The Annals of Applied Statistics
2016, Vol. 10, No. 4, 1958–1976
DOI: 10.1214/16-AOAS959

IMPROVING COVARIATE BALANCE IN 2^K FACTORIAL DESIGNS VIA RERANDOMIZATION WITH AN APPLICATION TO A NEW YORK CITY DEPARTMENT OF EDUCATION HIGH SCHOOL STUDY

BY ZACH BRANSON, TIRTHANKAR DASGUPTA AND DONALD B. RUBIN

Harvard University

A few years ago, the New York Department of Education (NYDE) was planning to conduct an experiment involving five new intervention programs for a selected set of New York City high schools. The goal was to estimate the causal effects of these programs and their interactions on the schools' performance. For each of the schools, about 50 premeasured covariates were available. The schools could be randomly assigned to the 32 treatment combinations of this 2^5 factorial experiment, but such an allocation could have resulted in a huge covariate imbalance across treatment groups. Standard methods used to prevent confounding of treatment effects with covariate effects (e.g., blocking) were not intuitive due to the large number of covariates. In this paper, we explore how the recently proposed and studied method of rerandomization can be applied to this problem and other factorial experiments. We propose how to implement rerandomization in factorial experiments, extend the theoretical properties of rerandomization from single-factor experiments to 2^K factorial designs, and demonstrate, using the NYDE data, how such a designed experiment can improve precision of estimated factorial effects.

1. Introduction. 2^K factorial designs involve K factors each with two levels, often denoted as the "high level" and "low level" of the factor [Yates (1937), Fisher (1942)]. With K factors, there are 2^K unique treatment combinations to which units can be assigned, and often the same number of units are assigned to each combination. Factorial designs are often discussed in an industrial setting, where units are essentially identical and the assignment of units to treatments is arbitrary. However, in recent years factorial designs have become more prevelant in fields where pretreatment covariates are available and units typically differ, including clinical trials [Apfel et al. (2002), Bays et al. (2004)], education [Kollar, Fischer and Slotta (2005)], health sciences [Ravaud et al. (2004)] and psychology [Ahluwalia et al. (2006), Kasari et al. (2012)]. Here, we focus on an experiment that was considered by the New York Department of Education (NYDE). The NYDE considered a 2^5 factorial experiment, where combinations of five "incentive programs" would be introduced to New York City high schools. The goal

Received April 2016; revised June 2016.

Key words and phrases. Experimental design, treatment allocation, randomization, Mahalanobis distance, factorial effects.

of the experiment was to estimate the effect of these programs and their combinations on schools' performance. Dasgupta, Pillai and Rubin (2015) discuss how to analyze such an experiment, but here we focus on the design of the experiment. The NYDE provided a dataset that includes 50 pretreatment covariates for 1376 schools. How should the department allocate the schools to the 32 treatment combinations such that the effects of the incentive programs and their combinations are well estimated?

An initial idea is to randomize the schools to the 32 treatment combinations. Randomized experiments are considered the "gold standard" because randomization balances all potential confounders *on average* [Krause and Howard (2003), Morgan and Rubin (2012)]. However, many have noted that randomized experiments can yield "bad allocations," where some covariates are not well balanced across treatment groups [Seidenfeld (1982), Lindley (1982), Papineau (1994) and Rosenberger and Sverdlov (2008)]. Bad allocations are a concern because covariate imbalance among different treatment groups complicates the interpretation of estimated treatment effects.

Bad allocations are particularly concerning for the NYDE experiment because any randomization may create covariate imbalance across some of the 32 treatment combinations. While covariate imbalance in randomized clinical trials has been addressed, often only treatment-versus-control experiments with categorical covariates are considered [e.g., the covariate-adaptive design of Hu and Hu (2012)]. Likewise, the factorial design literature has focused on categorical covariates. Classic experimental design textbooks like Box, Hunter and Hunter (2005) and Wu and Hamada (2009) suggest using blocking to balance important covariates in factorial designs, which is appropriate when there are few covariates to consider or when covariates are categorical. In contrast, the NYDE dataset includes 50 covariates, most of which are not categorical, and thus how to implement blocking is not obvious.

Instead, we propose a rerandomization algorithm for balanced 2^K factorial designs based on Morgan and Rubin (2012), which developed a framework for rerandomization in the treatment-versus-control case. The motivation behind rerandomization is that "bad allocations" where covariate imbalance is a concern should be avoided, and thus one should keep randomizing—that is, rerandomize—until an acceptable allocation is found. Here we establish several theoretical properties of rerandomization in balanced 2^K factorial designs that increase the precision of factorial effect estimators, which makes rerandomization particularly appealing for designing the NYDE experiment.

Both rerandomization and factorial designs have been explored since Fisher in the 1920s; however, to our knowledge, no one has laid out the framework for implementing rerandomization for factorial designs. Rubin (2008) noted that many did not implement rerandomization because it was computationally intensive; however, with recent improvements in computational power, some have revisited rerandomization. For example, Cox (2009), Bruhn and McKenzie (2009)

and Worrall (2010) all recommend rerandomization, and Morgan and Rubin (2012) formalized these recommendations in treatment-versus-control settings.

Our proposed rerandomization algorithm is not the first procedure that attempts to balance noncategorical covariates for experiments with multiple treatments. The Finite Selection Model (FSM) developed by Morris (1979) assigns units to multiple treatment groups such that covariates are relatively balanced among the groups. Morgan and Rubin (2012) noted that rerandomization and the FSM both attempt to ensure covariate balance, but the FSM does not maintain the correlation structure among covariates, whereas rerandomization can. Xu and Kalbfleisch (2013) proposed the "balance match weighted design" for multiple treatment groups, which performs many randomizations and then selects the randomization that yields the best covariate balance. This is similar to rerandomization, but rerandomization's theoretical guarantees, such as balancing on unobserved covariates on average in addition to improving balance for observed covariates, is appealing. Our rerandomization algorithm can also incorporate various desiderata, such as factorial effects and covariates that vary in importance, which makes the procedure particularly flexible.

In Section 2 we review rerandomization for the treatment-versus-control case, and in Section 3 we establish notation for 2^K factorial designs using the potential outcomes framework. In Section 4 we outline the proposed rerandomization procedure, and in Section 5 we establish theoretical properties that formalize the ways rerandomization is preferable to standard randomization. In Section 6 we use our rerandomization procedure on data from the NYDE.

2. Review of rerandomization. Rubin (2008) recalled a conversation with Bill Cochran, who in turn recalled a conversation with R. A. Fisher, who asserted that a way to protect ourselves against particularly bad randomizations is to rerandomize until a randomization is "acceptable." Morgan and Rubin (2012) suggested implementing rerandomization for a treatment-versus-control experiment as follows:

1. Collect covariate data.
2. Specify a balance criterion determining when a randomization is acceptable.
3. Randomize units to treatment and control groups.
4. Check the balance criterion. If the criterion is met, go to Step 5. Otherwise, return to Step 3.
5. Conduct the experiment using the final randomization obtained in Step 4.
6. Analyze the results using a randomization test, keeping only simulated randomizations that satisfy the balance criteria specified in Step 2.

