To be read before the Royal Statistical Society at a meeting on October 22th 2025, from 15.00, at Hallam Conference Centre, with Rute Vieira in the Chair

Balanced and Robust Randomized Treatment Assignments: The Finite Selection Model for the Health Insurance Experiment and Beyond*

Ambarish Chattopadhyay[†], Carl N. Morris[‡], José R. Zubizarreta[§]

Abstract

The Finite Selection Model (FSM) was developed by Carl Morris in the 1970s for the design of the RAND Health Insurance Experiment (HIE) (Morris 1979, Newhouse et al. 1993), one of the largest and most comprehensive social science experiments conducted in the U.S. The idea behind the FSM is that each treatment group takes turns selecting units in a fair and random order to optimize a common assignment criterion. At each of its turns, a treatment group selects the available unit that maximally improves the combined quality of its resulting group of units in terms of the criterion. In the HIE and beyond, we revisit, formalize, and extend the FSM as a general tool for experimental design.

Leveraging the idea of D-optimality, we propose and analyze a new selection criterion in the FSM. The FSM using the D-optimal selection function has no tuning parameters for covariate balance, is affine invariant, and when appropriate, retrieves several classical designs such as randomized block and matched-pair designs. For multi-arm experiments, we propose algorithms to generate a fair and random selection order of treatments. We demonstrate FSM's performance in a case study based on the HIE and in ten randomized studies from the health and social sciences. On average, the FSM achieves 68% better covariate balance than complete randomization and 56% better covariate balance than rerandomization in a typical study. We recommend the FSM be considered in experimental design for its conceptual simplicity, efficiency, and robustness.

Keywords: Causal inference; Covariate balance; Experimental design; Multi-valued treatments

*We thank John Golden, Angela Lee, and Bijan Niknam for helpful research assistance and comments. We also thank participants at Euro-CIM 2023 for their valuable comments. This work was supported through a grant from the Alfred P. Sloan Foundation (G-2020-13946).

[†]Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata, West Bengal, 700108; email: ambarish@isical.ac.in.

[‡]Department of Statistics, Harvard University, 1 Oxford Street Cambridge, MA 02138; email: carl.morris@comcast.net.

[§]Departments of Health Care Policy, Biostatistics, and Statistics, Harvard University, 180 Longwood Avenue, Office 307-D, Boston, MA 02115; email: zubizarreta@hcp.med.harvard.edu.

1 Introduction

1.1 The RAND Health Insurance Experiment

In the 1970's, the challenge of financing and delivering high-quality and affordable health care to all Americans was at the center of national policy debate. At the time, two central questions were "How much more medical care would people use if it is provided free of charge?" and "What are the consequences of using more medical care on their health?" To address these and other related questions, an interdisciplinary team of researchers led by Joseph P. Newhouse at RAND designed and conducted the Health Insurance Experiment (HIE), a large-scale, multi-year, randomized public policy experiment developed and completed between 1971 and 1982. To this day, the HIE is one of the largest and most comprehensive social science experiments ever conducted in the U.S. Even now, four decades after its completion, evidence from the HIE is still fundamental to the national discussion on health care cost sharing and health care reform.

In the HIE, a representative sample of 2,750 families comprising more than 7,700 individuals was chosen from six urban and rural sites across the United States. At the beginning of the study, participants completed a baseline survey providing numerous demographic, medical, and socioeconomic measurements. Families were then assigned to health insurance plans that varied substantially in their coinsurance rates and out-of-pocket expenditure maxima, for a total of 13 possible treatment groups. The goal of the study was to estimate the marginal averages of health and utilization in each of the six sites under each plan.

To provide the strongest possible evidence on health utilization and outcomes, the study had to be randomized. However, achieving balance for numerous continuous and categorical baseline covariates through randomization is challenging in experiments with so many treatment groups and different implementation sites. In the HIE the groups had to be balanced and representative of the sites. In the health and social sciences, there is an ever-increasing

need for methods for random assignment of units into multiple treatment groups that are balanced, efficient, and robust.

1.2 Toward balanced, efficient, and robust experimental designs

Randomized experiments are considered to be the gold standard for causal inference, as randomization provides an unequivocal basis for inference and control. In randomized experiments, the act of randomization ensures balance on both observed and unobserved covariates *on average*. However, a given realization of the random assignment mechanism may produce substantial imbalances on one or more covariates. This imbalance problem can be exacerbated in settings like the HIE, where treatments are multi-valued and many baseline covariates exist, leading to loss in efficiency of the effect estimates.

A variety of methods have been proposed in the literature to address this problem, such as blocking (Fisher 1925, Fisher 1935, Cochran and Cox 1957), optimal pair-matching (Greevy et al. 2004), greedy pair-switching (Krieger et al. 2019), and designs using mixed-integer programming (Bertsimas et al. 2015). In particular, rerandomization (Morgan and Rubin 2012) has gained popularity over the last few years and has become commonplace in experiments. However, rerandomization may not protect against and be robust to chance imbalances in functions of the covariates that are not explicitly addressed by the rerandomization criterion (Banerjee et al. 2017), especially in experiments with multi-valued or multiple (>2) treatments.[1] Moreover, defining the rerandomization criterion requires the selection of a tuning parameter governing the acceptable degree of imbalance, which may be difficult to choose and require iteration in practice.

To address these and other related challenges, we revisit and extend the Finite Selection Model (FSM) for experimental design. The original version of the FSM was proposed and developed by Carl N. Morris in the design of the HIE (Morris 1979, Newhouse et al. 1993,

[1] Throughout the paper, we use the terms "multi-valued treatment" and "multiple treatment groups" to refer to the same experimental situation.

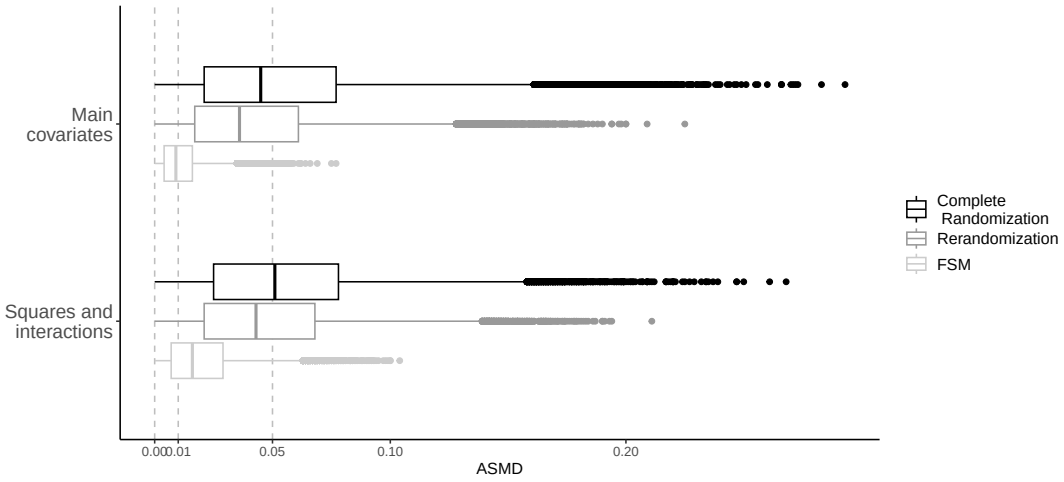
Morris and Hill 2000). The core concept of the FSM is to address the tension between randomization and optimization in treatment assignment (see, e.g., Harshaw et al. 2024). To achieve this, in the FSM each treatment group takes turns in a fair and random order to select units from a pool of available units such that, at each stage, each treatment group selects the unit that maximally improves the combined quality of its current group of units. The random order for selecting units introduces randomness into the assignment mechanism, thereby balancing unobserved covariates in expectation, and facilitating randomization-based inference. The criterion for measuring quality is flexible. Among other contributions, in this paper we develop a new criterion based on D-optimality, which does not require tuning parameters for covariate balance.

To illustrate, Figure 1 exhibits the performance of complete randomization, rerandomization, and the FSM in a version of the HIE data with four treatment groups and 20 covariates. For rerandomization, we compute the maximum Mahalanobis distance (across all pairs of treatment groups) based on the 20 covariates and their squares and pairwise products (i.e., all second-order transformations), and following Lock (2011), accept 0.1% of the assignments with the smallest covariate distance (see Sections 6.1 and 6.4 for details). The figure displays the distribution of absolute standardized mean differences (ASMD; Rosenbaum and Rubin 1985)[2] in covariates and the second-order transformations across multiple realizations of the randomization mechanisms for the three designs. Lower values of ASMD indicate better balance on the covariates or their transformations. Better balance can improve the validity and credibility of a study, and can also translate into increased efficiency and robustness.

We observe that, as expected, rerandomization outperforms complete randomization in terms of imbalances on the main covariates and the second-order transformations. The FSM, however, markedly outperforms both methods for both types of covariates without requiring

[2]The absolute standardized mean difference for a single covariate X between treatment groups g and g' is $\text{ASMD}(X) = |\bar{X}_g - \bar{X}_{g'}|/\sqrt{(s_g^2 + s_{g'}^2)/2}$, where $\bar{X}_g$ and s_g^2 are the mean and variance of X in treatment group g, respectively. Please see Rosenbaum and Rubin (1985)) for details.

Figure 1: Distributions of ASMD for complete randomization, rerandomization, and the FSM, for 20 baseline covariates in the HIE data. Without tuning parameters for balance, the FSM handles multiple (>2) treatment groups and substantially improves covariate balance and, thereby, statistical efficiency.



tuning parameters for covariate balance. This analysis reveals that, while rerandomization performs well by common covariate balance standards (the majority of the ASMD is smaller than 0.1), there is room for improvement. As we explain in Section 6, in experiments like the HIE, the space of possible assignments is vast, and the FSM can meaningfully improve the assignment of units into treatment groups to achieve better balance and efficiency.

In a nutshell, the FSM does better because it progressively randomizes units into treatment groups in a controlled manner towards a criterion that is common to all groups and robust against general outcome models. As we show in theory and in practice, the FSM is a flexible tool for random assignment in various settings.

1.3 Contribution and outline

In this paper, we revisit, formalize, and extend the FSM for experimental design. In particular, we reexamine the FSM under the potential outcomes framework (Neyman 1923, 1990, Rubin 1974). While the FSM was proposed in the context of the HIE several decades ago, its properties and performance for experimental design have not yet been thoroughly investigated. In this paper, from both theoretical and practical standpoints, we demonstrate

that the FSM can be used for balanced, efficient, and robust random treatment assignments, outperforming common assignment methods on these three dimensions.

From a methodological standpoint, we offer several extensions of the FSM. First, we propose a new selection criterion for treatments based on the idea of D-optimality and discuss its theoretical properties. We show that the FSM using this selection criterion is invariant with respect to affine transformations of the covariates. Under suitable conditions, it also retrieves several classical experimental designs, such as randomized block and matched-pair designs. Second, for experiments with multiple treatment groups, we propose and justify new algorithms to determine the selection order of treatments, building on the sequentially controlled Markovian random sampling (SCOMARS, Morris 1983) algorithm for experiments with two groups. Third, we discuss extensions of the FSM to more complex experimental design settings, such as stratified experiments and experiments with sequential arrival of units. Fourth, we discuss both model- and randomization-based inference under the FSM. While model-based inference was previously conducted in the HIE, randomization-based inference has not yet been explored. Regarding the latter, we show that the FSM can be used to conduct both Fisherian inference for unit-level treatment effects (Fisher 1935) and Neymanian inference for average treatment effects (Neyman 1923, 1990). Finally, we analyze the FSM's performance empirically and compare it to common assignment methods. In an accompanying paper (Chattopadhyay et al. 2021), we describe how these methods can be implemented in the new `FSM` package for `R`, which is publicly available on CRAN.

The paper proceeds as follows. In Section 2, we describe the design of the RAND Health Insurance Experiment, focusing on the assignment of each family to one of 13 health insurance plans. In Section 3, we present the setup, notation, and main components of the FSM. In Section 4, we propose a selection criterion based on D-optimality and analyze its properties. In Section 5, we discuss inference under the FSM. In Section 6, we evaluate the performance of the FSM and compare it to standard methods such as complete randomiza-

tion and rerandomization using the HIE data. In Section 7, we perform a similar comparison using the data from ten experimental studies from the health and social sciences. Finally, in Section 8 we consider extensions of the FSM to other settings such as multi-group, stratified, and sequential experiments. In Section 9, we conclude with a summary and remarks. In the Online Supplementary Materials, we present all the proofs of the propositions and theorems, extended theoretical results, further empirical results based on a simulation study, and supplemental experimenal results on the HIE study and the ten case studies.

2 Design of the Health Insurance Experiment

In the HIE, families were assigned to different health insurance plans using the original version of the FSM. Initially, assignments were made in each of the six HIE sites to 12 or 13 fee-for-service plans with varying combinations of coinsurance (cost sharing) rates and income-related deductibles. Coinsurance plans consisted of 0% (free care), 25%, 50%, or 95% coinsurance rates, plus a plan with mixed coinsurance rates, and an individual deductible plan. Within the cost sharing plans, families were further assigned to different out-of-pocket maxima where the out-of-pocket expenditures were capped at 5%, 10%, or 15% of family income, with an annual maximum of $1,000 (Brook et al. 2006). To ensure that the resulting treatment groups were balanced relative to the population of each site, the FSM considered a discard group of study non-participants as an additional treatment group.

Listed in chronological order of study initiation, the following sites were tracked for several years: Dayton, OH; Seattle, WA; Fitchburg, MA; Franklin County, MA; Charleston, SC; and Georgetown County, SC. The FSM was used, independently in each of the sites, to make random assignments to improve balance on up to 22 family-level baseline covariates across treatment groups. In each of the first two sites, the FSM was used multiple times for separate independent subsets of families to maintain baseline data schedules. In addition to estimating the overall marginal effects of health insurance plan design on healthcare utilization and outcomes, the HIE team also sought to understand how the experimental results were affected

by particular design choices, e.g., longer versus shorter enrollment duration, receiving versus not receiving participation incentives, higher versus lower interviewing frequency. To this end, four additional sub-experiments were conducted, and the FSM was used to randomize families to the sub-treatment groups.

3 Foundations and overview of the FSM

3.1 Setup and notation

Consider a sample of N units indexed by $i = 1, ..., N$. Each of these units is to be assigned into one of G treatment groups labeled by g, with $g = 1, ..., G$. Write n_g for the pre-specified size of group g. Denote $Z_i \in \{1, 2, ..., G\}$ as the assigned treatment group label of unit i and $\boldsymbol{Z} = (Z_1, ..., Z_N)^\top$ as the vector of treatment group labels. Following the potential outcomes framework for causal inference (Neyman 1923, 1990; Rubin 1974), each unit i has a potential outcome under each treatment g, $Y_i(g)$, but only one of these outcomes is observed: $Y_i^{\text{obs}} = \sum_{g=1}^{G} \mathbb{1}(Z_i = g) Y_i(g)$. Denote $\boldsymbol{Y}(g) = (Y_1(g), ..., Y_N(g))^\top$ as the vector of potential outcomes under treatment g. Each unit has a vector of K observed covariates, $\boldsymbol{X}_i$. We write $(\underline{\boldsymbol{X}}_{\text{full}})_{N\times k}$ for the matrix of observed covariates, and $\bar{\boldsymbol{X}}_{\text{full}}$ and $\underline{\boldsymbol{S}}_{\text{full}}$ for the mean vector and covariance matrix of these covariates in the full sample, respectively. Denote $(\underline{\tilde{\boldsymbol{X}}}_{\text{full}})_{N\times(k+1)}$ as the design matrix in the full sample.[3] We assume that $\underline{\tilde{\boldsymbol{X}}}_{\text{full}}$ has full column rank. In Table A1 of the Online Supplementary Materials we provide a list of the notation used in this paper.

Based on this notation, $Y_i(g') - Y_i(g'')$ is the causal effect of treatment g' relative to treatment g'' for unit i. We are interested in estimating the sample average treatment effect $\text{SATE}_{g',g''} = \frac{1}{N}\sum_{i=1}^{N}\{Y_i(g') - Y_i(g'')\}$ and the population average treatment effect $\text{PATE}_{g',g''} = \mathbb{E}\{Y_i(g') - Y_i(g'')\}$. For this, we will randomly assign the units into treatment groups using the FSM.

[3]The design matrix includes a column of all 1's (for the intercept) and k columns of covariates.

3.2 Components of the FSM

In the FSM, the G treatment groups take turns selecting units in a random but controlled order while optimizing a common criterion. This is accomplished by the two components of the FSM, namely, the *selection order matrix* and the *selection function*.

1. Selection order matrix (SOM): An SOM is a matrix that determines the order in which the treatment groups select the units. Typically, an SOM has two columns; the first specifies the stages of selection (from 1 to N), and the second specifies the treatment group that selects first at that stage.

2. Selection function: A selection function is a function that determines which unit gets selected by the choosing treatment group at each stage. Typically, a selection function is based on an optimality criterion that is common to all treatment groups.

A good SOM guarantees that the selection of units is fair, so that no single treatment group selects all the units of a given type, and random, so that both observed and unobserved covariates are balanced in expectation and there is a basis for inference. A good selection function will produce efficient and robust inferences under a wide class of possible outcome functions.

To illustrate, Table 1(a) presents an example data set with 12 observations and one covariate, age. We consider assigning these 12 units into two groups of equal sizes using the FSM. Table 1(b) shows an example of an SOM in this setting. The SOM determines the order in which each treatment selects a unit at each stage. In the example, treatment group 2 selects first in stage 1, treatment group 1 selects in stage 2, and so on. Treatment groups select units based on the selection function.

Table 1: (a) Example data set; (b) selection order matrix and an assignment using the FSM.

(a) Data set

Index	Age
1	24
2	30
3	34
4	36
5	40
6	41
7	45
8	46
9	50
10	54
11	56
12	60
Mean	43

(b) Selection order matrix and assignment

Selection order matrix		Unit selected	
Stage	Treatment	Index	Age
1	2	1	24
2	1	12	60
3	1	2	30
4	2	11	56
5	1	3	34
6	2	10	54
7	1	9	50
8	2	4	36
9	1	5	40
10	2	8	46
11	2	6	41
12	1	7	45

In general, it is crucial that the order of selection is random, but that no group chooses in a disproportionate manner. For two treatment groups of arbitrary sizes, this can be accomplished by means of the Sequentially Controlled Markovian Random Sampling (SCOMARS) algorithm (Morris 1983). In the FSM, SCOMARS specifies the probability of a treatment group selecting at stage r ($r \in \{1, 2, ..., N\}$), conditional on the number of selections made by that group up to stage $r-1$. See the Online Supplementary Materials for a formal description of the algorithm. SCOMARS satisfies the sequentially controlled condition (Morris 1983), which requires the deviation of the observed number of selections made by a treatment group up to stage r from its expectation to be strictly less than one. Intuitively, this condition ensures that throughout the selection process, no treatment group departs too much from its expected fair share of choices. Moreover, SCOMARS is Markovian because for each group, the probability of selection at stage r depends solely on the number of selections made up to stage $r-1$. For two groups of equal sizes (as in the example in Table 1), generating an SOM under SCOMARS boils down to successively generating $N/2$ independent random permutations of the treatment labels $(1, 2)$. In Section 8.1 and in the Online Supplementary Materials, we describe this and other extensions of SCOMARS to multi-group experiments.

Unless otherwise specified, in the rest of the paper, we will use SCOMARS to generate the SOM for experiments with two treatment groups.

The selection function gives a value to each of the units available for selection at each stage. This value depends on the characteristics of each available unit in addition to those already assigned to the treatment group that selects next. In principle, any criterion can be used in the selection function. For example, if the selection function is constant, then the treatment group selects a unit randomly from the available pool. Alternatively, the selection function can compute the contribution of each unit to a measure of the accuracy of the estimator. In this spirit, we propose the *D-optimal* selection function, which, at each stage, minimizes the generalized variance of the estimated regression coefficients in a linear potential outcome model (see Section 4 for details).

To build intuition, in Table 1(b) we discuss the special case of $k = 1$ covariate. With the D-optimal selection function, the choosing group, in its first choice, selects the unit whose covariate value is farthest from the full-sample mean of the covariate; and in the subsequent choices, selects the unit whose covariate value is farthest from its current mean of the covariate. In the example in Table 1, treatment 2 selects unit 1 with age 24, the farthest age from the full-sample mean 43. In the next stage, treatment 1 selects unit 12 with age 60, the farthest age from 43.[4] Next, treatment 1 selects unit 2 with age 30, the farthest age from its current mean age 60. The process continues until all the 12 units are selected.

In general, with multivariate data, the FSM first selects the units that are farthest from the full-sample mean of the covariates and successively approaches this target, ultimately selecting the units that are closest to it. In the FSM, the SOM produces balance out of an optimality criterion that is common to all the treatment groups. This is crucial so that all the choosers know the same, and as they choose, they produce groups that are balanced and

[4]Notice that for treatment 1's first selection, the mean of age remains 43 (i.e., the full-sample mean of age) and is not recalculated based on the 11 unselected units.

equally robust against the unknown outcome model.

Another important feature of the FSM is that, in addition to several treatment groups, it can accommodate a discard group of unassigned units. This is important, for example, in settings where the number of available units for assignment is greater than the total number of units that can feasibly be assigned (e.g., because of budgetary constraints). This feature of the FSM was used in the HIE to secure the representativeness of the treatment groups relative to the target populations.

4 The D-optimal selection function

Here, we formalize the D-optimal selection function and provide an equivalent, closed-form characterization that explains how this criterion governs the selection of units at each stage. Without loss of generality, assume that treatment 1 selects at stage r, $r \in \{1, 2, ..., N\}$. Let $\tilde{n}_{r-1}$, $\bar{\boldsymbol{X}}_{r-1}$, $\underline{\boldsymbol{S}}_{r-1}$, and $\underline{\tilde{\boldsymbol{X}}}_{r-1}$ be the number, mean vector, covariance matrix, and the design matrix of the units selected after the $(r-1)$th stage by treatment 1, respectively.

