Submitted to *Econometrica*

UNCONDITIONAL RANDOMIZATION TESTS FOR INTERFERENCE

LIANG ZHONG
Faculty of Business and Economics, The University of Hong Kong

Researchers are often interested in whether and how far interference extends across units, but complex dependence can invalidate inference. This paper introduces the pairwise imputation-based randomization test (PIRT), a design-based framework for testing partially sharp spillover nulls in experiments. PIRT uses unconditional randomization to compare each observed–reassigned assignment pair using only outcomes jointly imputable under that pair. The resulting procedure is straightforward and finite-sample valid without modeling network or spatial dependence. Beyond validity, I characterize when PIRT is powerful relative to conditional randomization tests: richer jointly imputable samples raise power by improving precision, whereas changes in sample composition across reassignments and the stricter cutoff required for validity can lower it. Simulations calibrated to the Bogotá hotspot-policing experiment show that PIRT can achieve competitive or higher power at substantially lower computational cost. Applied to the observed data, PIRT finds suggestive evidence of violent-crime displacement but no evidence of property-crime displacement.

KEYWORDS: Causal inference, non-sharp null hypothesis, dense networks.

1. INTRODUCTION

Many experiments evaluate costly place-based or network interventions whose effects need not stop at treated units. Police patrols may displace crime to nearby streets, public-health interventions may affect untreated contacts, and information treatments may diffuse

Liang Zhong: samzl@hku.hk

I am deeply grateful to my advisors, Hiroaki Kaido and Jean-Jacques Forneron, for their invaluable guidance and support. I also thank M. Daniele Paserman, Kevin Lang, Ivan Fernandez-Val, Pierre Perron, Marc Rysman, Xunkang Tian, Stella Hong, Zhongjun Qu, Benjamin Marx, Haoran Pan, Lei Ma, Justin Hong, Qingyuan Chai, Mohsen Bayati, Kirill Ponomarev, Guillaume Pouliot, Jizhou Liu, Julius Owusu, Panos Toulis, Doosoo Kim, Shakeeb Khan, Dmitry Arkhangelsky, Matias Cattaneo, José R. Zubizarreta, Demian Pouzo, Avi Feller, Xinran Li, Peter Deffebach, Juan Pantano, Toru Kitagawa, Federico Bugni, and participants at various workshops and seminars for their helpful comments. All errors are my own.

through social ties. In such settings, the outcome for one unit may depend not only on its own treatment, but also on the treatment assigned to others—a phenomenon known as interference. Understanding the extent of interference is crucial: it refines causal inference and model specification (Tella and Schargrodsky, 2004, Angrist, 2014), and it guides efficient resource allocation, especially when interventions are costly (Baird et al., 2018, Brollo et al., 2020).

However, testing for interference is econometrically challenging because complex dependence structures can make large-sample approximations difficult to justify (Kelly, 2021, Blattman et al., 2021), and even in randomized experiments, valid inference may require assumptions beyond random assignment alone (Aronow, 2012, Pollmann, 2020). This has led recent work (e.g., Bond et al., 2012, Blattman et al., 2021) to use randomization-based methods such as the Fisher Randomization Test (FRT). But the FRT is generally valid only for the *sharp null* of no treatment effect, under which all potential outcomes are imputable (Rosenbaum, 2007, Athey et al., 2018).[1] In interference settings, this null excludes both direct effects and spillovers, so its rejection cannot distinguish between them.

In this paper, I define *partially sharp null hypotheses*, using the terminology of Zhang and Zhao (2023), and introduce the *pairwise imputation-based randomization test* (PIRT), an unconditional design-based framework for testing such hypotheses. A null is partially sharp when it restricts equality of potential outcomes only over selected subsets of treatment assignments, rather than over the full assignment space. The leading application is to test whether interference extends beyond a prespecified distance threshold ϵ: the null allows direct treatment effects and nearby spillovers, while restricting the effect of more distant treatments. PIRT treats the potential outcome schedule as fixed and uses random treatment assignment as the sole source of randomness (Abadie et al., 2020, 2023). Because it relies on the experimental design rather than a model for network dependence, it remains applicable in dense or otherwise complex settings.[2]

[1] Here, "imputable" means that the missing potential outcomes can be inferred from the observed data under the null hypothesis. Under a sharp null, this is true for all units and all assignments.