Morgan and Rubin (2012) used the squared Mahalanobis distance [Mahalanobis (1936)] as a measure for covariate balance. With n units, half assigned to treatment and half assigned to control, and p observed covariates for each unit, the squared

Mahalanobis distance for the treatment-versus-control situation is defined as

$$M \equiv (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^T \operatorname{cov}\big[(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)\big]^{-1}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$$
$$= \frac{n}{4}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^T \operatorname{cov}[\mathbf{x}]^{-1}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C),$$

where $\operatorname{cov}[\mathbf{x}]$ is the sample covariance of the $n \times p$ covariate matrix $\mathbf{x}$, $\bar{\mathbf{x}}_T$ is the p-component column vector of covariate means for units assigned to treatment, and $\bar{\mathbf{x}}_C$ is analogously defined for the control. Here, $\mathbf{x}$ is fixed across randomizations, and thus the only stochastic component of M is the mechanism that assigns treatment and control.

A randomization is declared acceptable if $M \leq a$ for some threshold a. The Mahalanobis distance is well known within the matching and observational study literature where it is used to find subsets of the treatment and control that are similar [Rubin (1976), Rosenbaum and Rubin (1985), Gu and Rosenbaum (1993), Rubin and Thomas (2000)]. Constraining $M \leq a$ can be viewed as finding allocations where the treatment and control covariate means are "similar enough," where the "enough" is determined by the threshold a. Morgan and Rubin (2012) note that, similar to Rosenbaum and Rubin's (1985) argument that matching using the Mahalanobis distance reduces bias due to imbalances in covariates from observational studies, rerandomization using M reduces the sampling *variance* of the standard treatment effect estimator when outcome variables are correlated with covariates.

When $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ is multivariate normal, $M \sim \chi^2_p$ [Mardia, Kent and Bibby (1979)]. In this case, a can be selected by first deciding the percentage, p_a, of randomizations that will be "acceptably well balanced," and then setting a to the p_ath percentile of the χ^2_p distribution. If the multivariate normal assumption is suspect, the empirical distribution of M can be constructed by simulating many randomizations; then the quantile of this empirical distribution can be used instead of the quantile of the chi-squared distribution.

However, we would like to ensure covariate balance across the 2^K treatment combinations of a factorial design, not just between a treatment and control. Morgan and Rubin (2012) mention two options for balancing covariates among multiple treatment groups:

1. Create a criterion for each pairwise comparison among the treatment groups, and then rerandomize if any comparison does not satisfy the criterion.
2. Use a statistic that measures multivariate balance, such as those used in standard MANOVA analyses.

To implement Option 1, a criterion for each $\binom{2^K}{2} = 2^{K-1}(2^K - 1)$ pairwise comparison must be chosen, which may be computationally burdensome. To implement Option 2, there must be a notion of "within-group" variance, which is not immediate for unreplicated 2^K factorial designs where only one unit is assigned to each treatment combination. Furthermore, we may not want to estimate all factorial effects with the same level of precision; for instance, typically we want to estimate

main effects more precisely than high-order interactions, and it is not clear how to incorporate this desideratum into Options 1 or 2. We propose an intuitive adjustment to Option 1 for balanced 2^K factorial designs because Option 1 is more amenable to adjustment than Option 2 and also easier to interpret physically. The proposed adjustment also allows for different levels of precision for estimation of different factorial effects.

3. Notation for 2^K designs under the potential outcomes framework. Consider a balanced 2^K factorial design with $n = r2^K$ units and r replicates assigned to each of the 2^K treatment combinations. In a 2^K factorial design there are $2^K - 1$ factorial effects: K main effects, $\binom{K}{2}$ two-way interactions, $\binom{K}{3}$ three-way interactions and so on. In the NYDE experiment, $r = 43$ and $K = 5$.

The 2^K treatment combinations of a 2^K factorial design are often arranged in a specific order and represented as a $2^K \times K$ *design matrix* whose elements are either -1 (representing the "low level" of a factor) or $+1$ (representing the "high level" of a factor), and thus each row indicates a unique combination of factor assignments [Wu and Hamada (2009)]. We denote the design matrix by $\mathbf{G}$ and write it in the lexicographic order [Espinosa, Dasgupta and Rubin (2016)]. The columns labeled "main effect columns" in Table 1 show $\mathbf{G}$ for a 2^3 factorial design.

To define the interaction effects, we expand $\mathbf{G}$ by augmenting its columns. The column for a specific interaction is created using component-wise multiplication of the corresponding main-effects columns. For example, the last column in Table 1 represents the three-way interaction among factors A, B and C, and is obtained by multiplying the components in the three columns of $\mathbf{G}$. Having appended the interaction columns to the right of $\mathbf{G}$ (columns 5–8 of Table 1) to define the interaction effects, a column of $+1$'s is appended to the left of $\mathbf{G}$ (first column of Table 1) which defines the mean effect. The result is a $2^K \times 2^K$ matrix, denoted by $\widetilde{\mathbf{G}}$. The rows of $\widetilde{\mathbf{G}}$ are indexed by $j = 1, \ldots, 2^K$, one row for each treatment combination,

TABLE 1
$\widetilde{\mathbf{G}}$ *for a* 2^3 *design* (*columns* 2–4 *represent the design matrix* $\mathbf{G}$)

Mean column	Main effect columns			Interaction columns			
	A	*B*	*C*	*AB*	*AC*	*BC*	*ABC*
+1	−1	−1	−1	+1	+1	+1	−1
+1	−1	−1	+1	+1	−1	−1	+1
+1	−1	+1	−1	−1	+1	−1	+1
+1	−1	+1	+1	−1	−1	+1	−1
+1	+1	−1	−1	−1	−1	+1	+1
+1	+1	−1	+1	−1	+1	−1	−1
+1	+1	+1	−1	+1	−1	−1	−1
+1	+1	+1	+1	+1	+1	+1	+1

TABLE 2
Unit-level and population-level factorial effects for a 2^K design

Unit (i)	**Potential outcomes ($\mathbf{Y}_i$)**	**Mean of unit i (θ_{i0})**	**Factorial effect θ_{if}**
1	$\mathbf{Y}_1 = (Y_1(1), \ldots, Y_1(2^K))$	$\frac{1}{2^K}\mathbf{Y}_1\widetilde{\mathbf{G}}_{\cdot 0}$	$\frac{1}{2^{K-1}}\mathbf{Y}_1\widetilde{\mathbf{G}}_{\cdot f}$
2	$\mathbf{Y}_2 = (Y_2(1), \ldots, Y_2(2^K))$	$\frac{1}{2^K}\mathbf{Y}_2\widetilde{\mathbf{G}}_{\cdot 0}$	$\frac{1}{2^{K-1}}\mathbf{Y}_2\widetilde{\mathbf{G}}_{\cdot f}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$\mathbf{Y}_n = (Y_n(1), \ldots, Y_n(2^K))$	$\frac{1}{2^K}\mathbf{Y}_n\widetilde{\mathbf{G}}_{\cdot 0}$	$\frac{1}{2^{K-1}}\mathbf{Y}_n\widetilde{\mathbf{G}}_{\cdot f}$
Average	$\bar{\mathbf{Y}} = \frac{1}{n}(\sum_i Y_i(1), \ldots, \sum_i Y_i(2^K))$	$\bar{\theta}_0 = \frac{1}{2^K}\bar{\mathbf{Y}}\widetilde{\mathbf{G}}_{\cdot 0}$	$\bar{\theta}_f = \frac{1}{2^{K-1}}\bar{\mathbf{Y}}\widetilde{\mathbf{G}}_{\cdot f}$

as indicated by $\mathbf{G}$, and the columns are indexed by $f = 0, 1, \ldots, 2^K - 1$; "$f$" is for factorial effects. Let $\widetilde{\mathbf{G}}_{j\cdot}$ and $\widetilde{\mathbf{G}}_{\cdot f}$ denote the jth row and fth column of $\widetilde{\mathbf{G}}$, respectively.