To motivate the form of the selection function, we consider a linear potential outcome model of $Y_i(1)$ on $\boldsymbol{X}_i$, i.e., $Y_i(1) = \boldsymbol{\beta}^\top (1, \boldsymbol{X}_i^\top)^\top + \eta_i$, where η_i is an error term satisfying $\mathbb{E}\{\eta_i | \boldsymbol{X}_i\} = 0$.[5] Denote $\mathcal{R}_{r-1}$ as the set of unselected units after stage $r-1$. For unit $i \in \mathcal{R}_{r-1}$, let $\underline{\tilde{\boldsymbol{X}}}_{r,i}$ be the resulting design matrix in treatment group 1 if unit i is selected. We first consider the case where $\underline{\tilde{\boldsymbol{X}}}_{r-1}$ has full column rank. In this case, following standard notions of D-optimality in experimental design (see, e.g., Cox and Reid 2000, Chapter 7), the D-optimal selection function selects unit $i' \in \mathcal{R}_{r-1}$, where $i' \in \arg\max_{i \in \mathcal{R}_{r-1}} \det(\underline{\tilde{\boldsymbol{X}}}_{r,i}^\top \underline{\tilde{\boldsymbol{X}}}_{r,i})$. In other words, at the rth stage, the D-optimal selection function chooses the unit in $\mathcal{R}_{r-1}$ that optimally decreases the generalized variance of the estimated regression coefficients of the fitted linear model in

[5] More generally, one can consider a linear model of $Y_i(1)$ on a vector of basis functions $\boldsymbol{B}(\boldsymbol{X}_i)$ of the covariates. In principle, the functions $\boldsymbol{B}(\boldsymbol{X}_i)$ can encompass a wide range, from the components of the covariates to their higher order interactions, and extending to more complex functions such as Wavelet or Fourier bases. Furthermore, it could incorporate basis functions corresponding to a specific kernel, similar to those for a reproducing kernel Hilbert space. In common applications, the choice of $\boldsymbol{B}(\boldsymbol{X}_i)$ is often guided by substantive knowledge about the covariate functions that could be associated with the potential outcomes.

treatment 1. Ties in the values of the generalized variances are resolved randomly.

Next, we consider the case where $\underline{\tilde{\boldsymbol{X}}}_{r-1}$ does not have full column rank. In this case, however, $\underline{\tilde{\boldsymbol{X}}}_{r,i}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r,i}$ may be singular, implying that the resulting determinant may be zero and hence, non-informative. To address this issue, we revert to the preceding case and note that the criteria $\arg\max_{i\in\mathcal{R}_{r-1}} \det(\underline{\tilde{\boldsymbol{X}}}_{r,i}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r,i})$ is equivalent to

$$i' \in \arg\max_{i\in\mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}.$$

See Lemma A1 in the Online Supplementary Materials for a proof. Using this representation, we formally define the D-optimal selection function as follows.

Definition 1 (D-optimal selection function)**.** At stage r of the selection process, for unit $i \in \mathcal{R}_{r-1}$, the D-optimal selection function is given by

$$\begin{aligned}
&(1, \boldsymbol{X}_i^{\top})\Big(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^{\top}\underline{\tilde{\boldsymbol{X}}}_{\text{full}}\Big)^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}, \text{ if } \tilde{n}_{r-1} = 0 \\
(1, \boldsymbol{X}_i^{\top})\Big(\frac{1}{\tilde{n}_{r-1}}\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1} + \frac{\epsilon}{N}\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^{\top}\underline{\tilde{\boldsymbol{X}}}_{\text{full}}\Big)^{-1} &\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}, \text{ if } \tilde{n}_{r-1} \geq 1 \text{ and } \operatorname{rank}(\underline{\tilde{\boldsymbol{X}}}_{r-1}) < k+1 \\
&(1, \boldsymbol{X}_i^{\top})\Big(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1}\Big)^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}, \text{ if } \tilde{n}_{r-1} \geq 1 \text{ and } \operatorname{rank}(\underline{\tilde{\boldsymbol{X}}}_{r-1}) = k+1.
\end{aligned}$$

When $\underline{\tilde{\boldsymbol{X}}}_{r-1}$ is not of full rank (i.e., when $\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1}$ is not invertible), we augment $\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1}$ in the above definition with the invertible matrix $\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^{\top}\underline{\tilde{\boldsymbol{X}}}_{\text{full}}$. This process is similar to Ridge augmentation and is controlled by the augmentation parameter ϵ, which is strictly greater than zero. In practice, we opt for a small value of ϵ, such as 0.001, which in turn augments using 0.1% of the average covariate information in the data, as measured by $\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^{\top}\underline{\tilde{\boldsymbol{X}}}_{\text{full}}/N$. The choice of ϵ is generally inconsequential to the overall performance of the FSM, as it primarily influences the initial selection stages when $\underline{\tilde{\boldsymbol{X}}}_{r-1}$ is not of full rank. To further investigate this, we have conducted a simulation study to evaluate the performance of the

FSM across various ϵ values. See Appendix H.5 for details. The results from this study indicate that variations in the value of ϵ do not meaningfully alter the overall performance of the FSM.

The following theorem provides an equivalent characterization of the D-optimal selection function that elucidates the selection made by the choosing treatment group at each stage.

Theorem 4.1. Consider stage r of the selection process. The D-optimal selection function chooses unit i' such that

$$i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}^*_{r-1})^\top (\underline{\boldsymbol{S}}^*_{r-1})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}^*_{r-1}),$$

where

$$\bar{\boldsymbol{X}}^*_{r-1} = \begin{cases} \bar{\boldsymbol{X}}_{\text{full}} & \text{if } \tilde{n}_{r-1} = 0 \\ \frac{\bar{\boldsymbol{X}}_{r-1} + \epsilon \bar{\boldsymbol{X}}_{\text{full}}}{1+\epsilon} & \text{if } \tilde{n}_{r-1} \geq 1 \text{ and } \operatorname{rank}(\underline{\tilde{\boldsymbol{X}}}_{r-1}) < k+1 \\ \bar{\boldsymbol{X}}_{r-1} & \text{if } \tilde{n}_{r-1} \geq 1 \text{ and } \operatorname{rank}(\underline{\tilde{\boldsymbol{X}}}_{r-1}) = k+1 \end{cases}$$

and

$$\underline{\boldsymbol{S}}^*_{r-1} = \begin{cases} \underline{\boldsymbol{S}}_{\text{full}} & \text{if } \tilde{n}_{r-1} = 0 \\ (\frac{1}{\tilde{n}_{r-1}} \underline{\boldsymbol{X}}^\top_{r-1} \underline{\boldsymbol{X}}_{r-1} + \frac{\epsilon}{N} \underline{\boldsymbol{X}}^\top_{\text{full}} \underline{\boldsymbol{X}}_{\text{full}}) - (1+\epsilon) \bar{\boldsymbol{X}}^*_{r-1} \bar{\boldsymbol{X}}^{*\top}_{r-1} & \text{if } \tilde{n}_{r-1} \geq 1 \text{ and } \operatorname{rank}(\underline{\tilde{\boldsymbol{X}}}_{r-1}) < k+1 \\ \underline{\boldsymbol{S}}_{r-1} & \text{if } \tilde{n}_{r-1} \geq 1 \text{ and } \operatorname{rank}(\underline{\tilde{\boldsymbol{X}}}_{r-1}) = k+1. \end{cases}$$

Theorem 4.1 shows that at every stage, the D-optimal selection function selects the unit among the remaining pool of available units whose covariate vector maximizes a type of Mahalanobis distance. In its first choice, treatment 1 maximizes the Mahalanobis distance from the covariate distribution in the full sample (in particular, from $\bar{\boldsymbol{X}}_{\text{full}}$), thereby choosing the most outlying unit available in the full sample. For the subsequent stages where $\underline{\tilde{\boldsymbol{X}}}^\top_{r-1} \underline{\tilde{\boldsymbol{X}}}_{r-1}$ is not invertible, treatment 1 maximizes the Mahalanobis distance from a mixture covariate distribution between treatment group 1 and the full sample, where ϵ determines the mixing rate. Finally, the latter selections by treatment 1 maximize the Mahalanobis distance from the covariate distribution in treatment group 1. Therefore, with every selection, treatment

1 maximizes the overall separation of the covariates from its current mean, which increases the efficiency of the estimated regression coefficients.

By definition, the D-optimal selection function improves the accuracy of the fitted linear model in each treatment group by sequentially minimizing the generalized variance of the estimated regression coefficients, where the sequence is governed by the randomized SOM. Alternatively, one can simultaneously minimize the generalized variance among all possible assignments to optimize, in a global but deterministic manner, the D-optimal criterion. In fact, with a single covariate and two treatment groups, we show that the global D-optimal design aims to exactly balance the mean of the covariate across the two groups (see Proposition A5 in the Online Supplementary Materials). While such direct and global optimization is conceptually appealing, the FSM optimizes towards the D-optimal criterion progressively in order to build a randomization set for inference (see Section 5) and, in expectation, balance other functions of observed and unobserved covariates. As a result, the FSM provides robustness against adversarial outcome models (see Section 6, Appendices D and H; see also Harshaw et al. 2024). To explore further, we compare the FSM with the global D-optimal design using a simulation study in Appendix H.4. The results show that, although the global D-optimal design has better balance on the functions targeted by the D-optimal criterion, it displays inferior balance on other untargeted transformations of the covariates relative to the FSM.

With the D-optimal selection function, we can also establish several additional desirable properties of the FSM. In particular, leveraging the connection between D-optimality and Mahalanobis distance, we can show that FSM with the D-optimal selection function is affine invariant, i.e., the selections of units by the treatment groups remain unchanged even if the covariates are transformed linearly. See Section C in the Online Supplementary Materials for a proof. An implication of this property is that the FSM is invariant with respect to changes in the location and scale of the covariates.

The FSM with the D-optimal selection function is appealing also because it can encompass several classical designs, such as randomized blocked and matched-pair designs. Theorem 4.2 formalizes this result. In the traditional randomized block design (RBD), the units are grouped into blocks of size G according to a categorical, blocking variable, and each treatment is randomly applied to exactly one unit within each block (see, e.g., Cox and Reid 2000, Section 3.4). Here we consider a more general version of an RBD where the blocks are of size $c \times G$ (where c is a fixed positive integer) and each treatment is applied to c units within each block. This is a special case of a stratified randomized experiment with strata of equal size and equal allocation among treatments per stratum. In a matched-pair design with $G = 2$ treatments, similar units are grouped into pairs, and each treatment is randomly applied to one unit within each pair. This is also a special case of a stratified randomized experiment with equal allocation per strata, where the size of each stratum equals two.

Theorem 4.2. (a) Consider $N = cBG$ units belonging to B blocks of equal size that are to be randomly assigned into G treatment groups of equal size, where c is a fixed positive integer. Then, if the linear model in the FSM consists of an intercept and indicators of any $B - 1$ levels of the blocking variable, the FSM with the D-optimal selection function produces the same assignment as an RBD.

(b) Consider $N/2$ identical pairs of units in terms of baseline covariates $\boldsymbol{X}_i$ that are to be assigned into $G = 2$ treatment groups of equal size. Assume $\boldsymbol{X}_i$ is drawn from a continuous distribution. Then, if the linear model in the FSM consists of the intercept and the covariates $\boldsymbol{X}_i$, then the FSM almost surely produces the same assignment mechanism as a matched-pair design.

In the first case, Theorem 4.2(a) states that, by including the levels of a blocking variable as regressors, the FSM with the D-optimal selection function automatically blocks on that variable. Thus, the FSM retrieves an RBD without explicitly performing separate randomizations within each block. In the second case, Theorem 4.2(b) states that, by including the

covariates as regressors, the FSM with the D-optimal selection function produces the same assignment as a matched-pair experiment, without explicitly performing separate randomizations in each pair. This phenomenon is particularly useful when the sample consists of near-identical twins but that are difficult to identify a priori due to multiple covariates.

The properties presented in Theorem 4.2(a) and 4.2(b) are relevant because blocking and pair-matching are two fundamental and well-known techniques used in experimental design (see, e.g., Cox and Reid 2000, Chapter 3, and Imbens and Rubin 2015, Chapter 4). The FSM can also retrieve standard designs with appropriate choices of the SOM and the selection function. For instance, when the selection function is constant, and the SOM is constructed by randomly permuting the vector $(1, ..., 1, ..., G, ..., G)^\top$ where each entry $g = 1, ..., G$ is repeated n_g times, the FSM replicates complete randomization. Similarly, the FSM can replicate various stratified randomized designs with unequal stratum sizes and unequal group sizes within strata (see Section 8.2 and Appendix F in the Online Supplementary Materials). Examining whether the FSM can reproduce a wider spectrum of established experimental designs is an exciting direction for future research.

In summary, the D-optimal selection function provides a simple, interpretable, and effective selection criterion for the FSM. It is simple because it is easy to compute, interpretable because it connects to the well-known Mahalanobis distance, and effective because it requires no tuning parameters for balance while also being affine invariant and capable of recovering desirable classical designs. While other criteria, such as A-optimality, can serve as selection functions under the FSM, they may not offer the same advantages. As discussed in Appendix C.2 in the Online Supplementary Materials, unlike D-optimality, the A-optimality criterion typically entails numerous tuning parameters and may not be affine invariant.

5 Inference under the FSM

Using the FSM we can make model- and randomization-based inferences. Both modes of inference are feasible for any selection function and any randomized selection order matrix.

In model-based inference, the sample is typically assumed to be drawn randomly from some superpopulation, and inference for the PATE is done by modeling the observed outcome distribution conditional on the treatment indicators and the covariates. Thus, model-based inference is generally applicable to any design, both deterministic and randomized.

To formalize, let the potential outcome model under treatment g be $Y_i(g) = \boldsymbol{\beta}_g^\top \boldsymbol{B}(\boldsymbol{X}_i) + \epsilon_{ig}$, where $\boldsymbol{B}(\boldsymbol{X}_i) = (B_1(\boldsymbol{X}_i), ..., B_b(\boldsymbol{X}_i))^\top$ is a vector of b basis functions of the covariates, and ϵ_{ig}, $i \in \{1, 2, ..., N\}$ are mutually independent errors, independent of the covariates. Under this model, $\text{PATE}_{g',g''}$ can be unbiasedly estimated by $\widehat{\text{PATE}}_{g',g''} = \hat{\boldsymbol{\beta}}_{g'}^\top \overline{\boldsymbol{B}(\boldsymbol{X})} - \hat{\boldsymbol{\beta}}_{g''}^\top \overline{\boldsymbol{B}(\boldsymbol{X})}$, where $\overline{\boldsymbol{B}(\boldsymbol{X})} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{B}(\boldsymbol{X}_i)$ and $\hat{\boldsymbol{\beta}}_g$ is the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}_g$ obtained by fitting a linear regression of Y_i^{obs} on $\boldsymbol{B}(\boldsymbol{X}_i)$ in treatment group $g = g', g''$. We call this the regression imputation estimator of $\text{PATE}_{g',g''}$. Under the given model, this estimator is the best linear unbiased estimator for $\text{PATE}_{g',g''}$. The standard error of this estimator and the corresponding confidence interval for $\text{PATE}_{g',g''}$ can be obtained using standard OLS theory. We note that, in model-based inference, the standard errors and confidence intervals do not take into account the randomness stemming from the assignment mechanism.

In randomization-based inference, the potential outcomes and the covariates are typically considered fixed and the assignment mechanism is the only source of randomness (see Chapter 2 of Rosenbaum 2002 and chapters 5–7 of Imbens and Rubin 2015 for overviews). Thus, randomization-based inference is generally applicable to any randomized design. Inference for causal effects can be done via exact randomization tests for sharp null hypotheses on unit-level causal effects (Fisher 1935), or via estimation under Neyman's repeated sampling approach (Neyman 1923, 1990).

Under the FSM, randomness primarily arises from the selection order matrix, which in turn facilitates randomization-based inference. In particular, under the FSM, randomization tests for sharp null hypotheses can be performed by approximating the distribution of the test statistic through repeated realizations of the FSM. To illustrate, consider testing the

sharp null hypothesis of zero unit-level causal effects, i.e., $H_0 : Y_i(2) - Y_i(1) = 0$ for all i, at level α using the FSM. While any choice of test statistic preserves the validity of the test, a common choice is the absolute difference-in-means statistic $|\frac{1}{n_2}\sum_{i:Z_i=2} Y_i^{\text{obs}} - \frac{1}{n_1}\sum_{i:Z_i=1} Y_i^{\text{obs}}| = |\frac{1}{n_2}\sum_{i:Z_i=2} Y_i(2) - \frac{1}{n_1}\sum_{i:Z_i=1} Y_i(1)| =: T\{\boldsymbol{Z}, \boldsymbol{Y}(1), \boldsymbol{Y}(2)\}$. Large values of $T\{\boldsymbol{Z}, \boldsymbol{Y}(1), \boldsymbol{Y}(2)\}$ are considered evidence against H_0. Under H_0, $Y_i(2) = Y_i(1) = Y_i^{\text{obs}}$ and the vectors of potential outcomes $\boldsymbol{Y}(1)$ and $\boldsymbol{Y}(2)$ are known and fixed. The p-value of the test is given by $p = P_{H_0}[T\{\boldsymbol{Z}, \boldsymbol{Y}(1), \boldsymbol{Y}(2)\} \geq t_{\text{obs}}]$, where t_{obs} is the value of the test statistic for the observed realization of $\boldsymbol{Z}$ under the FSM. We can compute this p-value by Monte Carlo approximation, i.e., we generate independent vectors of assignments $\boldsymbol{Z}^{(m)} = (Z_1^{(m)}, ..., Z_N^{(m)})^\top$, $m \in \{1, 2, ..., M\}$ using the FSM and approximate the p-value as $\hat{p} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\big[T\{\boldsymbol{Z}^{(m)}, \boldsymbol{Y}(1), \boldsymbol{Y}(2)\} \geq t_{\text{obs}}\big]$. We reject H_0 at level α if $\hat{p} \leq \alpha$.

The test for the sharp null hypothesis described above is distribution-free, relying solely on repeated randomizations under the FSM. Thus, it is valid without any distributional assumptions on the outcomes. Similar tests can be applied for more general sharp hypotheses of treatment effects (e.g., dilated and tobit effects; Rosenbaum 2002, 2010). We can invert these tests to obtain a confidence interval for the hypothesized effect (Rosenbaum 2002, Section 2.6.1). Moreover, we can get a point estimate of the effect by solving a Hodges-Lehmann estimating equation corresponding to these tests (Rosenbaum 2002, Section 2.7.2).

Finally, using the FSM, we can perform randomization-based inference for the sample average treatment effect, similar to Neyman's repeated sampling approach. To formalize, we consider estimating an arbitrary linear combination of the average potential outcomes $\mu(\boldsymbol{l}) = \sum_{g=1}^{G} l_g\{\frac{1}{N}\sum_{i=1}^{N} Y_i(g)\}$, where $\boldsymbol{l} = (l_1, ..., l_G)^\top \in \mathbb{R}^G$ are constants specified by the investigator. As a special case, when $l_{g'} = 1$, $l_{g''} = -1$, and $l_g = 0$ for $g \neq g', g''$, $\mu(\boldsymbol{l})$ equals the sample average treatment effect $\text{SATE}_{g',g''}$. We estimate $\mu(\boldsymbol{l})$ by

$$\hat{\mu}(\boldsymbol{l}) = \sum_{g=1}^{G} l_g\{\frac{1}{n_g}\sum_{i:Z_i=g}^{N} Y_i^{\text{obs}}\}.$$

Here, the population average of the potential outcomes under treatment g is estimated by the sample average of the observed outcomes in treatment group g. In Proposition 5.1, we show that under the FSM with equal group sizes, $\hat{\mu}(\boldsymbol{l})$ is unbiased for $\mu(\boldsymbol{l})$. Here, the expectation is taken with respect to the randomization distribution of the FSM, and thus, this property does not rely on any distributional assumptions on the outcome.

Proposition 5.1. Let $n_1 = n_2 = ... = n_G$, and consider an arbitrary linear combination of the average potential outcomes $\mu(\boldsymbol{l}) = \sum_{g=1}^{G} l_g\{\frac{1}{N}\sum_{i=1}^{N} Y_i(g)\}$, where $\boldsymbol{l} = (l_1, ..., l_G)^\top \in \mathbb{R}^G$. Then, the estimator $\hat{\mu}(\boldsymbol{l}) = \sum_{g=1}^{G} l_g\{\frac{1}{n_g}\sum_{i:Z_i=g}^{N} Y_i^{\text{obs}}\}$ is unbiased for $\mu(\boldsymbol{l})$ under the FSM, $\mathbb{E}\{\hat{\mu}(\boldsymbol{l})\} = \mu(\boldsymbol{l})$.

Thus, under the FSM, we can unbiasedly estimate all possible linear combinations of the average potential outcomes, including different treatment contrasts. In particular, an immediate consequence of Proposition 5.1 is that for all $g, g' \in \{1, 2, ..., G\}$, $\mathbb{E}(\hat{\tau}_{g',g''}) = \text{SATE}_{g',g''}$, where $\hat{\tau}_{g',g''} = \frac{1}{n_{g'}}\sum_{i:Z_i=g'} Y_i^{\text{obs}} - \frac{1}{n_{g''}}\sum_{i:Z_i=g''} Y_i^{\text{obs}}$. Thus, in this case, the difference-in-means statistics between treatments g' and g'' is unbiased for $\text{SATE}_{g',g''}$. As is usual in most settings, the randomization-based variance of this unbiased estimator is not identifiable in general (see, e.g., Chapter 6 of Imbens and Rubin 2015); however, leveraging advances on design-based inference in randomized experiments (e.g., Mukerjee et al. 2018), we can obtain conservative estimators for this variance. Proposition A1 in the Online Supplementary Materials provides closed-form expressions of the variance and its corresponding estimator.

6 The Health Insurance Experiment

6.1 Data

We evaluate the performance of the FSM relative to other common treatment assignment approaches using the baseline data of the HIE. To this end, we consider a version of the HIE data presented in Aron-Dine et al. (2013). This dataset comprises the six cost-sharing plans described in Section 2. To make the group sizes more homogeneous, we combine

the groups with 25%, 50%, and mixed coinsurance plans. Thus, in our analysis, we have $G = 4$ treatment groups corresponding to $g = 1$, "free care" ($n_1 = 564$); $g = 2$, "25%, 50%, or mixed coinsurance" ($n_2 = 456$); $g = 3$, "95% coinsurance" ($n_3 = 372$); and $g = 4$, "individual deductible" ($n_4 = 495$). In total, there are $N = n_1 + ... + n_4 = 1,887$ families. We assign all N families to the four treatment groups (i.e., without a discard group of non-participants). In this version of the HIE data, we pool the data across five of the six sites, and we randomly assign all the families to the four treatment groups. Due to loss of data, the Dayton site is excluded from this analysis.