[2] The proposed method can also be used directly for causal inference by inverting the test, as in a traditional randomization test.

Partially sharp null hypotheses create two related difficulties for randomization inference. First, only part of the potential outcome schedule is imputable under the null. Second, the set of imputable units depends on the assignment. For example, under the null of no peer effects on untreated units, outcomes for untreated units may be imputable, whereas outcomes for treated units are not; as the treatment assignment changes, the identities of untreated units also change. Together, these challenges complicate the direct application of traditional randomization inference methods and underscore the need for specialized approaches.

I address the first difficulty by introducing *pairwise imputable statistics*. These statistics are defined on pairs of assignments: one assignment determines the relevant grouping or comparison, while the pair of assignments together determines which units are jointly imputable under the null. They remain close in spirit to familiar quantities such as differences in means or regression-based contrasts (Imbens and Rubin, 2015), but are evaluated only on units whose outcomes are jointly imputable under the null for both assignments. For example, a difference in means may compare units near treated locations to units farther away, while restricting the calculation to units jointly imputable under the observed and reassigned treatments.

I address the second difficulty by establishing a novel connection between randomization tests for non-sharp null hypotheses and the selective inference literature (Vovk, 2015, Barber et al., 2021, Guan, 2024). For each observed–reassigned assignment pair, PIRT computes the same pairwise imputable contrast twice, using the same observed outcomes from the same jointly imputable units: once grouping those units according to the reassignment and once according to the observed assignment. The p-value records how often grouping by the reassignment produces a contrast at least as large as grouping by the observed assignment. Under the null, exchanging the roles of the two assignments reverses the same numerical comparison. This pairwise symmetry yields unconditional finite-sample size control.

The paper then studies when this broadly applicable procedure is powerful. Because PIRT does not require fixed focal units or assignments, it can exploit assignment pairs that retain a rich jointly imputable set; this effective-overlap gain is valuable when spillovers are subtle and data collection is costly (Taylor and Eckles, 2018, Breza et al., 2020). The price of this generality is that reassignment can change which units enter the pairwise com-

parison, adding leave-one-out variation to the reference distribution, while finite-sample validity requires a conservative cutoff. Relative to conditional randomization tests, PIRT is therefore most useful when its effective-overlap gain outweighs the added reference variation and stricter cutoff. To show that PIRT's favorable region is empirically relevant, I use the large-scale hotspot-policing experiment of Blattman et al. (2021) in Bogotá, both as the basis for calibrated simulations and as an empirical application. In this spatial setting, the simulations show that PIRT provides valid inference, good power, and substantially lower computational burden than the conditioning-based alternative. Applying PIRT to the observed data, I find suggestive evidence of violent-crime displacement, but not property-crime displacement.

This paper contributes to three strands of literature. First, it contributes to the study of interference and network effects. Following Manski (1993), a large literature has developed model-based approaches that rely on parametric structure; see Graham (2020) for a review. These methods can be informative, but they often require restrictive assumptions on the form of interaction and must confront the high dimensionality of network dependence. By contrast, the present paper uses the null hypothesis itself to reduce dimensionality, yielding a nonparametric design-based approach that is flexible with respect to network or spatial structure.

Second, the paper contributes to the literature on randomization-based inference under interference (Aronow, 2012, Athey et al., 2018). Because classical Fisher randomization tests require sharp nulls, recent work has proposed conditional randomization tests that restore sharpness by restricting attention to carefully chosen subsets of units or assignments. These procedures can be powerful, but they often rely on conditioning choices that are tailored to particular interference structures, can be difficult to develop in general settings, and leave scope for implementation discretion (Basse et al., 2019, 2024, Puelz et al., 2022). The present paper develops an alternative route: rather than restoring sharpness through conditioning, it works directly with pairwise imputability. The resulting procedure remains finite-sample valid and straightforward to implement even when informative conditioning strategies are hard to construct.