Let $Y_i(j)$, $i = 1, \ldots, n$, $j = 1, \ldots, 2^K$ denote the potential outcome for the ith unit when exposed to the jth treatment combination, and let $\mathbf{Y}_i = (Y_i(1), \ldots, Y_i(2^K))$ denote the row vector of the 2^K potential outcomes for unit i. The ith row of the left part of Table 2 shows $\mathbf{Y}_i$ for a 2^K design.

Following Dasgupta, Pillai and Rubin (2015) and the standard definition of factorial effects in the experimental design literature [e.g., Wu and Hamada (2009)], the fth linear factorial effect for unit i is the difference of averages of potential outcomes that correspond to levels $+1$ and -1 in the fth column of $\widetilde{\mathbf{G}}$, and is given by

$$\theta_{if} = \frac{1}{2^{K-1}}\mathbf{Y}_i\widetilde{\mathbf{G}}_{\cdot f}, \qquad i = 1, \ldots, n,\ f \in F,$$

where $F \equiv \{1, \ldots, 2^K - 1\}$ indexes all factorial effects. The fth factorial effect for the finite population of n units is defined as

$$\bar{\theta}_f = \frac{1}{n}\sum_{i=1}^{n}\theta_{if} = \frac{1}{2^{K-1}}\bar{\mathbf{Y}}\widetilde{\mathbf{G}}_{\cdot f}, \tag{3.1}$$

where $\bar{\mathbf{Y}} = \frac{1}{n}(\sum_i Y_i(1), \ldots, \sum_i Y_i(2^K))$. The fth factorial effect at the unit level and the population level, represented as functions of the potential outcomes, are shown in the last column of Table 2. The second-to-last column of Table 2 shows the unit-level mean of the potential outcomes

$$\theta_{i0} = \frac{1}{2^K}\mathbf{Y}_i\widetilde{\mathbf{G}}_{\cdot 0}$$

and their grand mean $\bar{\theta}_0$. The population-level grand mean $\bar{\theta}_0$ and the linear factorial effects $\bar{\theta}_1, \ldots, \bar{\theta}_{2^K-1}$ are the estimands (objects of interest) in the standard

linear finite-population framework described here. They need to be estimated because only one element of $\mathbf{Y}_i$ can be observed for each i. We discuss unbiased estimators of these estimands in Section 4.

The vector $(\theta_{i0}, \ldots, \theta_{i(2^K-1)})$ of estimands for unit i is a linear transformation of the vector $\mathbf{Y}_i$ of potential outcomes. Letting the factorial effects vector for unit i be

$$\boldsymbol{\theta}_i = \left(\theta_{i0}, \frac{\theta_{i1}}{2}, \ldots, \frac{\theta_{i(2^K-1)}}{2}\right), \qquad i = 1, \ldots, n, \tag{3.2}$$

straightforward algebra shows that the potential outcomes for unit i can be written as

$$\mathbf{Y}_i = \boldsymbol{\theta}_{i\cdot}\widetilde{\mathbf{G}}^T$$

so that the jth component of $\mathbf{Y}_i$ is

$$Y_i(j) = \boldsymbol{\theta}_i \widetilde{\mathbf{G}}^T_{j\cdot}. \tag{3.3}$$

Having defined the factorial effects as estimands, we now introduce notation for the assignment mechanism of factorial designs and estimators of factorial effects.

Let $\mathbf{W}$ be a $n \times K$ random matrix where the ith row of $\mathbf{W}$, $\mathbf{W}_{i\cdot}$, indicates the treatment assignment for unit i. In randomized balanced factorial designs, a random r rows of $\mathbf{W}$ are equal to a particular row of $\mathbf{G}$, and thus the probability of a unit receiving a particular treatment combination is $\frac{1}{2^K}$, that is, $P(\mathbf{W}_{i\cdot} = \mathbf{G}_{j\cdot}) = \frac{1}{2^K}$ for $i = 1, \ldots, n$, $j = 1, \ldots, 2^K$. In the spirit of expanding $\mathbf{G}$ to $\widetilde{\mathbf{G}}$, we expand the $n \times K$ matrix $\mathbf{W}$ to the $n \times 2^K$ matrix $\widetilde{\mathbf{W}}$ such that $P(\widetilde{\mathbf{W}}_{i\cdot} = \widetilde{\mathbf{G}}_{j\cdot}) = \frac{1}{2^K}$. The first column of $\widetilde{\mathbf{W}}$, $\widetilde{\mathbf{W}}_{\cdot 0}$, is $+1$'s, as in $\widetilde{\mathbf{G}}$; every other element of $\widetilde{\mathbf{W}}$ for $i = 1, \ldots, n$ and $f \in F \equiv \{1, \ldots, 2^K - 1\}$ is defined as

$$\widetilde{W}_{if} = \begin{cases} +1, & \text{if the } i\text{th unit is assigned to } f^+, \\ -1, & \text{if the } i\text{th unit is assigned to } f^-. \end{cases} \tag{3.4}$$

Because the notion of "high" and "low" levels is not necessarily sensible for interactions among factorial effects, for ease of discussion we denote f^+ as the "treatment" for the fth factorial effect and f^- as the "control." For example, a unit is said to be assigned to f^+ (f^-) if that unit is assigned to a treatment combination with $+1$ (-1) in the fth column of $\widetilde{\mathbf{G}}$.

Let $\widetilde{\mathbf{W}}_{\cdot f}$ be the $n \times 1$ column vector denoting the assigned level of some $f \in F$ for all units. A particular random allocation of units in a 2^K design corresponds to one realization of $\widetilde{\mathbf{W}}$, the observed one, $\widetilde{\mathbf{W}}^{\text{obs}}$. The observed outcome for the ith unit will be the potential outcome $Y_i(j)$ when $\widetilde{\mathbf{W}}^{\text{obs}}_{i\cdot} = \widetilde{\mathbf{G}}_{j\cdot}$. Let $\mathbf{y}_{\text{obs}}$ be the n-component column vector of observed outcomes for the n units. The standard estimator of the factorial effect $\bar{\theta}_f$ defined in (3.1) can be written in terms of the observed outcomes and $\widetilde{\mathbf{W}}$:

$$\hat{\theta}_f = \bar{y}_{f^+} - \bar{y}_{f^-} = \frac{\mathbf{y}^T_{\text{obs}}\widetilde{\mathbf{W}}_{\cdot f}}{n/2}, \tag{3.5}$$

where $\bar{y}_{f+}$ is the mean outcome for units assigned to f^+ (the "treatment" for f), and $\bar{y}_{f-}$ is analogously defined for f^- (the "control" for f).

4. The rerandomization algorithm. Rerandomization involves randomizing until an allocation is declared "acceptable," using an acceptance criterion $\phi(\mathbf{x}, \widetilde{\mathbf{W}})$, where ϕ equals one if an allocation is "acceptable" and zero otherwise. Thus, randomization draws from $\{\widetilde{\mathbf{W}}\}$, the set of all allocations in a balanced factorial design, and rerandomization draws from $\{\widetilde{\mathbf{W}}|\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\}$, the set of all "acceptable" allocations. Consider an acceptance criterion that is symmetric in $\widetilde{\mathbf{W}}$, that is, a ϕ such that $\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = \phi(\mathbf{x}, -\widetilde{\mathbf{W}})$. We have the following theorem.

