Article

Journal of Educational and Behavioral Statistics
2026, Vol. 51, No. 3, pp. 667–695
DOI: 10.3102/10769986251342292
Article reuse guidelines: sagepub.com/journals-permissions

Inspection-Guided Randomization: A Flexible and Transparent Restricted Randomization Framework for Better Experimental Design

Maggie Wang
Stanford University

René F. Kizilcec
Cornell University

Michael Baiocchi
Stanford University

Randomized experiments are considered the gold standard for estimating causal effects. However, out of the set of possible randomized assignments, some may be more likely to produce poor effect estimates and misleading conclusions. Restricted randomization is an experimental design strategy that filters out undesirable treatment assignments, but its application has primarily been limited to ensuring covariate balance in two-arm studies where the target estimand is the average treatment effect. Other experimental settings with different design desiderata and target effect estimands could also stand to benefit from a restricted randomization approach. We introduce inspection-guided randomization (IGR), a transparent and flexible framework for restricted randomization that filters out undesirable treatment assignments by inspecting assignments against analyst-specified, domain-informed design desiderata. In IGR, the acceptable treatment assignments are locked in ex ante and preregistered in the trial protocol, thus safeguarding against p-hacking and promoting reproducibility. Through illustrative simulation studies motivated by behavioral health and education interventions, we demonstrate how IGR can improve effect estimates compared to benchmark designs in experiments involving interference and group formation.

Keyword: *Experimental design, restricted randomization, interference, context-dependence*

1. Introduction

Randomized experiments are considered the gold standard for estimating causal effects of interventions. By randomizing the treatment, the distribution of

baseline covariates is similar in expectation across treatment arms, and observed differences in the outcome can be attributed to the intervention. However, randomization alone does not guarantee covariate balance for every realized allocation of treatments (Fisher, 1935; Rubin, 2008). Suppose the treatment allocation chosen for an experiment, by chance, exhibits imbalance on covariates. At best, the analyst makes careful *ex post* adjustments for imbalance, usually according to a regression model that relies on assumptions that may or may not hold (Freedman, 2008). At worst, covariate imbalance is overlooked or handled improperly (e.g., in an attempt to *p*-hack), leading to unreliable conclusions about treatment effects (e.g., false positives). Restricted randomization helps prevent imbalanced treatment allocations. In restricted randomization, the assignment mechanism is modified such that the probability of choosing an undesirable allocation is set to zero. While performing restricted randomization is rather common, it is primarily applied to the canonical use case of balancing two-arm trials, where the target effect estimand is the average treatment effect.

Experimental settings with different design desiderata and estimands could also stand to benefit from a restricted randomization design. A natural extension of balancing a two-arm trial is to balance a multi-arm trial, where the goal is to simultaneously estimate treatment effects for multiple interventions. Additionally, in experiments where we want to explore treatment effect moderators (VanderWeele, 2015), we may want to construct treatment arms with intentional ''imbalance'' on a hypothesized moderating covariate, while maintaining balance on other covariates. We may also be interested in controlling factors beyond balance that could jeopardize the credibility of effect estimates. Causal inference in the presence of interference (Cox, 1958; Rosenbaum, 2007), when one individual's treatment affects a different individual's outcome, has been an area of growing interest (Aronow & Samii, 2017; Athey et al., 2018; Eckles et al., 2017; Hudgens & Halloran, 2008). For many target effect estimands, interference should be minimized to obtain unbiased effect estimates. Here, if we have some *a priori* knowledge of how interference is likely to occur, we can use restricted randomization to rule out allocations likely to generate large amounts of interference.

More broadly, the core task in experimental design is to set up the study such that it generates a trustworthy answer to the study's motivating causal question. We introduce inspection-guided randomization (IGR), a flexible and transparent framework for restricted randomization in which the study analyst *inspects* the desirability of candidate randomization allocations, then constrains the candidate space to produce less biased and more precise estimates of the target estimand. The flexibility in IGR exists along two fronts. First, IGR permits targeting diverse design desiderata beyond those related to balance, thus expanding the scope of use cases for restricted randomization. Second, IGR incorporates a checkpoint that encourages the study analyst to carefully evaluate and

potentially adapt the construction of the restricted assignment mechanism. IGR promotes transparency by enforcing that the set of acceptable treatment allocations be preregistered in the trial protocol prior to running the study.

This work is organized as follows. In Section 2, we review prior work on restricted randomization and experimental design, then highlight the contributions we make through IGR. In Section 3, we introduce notation and an overview of the IGR framework. In Section 4, we walk through each step of IGR in detail and use a simulation to illustrate how it can be applied to the design of an experiment with interference. In Sections 5 and 6, we discuss the analysis and inference for IGR-designed experiments and show how IGR can improve effect estimates. We conclude in Section 7 with a summary of the benefits of using IGR and directions for future work.

2. Related Work

Blocking and Matching. Blocking is a form of restricted randomization that rules out allocations that are imbalanced on one or more blocked covariates and has been used widely in the design of experiments (Box et al., 2005; Fisher, 1992). Matched pair designs use matching algorithms to pair experimental units on the basis of a metric of covariate similarity, thus facilitating balance on multiple categorical and continuous covariates at once (Greevy et al., 2004; Imai et al., 2009; Xu & Kalbfleisch, 2010). More recently, efficient blocking algorithms have been developed that extend matched-pairs designs and allow for balanced blocks of flexible sizes (Higgins et al., 2016).

Tightening, Constraining, and Re-Randomizing. An alternative approach to blocking and matching involves first generating a candidate set of allocations according to a prespecified mechanism, then selecting one allocation from this set as the official randomization to deploy. Tukey referred to this procedure as ''tightening'' the clinical trial (Tukey, 1993), and Moulton later formalized a ''covariate-constrained randomization'' design for cluster-randomized trials (Moulton, 2004). There has also been work in using constrained randomization in multi-arm settings with the use of multi-arm balance metrics (Ciolino et al., 2019; Watson et al., 2021; Zhou et al., 2021). With constrained designs, inference can be done via a randomization test over the candidate allocations.

Instead of listing all possible allocations at once, re-randomization repeatedly randomizes until producing the first allocation that meets prespecified criteria (Cox, 2009; Morgan & Rubin, 2012; Rubin, 2008). Re-randomization has been used beyond two-arm study designs in experiments with factorial designs (Branson et al., 2016) and in experiments on networks where units have correlated potential outcomes (Basse & Airoldi, 2018). Although re-randomization can save computing power, it opens the door to *p*-hacking because the set of criteria-meeting allocations used to conduct a randomization test is sampled

after the study outcomes have been observed (see further discussion in Supplemental Appendix G [available in the online version of this article] on the risk of Type I error rate inflation when the accepted allocations in the analysis phase are different from those in the design phase). Additionally, re-randomization assumes that the analyst's first choice of restriction criteria yields a satisfactory design, when in reality, the criteria may permit allocations with poor properties for estimation and inference.

Optimal and Model-Assisted Design. Kasy (2016) argues that if the goal of the experiment is to minimize ''disparity'' between the estimated and true treatment effect (e.g., mean squared error), one should deploy the single allocation that optimizes the expected loss over possible data-generating processes. Identifying an optimal allocation in this way, however, requires specifying a prior distribution for the loss function. Model-assisted design is similarly motivated, but optimizes the expected mean squared error over random data samples from a specific data-generating process rather than for a distribution of data-generating processes (Basse & Airoldi, 2018). Unlike with constrained randomization and re-randomization, when only a single optimal allocation is selected, randomization inference is no longer warranted.

Reproducible Science. Results of published studies are often not reproducible because they were generated through the use of questionable research practices aimed at ''publication-worthiness'' rather than scientific correctness and robustness (Baker, 2016; Ioannidis, 2005; Munafò et al., 2017). These practices include *p*-hacking, reworking research hypotheses after the results of a study have already been collected, and overinterpreting the data (Ioannidis, 2005; Munafò et al., 2017). To promote reproducible and open science, there has been a push to preregister study protocols that hold researchers accountable to a planned study design and set of analyses (Munafò et al., 2017).

Our Contributions. IGR unifies and builds upon ideas across several works above. Tightening, re-randomization, and covariate-constrained randomization were developed to target balance criteria only, and subsequent literature that applies or expands these randomization strategies also only considers balance (see Supplemental Appendix Table A.1 in the online version of the journal). While IGR resembles covariate-constrained randomization, we make an important extension by recognizing that it can be beneficial to inspect allocations on the basis of multiple properties, particularly properties other than balance. IGR is also related to model-assisted design (Basse & Airoldi, 2018), but whereas model-assisted design requires deriving a closed-form expression for mean squared error, IGR can use domain-informed, approximate metrics when the expression for mean squared error is unknown or difficult to derive.