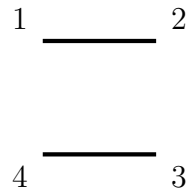
Third, the paper contributes more broadly to randomization testing under hypotheses that are not fully sharp. While the main focus is on partially sharp null hypotheses defined through distance, the logic of PIRT is more general and may be useful beyond network

settings, including staggered adoption designs such as those considered in Zhang and Zhao (2025).

2. SETUP AND NULL HYPOTHESIS OF INTEREST

Consider N units indexed by $i \in \{1, \ldots, N\}$, linked through a pre-treatment network or spatial structure. I study the extent of interference as a function of proximity, summarized by an $N \times N$ matrix G, where smaller G_{ij} means that units i and j are closer. The measure may be continuous or discrete, depending on the application; I normalize $G_{ii} = 0$ for all i and assume $G_{ij} > 0$ for all $i \neq j$. The precise interpretation of proximity is context specific.[3] The following running example illustrates the notation.

Running Example. Consider four street segments, labeled 1, 2, 3, and 4, where two segments are adjacent if they are directly connected, as shown in Figure 1. Units 1 and 2 form one area, and units 3 and 4 form another. For simplicity, I set the within-area distance to 1 and the across-area distance to 2. In practice, the distance measure should be chosen to reflect the economic or institutional channel through which interference may operate.



(a) Network structure

$$G = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 2 & 2 \\ 2 & 2 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{pmatrix}$$

(b) Distance matrix

FIGURE 1.—Running example: network structure and distance matrix.

I focus on experimental settings in which treatment assignment is random and follows a known probability distribution P, where $P(d) = \Pr(D = d)$ denotes the probability that the treatment assignment vector D equals d. Let X denote observed pre-treatment covariates. The assignment distribution P may incorporate covariate-based restrictions, such as strata,

[3] The same notation covers geographic distance, adjacency, network ties, travel time, interaction intensity, or product-space similarity, after any pre-specified transformation. The tests are design-valid for the null defined by the fixed G. If G is only a proxy for the channel through which spillovers operate, the test still answers the stated G-based question, but power and substantive interpretation may change.

blocks, or matched pairs; complete and cluster randomization are special cases in which no such restrictions are imposed.

I adopt the potential outcomes framework with binary treatment assignment vector $D = (D_1, \ldots, D_N) \sim P$, where $D \in \{0,1\}^N$ and D_i indicates whether unit i is treated. Let $Y(d) = (Y_1(d), \ldots, Y_N(d))$ denote the vector of potential outcomes under assignment d. Since $Y_i(d)$ may depend on the treatment assignments of all units, this setup allows for interference and relaxes the usual SUTVA restriction (Cox, 1958).

The observed data consist of the realized assignment D^{obs}, the realized outcome vector $Y^{\text{obs}} \equiv Y(D^{\text{obs}})$, the proximity matrix G, the covariates X, and the design P. I adopt a design-based perspective: D is random, while G, X, P, and the potential outcome schedule $Y(\cdot)$ are treated as fixed. For notational simplicity, these objects will usually be suppressed from function arguments.

Running Example Continued. Suppose the outcome of interest is the total number of crimes recorded over a year. In the running example, exactly one unit receives treatment, so $P(d) = 1/4$ for each assignment in the support. Let the observed assignment be $D^{\text{obs}} = (1,0,0,0)$, with observed outcomes $Y^{\text{obs}} = (2,4,3,2)$.

Table I displays the corresponding potential outcome schedule under the design-based perspective. The first row is the realized dataset. Although the full schedule of potential outcomes is fixed, only the row corresponding to D^{obs} is observed. In general, once interference is allowed, each unit may have up to 2^N potential outcomes.

