research was supported by Grant No. 2018112 from the United States–Israel Binational Science Foundation (BSF).

REFERENCES
Abadie, A. & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics, 29, 1–11.
Anscombe, F. J. (1948). The validity of comparative experiments. Journal of the Royal Statistical Society Series A, 111, 181–211.
Beck, C., Lu, B., & Greevy, R. (2016). nbpMatching: Functions for optimal non-bipartite matching. R package version 1.5.1.
Bertsimas, D., Johnson, M., & Kallus, N. (2015). The power of optimization over randomization in designing experiments involving small samples. Operations Research, 63, 868–876.
Branson, Z. & Shao, S. (2021). Ridge rerandomization: An experimental design strategy in the presence of covariate collinearity. Journal of Statistical Planning and Inference, 211, 287–314.
Cornfield, J. (1959). Principles of research. *American Journal of Mental Deficiency*, 64, 240–252.

DasGupta, A. (2011). Finite sample theory of order statistics and extremes. In *Probability for Statistics and Machine Learning*. Springer, Berlin, 221–248.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58, 403–417.

Fisher, R. A. (1935). *The Design of Experiments*, 1st ed. Oliver & Boyd, Edinburgh & London.

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

Groh, M. & McKenzie, D. (2016). Macroinsurance for microenterprises: A randomized experiment in post-revolution Egypt. *Journal of Development Economics*, 118, 13–25.

Gu, X. S. & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2, 405–420.

Harshaw, C., Savje, F., Spielman, D., & Zhang, P. 2021. Balancing covariates in randomized experiments with the Gram–Schmidt Walk design. arXiv preprint arXiv:1911.03071.

Hemerik, J. & Goeman, J. J. (2021). Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *International Statistical Review*, 89, 367–381.

Johansson, P. & Schultzberg, M. (2020). Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. *Journal of Computational and Graphical Statistics*, 29, 798–813.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society, Series B*, 80, 85–112.

Kapelner, A. & Krieger, A. (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, 70, 378–388.

Kapelner, A. & Krieger, A. (2021). A matching procedure for sequential experiments that iteratively learns which covariates improve power. *Biometrics* (in press).

Kapelner, A., Krieger, A. M., Sklar, M., Shalit, U., & Azriel, D. (2021). Harmonizing optimized designs with classic randomization in experiments. *The American Statistician*, 75, 195–206.

Krieger, A. M., Azriel, D., & Kapelner, A. (2019). Nearly random designs with greatly improved balance. *Biometrika*, 106, 695–701.

Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer Science & Business Media, New York.

Li, X. & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112, 1759–1769.

Li, X., Ding, P., & Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 9157–9162.

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

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

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

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

Nordin, M. & Schultzberg, M. (2020). Properties of restricted randomization with implications for experimental design. arXiv preprint arXiv:2006.14888.

Rosenbaum, P. (2002). *Observational Studies*, 2nd ed. Springer, New York.

Rosenberger, W. F. & Lachin, J. M. (2016). *Randomization in clinical trials: Theory and practice*. John Wiley & Sons, Hoboken, NJ.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74, 318–328.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1–21.

Student. (1938). Comparison between balanced and random arrangements of field plots. *Biometrika*, 29, 363–378.

Wang, X., Wang, T., & Liu, H. (2020). Rerandomization in stratified randomized experiments. arXiv preprint arXiv:2011.05195.

DOI: 10.1002/cjs.11685
Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B., & Zhang, A. (2018). Sequential rerandomization. *Biometrika*, 105, 745–752.

Zhu, K. & Liu, H. (2021). Rerandomization in stratified randomized experiments. *arXiv preprint arXiv:2103.13051*.

*Received 1 March 2021
Accepted 27 September 2021*