Additionally, most prior restricted randomization strategies, for example, re-randomization, use a naive sampling approach to select an allocation. In contrast, in IGR, we can efficiently ''search'' through a large candidate allocation

space (e.g., with genetic algorithms) to enumerate a set of candidate allocations that score well according to specified metrics. Speaking more generally, in re-randomization, the process of searching for an allocation and the process of randomizing are conflated in the act of sampling. IGR, on the other hand, separates the enumeration step, which can be done without any randomization, from the randomization step, and thus draws a clear distinction between searching and randomizing. Some implementations of constrained randomization use a similar approach to IGR (Ciolino et al., 2019; de Hoop et al., 2012; Li et al., 2016; Moulton, 2004;Watson et al., 2021; Zhou et al., 2021), where a candidate set of allocations is first enumerated, and the final allocation is chosen randomly from among criterion-meeting allocations within the candidate set. However, these implementations use a random sampling procedure to produce the initial candidate set; they do not highlight that enumerating the candidate set can involve a directed search and does not necessarily need to be random. The randomness used for inference occurs later, in the actual assignment.

In line with efforts to make studies more reproducible, IGR includes and enforces a step where accepted allocations are preregistered. As long as the study analyst performs randomization inference over the preregistered set of accepted allocations, getting the nominal Type I error rate is guaranteed. Preregistration is not explicitly present in any existing restricted randomization framework, so study analysts are not held accountable for conducting correct analyses. We show, for instance, that in re-randomization, we can easily inflate the Type I error rate by using a more restrictive balance criterion in the analysis phase than in the design phase.

Finally, IGR places emphasis on deliberately interrogating the quality of the restricted design and making subsequent adaptations to the restriction as necessary. This adaptation step is either underemphasized or absent from prior approaches to restricted randomization. In practice, we suspect that two common forms of adaptation will be: (a) fixing inspection metrics that do not correctly target the design desiderata (e.g., a balance metric that does not capture the kind of imbalance concerning to the study stakeholders) and (b) adjusting the relative emphasis placed on different design desiderata to better navigate trade-offs (e.g., in trials with interference, homophily can create tension between achieving covariate balance and minimizing interference, and which of the two desiderata is prioritized may need to be adjusted).

Several prior real-world studies have been designed with restricted randomization techniques related to IGR that included some, but not all, of the elements of the IGR framework (Cho et al., 2021, 2022; Kizilcec et al., 2022; Murnane et al., 2023; Zahrt et al., 2023). These studies demonstrate the need for a formal framework like IGR that unifies design principles, and they also highlight the potential for IGR to be applicable and useful in practice. In this work, we present simulated vignettes of an experiment with interference and a group

formation experiment (Section 4 and Supplemental Appendix E [available in the online version of this article], respectively) to explain specific features of the IGR framework and to highlight how IGR can be used for diverse experimental settings, including and beyond those where related restricted randomization techniques have been used in the past.

3. An Overview of IGR

3.1 Setup and Notation

We consider a setting with N participants in a k-armed study, where the participants are indexed by $i = 1, ..., N$. Let $\mathbf{X}$ be the N-by-p dimensional matrix representing the participants' baseline covariates. For a given participant i, we use the notation Z_i to denote the participant's treatment assignment, where $Z_i \in \{0, ..., k-1\}$. Let the N-dimensional vector of treatments be denoted $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)$. Since treatment is assigned at random, Z_i is a random variable, and it follows that $\mathbf{Z}$ is a random vector. We denote a particular realization of a treatment assignment for individual i as lowercase z_i and the intervention assignment vector as lowercase $\mathbf{z}$. The probability that random vector $\mathbf{Z}$ takes on a particular value $\mathbf{z}$ is written $P(\mathbf{Z} = \mathbf{z}) = p_{\mathbf{z}}$, and we refer to $P(\mathbf{Z})$ as the assignment mechanism. We call a particular realization of the intervention assignment vector a treatment allocation. Letting $\mathcal{Z}$ denote the collection of all k^N possible values of $\mathbf{z}$, we have that $\sum_{\mathbf{z} \in \mathcal{Z}} p_{\mathbf{z}} = 1$.

3.2 Outline of the IGR Framework

For clarity and ease of use in practice, we organize the IGR design framework into distinct steps. We describe each in brief below and then elaborate further in the next section:

1. *Specification*. The analyst specifies the target effect estimand and the fitness function used to measure the desirability of each candidate treatment allocation.
2. *Enumeration*. The analyst enumerates a large pool of candidate allocations using a simple assignment mechanism.
3. *Restriction*. The analyst scores each candidate allocation with the specified fitness function and filters out those that are low scoring, according to a restriction rule.
4. *Evaluation and Adaptation*. The analyst evaluates their choice of fitness function and restriction rule, then makes adaptations if needed to better achieve design desiderata.
5. *Preregistration*. The analyst preregisters the IGR assignment mechanism.

6. *Randomization*. The analyst samples one allocation to deploy for the experiment.

4. IGR in Detail, With Application to a Simulated Experiment With Interference

In this section, we provide a full description and rationale for the steps of the IGR framework. To make the steps more concrete, we also illustrate how to apply them to a simulated experiment with interference. We begin this section by describing the domain and setup for the simulated experiment, then walk through the steps of IGR in detail, explaining how each can be applied in the context of the simulated experiment. Note that IGR can be used in many experimental settings, not just those with interference. In Supplemental Appendix E (available in the online version of this article), for example, we also highlight how IGR can be applied to a group formation experiment. All simulation codes are available at https://github.com/wangmagg/inspection-guided-randomization.

4.1 Simulated Experiment with Interference

The Challenge of Interference. Interference occurs when the treatment assigned to one observational unit affects the outcome of a different unit. Interference is often present with behavioral interventions, where information can diffuse or spill over along social ties (Cai et al., 2015; Kim et al., 2015; Valente, 2012). When interference is present, a common target treatment effect estimand is the global average treatment effect, that is, a comparison of the average outcome if everyone were to receive treatment versus if everyone were to receive control (Hudgens & Halloran, 2008). Failing to properly account for interference can make estimates of the global average treatment effect both biased and imprecise (Eckles et al., 2017). This, in turn, can result in misleading conclusions (e.g., false nulls) and misinformed recommendations for policy-making (e.g., failure to implement an effective intervention).

A common approach to controlling interference is to use cluster randomization, where randomization is done on the cluster level such that all individuals in the same cluster receive the same treatment (Hayes & Moulton, 2017). Given enough separation between clusters, there is likely no spillover between different treatment arms, and comparing treated to control clusters can yield an unbiased estimate of the global average treatment effect. However, it may sometimes be difficult to identify clusters that are sufficiently separated to avoid any inter-cluster interference. In other cases, it may be possible to identify a few large clusters that are well separated, such as villages that are far apart geographically. However, randomizing only a few clusters leads to low power, especially if there is strong intra-cluster correlation on prognostic covariates. Often, for clusters

that form organically, individuals within a cluster are more similar to each other than those in different clusters, so intra-cluster correlation is expected.

Within the IGR design framework, we can be more flexible in how we control interference. Leveraging prior knowledge, which may not necessarily be complete or fully correct, we can use an interference-targeting constraint to control the amount of anticipated interference more precisely than with cluster randomization. Even if we do not have perfect knowledge of the interference structure (e.g., which individuals interact to produce spillover), we can still develop ''coarse'' metrics that measure interference.

Simulation Set-Up. Our simulated experiment with interference is based on a trial of a sexual assault prevention intervention for adolescent girls conducted across schools in urban settlements in Kenya. During the intervention, the girls learned and practiced verbal and physical skills for resisting sexual assault (Baiocchi et al., 2017; Sarnquist et al., 2024).

An option for the design of this trial was to randomize the intervention at the settlement level, which would have strongly mitigated interference. However, the settlements were known to have a wide range of baseline rates of sexual assault, meaning randomizing on the settlement level would have made it difficult to achieve balance. Hence, the trial was instead randomized at the school level, and because the settlements were densely populated, it is likely that girls attending different schools interacted with one another. The skills learned by girls assigned to receive the intervention may have been passed on to girls assigned to the control arm, resulting in interference.

In our simulation, we generate synthetic schools in five Kenyan settlements: Kibera, Mukuru, Huruma, Korogocho, and Dandora. Baseline covariates are sampled using a nested hierarchical model, such that individuals attending the same schools are more similar than individuals attending different schools, and individuals within the same settlement are more similar than individuals across different settlements. We use a statistical network model to generate the interference network, where the probability that two individuals are connected by an edge is inversely proportional to the γth power of the Euclidean distance between the schools. To generate the observed outcome, we apply an additive treatment effect and assume that an individual is exposed to treatment if they directly receive treatment or if a certain fraction of their network neighbors receive treatment. Further details on the data-generating process and the parameter settings used in the simulations can be found in Supplemental Appendix C (available in the online version of this article).