TABLE I

POTENTIAL OUTCOME SCHEDULE IN THE RUNNING EXAMPLE.

Assignment D	Potential Outcome Y_i			
	1	2	3	4
$(1,0,0,0)$	2	4	3	2
$(0,1,0,0)$	?	?	?	?
$(0,0,1,0)$	?	?	?	?
$(0,0,0,1)$	?	?	?	?

Note: The first row corresponds to the observed assignment D^{obs}. Entries marked by ? are unobserved potential outcomes.

2.1. *Partially sharp null hypotheses*

To test localized spillover restrictions without imposing a global sharp null, I use the following class of partially sharp null hypotheses, adopting the terminology of Zhang and Zhao (2023).

DEFINITION 1—Partially sharp null hypothesis: A null hypothesis is partially sharp if there exists a collection of subsets $\{\mathcal{D}_i\}_{i=1}^N$, with each $\mathcal{D}_i \subsetneq \{0,1\}^N$, such that

$$H_0: \quad Y_i(d) = Y_i(d') \quad \text{for all } i \in \{1,\dots,N\} \text{ and all } d, d' \in \mathcal{D}_i.$$

A partially sharp null restricts potential outcomes only within selected proper subsets of the assignment space. The sharp null is outside this class as defined here; it is recovered by the degenerate extension $\mathcal{D}_i = \{0,1\}^N$ for all i. The sets $\mathcal{D}_i$ may vary across units. One example is a null defined through an exposure mapping, where outcomes are assumed constant within exposure states (Hoshino and Yanagi, 2026). More generally, a null may be expressed as an intersection of several such restrictions (Puelz et al., 2022); Appendix A shows how PIRT extends to that case. Here, I focus on the important special case in which $\mathcal{D}_i$ is defined through distance.

DEFINITION 2—Distance interval assignment set: For unit $i \in \{1,\dots,N\}$ and threshold ϵ, define

$$\mathcal{D}_i(\epsilon) := \left\{ d \in \{0,1\}^N : \sum_{j=1}^N 1\{G_{ij} \le \epsilon\} d_j = 0 \right\}.$$

When $d \in \mathcal{D}_i(\epsilon)$, unit i is said to lie in the distance interval (ϵ, ∞).

Thus, for $\epsilon \ge 0$, $\mathcal{D}_i(\epsilon)$ contains the assignments under which no treated unit lies within distance ϵ of unit i. Since $G_{ii} = 0$, any assignment in $\mathcal{D}_i(\epsilon)$ must satisfy $d_i = 0$. When $\epsilon = 0$, this is the only restriction, so $\mathcal{D}_i(0)$ consists of all assignments that leave unit i untreated.[4] More generally, for $0 \le a < b$, $\mathcal{D}_i(a) \setminus \mathcal{D}_i(b)$ identifies assignments under which unit i's nearest treated unit lies in $(a, b]$.

[4]For $\epsilon < 0$, since $G_{ij} \ge 0$ for all i, j, we have $\mathcal{D}_i(\epsilon) = \{0,1\}^N$.

Running Example continued. To make Definition 2 concrete, return to the four-unit running example in Figure 1, where exactly one unit is treated. In that case, the sets $\mathcal{D}_i(\epsilon)$ can be listed explicitly for different values of ϵ.

TABLE II

DISTANCE INTERVAL ASSIGNMENT SETS IN THE RUNNING EXAMPLE.