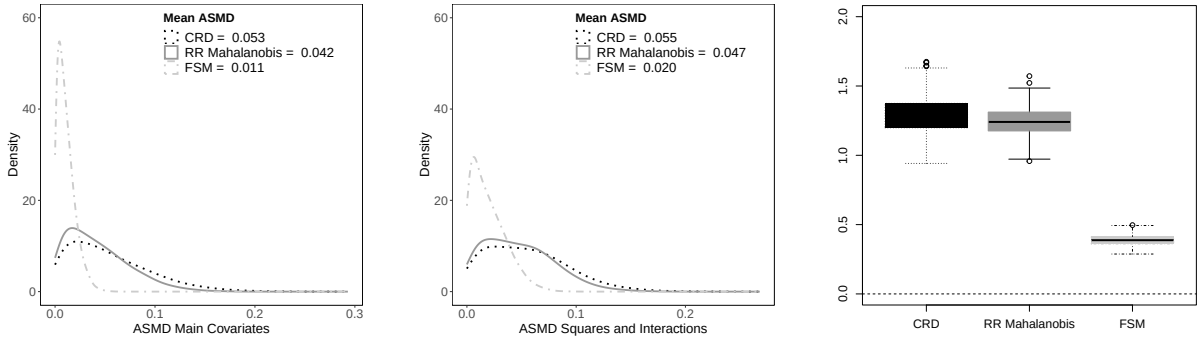
We consider $k = 20$ family-level baseline covariates, where $X_1, ..., X_5$ are scaled non-binary covariates, $X_6, ..., X_{14}$ are binary covariates, and $X_{15}, ..., X_{20}$ are binary covariates indicating missing data (see Table A14 for a description of each baseline covariate). Using this data, we compare complete randomization (CRD), rerandomization (RR), and the FSM in terms of balance and efficiency. For the FSM, we generate the SOM by first using SCOMARS on the combined groups $\{1, 2\}$ and $\{3, 4\}$, and then using SCOMARS again to split each combined group into its component groups. For the FSM, we also use the D-optimal selection function based on a linear potential outcome model on the main covariates. The assignments under the FSM are generated using the open source R package `FSM` available on CRAN. For rerandomization, we consider two balance criteria, one based on the Wilks' lambda statistic (RR Wilks; Lock 2011, Section 5.2) and the other based on the maximum pairwise Mahalanobis distance between any two treatment groups (RR Mahalanobis; Morgan and Rubin 2012). The balance criteria for both RR Wilks and RR Mahalanobis are based on all the main covariates and the squares and pairwise products of the scaled (non-binary) covariates. Finally, for both rerandomization methods, we use an acceptance rate of 0.001 (Lock 2011). We draw 400 independent assignments for each approach. The results under RR Wilks and RR Mahalanobis are roughly the same (see Section I in the Online Supplementary Materials), and hence, for conciseness, here we only discuss the results for RR Mahalanobis. The runtime of each of these assignments was approximately 78 seconds with RR Mahalanobis

and 28 seconds with the FSM on a Windows 64-bit laptop computer with an Intel(R) Core i7 processor.

6.2 Balance

Figures 2(a) and 2(b) display the distributions of ASMD across randomizations for the main covariates and their second-order transformations (squares and pairwise products). RR balances the main covariates and the second-order terms better than CRD. However, in both cases, the FSM improves considerably over CRD and RR. In fact, with the FSM, the average imbalance is less than half (0.02) of those under CRD and RR. Also, with both CRD and RR, it is common to see imbalances greater than 0.1 ASMD, whereas such extreme imbalances are non-existent with the FSM.

Figure 2: Distributions of absolute standardized mean differences (ASMD) of the main covariates (panel (a)) and their squares and pairwise products (panel (b)) across randomizations. For each plot, the legend presents the average ASMD across simulations for each method. Panel (c) shows the distributions of discrepancies between the correlation matrices of the covariates in treatment groups 1 and 2, as measured by the Frobenius norm, $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$. In terms of the main covariates, second-order transformations, and correlation matrices, the FSM substantially outperforms CRD and RR.



(a) Main covariates (b) Second order terms (c) Frobenius norm

A related question is how well the methods balance all second-order features of the joint distribution of the covariates. Figures 2(c) and A4 provide an answer to this question in the boxplots of the discrepancies between correlation matrices across randomizations. As a measure of discrepancy, we consider the Frobenius norm of the difference between correlation

matrices in two groups, i.e., $||\underline{\boldsymbol{R}}_g - \underline{\boldsymbol{R}}_{g'}||_F$, where $\underline{\boldsymbol{R}}_g$ is the sample correlation matrix in group g and $||\cdot||_F$ is the Frobenius norm.[6] Smaller values of $||\underline{\boldsymbol{R}}_g - \underline{\boldsymbol{R}}_{g'}||_F$ indicate better balance on the correlation matrix of the covariates between the groups g and g'. As in the aforementioned second-order transformations, we see a similar performance between complete randomization and rerandomization, which is considerably improved by the FSM with a median about three times smaller.

In Appendix I of the Online Supplementary Materials, we extend our evaluation of the FSM with a version of RR that uses the Mahalanobis distance solely on the main covariates as the imbalance criteria. The results show that the FSM outperforms this version of RR in terms of balancing both the main covariates and their second-order transformations. In fact, the average imbalance under the FSM is less than half of the average imbalance under RR.

6.3 Efficiency

In this section, we evaluate the estimation accuracy of the methods under model- and randomization-based approaches to inference. The main differences between the model- and randomization-based standard errors is that in the model-based approach, the variance calculation does not explicitly take into account the variability arising through the randomization distribution, whereas in the randomization-based approach it does. For illustration, here we consider estimating the average treatment effect of treatment 3 relative to treatment 2, i.e., $\text{SATE}_{3,2}$ and $\text{PATE}_{3,2}$. The results for the average treatment effects with other pairs of treatment groups are similar.

Under the model-based approach, we consider two potential outcome models, one that is linear on the main covariates (Model A1), and another that is linear on the main covariates and the second-order transformations of the scaled covariates (Model A2). The results are summarized in Table 2; see Appendix H.3 for computational details. While the performance of the three methods is similar under Model A1, under Model A2 there are substantial

[6]The Frobenius norm of a matrix is the square root of the sum of squares of all its elements.

differences, with the FSM outperforming both complete randomization and rerandomization. In fact, under Model A2, there is a 14-15% reduction in the average standard error, and a 53-64% reduction in the maximum standard error, with the FSM.

Table 2: Average and maximum model-based standard errors relative to the FSM across randomizations. In this setting, the SE is equivalent to the RMSE because the estimator is unbiased. Under Model A1 (linear model on the covariates), the FSM is slightly more efficient than RR and CRD. Under Model A2 (linear model on the covariates and their second-order transformations), the FSM is considerably more efficient than CRD and RR.

(a) Model A1

	Designs		
	CRD	RR Mahalanobis	FSM
Average SE	1.02	1.01	1.00
Maximum SE	1.04	1.02	1.00

(b) Model A2

	Designs		
	CRD	RR Mahalanobis	FSM
Average SE	1.15	1.14	1.00
Maximum SE	1.64	1.53	1.00

Under the randomization-based approach, we consider the generative models $Y(3) = 10 + 2X_1 + 3X_2 + 0.5X_3 + 0.3X_4 + \eta$ (Model B1) and $Y(3) = 10 + 2X_1 + 2X_2X_3 - X_4X_5 + \eta$ (Model B2) where $Y(3) = Y(2)$ and $\eta \sim \mathcal{N}(0, 1.5^2)$. Here, both the generative models satisfy the sharp-null hypothesis of zero treatment effect for every unit and hence, $\text{SATE}_{3,2} = 0$. Under each design, $\text{SATE}_{3,2}$ is estimated using the standard difference-in-means estimator. The corresponding randomization-based standard error is obtained by generating 400 randomizations and computing the standard deviation of the estimator across these 400 randomizations. Likewise, the corresponding root mean squared error is obtained by computing the square root of the average squared difference between the estimator and the estimand, across these 400 randomizations. We repeat this process 500 times, drawing a new set of potential outcomes each time. The average standard errors and root mean squared errors (across these 500 simulations) are presented in Table 3. See Appendix I for similar comparisons under a set of different generative models of the potential outcome. In terms of efficiency, we see again a clear advantage of the FSM. Under both Model B1 and Model B2, the average standard errors and root mean squared errors of complete randomization and

rerandomization are about twice of those under the FSM.

Table 3: Average randomization-based standard errors and root mean squared errors relative to the FSM. Both the average standard error (SE) and the average root mean squared error (RMSE) for the FSM are 0.12 under Model B1 (linear model on the covariates) and 0.63 under Model B2 (linear model on the covariates and their second-order transformations). Under both models, the FSM is considerably more efficient than both CRD and RR.

(a) Model B1

	Designs		
	CRD	RR Mahalanobis	FSM
SE	2.37	1.94	1
RMSE	2.36	1.94	1

(b) Model B2

	Designs		
	CRD	RR Mahalanobis	FSM
SE	2.70	2.31	1
RMSE	2.70	2.33	1

6.4 Intuition and further explorations

Our analysis illustrates some important differences between the FSM, CRD, and RR. With respect to RR, these differences pertain to the specification, role, and implementation of the assignment criterion. First, regarding the specification of the criterion, while RR uses the Mahalanobis distance, the FSM uses the D-optimality criterion, which, coupled with a suitable SOM, leads to robust assignments under a more general class of potential outcome models.

Second, regarding the role of this criterion, while RR essentially constrains the allowable treatment assignments, the FSM seeks to optimize them toward the criterion. In essence, while RR solves a feasibility problem by resampling, the FSM aims to solve a maximization problem by step-wise assignment. Furthermore, the feasibility problem solved by RR depends on the balance threshold, which can be difficult to select in practice. While a very high threshold can accept assignments with poor covariate balance, a very low one can be computationally onerous.

Third, regarding the implementation of the criterion, while RR assigns all units in one step and then discards imbalanced assignments, the FSM assigns units in multiple steps

(one at a time) in a random but optimal fashion determined by the selection order and the selection criterion. This difference is crucial because in experiments like the HIE with several treatment groups and many covariates, the space of possible treatment assignments is vast. As shown in our analyses, optimally selecting among these assignments in a step-wise manner can make a substantial improvement in terms of balance, efficiency, computational time, and, ultimately, in the use of scarce resources available for experimentation. [7]

To better see this, we asked how we would need to modify RR to achieve comparable performance to the FSM? Using the HIE data, we approximated the randomization distribution of the imbalance criterion of RR (i.e., the maximum Mahalanobis distance M across all pairs of treatment groups) by generating random assignments for 100 hours. See Table 4 for a summary of the results. The table displays summary statistics of the distribution of M under CRD, RR, and the FSM. As shown in Table 4, the highest (worst-case) value of M under the FSM is smaller than the smallest (best-case) value of M under CRD and RR. Importantly, even if we set the RR acceptance rate to 0.0000001 (i.e., 1 over 10 million), we still have imbalances higher than the worst-case imbalance of the FSM. In sum, even with an acceptance rate as low as 0.0000001, RR did not perform as well as the FSM, despite taking 100 hours on average to generate a single assignment, as opposed to the 30 seconds of running time of the FSM.

Table 4: Distribution of the maximum pairwise Mahalanobis distance across groups (M). For CRD, we obtain this distribution by generating over 10 million random assignments for 100 hours. For RR (0.001), we obtain this distribution using 0.1% of all these assignments with the smallest values of M. For the FSM, we obtain this distribution using the 400 random assignments from Section 6.1.

Design	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
CRD	18.5	39.5	43.9	44.4	48.7	96.1
RR (0.001)	18.5	25.4	26.2	25.9	26.7	27.1
FSM	2.8	4.7	5.3	5.4	6.0	10.6

[7] Figures 1 and 2 show that, although RR does well under common balance standards (the mean differences are systematically lower than the typical threshold of 0.1 ASMD), there is room to select better (more balanced) random treatment assignments, which is achieved by the FSM.

7 Ten further studies in the health and social sciences

In addition to the previous study, we evaluate the performance of the FSM in ten randomized studies from the health and social sciences. These ten studies are labelled (1) Crepon, which evaluates the impact of a microcredit program in rural Morocco on assets, profits, and consumption (Crépon et al. 2015); (2) Angrist, which evaluates the impact of cash incentives on certification rates among low-achievers in Israel (Angrist and Lavy 1999); (3) Finkelstein, which evaluates the impact of the Camden Coalition of Healthcare Providers' Hotspotting program on hospital readmission rates among patients with high use of healthcare services (Finkelstein et al. 2020); (4) Durocher, which evaluates the impact of intravenous infusion versus intramusculur oxytocin on postpartum blood loss and hemmorhage rates (Durocher et al. 2019); (5) Lalonde, which evaluates the impact of Nationally Supported Work program on earnings (LaLonde 1986); (6) Karlan, which evaluates the impact of loans with an indemnity component on demand for credit and investment decisions of farmers (Karlan et al. 2014); (7) Dupas, which evaluates the impact of different cost provisions for allocating dilute-chlorine water treatment solution on chlorine residuals in households' stored water (Dupas et al. 2016); (8) Blattman, which evaluates the impact of industrial job offers and entrepreneurial programs on health, income and other measures (Blattman and Dercon 2018); (9) Ambler, which evaluates the impact of offering Salvadoran migrant maching funds for educational remittances on educational investments and other outcomes Ambler et al. (2015); (10) Wantchekon, which evaluates the impact of townhall meeting based on programmatic, nonclientelist platforms on clientelism, voter turnout, and vote shares (Fujiwara and Wantchekon 2013). Table 5 provides details on the design parameters considered in these studies.

For each study, we generate 100 assignments of complete randomization (CRD), Rerandomization with Mahalanobis distance (based on the main covariates) and 0.001 acceptance rate

Table 5: Design parameters and balance results for ten case studies in the health and social sciences. The second average denoted with an asterisk (*) excludes the Wantchekon study because $\underline{\tilde{\boldsymbol{X}}}_r^\top \underline{\tilde{\boldsymbol{X}}}_r$ matrix is non-invertible for the first $r = 22$ selections.

Study	Design parameters				Main covariates					Second-order transformations				
	N	G	$(n_1, ..., n_G)$	k	CRD	RR	FSM	$\frac{\text{CRD}}{\text{FSM}}$	$\frac{\text{RR}}{\text{FSM}}$	CRD	RR	FSM	$\frac{\text{CRD}}{\text{FSM}}$	$\frac{\text{RR}}{\text{FSM}}$
Crepon	4465	2	(2266, 2199)	33	0.024	0.018	0.015	1.6	1.2	0.024	0.023	0.018	1.3	1.3
Angrist	3821	2	(1910,1911)	20	0.025	0.014	0.002	12.5	7.0	0.026	0.023	0.003	8.7	7.7
Finkelstein	782	2	(389,393)	10	0.062	0.020	0.010	6.2	2.0	0.059	0.048	0.013	4.5	3.7
Durocher	480	2	(239,241)	12	0.072	0.031	0.017	4.2	1.8	0.073	0.068	0.022	3.3	3.1
Lalonde	445	2	(222,223)	10	0.083	0.044	0.014	5.9	3.1	0.077	0.070	0.019	4.1	3.7
Karlan	169	2	(84, 85)	16	0.124	0.059	0.053	2.3	1.1	0.123	0.119	0.060	2.1	2.0
Dupas	1118	3	(351, 382, 385)	11	0.059	0.018	0.010	5.9	1.8	0.058	0.044	0.017	3.4	2.6
Blattman	947	3	(358,304,285)	34	0.064	0.048	0.026	2.5	1.8	0.065	0.064	0.036	1.8	1.8
Ambler	991	4	(360, 211, 203, 217)	16	0.073	0.053	0.015	4.9	3.5	0.073	0.071	0.017	4.3	4.2
Wantchekon	24	2	(12, 12)	10	0.334	0.170	0.245	1.4	0.7	0.333	0.289	0.237	1.4	1.2
Average					0.092	0.048	0.041			0.091	0.082	0.044		
Average*					0.065	0.034	0.018			0.064	0.059	0.023		

(RR), and the FSM (based on the main covariates). The mean ASMD of the main covariates and their squares and interactions under each method are presented in Table 5. See figures A7 and A8 in the Online Supplementary Materials for plots of the distributions of these imbalances, alongside the Frobenius norms of $\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2$.[8]

Table 5 shows that for each study, CRD achieves a similar mean balance on the main covariates and their squares and interactions. RR improves balance over CRD considerably for the main covariates, but only mildly for the squares and interactions. By contrast, for almost all the studies, the FSM substantially improves balance over CRD and RR in terms of both the main covariates and their transformations. The only exception is the Wantchekon study, where the group sizes barely exceed the number of covariates $k = 10$. The FSM is not designed for settings like this, where the number of covariates is greater than or close to the minimum treatment group size. In such settings, the matrix $\underline{\tilde{\boldsymbol{X}}}_r^\top \underline{\tilde{\boldsymbol{X}}}_r$ is non-invertible for almost every selection stage of the FSM, and therefore, the D-optimal selection function in the FSM relies almost entirely on ridge augmentation to feasibly select units (see Section

[8]Groups 1 and 2 are chosen haphazardly as a typical pair of groups. The results for the other pairs of groups are similar.

4), producing suboptimal selections.

Across the ten studies, the ASMD on the main covariates are 55% ($= \frac{0.092-0.041}{0.092}$) and 15% ($= \frac{0.048-0.041}{0.048}$) lower on average with the FSM than CRD and RR, respectively. If we exclude Wantchekon, then these percent reductions in ASMD are amplified to 72% and 47%. Similarly, across the ten studies, the ASMD on the squares and interactions of the covariates with the FSM are about 50% smaller than both CRD and RR, and without Wantchekon, they are at least 60% smaller. In fact, FSM has better balance on both the main covariates and their second-order transformations over CRD and RR uniformly across the first nine studies (as shown by the $\frac{\text{CRD}}{\text{FSM}}$ and $\frac{\text{RR}}{\text{FSM}}$ columns). For each study, the relative improvement in balance under the FSM over RR is larger for the second-order transformations than for the main covariates. In particular, for half of the ten studies, the mean ASMD of the second-order transformations under RR are at least three times larger than those under the FSM, implying substantial improvement in balance on these transformations under the FSM.

Overall, averaging the ASMD of the main covariates and their second-order transformations across the first nine studies, we see that the FSM achieves 68% better covariate balance than complete randomization and 56% better covariate balance than rerandomization in a typical study. Across these studies, the FSM's performance relative to CRD and RR is consistent with those in the HIE study in Section 6.[9] A similar pattern to the HIE study is also noted in the plots of $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$ in figures A7 and A8 in the Online Supplementary Materials, where for most studies, the worst (least balanced) assignment among all the draws of the FSM has a better balance on the correlation matrices than the best (most balanced) assignment among all the draws of CRD and RR. As discussed in Section 6.3, under both model- and randomization-based approaches to inference, better balance directly translates to more efficient estimates of treatment effects.

[9]Notably, the relative performances of the methods in the HIE study are comparable to those of the Ambler study, which involves roughly half the sample size of the HIE study and similar values of the other design parameters. For instance, the average ASMD of the main covariates under CRD, RR, and the FSM in the Ambler study are roughly $\sqrt{2}$ times those in the HIE study, where $\sqrt{2}$ is the factor that corrects for the difference in sample size.

Therefore, similar to the HIE case study, these results show that across a range of randomized experiments, the FSM is a flexible and robust approach to randomization.

8 Practical considerations and extensions

8.1 Multi-group experiments

As discussed, the FSM can readily handle experiments with multiple treatment groups. In so doing, the key methodological consideration is the choice of an SOM. As in two-group experiments, we would like to generate an SOM that is randomized and sequentially controlled, so that at every stage of the random selection process, the number of selections made by each treatment group up to that stage is close to its fair share. Constructing a sequentially controlled SOM for multi-group experiments with arbitrary group sizes is an open problem. However, such constructions are possible for several practically relevant configurations of the group sizes, namely (a) groups of equal size, (b) groups having one of two distinct sizes, and (c) groups of more than two distinct sizes such that when combined by groups of equal size they have the same total size. In the Online Supplementary Materials, we provide algorithms to construct an SOM for all three configurations and prove that the resulting SOM is sequentially controlled. In practice, for more general group size configurations, one strategy to generate an SOM is to first identify one of these three configurations that is structurally similar to the configuration at hand, and then use the corresponding SOM-generating algorithm. The resulting algorithm may not always be sequentially controlled, but is still likely to produce a well-controlled randomized selection order.

8.2 Stratified experiments

In stratified experiments, units are grouped into two or more strata, and within each stratum, units are randomly assigned to treatment. Here we propose a family of extensions of the FSM to such settings. Typically, in stratified experiments the treatment group sizes within each stratum are pre-specified by the investigator. The main challenge arises when the treatment

group sizes differ across strata. To address this challenge, we construct an augmented SOM with information of the treatment group that selects at each stage and the stratum that it selects from. This construction guarantees that each treatment group is assigned the pre-specified number of units in each stratum. In the Online Supplementary Materials, we discuss two approaches to construct such an SOM. At a high level, one approach generates a separate SOM for each stratum, while the other approach uses SCOMARS to determine the order of stratum labels for each treatment.

8.3 Sequential experiments

Sequential experiments are experiments where units progressively become available for random assignment, possibly in batches of varying sizes. Here we describe extensions of the FSM to such settings. The simplest approach is to run an independent FSM for each new batch of available units. However, in general, this approach fails to account for accrued covariate imbalances between the treatment groups. To address this issue, we propose an alternative approach that considers the new batch as a continuation of the previous one. More specifically, for each unit in the new batch we evaluate the value of the D-optimal selection function using all the units already assigned to the selecting treatment group. See the Online Supplementary Materials for technical details. Thus, the FSM tends to remove accrued covariate imbalances by adaptively updating its assignment mechanism from one batch to the next. In this sense, FSM connects to multi-arm bandits and other adaptive designs (see, e.g., (Villar et al. 2015)). However, adaptive designs and the FSM typically pursue different goals. Adaptive designs primarily aim to determine the optimal treatment by assigning as many units as possible to the superior treatments (exploitation) while allocating enough units to the inferior treatments (exploration) (Scott 2010). In contrast, the FSM primarily aims to efficiently estimate and conduct inferences on the average treatment effects by adequately balancing covariate distributions across groups. A promising direction of future research involves extending the FSM to contextual bandit settings, where information on

both covariates and interim outcomes are available.