4.2 Step-by-Step Application of IGR

We now go through the IGR step-by-step in detail. After describing each step, we show how it can be applied to designing the simulated experiment.

TABLE 1.
Common Desiderata Categories, Specific Examples of Desiderata Within Each Category, and Example Inspection Metrics That Can Be Used to Measure the Extent to Which an Allocation Satisfies a Particular Desideratum

Desideratum category	Example desideratum	Example inspection metric(s)
Estimation properties	Low variance	Standardized mean difference, Mahalanobis distance
	Low bias in the presence of network or spatial spillover	Connectivity in a graph network, geographic proximity between treated/control units
Study logistics	Ease of delivering a treatment for an infectious disease	Travel time or distance to reach all units assigned to the treatment arm
Conditions for testing a scientific theory	Placing individuals in certain social contexts to study the effect on academic achievement	Compositional characteristics of the individuals assigned to a given classroom (e.g., gender ratio)

4.2.1 Specification. In the *Specification* step, the study analyst first works with the domain experts on the research team to specify the goals of the study in terms of the causal question of interest and the target effect estimand. Based on these goals, the team compiles desiderata for the study such that if the deployed randomization allocation were guaranteed to satisfy all desiderata, then the data collected would likely be conducive to credible causal inference.

The analyst then constructs a set of inspection metrics used to assess candidate allocations, where each metric measures the extent to which an allocation satisfies a particular desideratum (e.g., standardized mean difference as a metric to measure balance). In Table 1, we list three broad categories of common design desiderata and examples of inspection metrics that might be used within each category.

The analyst also decides on an aggregator function that combines the separate metric scores and summarizes an allocation's desirability. In certain settings, the desiderata may conflict, where scoring better on some desiderata necessarily entails worse scores on others. The aggregator can be used to navigate desiderata trade-offs, for instance, by incorporating a weighting scheme that prioritizes certain desiderata over others. Let f be the fitness function that maps an allocation $\mathbf{z}$ to its score by combining the outputs of each of the individual inspection metrics via the aggregation function.

Finally, the analyst specifies a restriction rule, r, that constrains the space of candidate treatment allocations on the basis of each allocation's score, where $r(f(\mathbf{z}, \mathbf{X})) = 0$ means $\mathbf{z}$ is filtered out by setting $p_{\mathbf{z}}$ to zero and $r(f(\mathbf{z}, \mathbf{X})) = 1$ means $\mathbf{z}$ is deemed acceptable by setting $p_{\mathbf{z}}$ to a positive value between 0 and 1. For example, r could be a thresholding rule, where only allocations with scores below some threshold s^* are accepted.

Specification *in a Simulated Experiment*. We specify the target effect estimand to be the global average treatment effect, $\tau_{GATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i(\mathbf{1}) - Y_i(\mathbf{0})$ where 1 is the all-ones vector (everyone is treated) and $\mathbf{0}$ is the all-zeros vector (everyone is control). Our design goal is to assign treatments such that those in the treatment (control) arm experience a universe as close as possible to an all-treated (all-control) universe, while also balancing covariates across arms. Thus, we specify two inspection metrics, one to measure interference and one to measure balance.

For interventions that spread via person-to-person interaction, the likelihood of such spread or spillover may decrease with geographic distance between individuals. While we may not know precisely which individual(s) will interact with which other individual(s), we can try to allocate treatments in such a way that no two individuals assigned to treatment and control are too close to one another geographically. Closeness can be defined, for example, in terms of where they live, where they attend school, and where they work. This line of reasoning motivates an inspection metric that measures the minimum distance between any two individuals assigned to treatment and control, where a larger minimum distance is more desirable. To remain consistent with smaller scores being better, we use the reciprocal of the minimum distance. Formally, we define inspection metric $m^{(\texttt{InvMinEuclideanDist})}$ as

$$m^{(\texttt{InvMinEuclideanDist})} = \left(\min_{(i,j) \in [N] \times [N], z_i \neq z_j} \|\mathbf{l}_i - \mathbf{l}_j\|_2^2 \right)^{-1}, \tag{1}$$

where $z_i, z_j \in \{0, 1\}$ are the treatment assignments for units i and j, respectively, and $\mathbf{l}_i, \mathbf{l}_j \in \mathbb{R}^2$ are coordinates for i and j. While we use Euclidean distance here, other distance functions can also be used. It is also possible to expand upon a simple distance function by incorporating known information about ease of travel, such as the presence or absence of interconnecting roads. For balance, we use Mahalanobis distance,

$$m^{(\texttt{Mahalanobis})}(\mathbf{z}, \mathbf{g}, \mathbf{X}) \quad = \left(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(0)}\right)\mathbf{S}^{-1}\left(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(0)}\right)^T, \tag{2}$$

where $\bar{\mathbf{X}}^{(1)}$ and $\bar{\mathbf{X}}^{(0)}$ are the mean covariate vectors in the treated and control arms, respectively, and $\mathbf{S}$ is the sample covariance matrix. Combining the interference and balance metrics, we define the fitness function

$$f^{(\mathrm{I}+\mathrm{M})}(\mathbf{z},\mathbf{X}) \quad = 0.5 \cdot m^{(\texttt{InvMinEuclideanDist})}(\mathbf{z},\mathbf{X}) + 0.5 \cdot m^{(\texttt{Mahalanobis})} \cdot (\mathbf{z},\mathbf{X}). \tag{3}$$

To put the metrics on the same scale, we standardize the values across all enumerated candidate allocations (see the *Enumeration* step below) before combining the metrics in the fitness function.

4.2.2 Enumeration. In the *Enumeration* step, the analyst enumerates a pool of unique candidate allocations, $\mathcal{Z}_{\text{pool}} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_{\mathbf{M}}\}$. For a k-armed study, there are k^N total possible allocations. When N is large, it may be necessary to set $M << k^N$. The enumeration of $\mathcal{Z}_{\text{pool}}$ can be done by sampling from any choice of assignment mechanism P, including Bernoulli randomization and complete randomization. In some cases, sampling from an assignment mechanism that is already restricted, such as a block randomization, can increase the efficiency of the enumeration step, since the mechanism inherently rules out some of the undesirable allocations (see Supplemental Appendix E in the online version of the journal for an example).

Since $M << k^N$, it is possible that many desirable, high-scoring allocations are overlooked and not enumerated as part of $\mathcal{Z}_{\text{pool}}$. Instead of using simple enumeration strategies, we may instead want to efficiently search through the space of all k^N possible allocations to construct a more optimal $\mathcal{Z}^{(*)}_{\text{pool}}$. One approach to construct $\mathcal{Z}^{(*)}_{\text{pool}}$ is to use genetic algorithms (Mitchell, 1998). The initial $\mathcal{Z}_{\text{pool}}$ is viewed as a population of treatment allocations. These allocations ''mate'' in pairs to produce a new generation of allocations, where novel allocations are introduced through ''mutation'' (single entries of an allocation are changed) and ''crossover'' (segments of two parent allocations are swapped). The individuals in this generation are scored using the specified fitness function f, and those with the best scores survive and breed to produce the next generation. We can repeat this process over T generations to obtain $\mathcal{Z}^{(*)}_{\text{pool}}$.

Enumeration *in a Simulated Experiment*: We first sample candidate allocations from a complete randomization assignment mechanism that assigns an equal number of schools to the treatment and control arms. We enumerate a total of $M = 100,000$ allocations in the initial candidate pool, which is in line with prior work (Li et al., 2016; Nordin & Schultzberg, 2022; Watson et al., 2021). We then apply three generations of mutations and crossover to the initially enumerated pool to search for a pool with better fitness scores.

4.2.3 Restriction. The analyst selects a restriction rule r, then applies the fitness function and restriction rule to the enumerated pool of candidate allocations to produce the following assignment mechanism:

$$P_{IGR}(\mathbf{Z}=\mathbf{z})=\begin{cases}0 & \mathbf{z}\notin\mathcal{Z}_{pool}\\ 0 & \mathbf{z}\in\mathcal{Z}_{pool},(r\circ f)(\mathbf{z},\mathbf{X})=0\\ p_{\mathbf{z}} & \mathbf{z}\in\mathcal{Z}_{pool},(r\circ f)(\mathbf{z},\mathbf{X})=1\end{cases}.$$

We refer to allocations with a nonzero probability of being sampled as ''accepted'' allocations. At this step, the assignment mechanism P_{IGR} is proposed but not yet locked in.