Unit i	$\mathcal{D}_i(0)$	$\mathcal{D}_i(1)$	$\mathcal{D}_i(2)$
1	$\{d^{(2)}, d^{(3)}, d^{(4)}\}$	$\{d^{(3)}, d^{(4)}\}$	$\varnothing$
2	$\{d^{(1)}, d^{(3)}, d^{(4)}\}$	$\{d^{(3)}, d^{(4)}\}$	$\varnothing$
3	$\{d^{(1)}, d^{(2)}, d^{(4)}\}$	$\{d^{(1)}, d^{(2)}\}$	$\varnothing$
4	$\{d^{(1)}, d^{(2)}, d^{(3)}\}$	$\{d^{(1)}, d^{(2)}\}$	$\varnothing$

Note: Here $d^{(1)} = (1,0,0,0)$, $d^{(2)} = (0,1,0,0)$, $d^{(3)} = (0,0,1,0)$, and $d^{(4)} = (0,0,0,1)$.

Table II shows how the threshold ϵ changes the admissible assignment set. For example, $\mathcal{D}_1(0)$ consists of all assignments in which unit 1 is untreated, while $\mathcal{D}_1(1)$ further excludes the assignment in which its adjacent neighbor 2 is treated. Since every other unit is within distance 2 of unit 1 in the running example, $\mathcal{D}_1(2) = \varnothing$ under the one-treated design.

Using these assignment sets, I now define the null hypothesis of no interference beyond distance ϵ.

DEFINITION 3—Partially sharp null of no interference beyond distance $\epsilon \geq 0$: The partially sharp null hypothesis of no interference beyond distance $\epsilon \geq 0$ is

$$H_0^\epsilon : \quad Y_i(d) = Y_i(d') \quad \text{for all } i \in \{1, \ldots, N\} \text{ and all } d, d' \in \mathcal{D}_i(\epsilon).$$

Under H_0^ϵ, the outcome for unit i is unchanged across any two assignments that place all treated units farther than ϵ from i. Equivalently, H_0^ϵ leaves direct effects and spillovers within distance ϵ unrestricted, but rules out effects of more distant treatments on units outside the treated ϵ-neighborhood. This null is useful for assessing whether, and how far, spillovers extend in a network or spatial setting. When $\epsilon > 0$, it can be used to characterize the neighborhood within which spillovers operate, or to identify comparison groups for subsequent estimation. Broader range-of-interference nulls that also compare treated units are covered by the intersection-null framework in Appendix A.

Although the definition is written in terms of distance, it also nests clustered interference by assigning a small within-cluster distance and a larger between-cluster distance, as in

the running example. Under this null, some, but not all, missing potential outcomes can be uniquely determined from the observed outcomes and treatment assignment; we call such outcomes *imputable*.

Running Example Continued. Suppose researchers wish to test for spillover effects using Definition 3 with $\epsilon = 0$:

$$H_0^0: \quad Y_i(d) = Y_i(d') \text{ for all } i \in \{1, \ldots, N\} \text{ and all } d, d' \in \{0,1\}^N \text{ such that } d_i = d'_i = 0.$$

Throughout the paper, I use H_0^0 to illustrate the main ideas in the running example. Under this null, the potential outcome for unit i remains unchanged across assignments as long as unit i is untreated in both assignments. In the realized dataset, the observed assignment is $D^{\text{obs}} = (1,0,0,0)$. Therefore, when determining whether an entry in the potential outcome schedule is imputable, the relevant comparison is always between the counterfactual assignment and the fixed observed assignment D^{obs}.

Table III displays the resulting potential outcome schedule under H_0^0. The entries marked by question marks are those not pinned down by the null together with the realized dataset. More specifically, the outcome for unit i in row d is imputable if and only if unit i is untreated both under D^{obs} and under d. For example, in the row $d = (0,0,1,0)$, units 2 and 4 are untreated under both D^{obs} and d, so their outcomes are imputable, whereas units 1 and 3 are not. Similarly, in the row $d = (0,0,0,1)$, units 2 and 3 are imputable, whereas units 1 and 4 are not.

TABLE III

POTENTIAL OUTCOME SCHEDULE UNDER PARTIALLY SHARP NULL H_0^0.

Assignment D	Potential Outcome Y_i			
	1	2	3	4
$(1,0,0,0)$	2	4	3	2
$(0,1,0,0)$	?	?	3	2
$(0,0,1,0)$	?	4	?	2
$(0,0,0,1)$	?	4	3	?