9 Summary and remarks

We revisited, formalized, and extended the FSM for experimental design. We proposed a new selection function based on D-optimality that requires no tuning parameters for covariate balance. We showed that, equipped with this selection function, the FSM has a number of appealing properties. First, the FSM is affine invariant and hence, it self-standardizes covariates with possibly different units of measurement. Second, the FSM produces randomized block designs without explicitly randomizing in each block. Third, the FSM also produces matched-pair designs without explicitly constructing the matched pairs beforehand and randomizing within each pair. We described how both model-based and randomization-based inference on treatment effects can be conducted using the FSM. For a range of practically relevant configurations of group sizes in multi-group experiments, we proposed new algorithms to generate a fair and random selection order of treatments under the FSM. We also discussed potential extensions of the FSM to stratified and sequential experiments. In a case study on the RAND Health Insurance Experiment, and ten additional randomized studies from the health and social sciences, we showed that the FSM is a robust approach to randomization, exhibiting better performance than complete randomization and rerandomization in terms of balance and efficiency. While there are settings where complete randomization may perform better than the FSM in terms of efficiency, such settings are less common and involve jagged, i.e., highly non-smooth, potential outcome models. In settings where these models are reasonably smooth, the FSM is expected to perform well. Moreover, by virtue of randomization, the FSM ensures balance in unobserved covariates in expectation. Although it is possible that chance imbalances may arise with unobserved covariates under the FSM, it is important to note that this issue is an inherent risk with any randomized design. Regardless, by effectively balancing observed covariates and their transformations, the

FSM also balances any unobserved covariate that is highly correlated with them.[10] Overall, through our extensive explorations with real and simulated experimental data, the FSM has consistently stood out as a robust design that can handle multiple treatment groups and a fairly large number of categorical and continuous covariates without requiring tuning parameters for balance and nor coarsening covariates. We recommend giving strong considerations to the FSM in experimental design for its conceptual simplicity, practicality, balance, and robustness.

References

Ambler, K., Aycinena, D., and Yang, D. (2015), "Channeling remittances to education: A field experiment among migrants from El Salvador," *American Economic Journal: Applied Economics*, 7, 207–32.

Angrist, J. D. and Lavy, V. (1999), "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly Journal of Economics*, 114, 533–575.

Aron-Dine, A., Einav, L., and Finkelstein, A. (2013), "The RAND health insurance experiment, three decades later," *Journal of Economic Perspectives*, 27, 197–222.

Aronow, P. M. and Samii, C. (2013), "Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities," *Survey Methodology*, 39, 231–241.

Banerjee, A. V., Chassang, S., and Snowberg, E. (2017), "Decision theoretic approaches to experiment design and external validity," in *Handbook of Economic Field Experiments*, Elsevier, vol. 1, pp. 141–174.

Bertsimas, D., Johnson, M., and Kallus, N. (2015), "The power of optimization over randomization in designing experiments involving small samples," *Operations Research*, 63, 868–876.

Blattman, C. and Dercon, S. (2018), "The impacts of industrial and entrepreneurial work on income and health: Experimental evidence from Ethiopia," *American Economic Journal: Applied Economics*, 10, 1–38.

Brook, R. H., Keeler, E. B., Lohr, K. N., Newhouse, J. P., Ware, J. E., Rogers, W. H., Davies, A. R., Sherbourne, C. D., Goldberg, G. A., Camp, P., et al. (2006), "The health

[10]Moreover, since the FSM is used to design an experiment, i.e., to generate treatment assignments, it can also be used in conjunction with graphical causal models, e.g., the single world intervention graphs of Richardson and Robins (2013).

insurance experiment: a classic RAND study speaks to the current health care reform debate," *Santa Monica, CA: RAND Corporation*.

Chattopadhyay, A., Morris, C. N., and Zubizarreta, J. R. (2021), "Randomized and Balanced Allocation of Units into Treatment Groups Using the Finite Selection Model for R," *arXiv preprint arXiv:2105.02393*.

Cochran, W. and Cox, G. (1957), *Experimental Designs*, John Wiley & Sons New York.

Cox, D. R. and Reid, N. (2000), *The Theory of the Design of Experiments*, CRC Press.

Crépon, B., Devoto, F., Duflo, E., and Parienté, W. (2015), "Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco," *American Economic Journal: Applied Economics*, 7, 123–50.

Dupas, P., Hoffmann, V., Kremer, M., and Zwane, A. P. (2016), "Targeting health subsidies through a nonprice mechanism: A randomized controlled trial in Kenya," *Science*, 353, 889–895.

Durocher, J., Dzuba, I. G., Carroli, G., Morales, E. M., Aguirre, J. D., Martin, R., Esquivel, J., Carroli, B., and Winikoff, B. (2019), "Does route matter? Impact of route of oxytocin administration on postpartum bleeding: A double-blind, randomized controlled trial," *PloS one*, 14, e0222981.

Finkelstein, A., Zhou, A., Taubman, S., and Doyle, J. (2020), "Health care hotspotting—a randomized, controlled trial," *New England Journal of Medicine*, 382, 152–162.

Fisher, R. A. (1925), "Statistical methods for research workers, 13e," *London: Oliver and Loyd, Ltd*, 99–101.

— (1935), *The Design of Experiments*, London: Oliver & Boyd.

Fujiwara, T. and Wantchekon, L. (2013), "Can informed public deliberation overcome clientelism? Experimental evidence from Benin," *American Economic Journal: Applied Economics*, 5, 241–55.

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

Hainmueller, J. (2012), "Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, 20, 25–46.

Harshaw, C., Sävje, F., Spielman, D. A., and Zhang, P. (2024), "Balancing covariates in randomized experiments with the gram–schmidt walk design," *Journal of the American Statistical Association*, 1–13.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Karlan, D., Osei, R., Osei-Akoto, I., and Udry, C. (2014), "Agricultural decisions after relaxing credit and risk constraints," *The Quarterly Journal of Economics*, 129, 597–652.

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

LaLonde, R. J. (1986), "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, 604–620.

Lock, K. F. (2011), *Rerandomization to Improve Covariate Balance in Randomized Experiments*, Harvard University.

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

Morris, C. (1979), "A finite selection model for experimental design of the health insurance study," *Journal of Econometrics*, 11, 43–61.

— (1983), "Sequentially controlled Markovian random sampling (SCOMARS)," *Institute of Mathematical Statistics Bulletin*, 12, 237.

Morris, C. and Hill, J. (2000), "The health insurance experiment: design using the finite selection model," *Public Policy and Statistics: Case Studies from RAND*, Springer Science & Business Media, 29–53.

Mukerjee, R., Dasgupta, T., and Rubin, D. B. (2018), "Using standard tools from finite population sampling to improve causal inference for complex experiments," *Journal of the American Statistical Association*, 113, 868–881.

Newhouse, J. P. et al. (1993), *Free for All?*, Harvard University Press.

Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Statistical Science*, 5, 463–480.

Richardson, T. S. and Robins, J. M. (2013), "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality," *Center for the*

Statistics and the Social Sciences, University of Washington Series. Working Paper, 128, 2013.

Rosenbaum, P. R. (2002), *Observational Studies*, Springer.

— (2010), "Design sensitivity and efficiency in observational studies," *Journal of the American Statistical Association*, 105, 692–702.

Rosenbaum, P. R. and Rubin, D. B. (1985), "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician*, 39, 33–38.

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66, 688.

Scott, S. L. (2010), "A modern Bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, 26, 639–658.

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

Villar, S. S., Bowden, J., and Wason, J. (2015), "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges," *Statistical Science: a review journal of the Institute of Mathematical Statistics*, 30, 199.

Supplementary Materials

A Notation, estimands, and acronyms

Table A1: Notation

N	$\triangleq$	Full sample size
i	$\triangleq$	Index of unit, $i = 1, ..., N$
G	$\triangleq$	Number of treatments
g	$\triangleq$	Index of treatment group, $g = 1, 2, ..., G$
n_g	$\triangleq$	Size of treatment group g
k	$\triangleq$	Number of baseline covariates
$\boldsymbol{X}_i$	$\triangleq$	Observed vector of baseline covariates of unit i
$\underline{\boldsymbol{X}}_{\text{full}}$	$\triangleq$	$N \times k$ matrix of covariates in the full sample
$\underline{\tilde{\boldsymbol{X}}}_{\text{full}}$	$\triangleq$	$N \times k+1$ design matrix in the full sample
$\bar{\boldsymbol{X}}_{\text{full}}$	$\triangleq$	$k \times 1$ vector of means of the baseline covariates in the full sample
$\underline{\boldsymbol{S}}_{\text{full}}$	$\triangleq$	$k \times k$ covariance matrix of the baseline covariates in the full sample
$Y_i(g)$	$\triangleq$	Potential outcome of unit i under treatment g
$\boldsymbol{Y}(g)$	$\triangleq$	Vector of potential outcomes under treatment g, $(Y_1(g), ..., Y_N(g))^\top$
Z_i	$\triangleq$	Treatment assignment indicator of unit i, $Z_i \in \{1, 2, ..., G\}$
$\boldsymbol{Z}$	$\triangleq$	Vector of treatment assignment indicators, $(Z_1, ..., Z_N)^\top$
Y_i^{obs}	$\triangleq$	Observed outcome of unit i, $Y_i^{\text{obs}} = \sum_{g=1}^{G} \mathbb{1}(Z_i = g) Y_i(g)$

Table A2: Estimands

$Y_i(g') - Y_i(g'')$	$\triangleq$	Unit level causal effect of treatment g' relative to treatment g'' for unit i; $g', g'' \in \{1, 2, ..., G\}$
$\text{SATE}_{g',g''}$	$\triangleq$	$\frac{1}{N}\sum_{i=1}^{N}\{Y_i(g') - Y_i(g'')\}$, the Sample Average Treatment Effect of treatment g' relative to treatment g''
$\text{PATE}_{g',g''}$	$\triangleq$	$\mathbb{E}\{Y_i(g') - Y_i(g'')\}$, the Population Average Treatment Effect of treatment g' relative to treatment g''

Table A3: Acronyms

ASMD	Absolute Standardized Mean Difference
CRD	Completely Randomized Design
FSM	Finite Selection Model
HIE	Health Insurance Experiment
OLS	Ordinary Least Squares
PATE	Population Average Treatment Effect
RBD	Randomized Block Design
RR	Re-Randomization
SATE	Sample Average Treatment Effect
SCOMARS	Sequentially Controlled Markovian Random Sampling
SOM	Selection Order Matrix

B Proofs of theoretical results

Lemma A1. Let treatment 1 be the choosing group at the rth stage. Also, let $\underline{\tilde{\boldsymbol{X}}}_{r-1}$ be the $\tilde{n}_{r-1} \times (k+1)$ design matrix in treatment group 1 after the $(r-1)$th stage, where $\tilde{n}_{r-1} \geq 1$ and $\text{rank}(\underline{\tilde{\boldsymbol{X}}}_{r-1}) = k+1$. The D-optimal selection function chooses unit i' with covariate vector $\boldsymbol{X}_{i'} \in \mathbb{R}^k$, where

$$i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \tag{A1}$$

Proof. We follow the notations outlined in Section 4. At the rth stage, D-optimal selection function selects unit $i' \in \mathcal{R}_{r-1}$, where $i' \in \arg\max_{i \in \mathcal{R}_{r-1}} \det(\underline{\tilde{\boldsymbol{X}}}_{r,i}^\top \underline{\tilde{\boldsymbol{X}}}_{r,i})$. Now, for $i \in \mathcal{R}_{r-1}$,

$$\begin{aligned}
\det(\underline{\tilde{\boldsymbol{X}}}_{r,i}^\top \underline{\tilde{\boldsymbol{X}}}_{r,i}) &= \det\left\{ \underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1} + \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} (1, \boldsymbol{X}_i^\top) \right\} && \text{(A2)} \\
&= \det(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1}) \det\left\{ \underline{\boldsymbol{I}} + (\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-\frac{1}{2}} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} (1, \boldsymbol{X}_i^\top) (\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-\frac{1}{2}} \right\} && \text{(A3)} \\
&= \det(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1}) \Big\{ 1 + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \Big\}, && \text{(A4)}
\end{aligned}$$

where the final equality holds since for two matrices $\underline{\boldsymbol{A}}_{m \times n}$ and $\underline{\boldsymbol{B}}_{n \times m}$, $\det(\underline{\boldsymbol{I}}_m + \underline{\boldsymbol{AB}}) = \det(\underline{\boldsymbol{I}}_n + \underline{\boldsymbol{BA}})$. Equation A4 implies that the selected unit i' maximizes $(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}$. This completes the proof.

□

Proof of Theorem 4.1

Proof. We use the notations in Section 3.1 and Table A1. We first consider the case where $\tilde{n}_{r-1} = 0$. The selected unit i' satisfies,

$$i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}. \tag{A5}$$

Now, denoting $\boldsymbol{e}_1 = (1, 0, ..., 0)$ as the $k \times 1$ first standard unit vector, we have

$$\begin{aligned}
(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} &= (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 1 \\ \bar{\boldsymbol{X}}_{\text{full}} \end{pmatrix} + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 0 \\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}} \end{pmatrix} \\
&= (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}} \frac{\boldsymbol{e}_1}{N} + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 0 \\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}} \end{pmatrix} \\
&= \frac{1}{N} + \{0, (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}})^\top\}(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 0 \\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}} \end{pmatrix} \\
&= \frac{1}{N} + \frac{1}{N}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}})^\top (\underline{\boldsymbol{S}}_{\text{full}})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}}).
\end{aligned} \tag{A6}$$

Here the last equality holds since, by the formula for the inverse of a partitioned matrix, $(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} = \begin{pmatrix} \underline{\boldsymbol{B}}_{11} & \underline{\boldsymbol{B}}_{12} \\ \underline{\boldsymbol{B}}_{21} & \underline{\boldsymbol{B}}_{22} \end{pmatrix}$, where $\underline{\boldsymbol{B}}_{22}^{-1} = \underline{\boldsymbol{X}}_{\text{full}}^\top \underline{\boldsymbol{X}}_{\text{full}} - N\bar{\boldsymbol{X}}_{\text{full}} \bar{\boldsymbol{X}}_{\text{full}}^\top = N\underline{\boldsymbol{S}}_{\text{full}}$. This completes the proof of the $\tilde{n}_{r-1} = 0$ case. The proof for the case where $\tilde{n}_{r-1} \geq 1$ and $\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1}$ is invertible follows similar steps and hence is omitted.

We now consider the case where $\tilde{n}_{r-1} \geq 1$ and $\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1}$ is not invertible. We denote $\bar{\boldsymbol{X}}_{r-1}^* = \frac{\bar{\boldsymbol{X}}_{r-1} + \epsilon \bar{\boldsymbol{X}}_{\text{full}}}{1+\epsilon}$ and $\underline{\boldsymbol{S}}_{r-1}^* = (\frac{1}{\tilde{n}_{r-1}} \underline{\boldsymbol{X}}_{r-1}^\top \underline{\boldsymbol{X}}_{r-1} + \frac{\epsilon}{N} \underline{\boldsymbol{X}}_{\text{full}}^\top \underline{\boldsymbol{X}}_{\text{full}}) - (1+\epsilon)\bar{\boldsymbol{X}}_{r-1}^* \bar{\boldsymbol{X}}_{r-1}^{*\top}$. The selected unit i' satisfies,

$$i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^\top) \Big(\frac{1}{\tilde{n}_{r-1}} \underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1} + \frac{\epsilon}{N} \underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}} \Big)^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \tag{A7}$$

Denoting $\tilde{\underline{\boldsymbol{G}}} = \begin{pmatrix} \sqrt{\frac{1}{\tilde{n}_{r-1}}}\underline{\boldsymbol{X}}_{r-1} \\ \sqrt{\frac{\epsilon}{N}}\underline{\boldsymbol{X}}_{\text{full}} \end{pmatrix}$, we have

$$
\begin{aligned}
&(1, \boldsymbol{X}_i^\top)\Big(\frac{1}{\tilde{n}_{r-1}}\tilde{\underline{\boldsymbol{X}}}_{r-1}^\top\tilde{\underline{\boldsymbol{X}}}_{r-1} + \frac{\epsilon}{N}\tilde{\underline{\boldsymbol{X}}}_{\text{full}}^\top\tilde{\underline{\boldsymbol{X}}}_{\text{full}}\Big)^{-1}\begin{pmatrix}1\\ \boldsymbol{X}_i\end{pmatrix} \\
&= (1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1}\begin{pmatrix}1\\ \boldsymbol{X}_i\end{pmatrix} \\
&= (1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1}\begin{pmatrix}0\\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^*\end{pmatrix} + (1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1}\begin{pmatrix}1\\ \bar{\boldsymbol{X}}_{r-1}^*\end{pmatrix} \\
&= (0, (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^*)^\top)(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1}\begin{pmatrix}0\\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^*\end{pmatrix} + \frac{1}{1+\epsilon} \\
&= (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^*)^\top(\underline{\boldsymbol{S}}_{r-1}^*)^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^*) + \frac{1}{1+\epsilon}. \qquad \text{(A8)}
\end{aligned}
$$

Here, the third equality holds since $\begin{pmatrix}1\\ \bar{\boldsymbol{X}}_{r-1}^*\end{pmatrix} = \frac{1}{1+\epsilon}\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}}\boldsymbol{e}_1$ and the fourth equality holds since $(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1} = \begin{pmatrix}\underline{\boldsymbol{B}}_{11} & \underline{\boldsymbol{B}}_{12}\\ \underline{\boldsymbol{B}}_{21} & \underline{\boldsymbol{B}}_{22}\end{pmatrix}$, where $\underline{\boldsymbol{B}}_{22}^{-1} = (\frac{1}{\tilde{n}_{r-1}}\underline{\boldsymbol{X}}_{r-1}^\top\underline{\boldsymbol{X}}_{r-1} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}_{\text{full}}^\top\underline{\boldsymbol{X}}_{\text{full}}) - (1+\epsilon)\bar{\boldsymbol{X}}_{r-1}^*\bar{\boldsymbol{X}}_{r-1}^{*\top} = \underline{\boldsymbol{S}}_{r-1}^*$. This completes the proof. □

Proof of Theorem 4.2

Proof. (a) We first consider the setting of a standard block design where $N = BG$ (i.e., $c = 1$). The blocks are labelled $1, 2, ..., B$. Here, the SOM is constructed by stacking B independent random permutations of the 'chunk' $(1, 2, ..., G)$. We will show that the choices made by the treatment groups in the FSM follow the assignment mechanism of an RBD.

Consider the first randomized chunk of the SOM, i.e., a random permutation of $(1, 2, ..., G)$. At the first stage of this randomized chunk, the choosing treatment group aims to maximize $(1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{X}}}_{\text{full}}^\top\tilde{\underline{\boldsymbol{X}}}_{\text{full}})^{-1}\begin{pmatrix}1\\ \boldsymbol{X}_i\end{pmatrix}$. Note that we can write $\tilde{\underline{\boldsymbol{X}}}_{\text{full}}$ as $\tilde{\underline{\boldsymbol{X}}}_{\text{full}} = \begin{pmatrix}\underline{\boldsymbol{D}}\\ \vdots \\ \underline{\boldsymbol{D}}\end{pmatrix}$, where $\underline{\boldsymbol{D}}_{B\times B} = \begin{pmatrix}1&1&0&...&0&0\\ 1&0&1&...&0&0\\ \vdots&\vdots&\vdots&...&\vdots&\vdots\\ 1&0&0&...&0&1\\ 1&0&0&...&0&0\end{pmatrix}$. Now, consider a transformation of the rows of the design matrix given by $\tilde{\tilde{\boldsymbol{X}}}_i = (\underline{\boldsymbol{D}}^\top)^{-1}\begin{pmatrix}1\\ \boldsymbol{X}_i\end{pmatrix}$. The transformed design matrix is $\tilde{\tilde{\underline{\boldsymbol{X}}}}_{\text{full}} = \tilde{\underline{\boldsymbol{X}}}_{\text{full}}\underline{\boldsymbol{D}}^{-1} = \begin{pmatrix}\underline{\boldsymbol{I}}_B\\ \vdots \\ \underline{\boldsymbol{I}}_B\end{pmatrix}$. We note that the $\tilde{\tilde{\boldsymbol{X}}}_i$s nothing but standard unit vectors. Now,

$$
(1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{X}}}_{\text{full}}^\top\tilde{\underline{\boldsymbol{X}}}_{\text{full}})^{-1}\begin{pmatrix}1\\ \boldsymbol{X}_i\end{pmatrix} = \tilde{\tilde{\boldsymbol{X}}}_i^\top(\tilde{\tilde{\underline{\boldsymbol{X}}}}_{\text{full}}^\top\tilde{\tilde{\underline{\boldsymbol{X}}}}_{\text{full}})^{-1}\tilde{\tilde{\boldsymbol{X}}}_i. \qquad \text{(A9)}
$$

Therefore, the selection function remains the same under the above transformation. Now, $\tilde{\tilde{\boldsymbol{X}}}_i^\top(\tilde{\tilde{\underline{\boldsymbol{X}}}}_{\text{full}}^\top\tilde{\tilde{\underline{\boldsymbol{X}}}}_{\text{full}})^{-1}\tilde{\tilde{\boldsymbol{X}}}_i = \frac{1}{G}\tilde{\tilde{\boldsymbol{X}}}_i^\top\tilde{\tilde{\boldsymbol{X}}}_i = \frac{1}{G}$ for all i, which essentially implies that the choosing group has no preference among the units for selection and hence chooses any one of

the N units randomly. Similarly, at the subsequent stages of this randomized chunk, the corresponding choosing groups select one of the remaining units randomly.

Next, we consider the second randomized chunk of the SOM. Without loss of generality, suppose treatment 1 gets to choose first in this chunk. Also, without loss of generality, suppose that in its first choice, treatment 1 had selected a unit from block 1. We claim that in this selection, treatment 1 will choose one of the remaining units randomly from any block other than block 1, which respects the assignment mechanism of an RBD.