Restriction *in a Simulated Experiment*. We use the restriction rule $r(s)=0$ if $s\geqslant s^*, r(s)=1$ if $s<s^*$, where s^* is the 0.5th percentile of scores for the enumerated candidate allocations. This restriction rule filters out all but the top 0.5th percentile of the 100,000 enumerated allocations, resulting in $m=500$ accepted allocations.

4.2.4 Evaluation and Adaptation. Before proceeding and locking in the restricted assignment mechanism, the study analyst and domain experts pause to evaluate the fitness function f and restriction rule r, making adaptations and improvements to them if necessary. Even if f and r were carefully identified in *Specification*, they still may be flawed in unforeseen ways and could yield poor study designs. We highlight three dimensions along which f and r should be evaluated: discriminatory power, desiderata trade-offs, and over-restriction.

1. Discriminatory power: Fitness functions should provide meaningful differentiation between candidate allocations. If the fitness function used to score allocations produces a similar score for all candidates, then there would be no reason to expect that using a restricted assignment mechanism would confer any improvements in effect estimation. When confronted with a homogeneous set of scores, the fitness function may require redefinition, for instance, through introducing additional inspection metrics.
2. Desiderata trade-offs: When there are multiple competing design desiderata, it may not be immediately clear whether and how to best construct the fitness function to weigh desiderata trade-offs. Visualizing the space of all candidate allocations by plotting the score outputs of different inspection metrics on separate axes can assist in understanding the nature of desiderata trade-offs. Then, visualizing the space of accepted allocations can assist in understanding whether the fitness function navigates such trade-offs as desired. If necessary, the analyst can make informed adjustments to the fitness function by modifying the aggregator function that combines the inspection metric scores.
3. Over-restriction: We consider two types of over-restriction. In the first type of over-restriction, the combination of fitness function and restriction rule may result in too few accepted allocations to obtain p-values

with the desired level of granularity (see Section 5 for details on deriving *p*-values from randomization tests). When this occurs, the analyst may need to modify the restriction rule to be less aggressive or choose to apply a restriction rule that always selects the top *m* allocations. Too few accepted allocations could also be a consequence of using an enumeration strategy that does not ''hit'' the correct regions of the space of all possible allocations. This issue can be addressed by increasing the size of the enumerated pool or by using stochastic optimization routines. In the second type of over-restriction, the set of accepted allocations may be too similar to one another, such that multiple units or groups of units are assigned to the same treatment arm across most accepted allocations. If we rely on an asymptotic *p*-value in testing the null hypothesis, this kind of over-restriction can result in departures from the nominal Type I error rate (Moulton, 2004). An exact *p*-value obtained through randomization inference circumvents this issue, but highly correlated allocations can also lead to increased mean squared error and reduced power, regardless of whether randomization inference is used (Krieger et al., 2020; Nordin & Schultzberg, 2022). Relaxing the restriction rule *r* to permit additional allocations is one way to alleviate over-restriction issues.

Evaluation and Adaptation *in a Simulated Experiment*. We use visualizations to assist in evaluating our choice of fitness function and restriction rule. Additional toy examples and a more detailed discussion on the choice of visualizations can be found in Supplemental Appendix D (available in the online version of this article). Below, we show an example of the iterative process of evaluation and adaptation for our simulated experiment.

Iteration 1: To assess discriminatory power, we plot a histogram of the scores for the candidate allocations and check that the histogram has sufficient spread (Figure 1, left). A ''spiky'' histogram, where all fitness scores are concentrated around a few values, indicates that the fitness function has low discriminatory power.

To examine desiderata trade-offs in the pool of candidate allocations, we plot a two-dimensional histogram with the balance metric on the *x*-axis and the interference metric on the *y*-axis (Figure 1, middle). By looking at this heat map, we can check whether allocations with better values on one metric tend to have worse values on the other. To examine desiderata trade-offs within the accepted set of candidate allocations, we create a scatterplot with the balance metric on the *x*-axis and the interference metric on the *y*-axis. Each point in the scatterplot corresponds to a single accepted allocation. By comparing the positions of the accepted allocations to those of a benchmark randomization strategy, we can

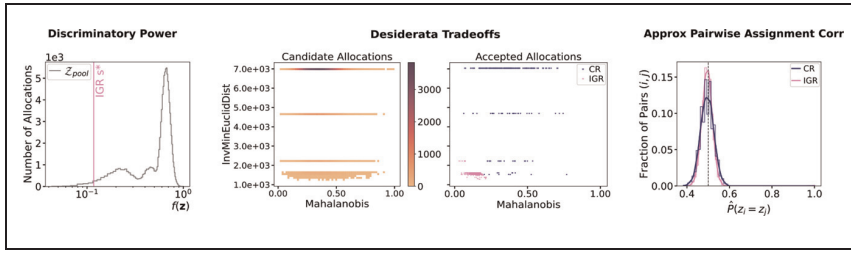

FIGURE 1. *Visual checks for discriminatory power, desiderata trade-offs, and over-restriction (via approximate pairwise assignment correlation) under fitness function* $\mathrm{f}^{(I+M)}$.

assess whether our restricted randomization scheme is navigating desiderata trade-offs as desired.

To check for over-restriction, we plot a histogram of the approximate pairwise assignment correlation (Figure 1, right). Specifically, for each pair of units, i, j, we compute the fraction of candidate allocations where i and j are assigned to the same arm. When there is no pairwise assignment correlation, each pair i, j should appear in the same arm in half of the candidate allocations.

Iteration 2: Based on the score histogram in Figure 1, the fitness function appears to provide sufficient differentiation between allocations. Based on the desiderata trade-off plots, however, we observe that the interference metric is extremely coarse, taking on only a handful of distinct values. At the lowest value of the interference metrics, fine-grained fitness score differentiation is driven primarily by the balance metric. This may be appropriate if we are only interested in filtering out the most interference-prone allocations. If, on the other hand, we want to be more selective against interference even in the low-interference regimes, then a different interference metric is necessary.

Suppose we are able to collect more detailed information about the network of interactions between individuals. For instance, we can ask participants to list the friends with whom they communicate regularly and build a network based on each participant's list of contacts (Cai et al., 2015; Kim et al., 2015). In the network interference literature, it is common to assume that an individual is ''exposed'' through the treatments of their neighbors in the network via a neighborhood exposure model (Aronow & Samii, 2017; Manski, 2013). We therefore propose a new interference metric that counts the fraction of control individuals exposed to treatment through their network neighbors,

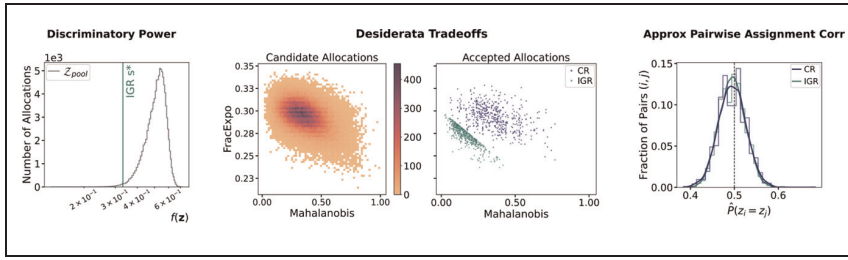

FIGURE 2. *Visual checks for discriminatory power, desiderata trade-offs, and over-restriction (via approximate pairwise assignment correlation) under fitness function* $\mathrm{f}^{(F+M)}$.

$$m^{(\mathrm{FracExpo})}(\mathbf{z}, \mathbf{A}) = \frac{\sum_{i=1}^{N} \mathbf{1}\left\{\sum_{j \in \mathcal{N}(i)} z_j > q \cdot |\mathcal{N}(i)|\right\} \cdot (1 - z_i)}{\sum_{i=1}^{N} (1 - z_i)}, \tag{4}$$

where $\mathbf{A}$ is an $N \times N$ adjacency matrix, $\mathcal{N}(i)$ is the set of nodes in the network that are connected to i by an edge, $\mathbf{z}_{\mathcal{N}(i)}$ is the vector of treatment assignments for the individuals in this neighborhood, and q is a scalar between 0 and 1 that defines an exposure threshold. In our simulation, we set q to 0.25, so that a control individual is considered exposed if at least 25% of their neighbors are treated. The new fitness function is

$$f^{(\mathrm{F+M})}(\mathbf{z}, \mathbf{X}) \quad = 0.5 \cdot m^{(\mathrm{FracExpo})}(\mathbf{z}, \mathbf{X}) + 0.5 \cdot m^{(\mathrm{Mahalanobis})}(\mathbf{z}, \mathbf{X}). \tag{5}$$

We recreate each of the plots from Iteration 1 (Figure 2). From the desiderata trade-off plots, we see that, as desired, the new interference metric based on network exposure is more granular than the metric based on distance. We also see that the new overall fitness function has adequate discriminatory power and has similar approximate pairwise assignment correlation to complete randomization.