Note: The table is constructed relative to the observed assignment $D^{\text{obs}} = (1,0,0,0)$. Potential outcomes marked by ? are not imputable under H_0^0 from the realized dataset.

The table highlights two difficulties for randomization inference under interference. First, only a subset of potential outcomes is imputable under the null. For example, in the row $(0, 1, 0, 0)$, the null provides no information about the outcomes of units 1 and 2, because each unit is treated in one of the two assignments being compared. Second, the set of imputable units is assignment-specific. In the row $(0, 0, 1, 0)$, the non-imputable units are 1 and 3, whereas in the row $(0, 0, 0, 1)$, they are 1 and 4. Thus, a standard FRT cannot simply reuse one fixed set of observed outcomes over all reassignments: which outcomes are imputable under the null changes with the reassignment. The remainder of the paper develops a framework that addresses these two challenges in turn.

2.2. *A sketch of the pairwise comparison idea*

Before turning to the formal construction, it is useful to give a brief implementation sketch. Like a standard randomization test, PIRT redraws treatment assignments from the design. The difference is that a partially sharp null does not make the full potential outcome schedule imputable. Hence each reassignment must be evaluated on the units jointly imputable under both the observed and reassigned treatments, and each reassignment generates a pairwise comparison rather than a single randomized statistic.

For a reassignment D, PIRT first forms the common imputable subsample for the pair (D^{obs}, D). Under H_0^0, this consists of units untreated under both assignments; under H_0^ϵ, it consists of units outside the treated ϵ-neighborhood under both assignments. On this common subsample, PIRT computes the same contrast twice. In one calculation, the units are grouped according to the reassignment D; in the other, the same units and outcomes are grouped according to the observed assignment D^{obs}. Thus, holding the outcomes and common subsample fixed, the two calculations differ only in the assignment used to form the groups.[5]

For a one-sided alternative in which larger values indicate stronger spillovers, the PIRT p-value is the frequency, over reassignments, with which the contrast formed using D is at least as large as the corresponding contrast formed using D^{obs}. For a two-sided alternative, the same construction can be applied to absolute values. The next section formalizes these objects and establishes finite-sample validity.

[5] For a concrete implementation in the Bogotá application, see Zhong (2024, Section 4.2).

3. PAIRWISE IMPUTATION-BASED RANDOMIZATION TEST (PIRT)

Throughout this section, ϵ and, when used, a control threshold $\epsilon_c > \epsilon$ are treated as fixed before testing. The threshold ϵ indexes the partially sharp null H_0^ϵ, while ϵ_c is part of the test statistic, partitioning jointly imputable units into the comparison interval $(\epsilon, \epsilon_c]$ and the control region (ϵ_c, ∞). Once fixed, these choices affect power but not finite-sample validity, even if ϵ_c does not define a perfectly clean control group (Basse et al., 2024). Appendix B of the Supplemental Material discusses sequential testing over multiple values of ϵ and design-stage guidance for choosing ϵ_c.

DEFINITION 4—Imputable units: Given a treatment assignment $d \in \{0,1\}^N$ and a partially sharp null hypothesis H_0^ϵ,

$$\mathbb{I}(d) \equiv \{i \in \{1,\dots,N\} : d \in \mathcal{D}_i(\epsilon)\}$$

is called the set of *imputable units* under treatment assignment d and null hypothesis H_0^ϵ.

The set $\mathbb{I}(d)$ is the largest assignment-specific candidate set of units whose outcomes are potentially imputable under the partially sharp null H_0^ϵ. Equivalently, it contains the units that lie in the distance interval (ϵ, ∞) under assignment d. This notion is related in spirit to the "super focal units" in Owusu (2023), but here the candidate set is defined for every assignment rather than fixed once and for all. For example, when $\epsilon = 0$, $\mathcal{D}_i(\epsilon) = \{d : d_i = 0\}$, so $\mathbb{I}(d) = \{i : d_i = 0\}$. Hence $\mathbb{I}(d) \neq \mathbb{I}(d')$ in general when different units are treated. Under the sharp null, by contrast, $\mathbb{I}(d) = \{1,\dots,N\}$ for every assignment d, so all units are imputable.[6]

3.1. *Pairwise Imputable Statistics*

To address the presence of missing potential outcomes, I restrict attention to test statistics that depend only on outcomes jointly imputable under the assignment pair.