To prove the claim, we first consider the objective function at this stage. Treatment 1 aims to maximize $(1, \boldsymbol{X}_i^\top)\Big(\frac{1}{\tilde{n}_{r-1}}\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top\underline{\tilde{\boldsymbol{X}}}_{r-1} + \frac{\epsilon}{N}\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top\underline{\tilde{\boldsymbol{X}}}_{\text{full}}\Big)^{-1}\begin{pmatrix}1\\ \boldsymbol{X}_i\end{pmatrix}$. Here, we denote the current stage by r. Using the same transformation as in the case of the first chunk, we can write the objective function as $\tilde{\tilde{\boldsymbol{X}}}_i^\top\Big(\frac{1}{\tilde{n}_{r-1}}\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} + \frac{\epsilon}{N}\underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}}^\top\underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}}\Big)^{-1}\tilde{\tilde{\boldsymbol{X}}}_i$, where $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = \underline{\tilde{\boldsymbol{X}}}_{r-1}\underline{\boldsymbol{D}}^{-1}$. Since $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}}^\top\underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}} = G\underline{\boldsymbol{I}}_B$, it is equivalent to maximize

$$\tilde{\tilde{\boldsymbol{X}}}_i^\top\Big(\underline{\boldsymbol{I}}_b + \frac{B}{\tilde{n}_{r-1}\epsilon G}\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}\Big)^{-1}\tilde{\tilde{\boldsymbol{X}}}_i = \tilde{\tilde{\boldsymbol{X}}}_i^\top\Big(\underline{\boldsymbol{I}}_b + \delta\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}\Big)^{-1}\tilde{\tilde{\boldsymbol{X}}}_i \tag{A10}$$

$$= \tilde{\tilde{\boldsymbol{X}}}_i^\top\Big\{\underline{\boldsymbol{I}}_b - \delta\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top(\underline{\boldsymbol{I}}_{\tilde{n}_{r-1}} + \delta\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top)^{-1}\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}\Big\}\tilde{\tilde{\boldsymbol{X}}}_i. \tag{A11}$$

Here, $\delta = \frac{B}{\tilde{n}_{r-1}\epsilon G}$. The final equality holds by the Woodbury matrix identity. Now, in this case, $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = (1, 0, ..., 0)$ (since treatment 1 has only selected one unit from block 1 up to this stage). So, the objective function in Equation A11 equals $1 - \frac{\delta}{1+\delta}\tilde{\tilde{\boldsymbol{X}}}_i^\top(\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1})\tilde{\tilde{\boldsymbol{X}}}_i^\top$. Since $\delta > 0$, it is equivalent to minimize $\tilde{\tilde{\boldsymbol{X}}}_i^\top(\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1})\tilde{\tilde{\boldsymbol{X}}}_i^\top = \tilde{\tilde{\boldsymbol{X}}}_i^\top\begin{pmatrix}1 & \mathbf{0}_{1\times(B-1)}^\top\\ \mathbf{0}_{(B-1)\times 1} & \underline{\mathbf{0}}_{(B-1)\times(B-1)}\end{pmatrix}\tilde{\tilde{\boldsymbol{X}}}_i$, which takes the value 0 for a unit in any block other than block 1 and 1 for a unit in block 1. This proves the claim for treatment 1. Moreover, by similar reasoning, the claim holds for all the other treatment groups in this randomized chunk.

Next, we consider a general randomized chunk of the SOM. Once again, without loss of generality, suppose treatment 1 gets to choose first in this chunk. Also, for simplicity of exposition and without loss of generality, suppose treatment 1 has already selected from blocks $1, 2, ..., b$, implying that $\tilde{n}_{r-1} = b$ and $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = (\underline{\boldsymbol{I}}_b\ \underline{\mathbf{0}}_{b\times(B-b)})$. This form of $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}$, along with Equation A11 implies that it is equivalent to minimize $\tilde{\tilde{\boldsymbol{X}}}_i^\top(\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1})\tilde{\tilde{\boldsymbol{X}}}_i^\top = \tilde{\tilde{\boldsymbol{X}}}_i^\top\begin{pmatrix}\underline{\boldsymbol{I}}_b & \underline{\mathbf{0}}_{b\times(B-b)}\\ \underline{\mathbf{0}}_{(B-b)\times b}^\top & \underline{\mathbf{0}}_{(B-b)\times(B-b)}\end{pmatrix}\tilde{\tilde{\boldsymbol{X}}}_i$, which is minimized for any unit i belonging to the blocks $b+1, ..., B$. This shows that at this stage, treatment 1 randomly chooses a unit from a block other than the blocks it has already chosen from. By similar reasoning, at subsequent stages of this randomized chunk, the choosing group follows the same selection strategy for their

own group. This completes the proof of the theorem for the setting of a standard block design.

We now prove the theorem for the general block design setting with $N = cBG$, $c > 1$. The proof strategy is exactly the same as the $c = 1$ setting. Here the SOM is generated by randomly permuting the chunk $(1, 2, ..., G)$ $B \times c$ times. Once the selections are completed for the the first B chunks, the resulting assignment resembles that of a standard RBD (by the previous proof), where each treatment group randomly chooses exactly one unit from each block. For the $(B + 1)$th chunk, suppose, without loss of generality, that treatment 1 gets to choose first. At this stage (denoted by stage r), treatment 1 tries to maximize,

$$\begin{aligned}(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} &= \tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1})^{-1} \tilde{\tilde{\boldsymbol{X}}}_i \\ &= \tilde{\tilde{\boldsymbol{X}}}_i^\top \tilde{\tilde{\boldsymbol{X}}}_i = 1, \end{aligned} \tag{A12}$$

where the penultimate equality holds since $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = \underline{\boldsymbol{I}}_B$. Thus, similar to the first randomized chunk in the setting of $c = 1$, treatment 1 (and the other treatments) randomly chooses one of the available units.

Finally, we consider a general chunk. Without loss of generality, suppose treatment 1 gets to choose first in this chunk. We can write the corresponding transformed design matrix $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}$ as

$$\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = \begin{pmatrix} \underline{\boldsymbol{I}}_B & \\ \underline{\boldsymbol{I}}_B & \\ \vdots & \\ \underline{\boldsymbol{I}}_B & \\ \underline{\boldsymbol{I}}_b & \underline{\mathbf{0}}_{b \times (B-b)} \end{pmatrix}. \tag{A13}$$

Here, without loss of generality, we have assumed that treatment 1 has chosen $c_0 + b$ times from the first b blocks and c_0 times from the remaining blocks, where $c_0 < c$. This implies that treatment 1 aims to maximize.

$$\tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1})^{-1} \tilde{\tilde{\boldsymbol{X}}}_i = \tilde{\tilde{\boldsymbol{X}}}_i^\top \left\{ c_0 \underline{\boldsymbol{I}}_B + \begin{pmatrix} \underline{\boldsymbol{I}}_b \\ \underline{\mathbf{0}}_{(B-b) \times b}^\top \end{pmatrix} (\underline{\boldsymbol{I}}_b \ \underline{\mathbf{0}}_{b \times (B-b)}) \right\}^{-1} \tilde{\tilde{\boldsymbol{X}}}_i, \tag{A14}$$

which has the same form as the objective function in Equation A10 in the $c = 1$ setting. Thus, following similar arguments as in the $c = 1$ setting, we conclude that at this stage, treatment 1 selects a unit randomly from blocks $b+1, ..., B$, which conforms to the assignment mechanism of an RBD. Also, at subsequent stages of the randomized chunk, the choosing group follows the same selection strategy for their own group. This completes the proof of the theorem.

(b) With two groups of equal sizes, the SOM consists of successive random permutations of the 'chunk' $(1, 2)$. By Theorem 4.1, for the first pair of stages of selection, the objective function (to maximize) is given by

$$(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}})^\top (\underline{\boldsymbol{S}}_{\text{full}})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}}). \tag{A15}$$

Under the assumption of identical twins and continuous data generating distributions, with probability 1, there are exactly two units (one being a twin of the other), whose common covariate value $\boldsymbol{X}^{(1)}$ (say) maximizes the objective function in Equation A15. Therefore, the choosing group at the first stage selects one of these two identical twins randomly, and in the next stage, the other treatment selects the remaining twin. This respects the assignment mechanism of a matched-pair design.

Consider the next pairs of stages. The objective function of the choosing treatment group is given by:

$$(\boldsymbol{X}_i - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)})^\top \Big\{ \boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}_{\text{full}}^\top \underline{\boldsymbol{X}}_{\text{full}} - (1+\epsilon)\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} \Big\}^{-1} (\boldsymbol{X}_i - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}) \tag{A16}$$

Similar to the previous case, here also we have (with probability 1) exactly two units, one being a twin of the other, whose common covariate value $\boldsymbol{X}^{(2)}$ maximizes the objective function in Equation A16. Thus, the choosing group at the first stage of this pair selects one of these two twins randomly, and in the next stage, the other treatment chooses the remaining twin. Proceeding in this manner, it follows that, at the end of the selection process, each treatment group ends up selecting one twin randomly from $\frac{N}{2}$ identical twins, which is equivalent to a matched-pair design. This completes the proof.

□

Proof of Proposition 5.1

With equal-sized groups, by symmetry, every unit has an equal chance of belonging to one of the G treatment groups. That is, $P(Z_i = g) = \frac{1}{G}$ for all $g \in \{1, 2, ..., G\}$. Therefore,

$$\begin{aligned}\mathbb{E}\Big\{\frac{1}{n_g}\sum_{i:Z_i=g} Y_i^{\text{obs}}\Big|\boldsymbol{Y}(g)\Big\} &= \mathbb{E}\Big\{\frac{G}{N}\sum_{i=1}^{N}\mathbb{1}(Z_i = g)Y_i(g)\Big|\boldsymbol{Y}(g)\Big\} \\ &= \frac{G}{N}\sum_{i=1}^{N} P(Z_i = g)Y_i(g) \\ &= \frac{1}{N}\sum_{i=1}^{N} Y_i(g). \qquad \text{(A17)}\end{aligned}$$

Using linearity of expectations, the proposition follows from Equation A17.

Next, we derive the randomization-based variance of the estimated SATE. For simplicity, and without loss of generality, we consider the case with $G = 2$ treatment groups of equal size, and focus on the estimand $\text{SATE}_{2,1}$. Let the corresponding unbiased estimator be denoted by $\widehat{\text{SATE}}_{2,1}$. Let $W_i = \mathbb{1}(Z_i = 2)$ be the indicator that unit i belongs to group 2. Following the Neymanian decomposition in Mukerjee et al. (2018), Proposition A1 presents the closed-form expression of the variance of $\widehat{\text{SATE}}_{2,1}$.

Proposition A1.

$$\begin{aligned}&\text{Var}(\widehat{\text{SATE}}_{2,1}) \\ &= -\frac{1}{N(N-1)}\sum_{i=1}^{N}(Y_i(2) - Y_i(1) - \tau)^2 + \frac{1}{N^2}\left(\sum_{i=1}^{N} 2\{Y_i^2(1) + Y_i^2(2)\} + \right. \\ &\quad + 2\sum_{i<i'}\sum\left[Y_i(2)Y_{i'}(2)\left\{4\pi_{ii'}(2,2) - \frac{N}{N-1}\right\} + Y_i(1)Y_{i'}(1)\left\{4\pi_{ii'}(1,1) - \frac{N}{N-1}\right\}\right] \\ &\quad \left. - 2\sum_{i<i'}\sum\left[Y_i(2)Y_{i'}(1)\left\{4\pi_{ii'}(2,1) - \frac{N}{N-1}\right\} + Y_i(1)Y_{i'}(2)\left\{4\pi_{ii'}(1,2) - \frac{N}{N-1}\right\}\right]\right),\end{aligned}$$

where $\pi_{i,i'}(z, z') = P(Z_i = z, Z_{i'} = z')$, for $z, z' \in \{1, 2\}$.
Moreover, if $\pi_{i,i'}(z, z') > 0$ for all i, i' and z, z', then a conservative estimator of this variance

is given by,

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{\mathrm{SATE}}_{2,1}) = \frac{1}{N^2}\Bigg(&\sum_{i=1}^{N} 4(Y_i^{\mathrm{obs}})^2 \\ &+ 2\sum_{i<i'}\sum \left[\frac{W_i W_{i'} Y_i^{\mathrm{obs}} Y_{i'}^{\mathrm{obs}}}{\pi_{ii'}(2,2)}\left\{4\pi_{ii'}(2,2) - \frac{N}{N-1}\right\}\right. \\ &\qquad \left. + \frac{(1-W_i)(1-W_{i'})Y_i^{\mathrm{obs}} Y_{i'}^{\mathrm{obs}}}{\pi_{ii'}(1,1)}\left\{4\pi_{ii'}(1,1) - \frac{N}{N-1}\right\}\right] \\ &- 2\sum_{i<i'}\sum \left[\frac{W_i(1-W_{i'})Y_i^{\mathrm{obs}} Y_{i'}^{\mathrm{obs}}}{\pi_{ii'}(2,1)}\left\{4\pi_{ii'}(2,1) - \frac{N}{N-1}\right\}\right. \\ &\qquad \left. + \frac{(1-W_i)W_{i'}Y_i^{\mathrm{obs}} Y_{i'}^{\mathrm{obs}}}{\pi_{ii'}(1,2)}\left\{4\pi_{ii'}(1,2) - \frac{N}{N-1}\right\}\right]\Bigg),\end{aligned}$$

This estimator is unbiased when treatment effect is constant across units, i.e., $Y_i(2) - Y_i(1) = c$ for all $i \in \{1, 2, ..., N\}$, where c is a constant.

When the condition $\pi_{i,i'}(z, z') > 0$ is violated for some i, i', z, z', we can still obtain a conservative variance estimator. For instance, suppose $\pi_{ii'}(1,1) = 0$. In this case, following Aronow and Samii (2013), we can upper bound the term $Y_i(2)Y_{i'}(2)\left\{4\pi_{ii'}(2,2) - \frac{N}{N-1}\right\} = -Y_i(2)Y_{i'}(2)\frac{N}{N-1}$ by $\frac{N}{2(N-1)}\{Y_i^2(2) + Y_{i'}^2(2)\}$, which admits an unbiased estimator given by $\frac{N}{N-1}W_i\{(Y_i^{\mathrm{obs}})^2 + (Y_{i'}^{\mathrm{obs}})^2\}$.

C Properties of D-optimal selection function

C.1 Affine invariance and covariate balance

Theorem A2. (a) The FSM with the D-optimal selection function is invariant under affine transformations of the covariate vector.

(b) For continuous, symmetrically distributed covariates and two groups of equal size, the FSM with the D-optimal selection function almost surely produces exact mean-balance on all even transformations of the centered covariate vector.

Proof of Theorem A2

Proof. (a) We consider the case where $\tilde{n}_{r-1} \geq 1$ and $\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1}$ is invertible. The proofs for the other two cases are similar. By Theorem 4.1, in this case, the chosen unit i' satisfies,

$$i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})^{\top}(\underline{\boldsymbol{S}}_{r-1})^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}). \tag{A18}$$

Consider an affine transformation of the covariate $\boldsymbol{X}$ given by $\boldsymbol{U} = \underline{\boldsymbol{A}}\boldsymbol{X} + \boldsymbol{b}$, where $\underline{\boldsymbol{A}}$ is a $k \times k$ invertible matrix and $\boldsymbol{b}$ is a vector of dimension k. Let the corresponding values of $\bar{\boldsymbol{X}}_{r-1}$ and $\underline{\boldsymbol{S}}_{r-1}$ be $\bar{\boldsymbol{U}}_{r-1}$ and $\underline{\boldsymbol{S}}_{U,r-1}$, respectively. We observe that,

$$\begin{aligned}(\boldsymbol{U}_i - \bar{\boldsymbol{U}}_{r-1})^\top (\underline{\boldsymbol{S}}_{U,r-1})^{-1} (\boldsymbol{U}_i - \bar{\boldsymbol{U}}_{r-1}) &= \{\underline{\boldsymbol{A}}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})\}^\top (\underline{\boldsymbol{A}}\underline{\boldsymbol{S}}_{r-1}\underline{\boldsymbol{A}}^\top)^{-1} \underline{\boldsymbol{A}}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}) \\ &= (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})^\top (\underline{\boldsymbol{S}}_{r-1})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}). \end{aligned} \tag{A19}$$

This shows that the D-optimal selection function remains unchanged under affine transformations and hence, FSM with the D-optimal selection function is affine invariant.

(b) The in-sample symmetry of the data essentially implies that if $\boldsymbol{X}$ belongs to the sample, then $-\boldsymbol{X}$ also belongs to the sample. Moreover, by the assumption of a continuous data generating distribution, with probability 1, the covariate values are different up to reflection. Now, consider an even transformation $g(\cdot)$, i.e., $g(-\boldsymbol{X}) = g(\boldsymbol{X})$. With two groups of equal sizes, the SOM consists of successive random permutations of the 'chunk' $(1, 2)$. By Theorem 4.1, for the first pair of stages of selection, the objective function (to maximize) is given by

$$(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}})^\top (\underline{\boldsymbol{S}}_{\text{full}})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}}) = \boldsymbol{X}_i^\top (\underline{\boldsymbol{S}}_{\text{full}})^{-1} \boldsymbol{X}_i. \tag{A20}$$

It follows that, if a unit in the sample with covariate $\boldsymbol{X}^{(1)}$ maximizes the objective function in Equation A20, then so does the unit with covariate $-\boldsymbol{X}^{(1)}$. Moreover, due to the continuous data generating distribution, with probability 1, these are the only two units that maximize this objective function. Therefore, if treatment 1 selects the unit with covariate $\boldsymbol{X}^{(1)}$, treatment 2 selects the unit with covariate $-\boldsymbol{X}^{(1)}$, and vice-versa. This preserves exact balance on $g(\boldsymbol{X})$.

Now, consider the next pair of stages. Without loss of generality, suppose treatment 1 had chosen a unit with covariate $\boldsymbol{X}^{(1)}$ and treatment 2 had chosen a unit with covariate $-\boldsymbol{X}^{(1)}$ in their respective previous choices. Also, without loss of generality, assume that in this pair of stages, treatment 1 gets to choose first. By Theorem 4.1, treatment 1 aims to maximize,

$$\begin{aligned}&(\boldsymbol{X}_i - \bar{\boldsymbol{X}}^*_{r-1})^\top (\underline{\boldsymbol{S}}^*_{r-1})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}^*_{r-1}) \\ &= \{\boldsymbol{X}_i - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}\}^\top \Big\{\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}^\top_{\text{full}}\underline{\boldsymbol{X}}_{\text{full}} - (1+\epsilon)\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top}\Big\}^{-1} \{\boldsymbol{X}_i - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}\}. \end{aligned} \tag{A21}$$

Also, during treatment 2's turn in this pair of stages, it tries to maximize

$$(\boldsymbol{X}_i + \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)})^\top \Big\{ \boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}_{\text{full}}^\top \underline{\boldsymbol{X}}_{\text{full}} - (1+\epsilon)\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} \Big\}^{-1} (\boldsymbol{X}_i + \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)})$$
$$= \{(-\boldsymbol{X}_i) - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}\}^\top \Big\{ \boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}_{\text{full}}^\top \underline{\boldsymbol{X}}_{\text{full}} - (1+\epsilon)\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} \Big\}^{-1} \{(-\boldsymbol{X}_i) - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}\}. \tag{A22}$$

Equations A21 and A22 imply that if treatment 1 chooses a unit with covariate value $\boldsymbol{X}^{(2)}$, then with probability 1, treatment 2 chooses the unit with covariate value $-\boldsymbol{X}^{(2)}$, and vice versa. This shows that, at the end of the second pair of stages in the SOM, exact mean balance on $g(\boldsymbol{X})$ is preserved. Proceeding in this manner it follows that, at the end of the selection process, with probability 1, both the treatment groups will have exact balance on $g(\boldsymbol{X})$. This completes the proof.

□

It follows from Theorem A2(a) that, for any SOM, the choices made by each treatment group remain unchanged even if the covariate vectors are transformed via an affine transformation (e.g., changing the units of measurement of the covariates). Therefore, the FSM with the D-optimal selection function self-standardizes the covariates. In addition, if the covariate vector is symmetrically distributed in the sample, then by Theorem A2(b), the FSM exactly balances even transformations such as the second, fourth order moments, and the pairwise products of the centerd covariates. An implication of Theorem A2(b) is that, for covariates drawn from symmetric continuous distributions (such as the Normal, t, and Laplace distributions), the FSM tends to balance all these transformations due to the approximate symmetry of the covariates in the sample. The choice of the D-optimal selection function is thus robust in the sense that it allows the FSM to balance a family of transformations of the covariate vector by design, without explicitly including them in the assumed linear model nor requiring the specification of tuning parameters.

C.2 Connection to A-optimality

The original FSM used a criterion based on A-optimality as the selection function (see Morris 1979). In this section, we compare the A-and D-optimal selection functions. The A-optimal selection function requires prespecifying a *policy matrix* $\underline{\boldsymbol{P}}_{p\times(k+1)}$ and a corresponding vector of *policy weights* $\boldsymbol{w}_{p\times 1}$. Here, $\underline{\boldsymbol{P}}$ transforms the original vector of regression coefficients to a vector of p linear combinations that are of policy interest, and $\boldsymbol{w}$ assigns weights to each combination according to their importance. Thus, the A-optimal selection function requires $p(k+2)$ tuning parameters.

If treatment 1 gets to choose at the rth stage, then this criterion selects the unit that minimizes the resulting trace$\Big\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r,i}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r,i})^{-1}\Big\}$, where $\underline{\boldsymbol{T}} = \underline{\boldsymbol{P}}^{\top}\text{diag}(\boldsymbol{w})\underline{\boldsymbol{P}}$. Proposition A3 shows an equivalent characterization of the A-optimal selection function.

Proposition A3. Let treatment 1 be the choosing group at the rth stage. Assume that $\tilde{n}_{r-1} \geq 1$ and $\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1}$ is invertible. The A-optimal selection function chooses unit i' with covariate vector $\boldsymbol{X}_{i'} \in \mathbb{R}^k$, where $i' \in \arg\max_{i\in\mathcal{R}_{r-1}} \frac{(1,\boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)}{1+(1,\boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)}$.

The A-optimality criterion provides a family of selection functions depending on $\underline{\boldsymbol{P}}$ and $\boldsymbol{w}$. For some choices of $\underline{\boldsymbol{P}}$ and $\boldsymbol{w}$, the selection function is not affine invariant , e.g., $\underline{\boldsymbol{P}} = \boldsymbol{I}$ and $\boldsymbol{w} = (1, 1, ..., 1)^{\top}$, while for other choices it is, e.g., $\underline{\boldsymbol{P}} = \underline{\tilde{\boldsymbol{X}}}_{\text{full}}$ and $\boldsymbol{w} = (1, 1, ..., 1)^{\top}$. In particular, the A-optimal selection function with $\underline{\boldsymbol{P}} = \underline{\tilde{\boldsymbol{X}}}_{\text{full}}$ and $\boldsymbol{w} = (1, 1, ..., 1)^{\top}$ is closely related to the D-optimal selection function. To see this, consider a case where in the selection process, the design matrices in each treatment group scale similarly relative to the design matrix in the full sample, i.e., $(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} = c_r(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^{\top}\underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1}$ for some constant $c_r > 0$. In this case, the A-optimal selection function chooses unit i' such that $i' \in \arg\max_{i\in\mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right) \iff i' \in \arg\max_{i\in\mathcal{R}_{r-1}} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})^{\top}(\underline{\boldsymbol{S}}_{r-1})^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})$, which is equivalent to the D-optimal selection function. Hence, in this case, the FSM under the D-optimal and A-optimal selection functions make similar choices of units.