Iteration 3: We decide to use the interference metric based on network exposure. In previous iterations, the interference and balance metrics were weighted equally. Now, we examine an alternative weighting scheme where more weight is applied to the interference metric. We therefore propose the fitness function,

$$f^{(\mathrm{F+M})}(\mathbf{z}, \mathbf{X}) \quad = 0.75 \cdot m^{(\mathrm{FracExpo})}(\mathbf{z}, \mathbf{X}) + 0.25 \cdot m^{(\mathrm{Mahalanobis})}(\mathbf{z}, \mathbf{X}). \tag{6}$$

Again, we construct plots to check discriminatory power, desiderata trade-offs, and over-restriction (Figure 3). From the desiderata trade-offs scatterplot, we see that a fitness function that upweights the interference metric accepts allocations with fewer exposed control units, at the cost of increased covariate imbalance. The pairwise assignment correlation remains similar to complete randomization.

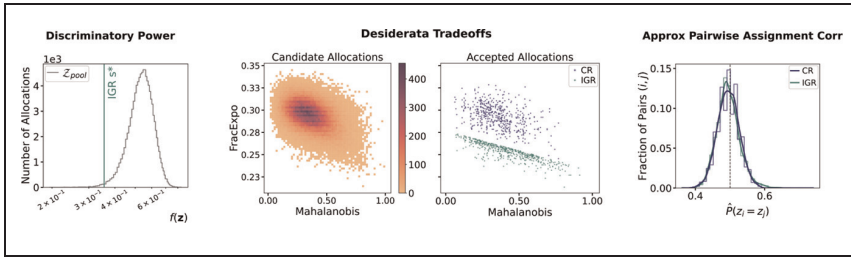

FIGURE 3. *Visual checks for discriminatory power, desiderata trade-offs, and over-restriction (via approximate pairwise assignment correlation) under fitness function* $\mathrm{f}^{(F+M)}$ *with an upweighted interference metric.*

Final decision: Ultimately, we proceed with the fitness function from Iteration 3 and the restriction rule used across iterations. We choose to prioritize mitigating interference over maximizing balance for two reasons. First, interference is typically very difficult, or even impossible, to measure and therefore hard to correct for during the analysis phase. Imbalance, on the other hand, can be addressed with standard covariate adjustment methods. Second, for the sexual assault prevention intervention in our simulated experiment, we particularly want to avoid false null findings because this can keep an intervention that helps protect girls' safety and well-being from being more widely implemented. Because interference often causes bias toward the null, we aim to prevent interference as much as possible. More generally, when deciding on a fitness function and restriction rule, study analysts may want to make methodological considerations, for example, whether statistical corrections can be made during the analysis phase, and policy considerations, for example, the consequences of a false negative or false positive finding.

4.2.5 Preregistration. When the research team is satisfied with the choice of fitness function and restriction rule, they lock in the assignment mechanism P_{IGR}. They then preregister the accepted allocations $\mathbf{z}^j$ in $\mathcal{Z}_{\text{pool}}$ that have $p_{\mathbf{z}_j} > 0$ and their associated probabilities $p_{\mathbf{z}}$. To guarantee valid statistical inference, we only need to preregister these accepted allocations and probabilities. However, it can also be helpful to preregister the fitness function f and the restriction rule r as an informative reference for the design of future trials.

Preregistration *for a Simulated Experiment.* We preregister the set of 500 accepted allocations, each with equal probability $\frac{1}{500}$ of being drawn. Additionally, we preregister the fitness function $f^{(\mathrm{F}+\mathrm{M})}$ and threshold restriction rule r.

4.2.6 Randomization. Finally, the observed allocation $\mathbf{z}^{\text{obs}} \sim P_{\text{IGR}}$ is drawn and deployed.

Randomization *in a Simulated Experiment*. We sample one of the 500 accepted allocations uniformly at random.

5. Analyzing IGR-Designed Experiments

In this section, we examine the bias properties of a difference-in-means estimate applied to an IGR-designed experiment. We first discuss bias under the usual no-interference assumption and then expand the discussion to settings where interference is present.

5.1 IGR Bias Properties

5.1.1 Unbiasedness Through the Mirror Property. We show that IGR produces an unbiased difference-in-means estimate of the sample average treatment effect if there is no interference, the IGR assignment mechanism satisfies the mirror property, and in every accepted allocation, the number of treated and control individuals are equal. For a binary intervention, the mirror property is satisfied if $P(Z_i = 0) = P(Z_i = 1)$ for all $i = 1, \ldots, N$. Suppose we enumerated *all* possible allocations in the *Enumeration* step. If we specify fitness function f and restriction rule r to be symmetric, such that $(r \circ f)(\mathbf{X}, \mathbf{z})) = (r \circ f)(\mathbf{X}, \mathbf{1} - \mathbf{z}))$, the mirror property will be satisfied by P_{IGR}. However, because we may not be able to enumerate all possible allocations, satisfaction of the mirror property is not guaranteed. Instead, we can update P_{IGR} by ensuring that, for every allocation that is accepted, its mirror allocation is also accepted. Let $\tilde{\mathcal{Z}}_{\text{pool}} = \{1 - \mathbf{z} \mid \mathbf{z} \in \mathcal{Z}_{\text{pool}}\}$ denote the set of mirror allocations for each allocation in $\mathcal{Z}_{\text{pool}}$. The assignment mechanism $\tilde{P}_{\text{IGR}}$ satisfies the mirror property, where

$$\tilde{P}_{\text{IGR}}(\mathbf{Z} = \mathbf{z}) = \begin{cases} 0 & \mathbf{z} \notin \mathcal{Z}_{\text{pool}} \cup \tilde{\mathcal{Z}}_{\text{pool}} \\ 0 & \mathbf{z} \in \mathcal{Z}_{\text{pool}} \cup \tilde{\mathcal{Z}}_{\text{pool}}, (r \circ f)(\mathbf{X}, \mathbf{z}) = 0 \\ p_z & \mathbf{z} \in \mathcal{Z}_{\text{pool}} \cup \tilde{\mathcal{Z}}_{\text{pool}}, (r \circ f)(\mathbf{X}, \mathbf{z}) = 1 \end{cases}.$$

Proposition 1. *Assume that there is no interference and that the consistency property holds, so* $\mathrm{Y_i} = \mathrm{Y_i}(1)\mathrm{Z_i} + \mathrm{Y_i}(0)(1 - \mathrm{Z_i})$. *If the assignment mechanism* $\tilde{\mathrm{P}}_{\text{IGR}}$ *satisfies the mirror property and assigns an equal number of individuals to treatment and control in every allocation, then the difference-in-means estimate for the sample average treatment effect is unbiased.*

Proof. See Supplemental Appendix F.1 (available in the online version of this article). □

5.1.2 Bias in the Presence of Interference. In an experiment with interference where the target estimand is the global average treatment effect, a design that satisfies the mirror property does not necessarily give an unbiased difference-in-means estimate because the interference itself introduces bias. To reduce interference bias, we may need to use a design that violates the mirror property. When the mirror property is violated, however, the individual's treatment assignment and potential outcome may no longer be independent, resulting in possible bias from confounding. It may nevertheless be beneficial to break the mirror property so long as doing so produces a sufficient reduction in interference bias.

In IGR, violation of the mirror property occurs when we choose an asymmetric fitness function and restriction rule pair to restrict the design. To understand why an asymmetric restriction can help to prevent interference, we consider a toy example. Suppose a subset of three or more individuals in the study is connected by a star-shaped network, such that there is a single central individual who is connected to every outer individual, and all outer individuals are only connected to this central individual. Suppose further that, for a unit assigned to the control level, exposure to treatment occurs if at least two neighbors are treated. Then, in the allocation where the center individual is treated and the outer individuals are control, none of the control individuals are exposed. Conversely, in the mirror allocation where the center individual is control and the outer individuals are treated, the center control individual becomes exposed. The fitness function and restriction rule could be specified so that, to rule out high-interference allocations, the latter allocation is rejected while the former is accepted. In Supplemental Appendix F.2 (available in the online version of this article), we derive conditions under which asymmetric restriction in IGR improves overall bias for a linear outcomes model with an additive direct effect and an additive neighborhood exposure spillover effect.