THEOREM 1. *Suppose a completely randomized balanced 2^K factorial design is rerandomized when $\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 0$ for some acceptance criterion symmetric in $\widetilde{\mathbf{W}}$. Then, for all $f \in F$,*

$$\mathbb{E}\big[\hat{\theta}_f | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] = \bar{\theta}_f,$$

where $\hat{\theta}_f$ is the estimator defined in (3.5) and $\bar{\theta}_f$ is the population-level estimand defined in (3.1).

Because ϕ is symmetric in $\widetilde{\mathbf{W}}$, the proof of the unbiasedness of $\hat{\theta}_f$ under rerandomization is analogous to that in Morgan and Rubin (2012) for the treatment-versus-control situation.

If the potential outcomes are correlated with pre-experiment covariates, then so will be the observed outcomes and the estimator $\hat{\theta}_f$ for any $f \in F$. Intuitively, we can increase the precision of $\hat{\theta}_f$ by ensuring covariates are "well balanced" over the two groups used to calculate $\hat{\theta}_f$: units assigned to f^+ and units assigned to f^-, which suggests a balance function that measures the covariate balance among all pairs of these groups.

One such balance function is the squared Mahalanobis distance. To measure the covariate balance between the "treatment" and "control" for a particular $f \in F$, define

$$M_f \equiv \frac{n}{4}(\bar{\mathbf{x}}_{f+} - \bar{\mathbf{x}}_{f-})^T \operatorname{cov}[\mathbf{x}]^{-1}(\bar{\mathbf{x}}_{f+} - \bar{\mathbf{x}}_{f-}), \tag{4.1}$$

where $\operatorname{cov}[\mathbf{x}]$ is the sample covariance of the $n \times p$ covariate matrix $\mathbf{x}$, $\bar{\mathbf{x}}_{f+}$ is the p-component vector of covariate means for units assigned to f^+ and $\bar{\mathbf{x}}_{f-}$ is analogously defined. Note that, analogous to (3.5), $\bar{\mathbf{x}}_{f+} - \bar{\mathbf{x}}_{f-} = \frac{\mathbf{x}^T \widetilde{\mathbf{W}}_{\cdot f}}{n/2}$.

The covariate balance between the "treatment" and the "control" for a particular f is declared "acceptable" by the acceptance criterion

$$\phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) = \begin{cases} 1, & \text{if } M_f \le a_f, \\ 0, & \text{if } M_f > a_f \end{cases} \tag{4.2}$$

for a predetermined threshold a_f. An intuitive procedure that parallels Morgan and Rubin (2012) is to randomize until $\phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) = 1$ in order to increase the covariate balance between the "treatment" and "control" for a particular f. We can do this for every $f \in F$, and thereby define the overall acceptance criterion as

$$\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = \prod_{f \in F} \phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) = \begin{cases} 1, & \text{if } M_f \leq a_f \ \forall f \in F, \\ 0, & \text{if } M_f > a_f \text{ for any } f \in F. \end{cases} \tag{4.3}$$

We thus propose the following rerandomization procedure for balanced 2^K factorial designs:

1. For each $f \in F$, create a squared Mahalanobis distance M_f and choose a threshold criterion a_f.
2. Randomize until $\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1$, where ϕ is defined as in (4.3).

Each threshold criterion a_f can be the same for all $f \in F$ or they can be different. For example, to place more importance on main effects than interactions, thresholds corresponding to main effects should be smaller than those corresponding to interactions. Furthermore, some factorial effects $f \in F$ could be considered unimportant, which corresponds to $a_f = \infty$. Thus, the algorithm can incorporate "tiers of importance" for factorial effects, which we discuss further in Section 5.

We have the following corollary.

COROLLARY 1. *Theorem 1 holds if $\phi(\mathbf{x}, \widetilde{\mathbf{W}})$ is defined as in (4.3); that is, the estimator (3.5) is unbiased under the above rerandomization algorithm.*

Section 5 establishes that the above rerandomization algorithm increases the precision of all factorial effect estimators compared to pure randomization, as long as the covariates $\mathbf{x}$ are correlated with the outcome.

5. Precision properties of rerandomization. The proposed rerandomization algorithm checks M_f for *all* $f \in F$, that is, $\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1$ iff $\phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) = 1$ for all $f \in F$. Thus, both the marginal and joint distributions of $\{\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-} : f \in F\}$ and $\{\hat{\theta}_f : f \in F\}$ need to be examined.

THEOREM 2. *Assume a completely randomized balanced 2^K factorial design is rerandomized using the algorithm proposed at the end of Section 4. Then*

$$\mathbb{E}\big[\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] = 0.$$

The proof of Theorem 2 follows immediately by symmetry of the acceptance criterion.

LEMMA 1. *Assume a completely randomized balanced 2^K factorial design is rerandomized using the algorithm proposed at the end of Section 4, and the covariate means are multivariate normal. Then the elements of $\{\phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) : f \in F\}$ defined in (4.2) are mutually independent.*

The proof of Lemma 1 is in the Appendix.

THEOREM 3. *Assume a completely randomized balanced 2^K factorial design is rerandomized using the algorithm proposed at the end of Section 4, and the covariate means are multivariate normal. Then*:

first, for all $f \in F$,

$$\operatorname{cov}\big[\overline{\mathbf{x}}_{f^+} - \overline{\mathbf{x}}_{f^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] = v_{a_f} \operatorname{cov}[\overline{\mathbf{x}}_{f^+} - \overline{\mathbf{x}}_{f^-}],$$

where

$$v_{a_f} = \frac{2}{p} \frac{\gamma(\frac{p}{2} + 1, \frac{a_f}{2})}{\gamma(\frac{p}{2}, \frac{a_f}{2})}, \tag{5.1}$$

and γ is the incomplete gamma function $\gamma(b, c) \equiv \int_0^c y^{b-1} e^{-y}\, dy$.

And, second, for $f_1, f_2 \in F$, $f_1 \neq f_2$,

$$\operatorname{cov}\big[\overline{\mathbf{x}}_{f_1^+} - \overline{\mathbf{x}}_{f_1^-}, \overline{\mathbf{x}}_{f_2^+} - \overline{\mathbf{x}}_{f_2^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] = \mathbf{0}.$$

The proof of Theorem 3 is in the Appendix.

Theorems 2 and 3 establish that rerandomization leads to unbiased estimators and reduces the variance of $(\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-})$. We define the *percent reduction in variance* for the jth covariate and fth factorial effect as

$$\begin{aligned} &100\left(\frac{\operatorname{var}[\bar{x}_{j,f^+} - \bar{x}_{j,f^-}] - \operatorname{var}[\overline{x}_{j,f^+} - \bar{x}_{j,f^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1]}{\operatorname{var}[\bar{x}_{j,f^+} - \bar{x}_{j,f^-}]}\right) \\ &\quad = 100(1 - v_{a_f}). \end{aligned} \tag{5.2}$$

Therefore, the rerandomization algorithm will reduce the variance of $(\bar{x}_{j,f^+} - \bar{x}_{j,f^-})$ in expectation by $100(1 - v_{a_f})\%$ compared to pure randomization.