Proof of Proposition A3

Proof. The A-optimal selection function aims to minimize

$$\begin{aligned}
&\text{trace}\left\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r,i}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r,i})^{-1}\right\} \qquad \text{(A23)}\\
&= \text{trace}\Big[\underline{\boldsymbol{T}}\{\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1} + \left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)(1, \boldsymbol{X}_i^{\top})\}^{-1}\Big]\\
&= \text{trace}\Big\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} - \underline{\boldsymbol{T}}\frac{(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)(1, \boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}}{1 + (1, \boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)}\Big\}\\
&= \text{trace}\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\} - \text{trace}\Big\{\underline{\boldsymbol{T}}\frac{(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)(1, \boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}}{1 + (1, \boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)}\Big\}\\
&= \text{trace}\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\} - \text{trace}\Big\{\frac{(1, \boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)}{1 + (1, \boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)}\Big\} \qquad \text{(A24)}
\end{aligned}$$

□

Here the second equality holds due to the Sherman-Morrison-Woodbury formula, the third and fourth equality hold due to the linearity and cyclicality of trace$(\cdot)$, respectively. Equation

A24 shows that it is equivalent to maximize trace$\left\{ \frac{(1,\boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\boldsymbol{T}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\binom{1}{\boldsymbol{X}_i}}{1+(1,\boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\binom{1}{\boldsymbol{X}_i}} \right\}$. This completes the proof.

D Optimal covariance design theorem and D-optimality

In this section, we focus on the setting with $G = 2$ treatment groups. Under a model-based approach, we first connect the notion of covariate balance to efficiency using the optimal covariance design theorem (Morris and Hill 2000, see also Chattopadhyay et al. 2021)

Theorem A4. Consider the linear regression model $Y_i^{\text{obs}} = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_i + \tau \mathbb{1}(Z_i = 2) + \epsilon_i$, where ϵ_is are the uncorrelated error terms with mean zero and variance σ^2. Let $\hat{\tau}_{\text{OLS}}$ be the ordinary least squares estimator of τ. Then,

$$\text{Var}(\hat{\tau}_{\text{OLS}}) = \frac{\sigma^2}{N s_2^2 (1 - R^2)},$$

where $s_2^2 = \frac{n_1 n_2}{N^2}$ and R^2 is the square of the multiple correlation coefficient of $\mathbb{1}(Z_i = 2)$ with the covariates.

Here, $\hat{\tau}_{\text{OLS}}$ is used to estimate the average treatment effect of treatment 2, relative to treatment 1. Theorem A4 implies that, under this model, the most efficient design minimizes R^2. In other words, the optimal design satisfies $R^2 = 0$ (if feasible), which equivalently means that the covariates $\boldsymbol{X}_i$ are exactly mean-balanced across the two treatment groups. Indeed, the optimality of this design is optimal depends heavily on the correctness of the outcome model. With model misspecification, this design may no longer be efficient. For instance, if the outcome model is linear in second-order transformations of the covariates, the design may perform poorly due to potential lack of balance on these transformations. In this sense, deterministic optimal designs lack robustness against model misspecification.

Next, we consider the global D-optimal design, i.e., the design that selects the D-optimal assignment among all possible assignments. If there are multiple D-optimal assignments, one of them is chosen randomly by the design. Proposition A5 shows that, with $k = 1$ covariate, the global D-optimal design aims to balance the means of the covariate exactly between the two treatment groups.

Proposition A5. Consider the linear model $Y_i^{\text{obs}} = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_i + \tau \mathbb{1}(Z_i = 2) + \epsilon_i$, where ϵ_is are the uncorrelated error terms with mean zero and variance σ^2. Under this model, the global D-optimal design minimizes $|\bar{X}_1 - \bar{X}_2|$, where $\bar{X}_1$ and $\bar{X}_2$ are the means of X_i in treatment groups 1 and 2, respectively.

Proposition A5 and Theorem A4 imply that, if the outcome model is linear in the covariates

and the treatment, then the global D-optimal design is the most efficient.

Proof of Proposition A5

By definition, the D-optimal design maximizes $\det(\underline{\boldsymbol{D}}^\top \underline{\boldsymbol{D}})$, where $\underline{\boldsymbol{D}} = (\mathbf{1}, \boldsymbol{X}, \boldsymbol{Z})$ is the design matrix. Without loss of generality, we assume that the covariates are scaled so that their variance in the full sample is 1, i.e., $\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}_{\text{full}})^2 = 1$. Then,

$$\begin{aligned}
\det(\underline{\boldsymbol{D}}^\top \underline{\boldsymbol{D}}) &= \det \begin{pmatrix} N & N\bar{X}_{\text{full}} & n_1 \\ N\bar{X}_{\text{full}} & N + N\bar{X}_{\text{full}} & n_1\bar{X}_1 \\ n_1 & n_1\bar{X}_1 & n_1 \end{pmatrix} \\
&= N^2 n_1 - N n_1^2 - N n_1^2(\bar{X}_{\text{full}}^2 + \bar{X}_1^2 - 2\bar{X}_{\text{full}}\bar{X}_1) \\
&= N n_1 n_2 - \frac{n_1^2 n_2^2}{N}(\bar{X}_1 - \bar{X}_2)^2, \qquad \text{(A25)}
\end{aligned}$$

where the last equality holds since $\bar{X}_{\text{full}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{N}$. Thus, maximizing $\det(\underline{\boldsymbol{D}}^\top \underline{\boldsymbol{D}})$ is equivalent to minimizing $|\bar{X}_1 - \bar{X}_2|$. This completes the proof.

E Algorithms for constructing an SOM

E.1 The SCOMARS algorithm

Consider a setting with $G = 2$ treatment groups of arbitrary sizes n_1 and n_2. Let W_r be the binary indicator for selection of group 1 stage r, $r \in \{1, 2, \ldots, N\}$, with $p_r := P(W_r = 1)$ being the marginal probability of selection at stage r. Write $S_r := \sum_{j=1}^{r} W_j$ and $F_r := \mathbb{E}(S_r) = \sum_{j=1}^{r} p_j$. A treatment assignment is sequentially controlled if $|S_r - F_r| < 1$ for all $r \in \{1, 2, \ldots, N\}$.

The SCOMARS algorithm proceeds as follows:

- Stage 1, $P(W_1 = 1) = p_1$.
- Stage $r \geq 2$, $P(W_r = 1 | S_{r-1} = s_{r-1}) = P\Big\{U \leq \frac{p_r - \max(0, s_{r-1} - F_{r-1})}{1 - |s_{r-1} - F_{r-1}|}\Big\}$, where $U \sim \text{Unif}(0, 1)$.

This algorithm satisfies the sequentially controlled condition, $|S_r - F_r| < 1$ for all $r \in \{1, 2, \ldots, N\}$ (Morris 1983). It is Markovian because the probability of selection at stage r depends solely on stage $r - 1$.

E.2 SOM for multi-group experiments

We first define the randomized chunk algorithm for generating an SOM for multi-group experiments with equal group sizes.

Definition 2 (Randomized chunk algorithm)**.** Suppose $n_1 = n_2 = ... = n_G$. The randomized chunk algorithm generates an SOM by generating and stacking $\frac{N}{G}$ independent random permutations of the 'chunk' $(1, 2, ..., G)$.

For example, with $N = 12$, $g = 3$, $n_1 = n_2 = n_3 = 4$, one instance of an SOM generated using randomized chunk is $(\underbrace{2,1,3}, \underbrace{1,2,3}, \underbrace{2,1,3}, \underbrace{2,3,1})^\top$.

The following proposition shows that the randomized chunk algorithm is sequentially controlled.

Proposition A6. For $G \geq 2$ and $n_1 = n_2 = ... = n_G$, the randomized chunk algorithm satisfies $|S_{ig} - F_{ig}| \leq \frac{G-1}{G} < 1$ for all $g \in \{1, 2, ..., G\}$.

Proof. Let S_{ig} and F_{ig} be the same as defined in Section 8.1 ($i \in \{1, 2, ..., N\}$, $g \in \{1, 2, ..., G\}$). For equal sized treatment groups, $F_{ig} = \frac{i}{G}$. Now, without loss of generality, it suffices to show that $|S_{i1} - F_{i1}| \leq \frac{G-1}{G}$ for all $i \in \{1, 2, ..., N\}$. Consider the first chunk in the SOM, which is a random permutation of $(1, 2, ..., G)$. If treatment 1 appears in position $i^* \in \{1, 2, ..., G\}$ the permutation ($j \in \{1, 2, ..., G\}$), then

$$|S_{i1} - F_{i1}| = \begin{cases} \frac{i}{G} & \text{if } i \in \{1, ..., i^* - 1\} \\ 1 - \frac{i}{G} & \text{if } i \in \{i^*, ..., G\}. \end{cases} \tag{A26}$$

In each case, $|S_{i1} - F_{i1}| \leq \frac{G-1}{G}$ for all $i \in \{1, 2, ..., G\}$. Moreover, since $|S_{G1} - F_{G1}| = 0$, the SOM restarts itself after the first chunk. Hence, we can conclude that $|S_{i1} - F_{i1}| \leq \frac{G-1}{G}$ for all $i \in \{1, 2, ..., G\}$. This completes the proof.

□

Below we describe two algorithms to generate an SOM for multi-group experiments and show that they are sequentially controlled. The key idea in these algorithms is the formation of 'supergroups', i.e., combination of one or more treatment groups. For example, with $g = 3$, $n_1 = 10, n_2 = 20, n_3 = 30$, one can consider two supergroups, namely $\{1, 2\}$ of size $10 + 20 = 30$ and $\{3\}$ of size 30.

Theorem A7. For $1 \leq G_1 \leq G - 1$, let $n_1 = n_2 = ... = n_{G_1} \neq n^{(1)}$, and $n_{G_1+1} = n_{G_1+2} = ... = n_G = n^{(2)}$, where $n^{(1)} \neq n^{(2)}$. Consider the following three-stage algorithm.

1. Run SCOMARS with supergroups $\{1, ..., G_1\}$ and $\{G_1 + 1, ..., G\}$ to generate an SOM at the supergroup level.

2. Consider the locations of the SOM in step 1 where supergroup $\{1, ..., G_1\}$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

3. Consider the locations of the SOM in step 1 where supergroup $\{G_1 + 1, ..., G\}$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

The above SOM generating algorithm is sequentially controlled.

We first prove a special case of Theorem A7, given below in Lemma A2

Lemma A2. The algorithm in Theorem A7 is sequntially controlled for the special case of $G_1 = 1$.

Proof. The first step of the algorithm in Theorem A7 runs SCOMARS with treatment group 1 and the supergroup $\{2, 3, ..., G\}$. Thus, the first step itself determines the locations of the SOM where treatment 1 gets to choose. Since SCOMARS is sequentially controlled, we immediately have $|S_{i1} - F_{i1}| < 1$ for all $i \in \{1, 2, ..., N\}$.

It remains to show that for $g \in \{2, 3, ..., G\}$, $|S_{ig} - F_{ig}| < 1$ for all $i \in \{1, 2, ..., N\}$. By symmetry, it suffices to show this for $g = 2$. Now, the randomized chunk algorithm on the supergroup $\{2, 3, ..., G\}$ determines the locations of the SOM where treatment 2 gets to choose. We will prove the result by first mapping this SOM to an SOM where treatment 1 is absent, and then by using the sequential controlled property of randomized chunk.

Let us first denote $1 \leq r_1 < r_2 < ... < r_{n_1-1} < r_{n_1} \leq N$ as the stages or locations of the SOM where treatment 1 gets to choose. We consider the following cases,

(i) Case-1: $i \in \{1, 2, ..., r_1 - 1\}$. In this case, by stage i, treatment 1 has not made any choices. Now,

$$\begin{aligned} |S_{i2} - F_{i2}| &= |S_{i2} - \frac{in^{(2)}}{N}| \\ &\leq |S_{i2} - \frac{i}{G-1}| + |\frac{i}{G-1} - \frac{in^{(2)}}{N}| \\ &\leq \frac{G-2}{G-1} + i\frac{n_1}{N(G-1)} \\ &< \frac{G-2}{G-1} + \frac{1}{G-1} = 1. \end{aligned} \tag{A27}$$

Here the first inequality holds due to triangle inequality. To see that second inequality,

consider a new experiment with treatment groups $\{2, ..., G\}$ of size $n^{(2)}$ each and an SOM generated by randomized chunk as in the second step of the algorithm in Theorem A7. Let $\tilde{S}_{i2}$ be the number of selections made by treatment 2 up to stage i in this new experiment and $\tilde{F}_{i2} = \frac{i}{G-1}$ be its expectation. By Proposition A6, $|\tilde{S}_{i2} - \tilde{F}_{i2}| \leq \frac{G-2}{G-1}$. Now, $|S_{i1} - \frac{i}{G-1}| = |\tilde{S}_{i1} - \frac{i}{G-1}|$, which gives us the second inequality. Finally, the last inequality holds since $\frac{in_1}{N} = F_{i1} < 1$.

(ii) Case-2: $i \in \{r_t, r_t + 1, ..., r_{t+1} - 1\}$ for some $t \in \{1, 2, ..., n_1 - 1\}$. In this case, by stage i, treatment 1 has made exactly t choices. Now,

$$\begin{aligned} |S_{i2} - F_{i2}| &= |S_{i2} - \frac{in^{(2)}}{N}| \\ &\leq |S_{i2} - \frac{i-t}{G-1}| + |\frac{i-t}{G-1} - \frac{in^{(2)}}{N}| \\ &\leq \frac{G-2}{G-1} + \frac{1}{G-1}|t - \frac{in_1}{N}| \\ &< \frac{G-2}{G-1} + \frac{1}{G-1} = 1. \end{aligned} \tag{A28}$$

Here, the first inequality is due to triangle inequality. To see the second inequality, we again consider the new experiment described in Case-1. Notice that, $|S_{i2} - \frac{i-t}{G-1}| = |\tilde{S}_{(i-t)2} - \tilde{F}_{(i-t)2}| \leq \frac{G-2}{G-1}$, where the last inequality holds by Proposition A6. Finally, the final inequality in Equation A28 holds since $|t - \frac{in_1}{N}| = |S_{i1} - F_{i1}| < 1$. This completes the proof of the lemma. □

We now prove Theorem A7.

Proof. We first show that, for $g \in \{1, 2, ..., G_1\}$,

$$|S_{ig} - F_{ig}| < 1 \quad \forall i \in \{1, 2, ..., N\}. \tag{A29}$$

To show this, we consider steps 1 and 2 of the algorithm as these two steps are sufficient to determine the location of treatments $1, ..., G_1$ in the SOM. We note that, steps 1 and 2 generate an SOM for an experiment with $G_1 + 1$ treatment groups, namely supergroup $\{G_1 + 1, ..., G\}$ (of size $(G - G_1)n^{(2)}$) and groups $1, 2, ..., G_1$ (each of size $n^{(1)}$). Thus, by Lemma A2, it follows that Equation A29 holds for $g \in \{1, 2, ..., G_1\}$.

To show that Equation A29 holds for $g \in \{G_1 + 1, ..., G\}$, we first notice that steps 2 and 3 of the algorithm are completely independent and hence can be performed in any order. Therefore, by changing the order of steps 2 and 3 and applying the same argument as before,

we get that Equation A29 holds for $g \in \{G_1 + 1, ..., G\}$. This completes the proof of the theorem. □

Theorem A8. Let $G_1, ..., G_m$ be such that $1 \leq G_j \leq G - 1$ for all $j \in \{1, 2, ..., m\}$ and $G_1 + G_2 + ... + G_m = G$. Moreover, for $j \in \{1, 2, ..., m\}$, let $n^{(j)}$ be the group size of G_j many treatment groups, with $n^{(1)}G_1 = n^{(2)}G_2 = ... = n^{(m)}G_m$. Denote the collection of G_j treatment groups with group sizes $n^{(j)}$ as supergroup $\mathcal{G}_j$. Consider the following multi-stage algorithm.

1. Run randomized chunk on supergroups $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_m$ to generate an SOM at the supergroup level.
2. For $j \in \{1, 2, ..., m\}$, consider the locations of the SOM in step 1 where supergroup $\mathcal{G}_j$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

The above SOM generating algorithm is sequentially controlled.

To prove this theorem, we first use the following Lemma.

Lemma A3. Let $n_1 = n_2 = ... = n_G = n$. Consider the following SOM generating algorithm.

1. Consider the supergroups $\{1\}$ (of size n) and $\{2, 3, ..., G\}$ (of size $(G-1)n$). Generate an SOM at the superpopulation level using SCOMARS.
2. Consider the locations of the SOM in step 1 where supergroup $\{2, 3, ..., G\}$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

This algorithm is equivalent to the randomized chunk algorithm.

Below we prove this lemma.

Proof. To show that the algorithm is equivalent to randomized chunk, we have to show that it generates a random permutation of $(1, 2, ..., G)$ for the first G stages, a fresh random permutation of $(1, 2, ..., G)$ for the next G stages, and so on. Since the locations of groups $\{2, ..., G\}$ are chosen using randomized chunk, it thus suffices to show that, treatment 1 gets to choose once (in a random location) in the first G stages, once in the next G stages, and so on.

We use the notation as in Section E.1. Now, suppose among the first G stages, treatment 1

gets to choose at stage r^* first. Notice that r^* cannot be greater than G as

$$P(W_G = 1|S_{G-1} = 0) = P\Big\{U \leq \frac{\frac{1}{G} - \max(0, 0 - F_{G-1})}{1 - |0 - F_{G-1}|}\Big\} = P\Big\{U \leq \frac{1}{G - (G-1)}\Big\} = 1. \tag{A30}$$

Now, for $r \in \{1, 2, ..., r^* - 1\}$ we have,

$$P(W_r = 1|S_{r-1} = 0) = P\Big\{U \leq \frac{\frac{1}{G} - \max(0, 0 - F_{r-1})}{1 - |0 - F_{r-1}|}\Big\} = P\Big\{U \leq \frac{1}{G - (r-1)}\Big\} = \frac{1}{G - (r-1)}. \tag{A31}$$

For $r^* + 1 \leq r \leq G$,

$$P(W_r = 1|S_{r-1} = 1) = P\Big\{U \leq \frac{p_r - \max(0, 1 - F_{r-1})}{1 - |1 - F_{r-1}|}\Big\} = P\Big\{U \leq \frac{\frac{1}{G} - 1 + \frac{r-1}{G}}{\frac{r-1}{G}}\Big\} = 0. \tag{A32}$$

Finally,

$$P(W_{G+1} = 1|S_G = 1) = P\Big\{U \leq \frac{p_{G+1} - \max(0, 1 - F_G)}{1 - |1 - F_G|}\Big\} = P\Big(U \leq \frac{1}{G}\Big) = \frac{1}{G}. \tag{A33}$$

Therefore, by Equation A32, if treatment 1 selects at the r^*th stage, it never selects again $2, 3, ..., G$. Also, by Equation A31, before the r^*th stage, the conditional probabilities of treatment 1 selecting are same as what it would have been under random permutation of the group labels. Finally, by Equation A33, the process restarts itself at the $(G+1)$th stage, which is equivalent to starting a fresh new random permutation of the group labels. This completes the proof of the lemma. □

We now prove Theorem A8.

Proof. By the symmetry of the problem, it suffices to show that $|S_{i1} - F_{i1}| < 1$ for all $i \in \{1, 2, ..., N\}$. Without loss of generality, we assume that $\mathcal{G}_1 = \{1, 2, ..., G_1\}$, which implies that treatment 1 belongs to supergroup $\mathcal{G}_1$. Now, it suffices to focus on the following to steps of the algorithm:

1. Run randomized chunk on supergroups $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_m$ to generate an SOM at the supergroup level.

2. Consider the locations of the SOM in step 1 where supergroup $\mathcal{G}_1$ chooses. Then, use

randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

This is because, these two steps completely determine the locations of treatment 1 in the SOM. By Lemma A3, these two steps can be equivalently performed as follows.

1. Consider the supergroups $\mathcal{G}_1$ (of size $n^{(1)}G_1$) and $\{\mathcal{G}_2, ..., \mathcal{G}_m\}$ (of size $(m-1)n^{(1)}G_1$). Generate an SOM at this supergroup level using SCOMARS.

2. Consider the locations of the SOM in step 1 where supergroup $\{\mathcal{G}_2, ..., \mathcal{G}_m\}$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of $\mathcal{G}_j$ in those locations.

3. Consider the locations of the SOM in step 1 where supergroup $\mathcal{G}_1$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

We note that this above algorithm is exactly equivalent to the SOM generating algorithm in Theorem A7 for an experiment with G_1+m-1 treatment groups, namely, $1, 2, ..., G_1, \mathcal{G}_2, \mathcal{G}_3, ..., \mathcal{G}_m$. Thus, by Theorem A7, we have $|S_{i1} - F_{i1}| < 1$ for all $i \in \{1, 2, ..., N\}$. □

F FSM for stratified experiments

In this section, we discuss two potential approaches to use an FSM for stratified experiments. We consider stratified experiments where the treatment group sizes within each stratum are set by the investigator beforehand. To accommodate the FSM to such experiments, we again need to carefully construct an SOM. In particular, we append the SOM with an additional column of stratum labels, indicating which stratum the treatment group selects from at each stage of the selection process. This column of stratum labels is specified in such a way that the resulting SOM satisfies the group size requirements within each stratum.

Conceptually, the most straightforward approach is to generate a separate SOM for each stratum. This is equivalent to setting the column of stratum labels as $(\underbrace{1, ..., 1}_{m_1}, \underbrace{2, ..., 2}_{m_2}, ..., \underbrace{S, ..., S}_{m_S})^\top$, where S is the number of strata and m_s is the size of sth stratum, $s \in \{1, 2, ..., S\}$. This approach is easy to implement and can be useful if, e.g., data on each stratum is available at different stages of the experiment, akin to a sequential experiment. However, in this approach, the treatment groups only get to explore the covariate space of a single stratum for a number of successive stages of selection and hence may not make the most efficient choices. We address this issue with an alternative approach. For ease of exposition, we consider two strata: 1 and 2. Let n_{1g} and n_{2g} be the (fixed) sizes of treatment group $g \in \{1, 2, ..., G\}$

in strata 1 and 2, respectively, where $n_{1g} + n_{2g} = n_g$. In this approach, we first generate a usual SOM with group sizes $n_1, ..., n_G$. For $g \in \{1, 2, ..., G\}$, we then select the order of the strata that treatment g chooses from by running a SCOMARS algorithm with group sizes n_{1g} and n_{2g}. By allowing the treatment groups to select units from different strata in a balanced manner, this approach mimics the unstratified FSM where the covariate space of the entire sample is explored for choosing units. Also, by design, this approach satisfies the size requirement of each treatment group within each stratum.