Note that, even if the fitness function and restriction rule are asymmetric, we can force the IGR assignment mechanism to satisfy the mirror property by incorporating all mirror allocations as ''accepted'' allocations, regardless of whether they actually satisfy the criteria for acceptance. The analyst may consider enforcing the mirror property if the difference between the scores $f(\mathbf{X}, \mathbf{z})$, and $f(\mathbf{X}, \mathbf{1} - \mathbf{z})$ is small for the $\mathbf{z}$s in the candidate pool of allocations. We empirically investigate the implications of including and excluding mirror allocations in Section 6.

5.2 Randomization Inference for IGR-Designed Experiments

We take a design-based approach to inference, meaning we view the potential outcomes as fixed conditional on the study sample, so that the only randomness in the study arises through the randomness in the treatment assignment

mechanism. In IGR specifically, randomness comes from the *Randomization* step, where the observed allocation is sampled from the restricted assignment mechanism, P_{IGR}. Randomization then forms the basis for inference, and we can perform a Fisher randomization test to evaluate null hypotheses. Specifically, we compute the value of the test statistic under the null hypothesis for different possible assignment vectors. With IGR, the empirical null distribution for the test statistic is derived over the set of accepted assignment vectors $\mathbf{z}^j$ in $\mathcal{Z}_{\text{pool}}$ that have $p_{\mathbf{z}_j}>0$. We compare the observed value of the test statistic for $\mathbf{z}^{\text{obs}}$ to the empirically derived null distribution of the test statistic to obtain a *p*-value.

6. How IGR Improves Effect Estimates

Using IGR to design an experiment can lead to less biased, more precise effect estimates. We demonstrate these benefits with results from our simulated experiment (refer to Section 4.1 for a description of the simulation). In the *Evaluation and Adaptation* step in Section 4.2, we chose the fitness function, $f^{(\mathtt{F+M})}(\mathbf{z},\mathbf{X}) = 0.75 \cdot m^{(\mathtt{FracExpo})}(\mathbf{z},\mathbf{X}) + 0.25 \cdot m^{(\mathtt{Mahalanobis})}(\mathbf{z},\mathbf{X})$, to apply to the design of the simulated experiment. Because the interference metric $m^{(\mathtt{FracExpo})}$ is asymmetric, the resulting restricted assignment mechanism does not generally satisfy the mirror property unless we incorporate all mirror allocations, including those not accepted by the restriction rule (refer to Section 5.1 for a discussion on the mirror property and estimation bias). To understand the implications of satisfying versus violating the mirror property, we examine effect estimates both with and without the inclusion of mirror allocations. We additionally assess how the estimation properties change when we apply different weights to the interference and balance metrics in the fitness function; below, we refer to the weight applied to the interference metric, $m^{(\mathtt{FracExpo})}$, as $w_{\text{Interference}}$. We also show how, when the interference metric is upweighted in the design phase, the incurred covariate imbalance can be corrected for in the analysis phase through covariate adjustment. Results are reported based on five repeatedly drawn samples of the covariate and outcome data.

When mirrors are included and $w_{\text{Interference}} = 0.75$, IGR reduces the absolute bias of the difference-in-means estimate (88.31% $\pm$ 0.56% of the absolute bias of an unrestricted design; Figure 4). As the magnitude of the weight applied to the interference metric increases, we observe more bias reduction (86.18% $\pm$ 0.37% of the absolute bias of an unrestricted design when $w_{\text{Interference}} = 1$). At small values of $w_{\text{Interference}}$, IGR decreases the variance of the difference-in-means estimate (e.g., variance at $w_{\text{Interference}} = 0.125$ is 0.003 ± 0.0007 compared to 0.61 ± 0.10 for an unrestricted design). Increasing $w_{\text{Interference}}$ leads to variance inflation (e.g., variance at $w_{\text{Interference}} = 0.875$ is 1.01 ± 0.28). When $w_{\text{Interference}} = 0.75$, leaving mirrors out leads to a large

reduction in absolute bias for some data samples (e.g., 26.67% of the absolute bias of the unrestricted design) and no reduction for others (e.g., 103.89% of the absolute bias of the unrestricted design). This suggests that, in our example, including mirror allocations helps to make the estimation bias less dependent on the particular sample of individuals in the study.

Instead of a difference-in-means estimator, we can also fit a linear regression model that adjusts for observed baseline covariates. As mentioned in the *Evaluation and Adaptation* step in Section 4.2, using this estimator can be helpful when we choose to upweight the interference metric and downweight the balance metric in the design phase. As expected, the regression estimator substantially reduces variance compared to the difference-in-means estimator, particularly when $w_{\text{Interference}}$ is large (Figure 5). When mirror allocations are excluded, the bias in the regression estimate is more stable across data samples compared to the difference-in-means estimate. On the other hand, when mirror allocations are included, the regression estimate generally provides less bias reduction than the difference-in-means estimate ($93.36\% \pm 0.67\%$ of the absolute bias of an unrestricted design at $w_{\text{Interference}} = 0.75$).

From the above, we reason that, when using a difference-in-means estimator, it is beneficial to include mirror allocations and not to overweight the interference metric, for example, above 0.5. If using the regression estimator, however, mirror allocations can be excluded, and a larger $w_{\text{Interference}}$ can be applied without needing to worry about the variance of the estimator. We note, however, that a difference-in-means estimator is generally less amenable to *p*-hacking, which is one factor to consider when choosing between these two estimators.

7. Discussion

We introduced IGR as a flexible framework for performing restricted randomization with diverse design desiderata and used illustrative examples to showcase its usefulness in different experimental settings. IGR is distinct from prior restricted randomization frameworks along several fronts. It allows for multiple design desiderata, including and beyond balance, thus expanding the use cases for restricted randomization; we showed this with an example of an experiment with interference in Section 4 and a group formation experiment in Supplemental Appendix E (available in the online version of this article). IGR also separates the process of enumerating candidate allocations—which can be done with an efficient, directed search—from the act of randomization; we explicitly make enumeration and randomization separate steps in the IGR framework. In addition, IGR has a built-in iterative loop for evaluating and adapting the restricted assignment mechanism; we show this in action in Section 4.2, using the visual tools described in Supplemental Appendix D (available in the online version of this article). Lastly, IGR enforces

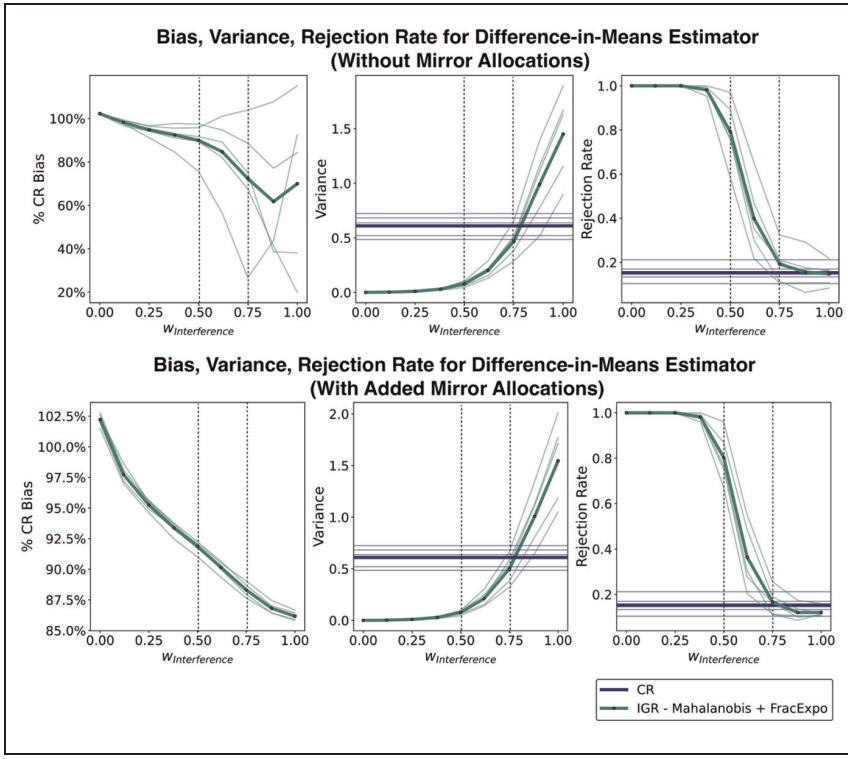

FIGURE 4. *Bias, variance, and rejection rate of the difference-in-means estimator in an IGR design versus an unrestricted CR design when mirror allocations are excluded (top) and when mirror allocations are included (bottom). Bias is plotted as a percentage relative to the bias of a CR design.*
Note. CR = complete randomization; IGR = inspection-guided randomization.

preregistration of the accepted allocations, thus promoting transparency and reproducibility; in Supplemental Appendix G (available in the online version of this article), we show that, without preregistration, it is easy to inflate the Type I error rate above the nominal level.