To state properties of the marginal and joint distributions of the factorial effect estimators $\{\hat{\theta}_f : f \in F\}$, assumptions must be made about the relationship between the potential outcomes and the factorial effects and covariates. Suppose the factorial effects $\boldsymbol{\theta}_i$ defined in (3.2) are constant across units and there is no interaction between factorial effects and covariate effects. Then the potential outcomes can be written using the following linear model:

$$Y_i(j) = \boldsymbol{\theta}_i \widetilde{\mathbf{G}}_{j\cdot}^T + \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,\ j = 1, \ldots, 2^K, \tag{5.3}$$

where $\widetilde{\mathbf{G}}_{j\cdot}$ is the jth row of $\widetilde{\mathbf{G}}$ defined in Section 3, $\boldsymbol{\beta}$ is the p-component column vector of fixed covariate coefficients, and ε_i indicates any deviations from the linear model. Then the standard unbiased estimator (3.5) can be written as

$$\hat{\theta}_f = \bar{\theta}_f + \boldsymbol{\beta}^T(\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-}) + (\bar{\boldsymbol{\varepsilon}}_{f^+} - \bar{\boldsymbol{\varepsilon}}_{f^-}), \tag{5.4}$$

and the theorem below follows.

THEOREM 4. *Assume* (a) *a completely randomized balanced* 2^K *factorial design is rerandomized using the algorithm proposed at the end of Section* 4, (b) *the covariate means are multivariate normal*, (c) *factorial effects are constant across units*, *and* (d) *there is no interaction between factorial effects and covariate effects*. *Then*, *for all* $f \in F$,

$$\operatorname{var}\big(\hat{\theta}_f | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big) = \big(1 - (1 - v_{a_f})R_f^2\big)\operatorname{var}(\hat{\theta}_f), \tag{5.5}$$

and for $f_1, f_2 \in F$, *such that* $f_1 \neq f_2$,

$$\operatorname{cov}\big(\hat{\theta}_{f_1}, \hat{\theta}_{f_2} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big) = 0,$$

where R_f^2 *is the squared multiple correlation coefficient between* $\mathbf{x}$ *and the group of potential outcomes that correspond to* f^+, *and* v_{a_f} *is defined in* (5.1).

Note that, by the additivity assumption in Theorem 4, the squared multiple correlation coefficient between $\mathbf{x}$ and the group of potential outcomes that correspond to f^+ is equivalent to that between $\mathbf{x}$ and the group of potential outcomes that correspond to f^-. The proof of Theorem 4 is in the Appendix.

Theorem 4 has several implications. First, precision gains for $\hat{\theta}_f$ are calibrated by the threshold a_f. Second, by Lemma 1 and the uncorrelated result of Theorem 4, constraining $M_f \leq a_f$ for some $f \in F$ does not affect another $M_{f'}$ or the precision of $\hat{\theta}_{f'}$. Thus, the algorithm incorporates tiers of importance for factorial effects, where more stringent thresholds lead to greater precision gains. The same can be done for covariates, analogous to Morgan and Rubin (2015), which shows how to adapt rerandomization according to tiers of importance for covariates in the treatment-versus-control case.

To conduct inference using rerandomization, the significance levels of hypotheses should be calculated using a permutation test [Fisher (1942)]. During the permutation test, the distribution of the test statistic under Fisher's sharp null must be created using only randomizations that would be accepted under rerandomization. Thus, thresholds a_f cannot be arbitrarily stringent because there must be enough "acceptable randomizations" such that a proper permutation test can be performed. For additive factorial effects, confidence intervals can be obtained by inverting these permutation tests; that is, a confidence interval can be constructed as the set of values for which the observed data would fail to reject Fisher's sharp null [Dasgupta, Pillai and Rubin (2015)]. Corrections for multiple testing and selection

of active versus inactive effects [as in Espinosa, Dasgupta and Rubin (2016)] are topics for future work.

Other constraints, such as cost or other practical constraints, can also be considered in rerandomization, as we will discuss for the NYDE experiment. For example, if it is unrealistic to fully randomize particular factors, it may be more practical to implement a split-plot design, where one or more factors are first randomized to "whole plots" and then other factors are randomized within plots [Fisher (1925)]. These designs induce a restricted randomization distribution which can be incorporated in the design and analysis stages of rerandomization.

The only computational concern for rerandomization is how long it takes to find acceptable randomizations—one to run the experiment, and many to perform a permutation test. By the independence of Lemma 1, the probability of a randomization being acceptable is $\prod_{f \in F} P(\chi_p^2 \leq a_f)$. This was not a concern for the NYDE experiment, but could be if the number of stringent thresholds is large.

Theorems 3 and 4 require n to be sufficiently large such that the covariate means are multivariate normal. If n is not large enough for the normality assumption to hold via the Central Limit Theorem, then (a) the Mahalanobis distance will not be χ_p^2, and (b) the independence in Lemma 1 will not hold. To address (a), the empirical distribution of each M_f can be used to select each corresponding threshold a_f. As for (b), the elements of $\{\bar{\mathbf{x}}_{f+} - \bar{\mathbf{x}}_{f-} : f \in F\}$ (and, as a consequence, the elements of $\{M_f : f \in F\}$) are always uncorrelated under our proposed rerandomization algorithm. This implies that, under mild regularity conditions, rerandomization will still increase the precision of factorial effect estimators; however, theoretical results found in Theorems 3 and 4 will not hold exactly.

6. Implementing rerandomization for the NYDE experiment. Dasgupta, Pillai and Rubin (2015) discuss how to analyze an educational experiment considered by the New York Department of Education (NYDE) with five "incentive programs" to be introduced to high schools "which desperately need performance improvement." The programs include a quality review, a periodic assessment, inquiry teams, a school-wide performance bonus program and an online resource program; details about these programs are found in Dasgupta, Pillai and Rubin (2015). The dataset includes 2008 information about 50 covariates on 1376 schools. While the NYDE has not yet run such an experiment, here we utilize rerandomization to hypothetically design the NYDE experiment. The NYDE dataset and the R code used to implement our rerandomization algorithm on this dataset are provided in our Supplementary Materials [Branson, Dasgupta and Rubin (2016)].

The NYDE measures schools' performance with a score in each school's *Progress Report*, and we consider nine covariates that will likely be correlated with this score: Total number of students, five different race variables (proportion of white, black, Asian, Native American and Latino students), proportion of female students, enrollment rate and poverty rate. This situation can be considered

an extreme case of a "tiers of covariates" framework, where a subset of nine covariates are considered "important" and the rest are considered "not important." The goal is to assign 43 schools to each of the 32 different treatment combinations such that the factorial effects of the experiment will be well estimated.

Interest usually focuses on main effects and possibly two-way interactions, and higher-order interactions are often considered negligible [Wu and Hamada (2009)]. Thus, we implement a rerandomization algorithm that considers main effects "most important," two-way interactions "less important" and higher-order interactions "not important." We created fifteen squared Mahalanobis distances: one for each of the five main effects and ten two-way interactions. The rerandomization algorithm involves randomizing until $\max(M_1, \ldots, M_5) \leq a_{\text{main}}$ and $\max(M_6, \ldots, M_{15}) \leq a_{\text{interaction}}$, where a_{main} is the $100(0.01^{1/5})$ percentile of the χ^2_9 distribution and $a_{\text{interaction}}$ is the $100(0.1^{1/10})\%$ percentile, so $P(M_1, \ldots, M_5 \leq a_{\text{main}}) = 1\%$ and $P(M_6, \ldots, M_{15} \leq a_{\text{interaction}}) = 10\%$, making the criterion corresponding to the interaction effects less stringent than that of the main effects.