G FSM for sequential experiments

In this section, we describe our approach to using the FSM for sequential experiments. Suppose treatment 1 gets to choose at the first stage of selection for the new batch. Let $\underline{\tilde{\boldsymbol{X}}}_{\text{old}}$ be the design matrix based on units already assigned to treatment 1. Also, for each unit i in the new batch, let $\underline{\tilde{\boldsymbol{X}}}_{\text{new},i} := \begin{pmatrix} \underline{\tilde{\boldsymbol{X}}}_{\text{old}} \\ (1, \boldsymbol{X}_i^\top) \end{pmatrix}$ be the resulting design matrix in treatment group 1 if unit i is selected. Treatment 1 selects the unit that maximizes $\det(\underline{\tilde{\boldsymbol{X}}}_{\text{new},i}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{new},i})$. In other words, we use the design matrix based on all the units already assigned to the choosing treatment group to evaluate the D-optimal selection function for each unit in the new batch, and select the unit that maximizes the selection function. By carrying over the existing design matrix to the new batch, this approach tends to correct for any existing covariate imbalances.

H A simulation study

H.1 Setup

We now compare the performance of the FSM to complete randomization and rerandomization in a simulation study. Here, $N = 120$, $G = 2$, $n_1 = n_2 = 60$, and $k = 6$. The covariates are generated following the design of Hainmueller (2012):

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}_3 \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{pmatrix} \right\}, \; X_4 \sim \text{Unif}(-3, 3), \; X_5 \sim \chi_1^2, \; X_6 \sim \text{Bernoulli}(0.5). \tag{A34}$$

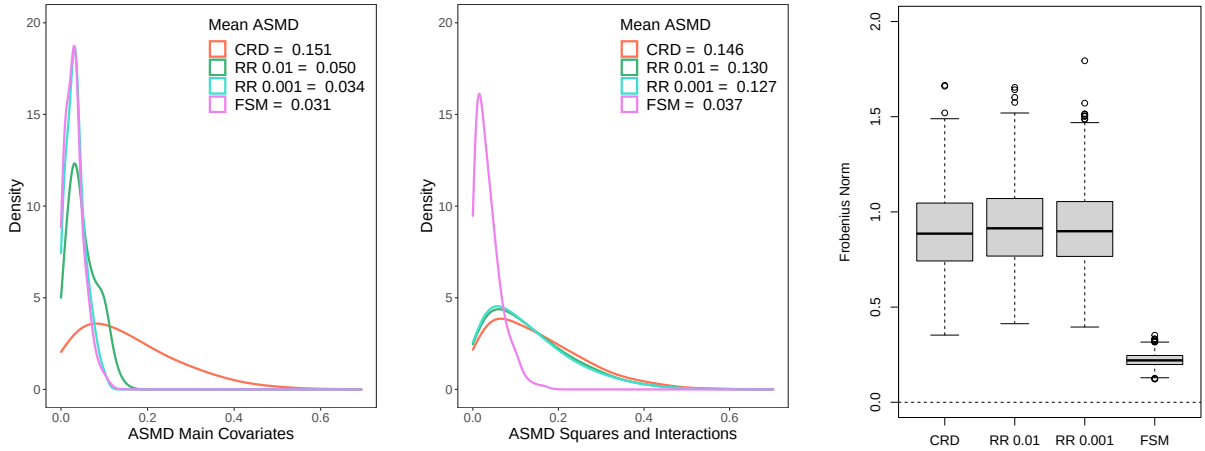
In this design, X_4, X_5, and X_6 are mutually independent and separately independent of $(X_1, X_2, X_3)^\top$. We draw a sample of 120 units once from the data generating mechanism in (A34). Conditional on this sample, we compare four different assignment methods, namely a completely randomized design (CRD), rerandomization with 0.01 acceptance rate (RR 0.01), rerandomization with 0.001 acceptance rate (RR 0.001), and the FSM. Both RR 0.01 and RR 0.001 use as rerandomization criteria the Mahalanobis distance between the two treatment groups on the original covariates. The FSM uses a linear potential outcome model on the original covariates and the D-optimal selection function. For each design we

draw 800 independent assignments. The assignments under the FSM are generated using the open source R package `FSM` available on CRAN. The total runtime of the FSM for the 800 simulated experiments was about one and a half minutes on a Windows 64-bit computer with an Intel(R) Core i7 processor. See Chattopadhyay et al. (2021) for detailed step-by-step instructions and vignettes on the use of FSM package.

H.2 Balance

We evaluate balance on the main and transformed covariates. Figures A1(a) and A1(b) show density plots of the Absolute Standardized Mean Differences (ASMD; Rosenbaum and Rubin 1985, Stuart 2010) of the six main covariates and their second-order transformations (including squares and pairwise products), respectively. A smaller ASMD for a covariate indicates better mean-balance on that covariate between the two treatment groups. Figure A1(a) indicates that both rerandomization methods improve balance on the means of the original covariates over CRD. As expected, the ASMD distribution under RR 0.001 is more concentrated than that of RR 0.01, with 32% smaller mean ASMD than RR 0.01. Both the FSM and RR 0.001 have similar distributions of the ASMD with FSM having moderately (9%) smaller mean ASMD. See Table A10 for a comparison of the average ASMD of each covariate.

Figure A1: Panels (a) and (b) show distributions of absolute standardized mean differences (ASMD) of the main covariates and all their second-order transformations across 800 randomizations. For each plot, the legend presents the average ASMD across simulations for the four methods. Panel (c) shows distributions of discrepancies between the correlation matrices of the covariates in the treatment and the control group (as measured by the Frobenius norm, $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$). On average the FSM achieves better covariate balance. In terms of the main covariates, the FSM marginally outperforms RR 0.001. In terms of the second-order transformations and correlation matrices, the FSM substantially outperforms RR 0.001.



(a) Main covariates (b) Squares and pairwise products (c) Frobenius norm

Figure A1(b) shows that the imbalances of covariate transformations are substantially smaller with the FSM than with CRD, RR 0.01, and RR 0.001. In fact, the FSM achieves a 70% reduction in the mean ASMD with respect to RR 0.001. Thus, although the FSM and RR 0.001 exhibit comparable balance in terms of the main covariates, the FSM balances these transformations of the covariates much better than RR 0.001. This highlights the improved robustness of the FSM against model misspecification. Moreover, reducing the tuning parameter of rerandomization from 0.01 to 0.001 yields only 2% improvement in the mean ASMD.[11] In Figure A1(b), both RR 0.01 and RR 0.001 often produce ASMD larger than 0.1, and in some cases, larger than 0.5, indicative of substantial imbalances on these covariate transformations.

For each method, we also compare balance in the overall correlation structure of the covariates. Figure A1(c) shows the boxplots of the distributions of $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$. The FSM outperforms the other three designs with at least 75% smaller average $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$. In particular, among the 800 randomizations, the highest value of $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$ under FSM is smaller than the corresponding lowest value under the other three designs, indicating that in terms of the correlation structure (and hence the interactions) of the covariates, the least balanced realization of the 800 FSMs exhibits better balance than the best balanced realization of the 800 complete randomizations and rerandomizations.

H.3 Efficiency

We now compare the efficiency of the methods under both model- and randomization-based approaches to inference. Under the model-based approach, we consider a potential outcome model where $\mathbb{E}\{Y_i(g)|\boldsymbol{X}_i\}$ is linear in $\boldsymbol{X}_i$ (Model A1), and another model where $\mathbb{E}\{Y_i(g)|\boldsymbol{X}_i\}$ is linear in $\boldsymbol{X}_i$ and all its second-order transformations (Model A2). In each case, we assume homoscedasticity, i.e., $\text{Var}\{Y_i(g)|\boldsymbol{X}_i\} = \sigma^2$. For each potential outcome model, we fit the corresponding observed outcome model by OLS and estimate $\text{PATE}_{2,1}$ using the regression imputation method described in Section 5.

More concretely, consider a specific treatment assignment vector $\boldsymbol{Z}$. Under Model A1, we fit a linear regression model $Y_i^{\text{obs}} = (1, \boldsymbol{X}_i)^\top \boldsymbol{\beta}_g + \epsilon_{ig}$ in treatment group $g \in \{1, 2\}$ and estimate $\text{PATE}_{2,1}$ by the regression imputation estimator $\widehat{\text{PATE}}_{2,1} = \hat{\boldsymbol{\beta}}_2^\top \bar{\tilde{\boldsymbol{X}}} - \hat{\boldsymbol{\beta}}_1^\top \bar{\tilde{\boldsymbol{X}}}$, where $\bar{\tilde{\boldsymbol{X}}}^\top = \frac{1}{N}\sum_{i=1}^N (1, \boldsymbol{X}_i^\top)$. The model-based standard error of this estimator is $\text{SE}_{\boldsymbol{Z}} = \sigma\sqrt{\bar{\tilde{\boldsymbol{X}}}^\top \{(\underline{\tilde{\boldsymbol{X}}}_{1,\boldsymbol{Z}}^\top \underline{\tilde{\boldsymbol{X}}}_{1,\boldsymbol{Z}})^{-1} + (\underline{\tilde{\boldsymbol{X}}}_{2,\boldsymbol{Z}}^\top \underline{\tilde{\boldsymbol{X}}}_{2,\boldsymbol{Z}})^{-1}\} \bar{\tilde{\boldsymbol{X}}}}$, where $\underline{\tilde{\boldsymbol{X}}}_{g,\boldsymbol{Z}}$ is the design matrix in group g, for the given treatment assignment $\boldsymbol{Z}$.

[11] In fact, for some covariate transformations, reducing this tuning parameter exacerbates imbalance (see Table A11).

Now, for a design d, the average and maximum model-based standard error relative to the FSM is given by $\frac{\frac{1}{M}\sum_{r=1}^{M} \mathrm{SE}_{\boldsymbol{Z}_d^{(r)}}}{\frac{1}{M}\sum_{r=1}^{M} \mathrm{SE}_{\boldsymbol{Z}_{\mathrm{FSM}}^{(r)}}}$ and $\frac{\max_r \mathrm{SE}_{\boldsymbol{Z}_d^{(r)}}}{\max_r \mathrm{SE}_{\boldsymbol{Z}_{\mathrm{FSM}}^{(r)}}}$, respectively, where $\boldsymbol{Z}_d^{(1)}, ..., \boldsymbol{Z}_d^{(M)}$ are M independent assignment vectors generated under design d, and $\boldsymbol{Z}_{\mathrm{FSM}}^{(1)}, ..., \boldsymbol{Z}_{\mathrm{FSM}}^{(M)}$ are generated under the FSM. These measures do not depend on σ^2 and can be computed exactly from the observed data. Tables A4(a) and A4(b) show the average and maximum model-based standard error (SE) of the regression imputation estimator relative to the FSM across $M = 800$ randomizations under the two models.

Table A4: Average and maximum model-based standard errors relative to the FSM across randomizations. Under Model A1 (linear model on the main covariates), the FSM and RR exhibit similar performance, improving over CRD. Under Model A2 (linear model on the main covariates and their second-order transformations), the FSM is considerably more efficient than both CRD and RR.

(a) Model A1

	Designs			
	CRD	RR 0.01	RR 0.001	FSM
Average SE	1.03	1.00	1.00	1.00
Maximum SE	1.13	1.00	1.00	1.00

(b) Model A2

	Designs			
	CRD	RR 0.01	RR 0.001	FSM
Average SE	1.39	1.27	1.26	1.00
Maximum SE	3.61	1.97	1.80	1.00

Under Model A1, since both rerandomization and the FSM are able to adequately balance the means of the original covariates, they lead to lower SE (hence, higher efficiency) than CRD. Across randomizations, the worst case SE under RR 0.01, RR 0.001, and the FSM are 13% smaller than under CRD. Under Model A1, the FSM has similar model-based SE as the two rerandomization methods. However, under Model A2, the FSM uniformly outperforms the other three designs, with a 26% reduction in average SE and an 80% reduction in maximum SE than RR 0.001. This improvement in efficiency can be attributed to the balance achieved by the FSM on the main covariates and their squares and pairwise products. In sum, when the model assumed at the design stage is correct and is used at the analysis stage, the FSM is as efficient as the two rerandomizations for estimating the treatment effect. However, when the model assumed at the design stage is misspecified and later corrected by augmenting transformations of the covariates (e.g., squares and pairwise products), the FSM is considerably more efficient and robust than the other designs.

Under the randomization-based approach, we compare the standard errors of the difference-in-means statistic under each design. Following Hainmueller (2012), the potential outcomes are generated using the models: $Y(1) = X_1 + X_2 + X_3 - X_4 + X_5 + X_6 + \eta$, $Y(2) = Y(1)$ (Model

B1) and $Y(1) = (X_1 + X_2 + X_5)^2 + \eta$, $Y(2) = Y(1)$ (Model B2), where $\eta \sim \mathcal{N}(0, 1)$. Both generative models satisfy the sharp-null hypothesis of zero treatment effect for every unit and hence, $\text{SATE}_{2,1} = 0$. Conditional on these potential outcomes, $\text{SATE}_{2,1}$ is estimated under each design using the standard difference-in-means estimator. The corresponding randomization-based SE of this estimator is obtained by generating 800 randomizations of the design and computing the standard deviation of the difference-in-means estimator across these 800 randomizations. Table A5 shows the randomization-based SE of the difference-in-means statistic for $\text{SATE}_{2,1}$ under each model.

Table A5: Randomization-based standard errors relative to the FSM. The standard error for the FSM is 0.2 under Model B1 (linear model on the main covariates) and 0.43 under Model B2 (linear model on the main covariates and their second-order transformations). Especially under Model B2, the FSM is considerably more efficient than both CRD and RR.

(a) Model B1

	Designs			
	CRD	RR 0.01	RR 0.001	FSM
SE	2.72	1.26	1.08	1

(b) Model B2

	Designs			
	CRD	RR 0.01	RR 0.001	FSM
SE	5.69	4.56	4.47	1

Under Model B1, the potential outcomes depend linearly on the covariates and therefore balancing the means of the covariates improves efficiency. This is reflected in Table A5 as the FSM has the smallest SE, closely followed by RR 0.001. Under Model B2, the potential outcomes depend linearly on the squares and pairwise products of the covariates. By better balancing these transformations, the FSM yields a considerably smaller SE than the other designs. In particular, under Model B2, the SE under the FSM is 67% smaller than the SE under RR 0.001. Therefore, as in the model-based approach, in the randomization-based approach the FSM exhibits comparable efficiency to rerandomization under correct-specification of the outcome model and considerable robustness under model misspecification.

H.4 Comparison with the global D-optimal design

In this section, we compare the performance of the FSM with the global D-optimal design (or simply, the D-optimal design), as defined in Section H.4. Obtaining the exact D-optimal assignment is an NP-hard problem in general, so we consider two alternatives. First, we randomly sample a large number (20000) of assignment vectors from the space of all possible assignments and obtain the D-optimal assignment among them. Due to random sampling, this assignment is expected to have similar properties (e.g., balance) as the exact D-optimal assignment. Second, we consider a random subsample of 20 units from the original sample

of 120 units. For this subsample, we compare FSM with the D-optimal assignment. In this case, both the designs assign units into two groups of size 10 each.

Tables A6 and A7 display the average ASMD values for the original covariates, as well as their squares and interactions, respectively, under the first scenario. Correspondingly, Tables A8 and A9 present these ASMD values under the second scenario.

Table A6: ASMD of the original covariates under the D-optimal design (D-opt), and the average ASMD of the original covariates under the FSM. The ASMD for the D-optimal design is approximated based on 20000 randomizations.

Covariates	Designs	
	D-opt	FSM
X_1	0.031	0.029
X_2	0.008	0.025
X_3	0.020	0.042
X_4	0.004	0.029
X_5	0.041	0.029
X_6	0.033	0.034
Mean	0.023	0.031

Table A7: ASMD of the squares and pairwise products of the covariates under the D-optimal design (D-opt), and the average ASMD of the same transformations under the FSM. The ASMD for the D-optimal design is approximated based on 20000 randomizations.

Covariate transformations	Designs	
	CRD	RR 0.01
X_1X_2	0.029	0.041
X_1X_2	0.038	0.041
X_1X_2	0.206	0.024
X_1X_2	0.074	0.035
X_1X_2	0.223	0.030
X_1X_2	0.057	0.051
X_1X_2	0.090	0.027
X_1X_2	0.027	0.030
X_1X_2	0.075	0.026
X_1X_2	0.329	0.032
X_1X_2	0.147	0.096
X_1X_2	0.087	0.035
X_1X_2	0.064	0.037
X_1X_2	0.091	0.027
X_1X_2	0.036	0.024
X_1^2	0.029	0.031
X_2^2	0.085	0.038
X_3^2	0.110	0.041
X_4^2	0.060	0.053
X_5^2	0.047	0.013
Mean	0.095	0.037

Table A8: ASMD of the original covariates in the sampled dataset under the D-optimal design (D-opt), and the average ASMD of the original covariates under the FSM.

Covariates	Designs	
	D-opt	FSM
X_1	0.022	0.191
X_2	0.071	0.142
X_3	0.036	0.213
X_4	0.016	0.147
X_5	0.054	0.194
X_6	0.000	0.051
Mean	0.033	0.156

Table A9: ASMD of the squares and pairwise products of the covariates in the sampled dataset under the D-optimal design (D-opt), and the average ASMD of the same transformations under the FSM.

Covariate transformations	Designs	
	CRD	RR 0.01
X_1X_2	0.825	0.353
X_1X_2	1.231	0.210
X_1X_2	0.484	0.162
X_1X_2	0.588	0.388
X_1X_2	0.727	0.230
X_1X_2	1.526	0.248
X_1X_2	0.765	0.095
X_1X_2	0.625	0.363
X_1X_2	0.477	0.264
X_1X_2	0.638	0.269
X_1X_2	0.740	0.392
X_1X_2	0.440	0.263
X_1X_2	0.559	0.404
X_1X_2	0.609	0.147
X_1X_2	0.238	0.063
X_1^2	0.952	0.200
X_2^2	0.116	0.365
X_3^2	0.833	0.313
X_4^2	0.566	0.167
X_5^2	0.019	0.233
Mean	0.648	0.256

From the above tables we observe that, on an average, the D-optimal design produces better balance on the main covariates. This observation is consistent with Proposition A5, which shows that with a single covariate, the D-optimal design aims to exactly balance its mean across the two groups. However, akin to randomization, it produces worse balance on the second-order transformations of the covariates.

H.5 Additional results from the simulation study

Table A10: Averages of the ASMD of the original covariates across 800 randomizations.

Covariates	Designs			
	CRD	RR 0.01	RR 0.001	FSM
X_1	0.162	0.051	0.035	0.029
X_2	0.156	0.048	0.033	0.025
X_3	0.158	0.049	0.033	0.042
X_4	0.150	0.049	0.034	0.029
X_5	0.140	0.052	0.034	0.029
X_6	0.141	0.052	0.036	0.034
Mean	0.151	0.050	0.034	0.031

Table A11: Averages of the ASMD of squares, pairwise products, and other transformations of the covariates across 800 randomizations.

Covariate transformations	Designs			
	CRD	RR 0.01	RR 0.001	FSM
X_1X_2	0.144	0.153	0.148	0.041
X_1X_3	0.144	0.140	0.137	0.041
X_1X_4	0.141	0.148	0.147	0.023
X_1X_5	0.150	0.135	0.134	0.035
X_1X_6	0.152	0.109	0.101	0.030
X_2X_3	0.147	0.147	0.146	0.051
X_2X_4	0.140	0.155	0.150	0.027
X_2X_5	0.147	0.143	0.136	0.030
X_2X_6	0.152	0.115	0.104	0.026
X_3X_4	0.141	0.143	0.152	0.032
X_3X_5	0.149	0.140	0.139	0.096
X_3X_6	0.148	0.099	0.091	0.035
X_4X_5	0.148	0.132	0.130	0.037
X_4X_6	0.152	0.100	0.095	0.027
X_5X_6	0.146	0.095	0.094	0.024
X_1^2	0.140	0.145	0.143	0.031
X_2^2	0.151	0.155	0.150	0.038
X_3^2	0.144	0.136	0.132	0.041
X_4^2	0.143	0.145	0.147	0.053
X_5^2	0.142	0.073	0.067	0.013
Mean	0.146	0.130	0.127	0.037
$X_5^{1.5}$	0.141	0.060	0.048	0.018
X_2^3	0.155	0.090	0.081	0.071
X_4^4	0.140	0.143	0.147	0.072
$\frac{1}{4+X_3}$	0.157	0.073	0.064	0.050
Mean	0.148	0.092	0.085	0.053

Figure A2: Boxplot of the distribution of $||\underline{\boldsymbol{S}}_1 - \underline{\boldsymbol{S}}_2||_F$ across 800 randomizations, where $\underline{\boldsymbol{S}}_g$ is the sample covariance matrix of the covariates in treatment group $g \in \{1, 2\}$.

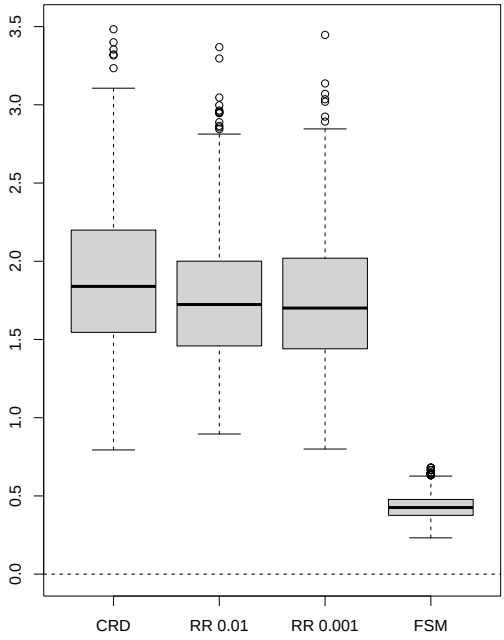


Table A12: Averages of the ASMD of the original covariates across 800 randomizations under the FSM with differences choices of ϵ.

Covariates	Choice of ϵ			
	0.1	0.01	0.001	0.0001
X_1	0.032	0.030	0.030	0.030
X_2	0.029	0.026	0.026	0.026
X_3	0.041	0.043	0.043	0.043
X_4	0.026	0.028	0.028	0.028
X_5	0.029	0.031	0.031	0.031
X_6	0.034	0.034	0.034	0.034
Mean	0.032	0.032	0.032	0.032

Table A13: Averages of the ASMD of squares and pairwise products of the covariates across 800 randomizations under the FSM with different choices of ϵ.