The core motivation for IGR, as for any restricted randomization, is to prevent the possibility of choosing a bad randomization allocation that leads to an effect estimate that deviates dramatically from the true value of the estimand. For obviously bad allocations, there is little reason not to discard them; this has been well understood since the early days of the experimental design literature. In a conversation between Cochran and Fisher (recounted by Rubin and

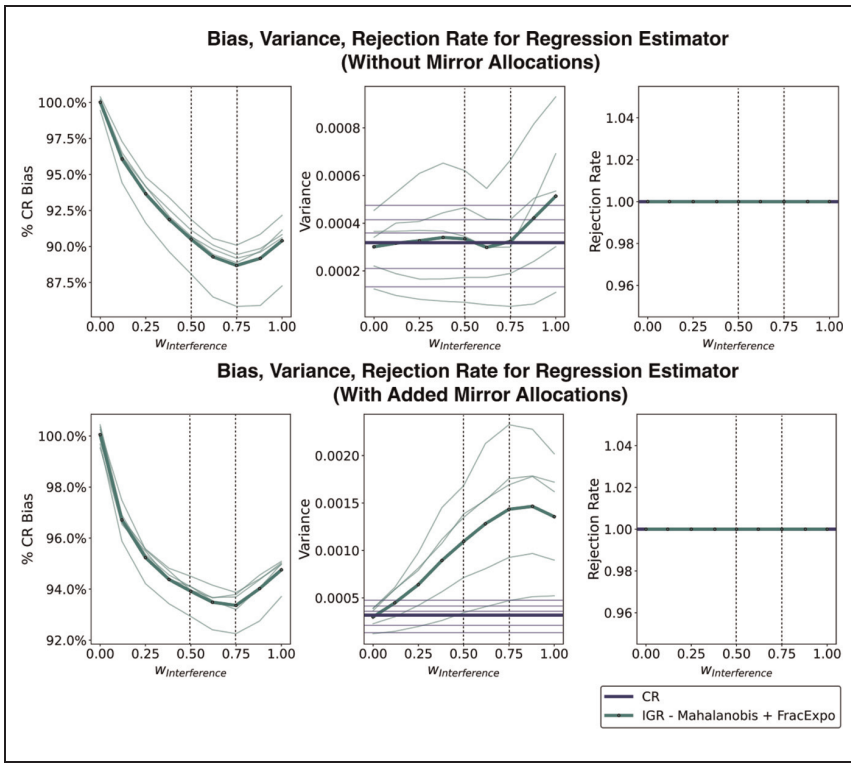

FIGURE 5. *Bias, variance, and rejection rate of the linear regression estimator in an IGR design versus an unrestricted CR design when mirror allocations are excluded (top) and when mirror allocations are included (bottom). Bias is plotted as a percentage relative to the bias of a CR design.*
Note. CR = complete randomization; IGR = inspection-guided randomization.

Cochran), Fisher is reported to have said that, confronted with a randomized allocation that ''exhibited substantial imbalance on a prognostically important baseline covariate,'' he would ''of course'' re-randomize if the experiment had not started yet (Rubin, 2008).

Historically, the focus of restricted randomization has been on achieving covariate balance in two-arm experiments that compare a single treatment intervention to a control or status quo. However, some interventions, often behavioral ones, need to be evaluated through more complicated restricted designs. Researchers across several disciplines, including medicine, education, psychology, and computer science, have recognized this need and independently

applied principles of the IGR framework to experiment design, albeit without the formal structure of IGR (Cho et al., 2021, 2022; Kizilcec et al., 2022; Murnane et al., 2023; Zahrt et al., 2023). Cho et al. (2021, 2022) for example, used a form of restricted randomization to achieve balance on prognostic covariates in trials with more than two arms. These existing real-life applications suggest that IGR is conceptually straightforward and intuitive, and they also point to how an explicit IGR framework can help to bring consistency to existing efforts to expand the scope of restricted randomization.

In this work, we showed how the IGR framework can be used to design experiments in two settings where existing restricted randomization frameworks are not immediately applicable. In our first example, we applied IGR to the design of experiments with interference. Experimental designs that seek to mitigate interference often do so through a form of cluster randomization, either over naturally existent clusters or over algorithmically detected ones (Ugander & Yin, 2023; Ugander et al., 2013). Without IGR, the study analyst chooses between randomizing at the more aggregated level to limit interference, thus sacrificing power and covariate balance, or randomizing at a more granular level to increase efficiency, thus permitting more interference bias. With IGR, the analyst can construct fitness functions that combine both interference and balance metrics, thus reducing interference bias while maintaining efficiency. IGR can also be used even if we have limited knowledge of the interference structure and need to rely on an approximate measure of interference. While we explored a specific setting where interference occurred via neighborhood exposure in a network of interactions, IGR can easily be applied to other interference structures by modifying the inspection metric used to measure the extent of anticipated interference.

In Supplemental Appendix E (available in the online version of this article), we further show how IGR can be applied to the design of group formation experiments, where effect sizes may exhibit dependence on group composition. Recent work has highlighted that peer context should be taken into consideration as a driver of treatment effect heterogeneity for behavioral and psychological interventions (Carroll et al., 2017; Grover et al., 2023; Kizilcec & Cohen, 2017; Kizilcec et al., 2020; Walton & Yeager, 2020; Yeager et al., 2019). Understanding how an intervention depends on peer context has important implications for understanding causal mechanisms, devising policies that target certain contexts, and designing concurrent interventions that can be used to manipulate contexts.

We call out the *Evaluation and Adaptation* step of IGR as an especially critical component of the framework. In some sense, the inclusion of the *Evaluation and Adaptation* step is a recognition that the study analyst is not all-knowing and hence may not select a good fitness function and restriction rule in their first attempt. In this work, we highlighted three issues to consider in *Evaluation and*

Adaptation: desiderata trade-offs, discriminatory power, and over-restriction. We introduced visual tools to help with assessing each of these issues (described in detail in Supplemental Appendix D [available in the online version of this article] and applied in Section 4.2). For example, in our simulated experiment, the analyst may initially choose to assign equal weight to the interference and balance metrics (Figure 2). After considering the risk of a false null, however, the analyst may choose to modify the weighting scheme to instead upweight the interference metric. These iterative refinements push the design toward one that is better and more likely to yield credible inference. As mentioned in Related Works (Section 2), an emphasis on detecting and correcting flawed restricted assignment mechanisms is missing from standard restricted randomization frameworks such as re-randomization.

Fundamentally, IGR compels the study analyst to invest time and care in the design phase of the experiment. The *Specification* and *Evaluation and Adaptation* steps in particular require considerable thoughtfulness, from specifying the study's design desiderata, to developing the inspection metrics used to measure desiderata satisfaction, to constructing and evaluating the fitness function and restriction rule that characterize the restricted assignment mechanism. Attentiveness to design can produce large payoffs in the bias and precision of effect estimates. While using a tailored estimator in the analysis phase can sometimes produce similar payoffs, focusing efforts on the design phase has additional merits. For one, even the most clever estimators may not be able to salvage a study that was performed under an extremely poor choice of randomization allocation. In addition, as previously mentioned, the *Preregistration* step precludes statistical hacking, as long as the analyst is held accountable to doing randomization inference over a preregistered space of randomization allocations (see Supplemental Appendix G [available in the online version of this article] for an illustration of Type I error rate inflation when the analyst is not held accountable to a preregistered design and analysis plan).

There are several interesting directions for future work on IGR. Since the fitness function in IGR is handcrafted, it may be useful to develop theory on how similar the specified fitness function needs to be to the true conditional mean squared error in order for IGR to produce better effect estimates than naive randomization strategies. When the fitness function is not a faithful proxy, then analysts may choose to use a non-IGR design. Alternatively, analysts could choose to improve the fitness function by using additional sources of data and domain knowledge. If experiments consist of multiple waves or are preceded by a pilot study, there is an opportunity to use information from the prior wave or pilot study to refine the fitness function for the next wave or the main trial. In an experiment with anticipated interference, for example, researchers could use a pilot study to estimate the interference network structure within a subsample of the study participants, then extrapolate this network to the entirety of the study

sample. Fitness functions for new experiments could potentially also leverage data from prior historical experiments that evaluated different but comparable interventions in similar settings. Lastly, to make IGR as accessible as possible to study analysts, good software for implementation will be useful; this will likely include a dashboard interface that allows users to easily perform each of the steps of IGR, including interacting with visualizations.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgement

The authors thank Timothy Sudijono for help with outlining the proof in Appendix G.3.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Maggie Wang is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2146755.