We performed pure randomization and rerandomization 10,000 times. Each rerandomization took about five seconds on a 2.5 GHz Intel Core i7 computer. For each (re)randomization, the covariate mean difference $(\bar{x}_{j,f^+} - \bar{x}_{j,f^-})$ was calculated for each covariate j and factor/interaction f. Figure 1 displays the empirical percent reduction in variance, which shows how much rerandomization reduced the variance of the covariate mean difference for various covariates and factors/interactions compared to pure randomization. Main effects are marked with circles, two-way interaction effects with squares, and three-way interaction effects with triangles. The percent reduction in variance expected given Theorem 3 is marked by a vertical line for each type of factorial effect.

The nine covariates we considered during rerandomization are displayed at the top of the vertical axis of Figure 1. Rerandomization reduced the variance of the covariate mean difference across factors and two-way interactions compared to pure randomization for these covariates, and the reduction varies around what we would expect given Theorem 3. There is more reduction for individual factors than for interactions, as is expected, because the threshold a_{main} was more stringent than $a_{\text{interaction}}$. The percent reduction in variance across three-way interactions is occasionally negative—implying randomization yielded better covariate balance in this case—but this reduction averages close to zero, as expected, because three-way interactions were not considered during rerandomization. Therefore, rerandomization on average increased the covariate balance across main effects and two-way interactions without sacrificing the covariate balance across higher-order interactions.

Figure 1 also displays the percent reduction in variance for two covariates not considered during rerandomization: "number of teachers" and "number of students in temporary housing." Rerandomization yielded more balance for "number of teachers" compared to pure randomization because "number of teachers" is highly correlated ($R^2 = 0.95$) with "number of students," which was considered during

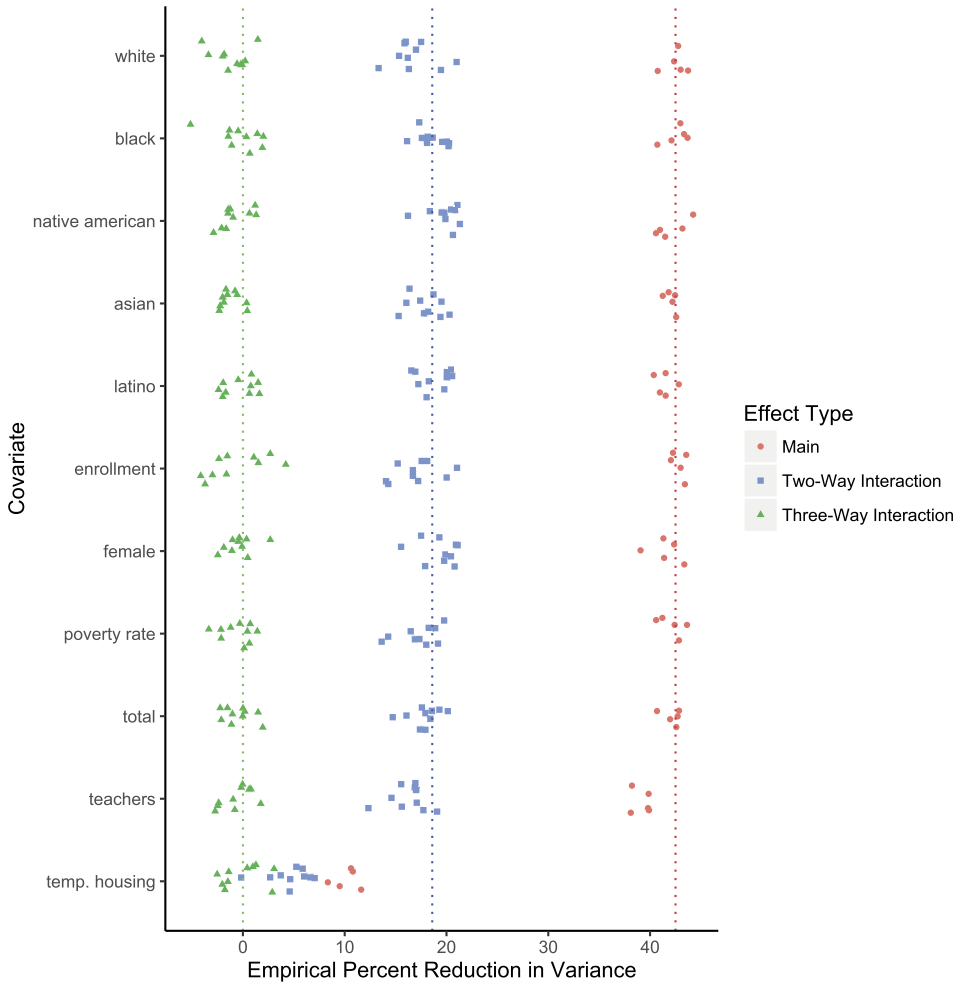

FIG. 1. *Percent reduction in variance in the covariate mean difference after rerandomization for various covariates and factorial effects. The expected percent reduction in variance given Theorem 3 for each type of factorial effect is marked by a vertical line. Displayed are the nine covariates considered during rerandomization as well as "number of teachers" and "number of students in temporary housing," which were not considered. Random vertical jitter is added to see the percent reduction corresponding to all factorial effects.*

rerandomization. Likewise, "number of students in temporary housing" was only mildly correlated with the covariates considered during rerandomization, and thus it did not benefit greatly from rerandomization. If the NYDE decided that these two covariates were important to balance, but less so than the nine covariates already considered, we could rerandomize efficiently by balancing only the functions of "number of teachers" and "number of students in temporary housing" that

are orthogonal to the nine "most important" covariates because the parts that are correlated will already be balanced [Morgan and Rubin (2015)].

If outcome variables of the NYDE experiment are correlated with these covariates, then a hypothetically designed experiment using rerandomization will yield more precise estimates of the main factorial effects and two-way interactions. Furthermore, the precision of higher-order factorial effects will not be worse compared to pure randomization.

7. Conclusion. The NYDE experiment is one example of a factorial design that incorporates pretreatment covariates, where randomization may have yielded unnecessary imbalance across some covariates and/or factorial effects. Here, we proposed a rerandomization methodology for the NYDE experiment and balanced 2^K factorial designs in general. Theoretical results under common assumptions show that rerandomization yields unbiased estimators and increases the precision of factorial effect estimators of interest without sacrificing the precision of other estimators. We demonstrated how rerandomization would be used to design the NYDE experiment and confirmed that our theoretical results held in practice. The rerandomization algorithm also incorporates tiers of importance for covariates and factorial effects.

Here we assumed that the NYDE can randomize any program combination to any school. However, the NYDE may not be able to fully randomize some programs, such as the school-wide bonus program, due to schools' strong opinions about the intervention and other practical constraints. For example, Dasgupta, Pillai and Rubin (2015) discuss how to analyze a "semi-observational study," where some programs are randomly assigned and other programs are assigned according to schools' preferences. Such constraints could have been incorporated during rerandomization, and the corresponding restricted randomization distribution would be used during analysis. However, further work needs to be done to determine the precision gains of such a rerandomized experiment. Extensions for more complex designs, such as unbalanced designs, fractional factorial designs and split-plot designs, will be future work.