Covariate transformations	Choice of ϵ			
	0.1	0.01	0.001	0.0001
X_1X_2	0.044	0.038	0.038	0.038
X_1X_3	0.040	0.041	0.041	0.041
X_1X_4	0.028	0.025	0.025	0.025
X_1X_5	0.039	0.037	0.037	0.037
X_1X_6	0.031	0.030	0.030	0.030
X_2X_3	0.045	0.048	0.048	0.048
X_2X_4	0.029	0.026	0.026	0.026
X_2X_5	0.040	0.029	0.029	0.029
X_2X_6	0.028	0.026	0.026	0.026
X_3X_4	0.038	0.033	0.033	0.033
X_3X_5	0.091	0.097	0.097	0.097
X_3X_6	0.028	0.033	0.033	0.033
X_4X_5	0.046	0.038	0.038	0.038
X_4X_6	0.026	0.027	0.027	0.027
X_5X_6	0.024	0.027	0.027	0.027
X_1^2	0.031	0.032	0.032	0.032
X_2^2	0.038	0.036	0.036	0.036
X_3^2	0.040	0.040	0.040	0.040
X_4^2	0.052	0.052	0.052	0.052
X_5^2	0.011	0.014	0.014	0.014
Mean	0.037	0.036	0.036	0.036

I Additional results from the Health Insurance Experiment

Table A14: Average ASMD of the main covariates between treatment groups 1 and 2 across 400 randomizations.

Covariates	Designs			
	CRD	RR Wilks	RR Mahalanobis	FSM
X_1 : Log family size	0.052	0.039	0.038	0.012
X_2 : Log family income	0.052	0.040	0.043	0.010
X_3 : Max hourly wage	0.051	0.042	0.047	0.017
X_4 : Adult med visits	0.049	0.043	0.041	0.014
X_5 : Kid med visits	0.048	0.039	0.040	0.010
X_6 : Female	0.047	0.039	0.040	0.010
X_7 : Age 0 to 5	0.053	0.038	0.039	0.010
X_8 : Age 6 to 17	0.051	0.041	0.039	0.011
X_9 : Age 18 to 44	0.053	0.038	0.040	0.010
X_{10} : Male HS Grad	0.051	0.038	0.041	0.006
X_{11} : Male more than HS	0.048	0.037	0.041	0.006
X_{12} : Insured	0.049	0.040	0.038	0.010
X_{13} : Excellent health	0.052	0.040	0.037	0.009
X_{14} : Good health	0.053	0.038	0.037	0.010
X_{15} : Family income mis	0.052	0.038	0.041	0.011
X_{16} : Max hourly wage mis	0.051	0.038	0.041	0.013
X_{17} : Adult med visits mis	0.054	0.040	0.040	0.011
X_{18} : Kid med visits mis	0.057	0.041	0.039	0.011
X_{19} : Education male mis	0.048	0.038	0.041	0.008
X_{20} : Insured mis	0.048	0.039	0.038	0.011
Mean	0.051	0.039	0.040	0.011

Table A15: Averages of the ASMD between each pair of treatment groups across the original covariates and across 400 randomizations.

Treatment group	Designs			
	CRD	RR Wilks	RR Mahalanobis	FSM
$1, 2$	0.051	0.039	0.040	0.011
$1, 3$	0.055	0.041	0.043	0.011
$1, 4$	0.049	0.038	0.039	0.010
$2, 3$	0.056	0.043	0.045	0.012
$2, 4$	0.053	0.040	0.041	0.010
$3, 4$	0.056	0.042	0.044	0.012
Mean	0.053	0.040	0.042	0.011

Table A16: Averages of the ASMD of the squares and pairwise products of the (demeaned) covariates X_1,..., X_5 between treatment groups 1 and 2 across 400 randomizations.

Covariates	Designs			
	CRD	RR Wilks	RR Mahalanobis	FSM
X_1X_2	0.053	0.039	0.041	0.020
X_1X_3	0.053	0.047	0.046	0.027
X_1X_4	0.054	0.045	0.045	0.020
X_1X_5	0.049	0.040	0.041	0.013
X_2X_3	0.054	0.049	0.053	0.038
X_2X_4	0.050	0.045	0.048	0.017
X_2X_5	0.052	0.039	0.039	0.015
X_3X_4	0.054	0.043	0.045	0.022
X_3X_5	0.050	0.042	0.046	0.022
X_4X_5	0.054	0.044	0.045	0.015
X_1^2	0.053	0.041	0.040	0.026
X_2^2	0.051	0.041	0.042	0.015
X_3^2	0.057	0.055	0.058	0.026
X_4^2	0.053	0.053	0.053	0.012
X_5^2	0.051	0.043	0.044	0.004
Mean	0.053	0.044	0.046	0.019

Table A17: Averages of the ASMD between each pair of treatment groups across the squares and pairwise products of the (demeaned) covariates X_1,..., X_5 and across 400 randomizations.

Treatment group	Designs			
	CRD	RR 0.01	RR 0.001	FSM
$1, 2$	0.053	0.044	0.046	0.019
$1, 3$	0.056	0.046	0.048	0.020
$1, 4$	0.051	0.043	0.044	0.017
$2, 3$	0.058	0.049	0.049	0.021
$2, 4$	0.054	0.046	0.046	0.018
$3, 4$	0.058	0.048	0.049	0.023
Mean	0.055	0.046	0.047	0.020

Figure A3: Distributions of ASMD of all cubes and three-way interactions of the non-binary covariates across randomizations.

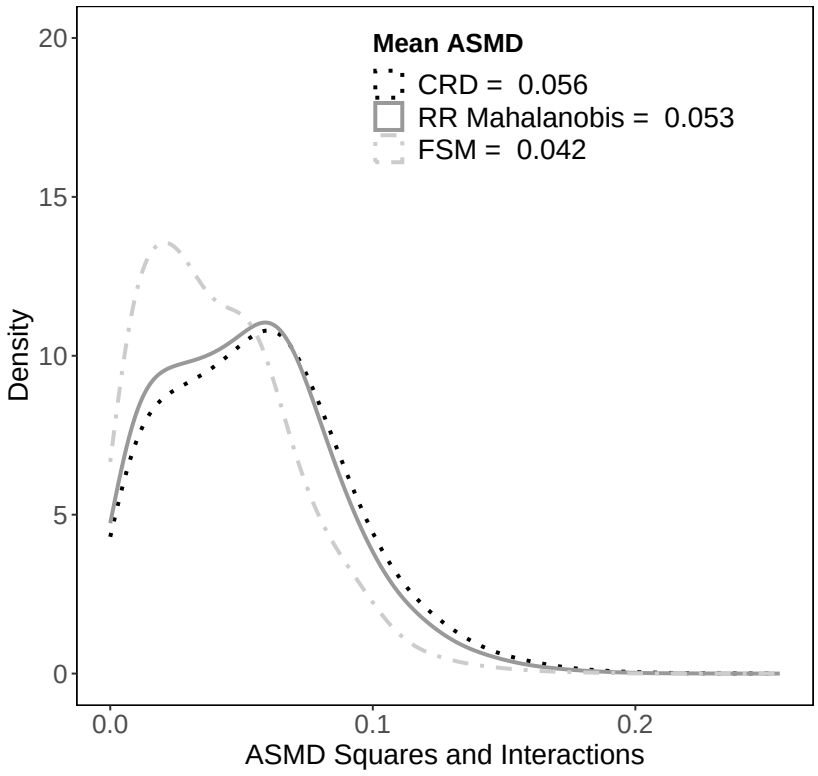
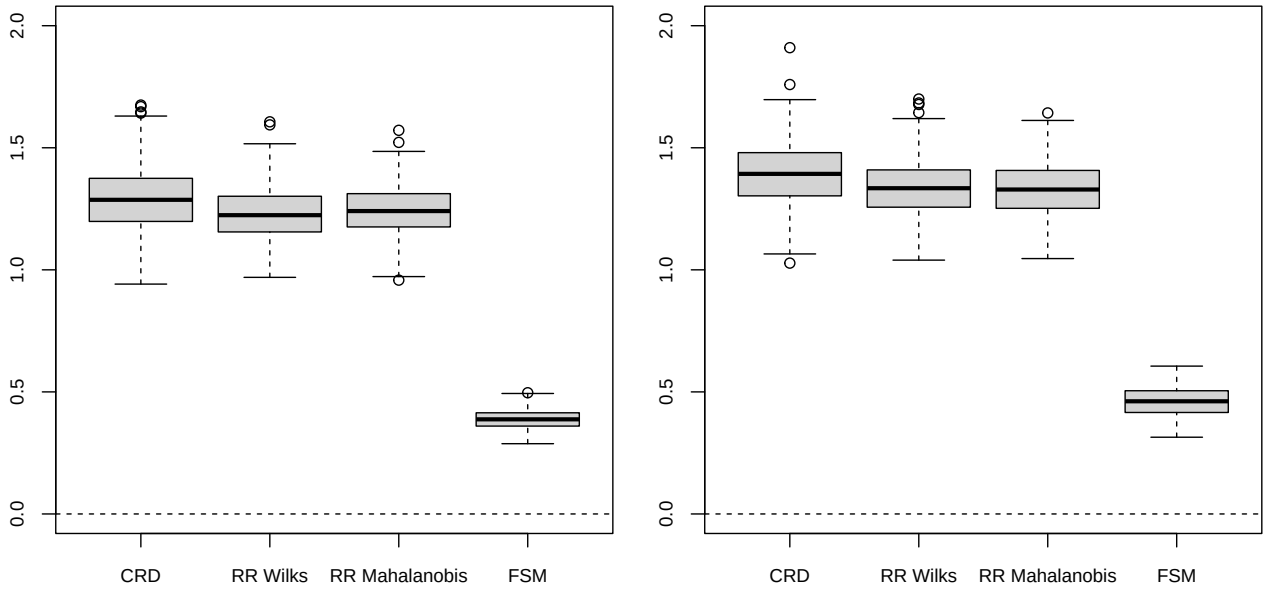
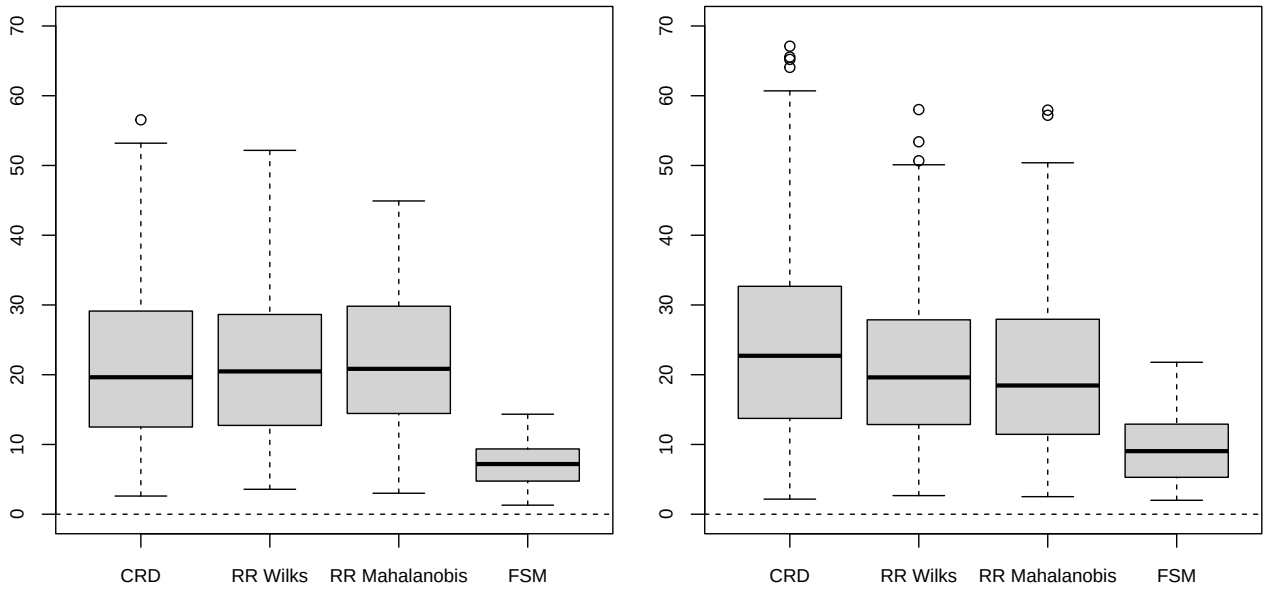


Figure A4: Distributions of discrepancies of the correlation matrices of the covariates in the treatment groups of the HIE data across randomizations. The discrepancies are measured by $||\underline{\boldsymbol{R}}_g - \underline{\boldsymbol{R}}_{g'}||_F$, where $\underline{\boldsymbol{R}}_g$ is the sample correlation matrix of the covariates in treatment group g and $||\cdot||_F$ is the Frobenius norm. The FSM systematically produces lower discrepancies than the other methods, exhibiting substantially improved balance on the correlations of the covariates.



(a) $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$ (b) $||\underline{\boldsymbol{R}}_3 - \underline{\boldsymbol{R}}_4||_F$

Figure A5: Boxplot of the distribution of $||\underline{\boldsymbol{S}}_g - \underline{\boldsymbol{S}}_{g'}||_F$ across 400 randomizations, where $\underline{\boldsymbol{S}}_g$ is the sample covariance matrix of the covariates in treatment group $g \in \{1, 2, 3, 4\}$.



(a) $||\underline{\boldsymbol{S}}_1 - \underline{\boldsymbol{S}}_2||_F$ (b) $||\underline{\boldsymbol{S}}_3 - \underline{\boldsymbol{S}}_4||_F$

We now compare the efficiency of the designs in the randomization-based approach with four additional potential outcome models given below.

- Model B3: $Y(3) = 5 - 3X_1 + X_2 + X_3 - 0.2X_4 + 0.8X_5 + \eta$, $Y(3) = Y(2)$.
- Model B4: $Y(3) = 5 - 2X_1^2 + 0.5X_3^2 + 0.5X_5^2 + 5X_1X_2 - 0.8X_3X_5 + \eta$, $Y(3) = Y(2)$.
- Model B5: $Y(3) = 10 + 8X_1X_2 + 3X_2X_5 - 0.5X_3X_5 + \eta$, $Y(3) = Y(2)$.
- Model B6: $Y(3) = 0.8X_1X_2 - 3X_3^2 + \frac{1}{1+X_4} - 4X_1^3 + \eta$

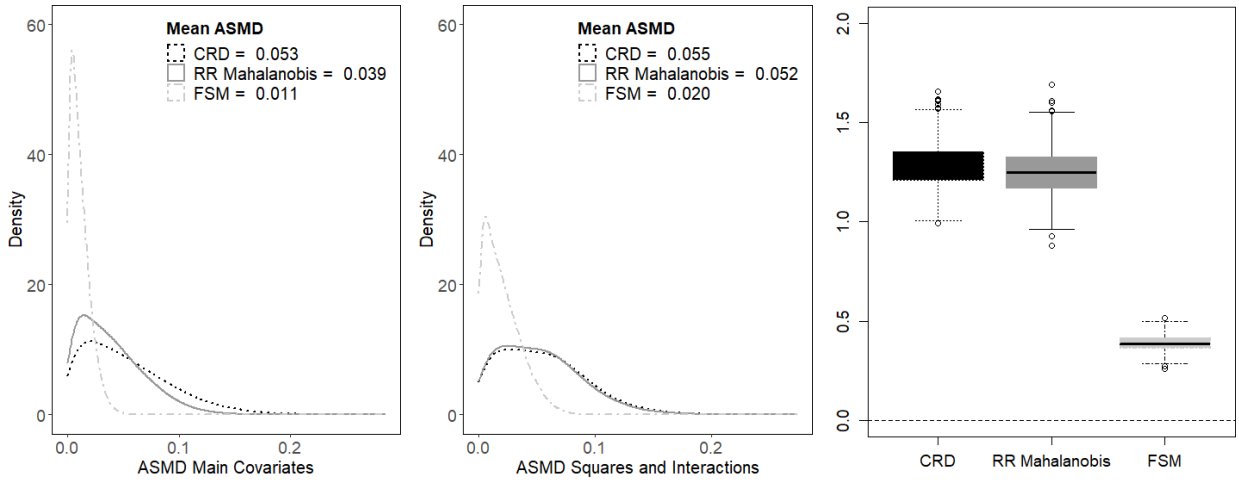
For each model, the error term $\eta \sim \mathcal{N}(0, 1.5^2)$. Under each design, $\text{SATE}_{3,2}$ is estimated using the standard difference-in-means estimator and the corresponding randomization-based SE is obtained by generating 400 randomizations and computing the standard deviation of the estimator across these 400 randomizations. The average randomization-based standard errors (across 500 simulations) are presented in Table A18.

Table A18: Average randomization-based standard errors relative to the FSM under Models B3, B4, B5, B6

	Designs			
	CRD	RR Wilks	RR Mahalanobis	FSM
Model B3	2.36	1.80	1.90	1
Model B4	2.14	1.75	1.81	1
Model B5	2.99	2.40	2.44	1
Model B6	1.61	1.42	1.44	1

We finish this section by evaluating and comparing the covariate balance on the main covariates and the second-order transformations, for CRD, RR, and the FSM, where RR uses the Mahalanobis distance based on the main covariates only and accepts 0.1% of the assignments.

Figure A6: Distributions of absolute standardized mean differences (ASMD) of the main covariates (panel (a)) and the squares and pairwise products of the scaled covariates (panel (b)) across randomizations. For each plot, the legend presents the average ASMD across simulations for the four methods. Panel (c) shows distributions of discrepancies between the correlation matrices of the covariates in treatment groups 1 and 2 (as measured by the Frobenius norm, $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$). In terms of the main covariates, second-order transformations, and correlation matrices, the FSM substantially outperforms CRD and RR.



(a) Main covariates (b) Squares and pairwise products (c) Frobenius norm

Figure A6 shows a pattern of performances of the designs akin to those illustrated in Section 6. The FSM, outperforms CRD and RR in terms of balancing both the main covariates and their second-order transformations. As compared to the previous version, this version of RR reduces the average imbalance on the main covariates, while increasing the average imbalance on the second-order transformations. This behavior aligns with our expectations, since this version of RR specifically targets balance on the main covariates, not on their second order transformations.

J Additional figures from the case studies

Figure A7: Distributions of the absolute standardized mean differences of the main covariates and their squares and interactions, and the Frobenius norms of $\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2$ under complete randomization, rerandomization, and the FSM, for the five studies: (1) Angrist, (2) Blattman, (3) Durocher, (4) Finkelstein, (5) Lalonde.

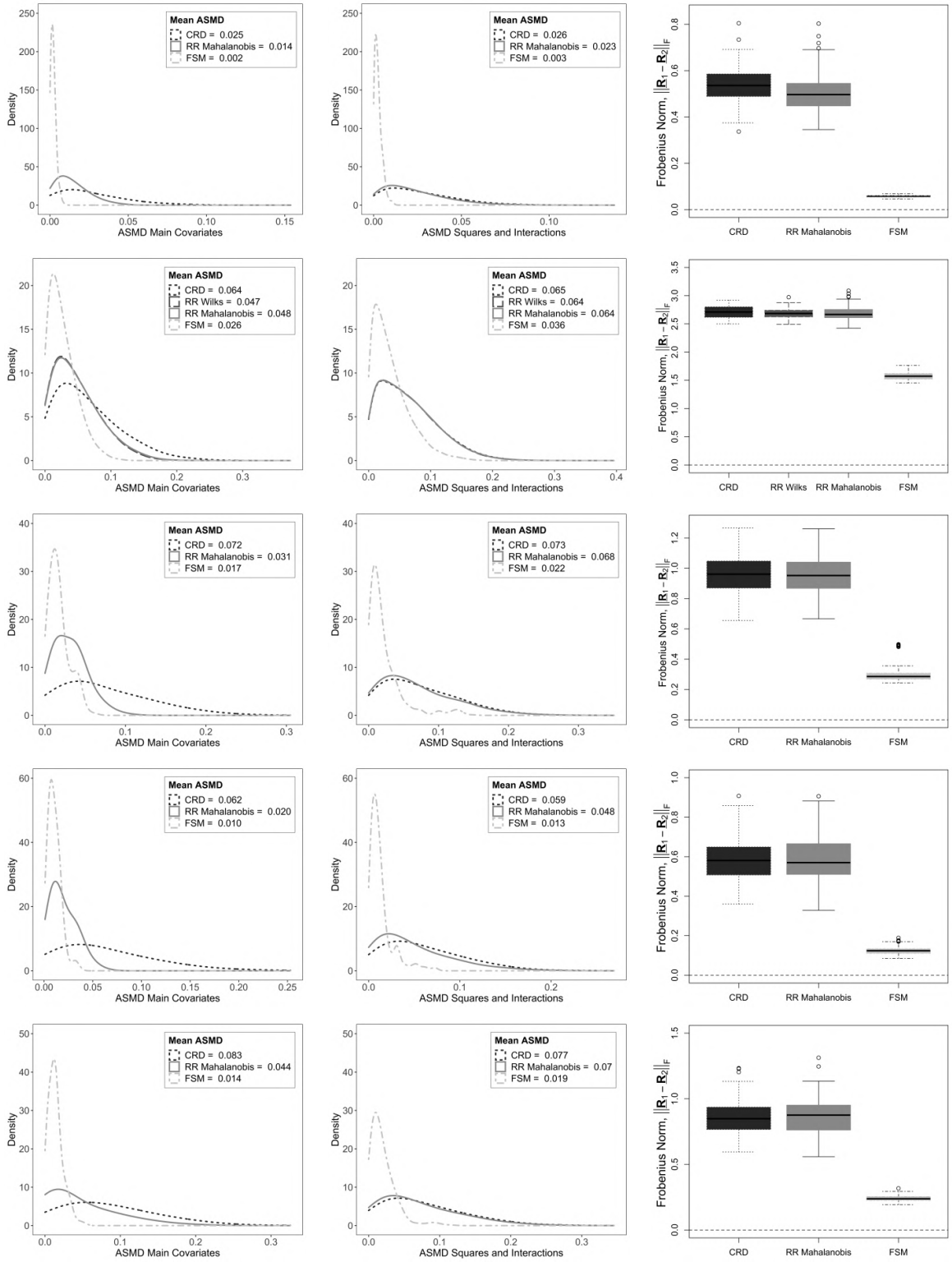


Figure A8: Distributions of the absolute standardized mean differences of the main covariates and their squares and interactions, and the Frobenius norms of $\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2$ under complete randomization, rerandomization, and the FSM, for the five studies: (6) Ambler, (7) Crepon, (8) Dupas, (9) Karlan, (10) Wantchekon.