ORCID iDs

Maggie Wang https://orcid.org/0000-0002-1102-0969
René F. Kizilcec https://orcid.org/0000-0001-6283-5546

References

Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, *11*(4), 1912–1947.

Athey, S., Eckles, D., & Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, *113*(521), 230–240.

Baiocchi, M., Omondi, B., Langat, N., Boothroyd, D. B., Sinclair, J., Pavia, L., Mulinge, M., Githua, O., Golden, N. H., & Sarnquist, C. (2017). A behavior-based intervention that prevents sexual assault: The results of a Matched-Pairs, Cluster-Randomized study in Nairobi, Kenya. *Prevention Science*, *18*(7), 818–827.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454.

Basse, G. W., & Airoldi, E. M. (2018). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, *105*(4), 849–858.

Basse, G. W., Ding, P., Feller, A., & Toulis, P. (2024). Randomization tests for peer effects in group formation experiments. *Econometrica*, *92*(2), 567–590.

Box, G. E. P., Stuart Hunter, J., & Hunter, W. G. (2005). *Statistics for experimenters: Design, innovation, and discovery*. Wiley.

Branson, Z., Dasgupta, T., & Rubin, D. B. (2016). Improving covariate balance in 2K factorial designs via rerandomization with an application to a New York City Department of Education high school study. *Annals of Applied Statistics*, *10*(4), 1958–1976.

Cai, J., De Janvry, A., & Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, *7*(2), 81–108.

Carroll, J. M., Yeager, D. S., Buontempo, J., Hecht, C., Cimpian, A., Mhatre, P., Muller, C., & Crosnoe, R. (2023). Mindset × context: Schools, classrooms, and the unequal translation of expectations into math achievement. *Monographs of the Society Research in Child Development*, *88*(2), 7–109.

Cho, J., Li, Y., Armstrong, A. K., Russ, A., Krasny, M. E., & Kizilcec, R. F. (2021). Using social norms to promote actions beyond the course. In *Proceedings of the eighth ACM conference on Learning @ Scale*, L@S '21 (pp. 161–172). Association for Computing Machinery.

Cho, J., Li, Y., Kudryavtsev, A., Kizilcec, R. F., & Krasny, M. (2022). Online learning and social norms: Evidence from a cross-cultural field experiment in a course for a cause.

Ciolino, J. D., Diebold, A., Jensen, J. K., Rouleau, G. W., Koloms, K. K., & Tandon, D. (2019). Choosing an imbalance metric for covariate-constrained randomization in multiple-arm cluster-randomized trials. *Trials*, *20*(1), 293.

Cox, D. R. (1958). *Planning of experiments*. Wiley.

Cox, D. R. (2009). Randomization in the design of experiments. *International Statistical Review*, *77*(3), 415–429.

de Hoop, E., Teerenstra, S., van Gaal, B. G. I., Moerbeek, M., & Borm, G. F. (2012). *The "best balance" allocation led to optimal balance in cluster-controlled trials*. *Journal of Clinical Epidemiology*, *65*(2), 132–137.

Eckles, D., Karrer, B., & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, *5*(1), 20150021.

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

Fisher, R. A. (1992). The arrangement of field experiments. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 82–91). Springer.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*(2), 180–193.

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

Grover, S. S., Ito, T. A., & Park, B. (2017). The effects of gender composition on women's experience in math work groups. *Journal of Personality and Social Psychology*, *112*(6), 877–900.

Hayes, R. J., & Moulton, L. H. (2017). *Cluster randomised trials* (2nd ed.). Chapman and Hall/CRC.

Higgins, M. J., Sävje, F., & Sekhon, J. S. (2016). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7369–7376.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842.

Imai, K., King, G., & Nall, C. (2009). The essential role of pair matching in Cluster-Randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *SSO Schweiz Monatsschr Zahnheilkd*, *24*(1), 29–53.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, *24*(3), 324–338.

Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H., & Christakis, N. A. (2015). Social network targeting to maximise population behaviour change: A cluster randomised controlled trial. *Lancet*, *386*(9989), 145–153.

Kizilcec, R. F., & Cohen, G. L. (2017). Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *Proceedings of the National Academy of Sciences*, *114*(17), 4348–4353.

Kizilcec, R. F., Mimno, J. A., & Karhan, A. J. (2022). Effects of framing professional development as a career growth opportunity on course completion. In *Proceedings of the Ninth ACM Conference on Learning @ Scale*, L@S '22 (pp. 335–339). Association for Computing Machinery.

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26), 14900–14905.

Krieger, A. M., Azriel, D., Sklar, M., & Kapelner, A. (2020). *Improving the power of the randomization test*. arXiv:2008.05980.

Li, F., Lokhnygina, Y., Murray, D. M., Heagerty, P. J., & DeLong, E. R. (2016). An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Statistics in Medicine*, *35*(10), 1565–1579.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Economic Journal*, *16*(1), S1–S23.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT Press.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, *40*(2), 1263–1282.

Moulton, L. H. (2004). Covariate-based constrained randomization of group-randomized trials. *Clinical Trials*, *1*(3), 297–305.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021.

Murnane, E. L., Glazko, Y. S., Costa, J., Yao, R., Zhao, G., Moya, P. M. L., & Landay, J. A. (2023). Narrative-Based visual feedback to encourage sustained physical activity: A field trial of the WhoIsZuki mobile health platform. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *7*(1), 1–36.

Nordin, M., & Schultzberg, M. (2022). Properties of restricted randomization with implications for experimental design. *Journal of Causal Inference*, *10*(1), 227–245.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, *102*(477), 191–200.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*(484), 1350–1356.

Sarnquist, C., Friedberg, R., Rosenman, E. T. R., Amuyunzu-Nyamongo, M., Nyairo, G., & Baiocchi, M. (2024). Sexual assault among young adolescents in informal settlements in Nairobi, Kenya: Findings from the IMPower and SOS Cluster-Randomized controlled trial. *Prevention Science*, *25*(4), 578–589.

Tukey, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials*, *14*(4), 266–285.

Ugander, J., Karrer, B., Backstrom, L., & Kleinberg, J. (2013). Graph cluster randomization: network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, KDD* (pp. 329–337). Association for Computing Machinery.

Ugander, J., & Yin, H. (2023). Randomized graph cluster randomization. *Journal of Causal Inference*, *11*(1), 20220014.

Valente, T. W. (2012). Network interventions. *Science*, *337*(6090), 49–53.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Walton, G. M., & Yeager, D. S. (2020). Seed and soil: Psychological affordances in contexts help to explain where wise interventions succeed or fail. *Current Directions in Psychological Science*, *29*(3), 219–226.

Watson, S. I., Girling, A., & Hemming, K. (2021). Design and analysis of three-arm parallel cluster randomized trials with small numbers of clusters. *Statistics in Medicine*, *40*(5), 1133–1146.

Xu, Z., & Kalbfleisch, J. D. (2010). Propensity score matching in randomized clinical trials. *Biometrics*, *66*(3), 813–823.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., & . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369.

Zahrt, O. H., Evans, K., Murnane, E., Santoro, E., Baiocchi, M., Landay, J., Delp, S., & Crum, A. (2023). Effects of wearable fitness trackers and activity adequacy mindsets on affect, behavior, and health: Longitudinal randomized controlled trial. *Journal of Medical Internet Research*, *25*, e40529.

Zhou, Y., Turner, E. L., Simmons, R. A., & Li, F. (2021). *Constrained randomization and statistical inference for multi-arm parallel cluster randomized controlled trials*. *Statistics in Medicine*, *41*(10), 1862–1883.

Authors

MAGGIE WANG is a PhD student in the Department of Biomedical Data Science at Stanford University, Stanford, CA, USA. Her interests are in causal inference and experimental design for behavioral and social interventions.

RENÉ F. KIZILCEC is an associate professor in the Ann S. Bowers College of Computing and Information Science at Cornell University, Ithaca, NY, USA. His interests are in the behavioral, psychological, and computational aspects of technology in education and how they can inform practices and policies that promote learning, equity, and academic and career success.

MICHAEL BAIOCCHI is an associate professor in the Department of Epidemiology and Population Health at Stanford University, Stanford, CA, USA. His interests are in creating and rigorously evaluating behavioral interventions. His research is tied together by the unifying goal of bringing rigorous, quantitative approaches to bear on messy, real-world questions in order to improve people's lives.

Manuscript received August 27, 2024
Revision received January 29, 2025
Accepted April 3, 2025