APPENDIX

PROOF OF LEMMA 1. Assume a completely randomized balanced 2^K factorial design is rerandomized using the algorithm proposed at the end of Section 4, and the covariate means are multivariate normal. Under both randomization and rerandomization, the columns of $\widetilde{\mathbf{W}}$ defined in (3.4) are orthogonal. Because the factorial design is balanced and the criterion function ϕ defined in (4.3) is symmetric in $\widetilde{\mathbf{W}}$, $\mathbb{E}[\widetilde{\mathbf{W}}_{\cdot f}|\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1] = \mathbf{0}$ for all $f \in F$. Therefore, for any $f_1, f_2 \in F$, $\operatorname{cov}(\widetilde{\mathbf{W}}_{f_1}, \widetilde{\mathbf{W}}_{f_2}|\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1) = \mathbf{0}$.

Therefore, $\operatorname{Cov}(\widetilde{\mathbf{W}}|\phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1)$ is a block-diagonal matrix. Because $\bar{\mathbf{x}}_{f+} - \bar{\mathbf{x}}_{f-} = \frac{\mathbf{x}^T \widetilde{\mathbf{W}}_{\cdot f}}{n/2}$, the covariance matrix of the elements of $\{\bar{\mathbf{x}}_{f+} - \bar{\mathbf{x}}_{f-} : f \in F\}$ is

block-diagonal under rerandomization. By assumption, the covariate means are multivariate normal, and thus this block-diagonal covariance matrix implies the elements of $\{\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-} : f \in F\}$ are mutually independent under rerandomization. Additionally, the elements of $\{M_f : f \in F\}$ are mutually independent because every M_f is a function of $\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-}$. Similarly, the elements of $\{\phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) : f \in F\}$ are mutually independent, where $\phi_f(\mathbf{x}, \widetilde{\mathbf{W}})$ is defined in (4.2). □

PROOF OF THEOREM 3. Assume a completely randomized balanced 2^K factorial design is rerandomized using the algorithm proposed at the end of Section 4, and the covariate means are multivariate normal. The elements of $\{\phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) : f \in F\}$ are mutually independent given Lemma 1. Therefore,

$$\begin{aligned}\mathbb{E}\big[\overline{\mathbf{x}}_{f^+} - \overline{\mathbf{x}}_{f^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] &= \mathbb{E}\big[\overline{\mathbf{x}}_{f^+} - \overline{\mathbf{x}}_{f^-} | \phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] \\ &= \mathbb{E}[\overline{\mathbf{x}}_{f^+} - \overline{\mathbf{x}}_{f^-} | M_f \le a],\end{aligned}$$

where $\phi(\mathbf{x}, \widetilde{\mathbf{W}})$ is defined in (4.3). Similarly, for $f_1 = f_2$,

$$\begin{aligned}&\operatorname{cov}\big[\overline{\mathbf{x}}_{f_1^+} - \overline{\mathbf{x}}_{f_1^-}, \overline{\mathbf{x}}_{f_2^+} - \overline{\mathbf{x}}_{f_2^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] \\ &\quad = \operatorname{cov}\big[\overline{\mathbf{x}}_{f_1^+} - \overline{\mathbf{x}}_{f_1^-}, \overline{\mathbf{x}}_{f_2^+} - \overline{\mathbf{x}}_{f_2^-} | \phi_f(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] \\ &\quad = \operatorname{cov}[\overline{\mathbf{x}}_{f_1^+} - \overline{\mathbf{x}}_{f_1^-}, \overline{\mathbf{x}}_{f_2^+} - \overline{\mathbf{x}}_{f_2^-} | M_f \le a],\end{aligned}$$

while, for $f_1 \neq f_2$,

$$\operatorname{cov}\big[\overline{\mathbf{x}}_{f_1^+} - \overline{\mathbf{x}}_{f_1^-}, \overline{\mathbf{x}}_{f_2^+} - \overline{\mathbf{x}}_{f_2^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] = \mathbf{0}$$

because the elements of $\{\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-} : f \in F\}$ are mutually independent. The remainder of the proof is identical to the treatment-versus-control case, where the units assigned to f^+ are the "treatment" and the units assigned to f^- are the "control." Thus, analogous to Morgan and Rubin (2012), for $f_1 = f_2$,

$$\operatorname{cov}[\overline{\mathbf{x}}_{f_1^+} - \overline{\mathbf{x}}_{f_1^-}, \overline{\mathbf{x}}_{f_2^+} - \overline{\mathbf{x}}_{f_2^-} | M_f \le a] = v_{a_f} \operatorname{cov}[\overline{\mathbf{x}}_{f_1^+} - \overline{\mathbf{x}}_{f_1^-}, \overline{\mathbf{x}}_{f_2^+} - \overline{\mathbf{x}}_{f_2^-}],$$

where v_{a_f} is defined as in (5.2). □

PROOF OF THEOREM 4. Assume (a) a completely randomized balanced 2^K factorial design is rerandomized using the algorithm proposed at the end of Section 4, (b) the covariate means are multivariate normal, (c) factorial effects are constant across units, and (d) there is no interaction between factorial effects and covariate effects. Because the factorial effects are constant across units, each factorial effect estimator $\hat{\theta}_f$ can be written as (5.4). By Lemma 1, for $f_1 \neq f_2$, $\operatorname{cov}(\overline{\mathbf{x}}_{f_1^+} - \overline{\mathbf{x}}_{f_1^-}, \overline{\mathbf{x}}_{f_2^+} - \overline{\mathbf{x}}_{f_2^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1) = \mathbf{0}$. Furthermore, the difference of the covariate means is orthogonal to the difference of the residual means, and therefore the covariance between them is zero. Therefore,

$$\operatorname{cov}\big(\hat{\theta}_{f_1}, \hat{\theta}_{f_2} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big) = \operatorname{cov}\big[(\bar{\varepsilon}_{f_1^+} - \bar{\varepsilon}_{f_1^-}), (\bar{\varepsilon}_{f_2^+} - \bar{\varepsilon}_{f_2^-}) | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big] = 0.$$

The final equality holds because, by the balance of the design, under both randomization and rerandomization,

$$\operatorname{cov}(\bar{\varepsilon}_{f_1^+}, \bar{\varepsilon}_{f_2^+}) = \operatorname{cov}(\bar{\varepsilon}_{f_1^+}, \bar{\varepsilon}_{f_2^-}) = \operatorname{cov}(\bar{\varepsilon}_{f_1^-}, \bar{\varepsilon}_{f_2^+}) = \operatorname{cov}(\bar{\varepsilon}_{f_1^-}, \bar{\varepsilon}_{f_2^-}),$$

and thus the covariance between any two factorial effect estimators under rerandomization is zero. Furthermore, for all $f \in F$,

$$\begin{aligned}
&\operatorname{var}\big(\hat{\theta}_f | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big) \\
&\quad = \boldsymbol{\beta}^T \operatorname{cov}\big(\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big)\boldsymbol{\beta} + \operatorname{var}\big(\bar{\boldsymbol{\varepsilon}}_{f^+} - \bar{\boldsymbol{\varepsilon}}_{f^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big) \\
&\quad = v_{a_f} \boldsymbol{\beta}^T \operatorname{cov}(\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-})\boldsymbol{\beta} + \operatorname{var}\big(\bar{\boldsymbol{\varepsilon}}_{f^+} - \bar{\boldsymbol{\varepsilon}}_{f^-} | \phi(\mathbf{x}, \widetilde{\mathbf{W}}) = 1\big) \\
&\quad = v_{a_f} \boldsymbol{\beta}^T \operatorname{cov}(\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-})\boldsymbol{\beta} + \operatorname{var}(\bar{\boldsymbol{\varepsilon}}_{f^+} - \bar{\boldsymbol{\varepsilon}}_{f^-}).
\end{aligned}$$

The second equality is a result of Theorem 3. By assumption, n is large enough that $\bar{\mathbf{x}}_{f^+} - \bar{\mathbf{x}}_{f^-}$ and $\bar{\boldsymbol{\varepsilon}}_{f^+} - \bar{\boldsymbol{\varepsilon}}_{f^-}$ are normally distributed, and thus orthogonality implies independence. Thus, rerandomization does not affect the variance of $\bar{\boldsymbol{\varepsilon}}_{f^+} - \bar{\boldsymbol{\varepsilon}}_{f^-}$, and the final equality holds. The remainder of the proof is analogous to Morgan and Rubin (2012) because it is identical to the treatment-versus-control case, as in the proof of Theorem 3. □

Acknowledgments. We would like to thank the Editor and referees for their insightful comments, which improved the quality of this paper.

SUPPLEMENTARY MATERIAL

Dataset and R Code for "Improving covariate balance in 2^K factorial designs via rerandomization with an application to a New York City Department of Education High School Study." (DOI: 10.1214/16-AOAS959SUPP; .zip). We provide the NYDE dataset discussed in the paper, as well as the R code used to implement the rerandomization algorithm for this dataset.

REFERENCES

AHLUWALIA, J. S., OKUYEMI, K., NOLLEN, N., CHOI, W. S., KAUR, H., PULVERS, K. and MAYO, M. S. (2006). The effects of nicotine gum and counseling among African American light smokers: A 2×2 factorial design. *Addiction* **101** 883–891.

APFEL, C. C., KRANKE, P., KATZ, M. H., GOEPFERT, C., PAPENFUSS, S., RAUCH, S., HEINECK, R., GREIM, C. A. and ROEWER, R. (2002). Volatile anaesthetics may be the main cause of early but not delayed postoperative vomiting: A randomized controlled trial of factorial design. *Br. J. Anaesth.* **88** 659–668.

BAYS, H. E., OSE, L., FRASER, N., TRIBBLE, D. L., QUINTO, K., REYES, R., JOHNSON-LEVONAS, A. O., SAPRE, A., DONAHUE, S. R. and EZETIMIBE STUDY GROUP (2004). A multicenter, randomized, double-blind, placebo-controlled, factorial design study to evaluate the lipid-altering efficacy and safety profile of the ezetimibe/simvastatin tablet compared with ezetimibe and simvastatin monotherapy in patients with primary hypercholesterolemia. *Clin. Ther.* **26** 1758–1773.

BOX, G. E. P., HUNTER, J. S. and HUNTER, W. G. (2005). *Statistics for Experimenters*: *Design*, *Innovation*, *and Discovery*, 2nd ed. Wiley, Hoboken, NJ. MR2140250

BRANSON, Z., DASGUPTA, T. and RUBIN, D. B. (2016). Supplement to "Improving covariate balance in 2^K factorial designs via rerandomization with an application to a New York City Department of Education High School Study." DOI:10.1214/16-AOAS959SUPP.

BRUHN, M. and MCKENZIE, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *Am. Econ. J. Appl. Econ.* **1** 200–232.

COX, D. R. (2009). Randomization in the design of experiments. *Int. Stat. Rev.* **77** 415–429.

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753. MR3382595

ESPINOSA, V., DASGUPTA, T. and RUBIN, D. B. (2016). A Bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. *Technometrics* **58** 62–73. MR3463157

FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1942). *The Design of Experiments*, 3rd ed. ed. Hafner-Publishing, New York.

GU, X. S. and ROSENBAUM, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *J. Comput. Graph. Statist.* **2** 405–420.

HU, Y. and HU, F. (2012). Asymptotic properties of covariate-adaptive randomization. *Ann. Statist.* **40** 1794–1815. MR3015044

KASARI, C., ROTHERAM-FULLER, E., LOCKE, J. and GULSRUD, A. (2012). Making the connection: Randomized controlled trial of social skills at school for children with autism spectrum disorders. *J. Child Psychol. Psychiatry* **53** 431–439.

KOLLAR, I., FISCHER, F. and SLOTTA, J. D. (2005). Internal and external collaboration scripts in web-based science learning at schools. In *Proceedings of the* 2005 *Conference on Computer Support for Collaborative Learning*: *Learning* 2005: *The Next* 10 *Years! CSCL* '05, *Taipei*, *Taiwan*, *May* 30–*June* 4, 2005. 331–340. International Society of the Learning Sciences.

KRAUSE, M. S. and HOWARD, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology* **59** 751–766.

LINDLEY, D. (1982). The role of randomization in inference. *PSA*: *Proceedings of the Biennial Meeting of the Philosophy of Science Association* **2** 431–446.

MAHALANOBIS, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences* (*Calcutta*) **2** 49–55.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR0560319

MORGAN, K. L. and RUBIN, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Ann. Statist.* **40** 1263–1282. MR2985950

MORGAN, K. L. and RUBIN, D. B. (2015). Rerandomization to balance tiers of covariates. *J. Amer. Statist. Assoc.* **110** 1412–1421. MR3449036

MORRIS, C. (1979). A finite selection model for experimental design of the Health Insurance study. *J. Econometrics* **11** 43–61.

PAPINEAU, D. (1994). The virtues of randomization. *British J. Philos. Sci.* **45** 437–450, 712–715. MR1292321

RAVAUD, P., GIRAUDEAU, B., LOGEART, I., LARGUIER, J. S., ROLLAND, D., TREVES, R., EULLER-ZIEGLER, L., BANNWARTH, B. and DOUGADOS, M. (2004). Management of osteoarthritis (OA) with an unsupervised home based exercise programme and/or patient administered assessment tools. A cluster randomised controlled trial with a 2×2 factorial design. *Ann. Rheum. Dis.* **63** 703–708.

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

ROSENBERGER, W. F. and SVERDLOV, O. (2008). Handling covariates in the design of clinical trials. *Statist. Sci.* **23** 404–419. MR2483911

RUBIN, D. B. (1976). Multivariate matching methods that are equal percent bias reducing. I. Some examples. *Biometrics* **32** 109–120. MR0400555

RUBIN, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments [MR2655714]. *J. Amer. Statist. Assoc.* **103** 1350–1353. MR2655717

RUBIN, D. B. and THOMAS, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *J. Amer. Statist. Assoc.* **95** 573–585.

SEIDENFELD, T. (1982). *Levi on the Dogma of Randomization in Experiments* (H. E. Kyburg, JR. and I. Levi, eds.) 263–291. Springer, Berlin.

WORRALL, J. (2010). Evidence: Philosophy of science meets medicine. *J. Eval. Clin. Pract.* **16** 356–362.

WU, C. F. J. and HAMADA, M. S. (2009). *Experiments*: *Planning*, *Analysis*, *and Optimization*, 2nd ed. Wiley, Hoboken, NJ. MR2583259

XU, Z. and KALBFLEISCH, J. D. (2013). Repeated randomization and matching in multi-arm trials. *Biometrics* **69** 949–959. MR3146790

YATES, F. (1937). The design and analysis of factorial experiments. Imperial Bureau of Soil Sciences—Technical Communication. No. 35, Harpenden.

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
1 OXFORD STREET
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: zbranson@g.harvard.edu
dasgupta@stat.harvard.edu
rubin@stat.harvard.edu