DEFINITION 5—Pairwise Imputable Statistic: Let

$$T : \mathbb{R}^N \times \{0,1\}^N \times \{0,1\}^N \to \mathbb{R} \cup \{\infty\}$$

[6] In some designs, $\mathbb{I}(D^{\text{obs}})$ may be empty. Size can still be maintained by rejecting with probability α, although power is necessarily limited. Appendix A discusses related sparse-overlap cases.

be a measurable function. We say that T is *pairwise imputable* if, for any $d, d' \in \{0,1\}^N$ and any $Y, Y' \in \mathbb{R}^N$, the following holds:

$$\text{if } Y_i = Y'_i \text{ for all } i \in \mathbb{I}(d) \cap \mathbb{I}(d'), \text{ then } T(Y, d, d') = T(Y', d, d').$$

That is, T depends only on the entries of Y indexed by the intersection $\mathbb{I}(d) \cap \mathbb{I}(d')$.

The set $\mathbb{I}(d) \cap \mathbb{I}(d')$ parallels the set H in Definition 1 of Zhang and Zhao (2023), capturing all units whose potential outcomes are jointly imputable under both assignments. Although it may appear restrictive, common statistics can be made pairwise imputable with minor modifications. For example, consider a difference in means comparing units in $(\epsilon, \epsilon_c]$ with units in (ϵ_c, ∞). In $T(Y, d, d')$, the second argument d determines the comparison and control groups, while the pair (d, d') determines which outcomes are jointly imputable. For the observed data and a reassignment D, this statistic can be written as

$$T(Y(D^{\text{obs}}), D, D^{\text{obs}}) = \underbrace{\bar{Y}_{\mathbb{I}(D^{\text{obs}})}(D^{\text{obs}})_{\{i:D\in\mathcal{D}_i(\epsilon)\setminus\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of imputable neighbors}} - \underbrace{\bar{Y}_{\mathbb{I}(D^{\text{obs}})}(D^{\text{obs}})_{\{i:D\in\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of imputable controls}},$$

where, for any collection of sets $A_i \subseteq \mathcal{D}_i(\epsilon)$,

$$\bar{Y}_{\mathbb{I}(D^{\text{obs}})}(D^{\text{obs}})_{\{i:D\in A_i\}} = \sum_{i\in\mathbb{I}(D^{\text{obs}})} 1\{D \in A_i\} Y_i(D^{\text{obs}}) \Big/ \sum_{i\in\mathbb{I}(D^{\text{obs}})} 1\{D \in A_i\}.$$

Because $A_i \subseteq \mathcal{D}_i(\epsilon)$, any unit receiving positive weight also belongs to $\mathbb{I}(D)$. Hence every outcome used by the statistic is indexed by a unit in $\mathbb{I}(D^{\text{obs}}) \cap \mathbb{I}(D)$, and changing outcomes outside this intersection cannot change the statistic. The difference in means is therefore pairwise imputable. Under the sharp null, $\mathbb{I}(d) \cap \mathbb{I}(d') = \{1, \ldots, N\}$ for all d, d', so ordinary difference-in-means statistics are recovered as special cases (Imbens and Rubin, 2015).

If, for a given assignment D, no jointly imputable unit falls into one of the required comparison cells, the corresponding sample mean is undefined. Formally, I set the statistic equal to $+\infty$ whenever a required cell is empty. The empty-cell event depends only on the assignment pair and the specified comparison intervals, not on outcomes, so the statistic remains pairwise imputable. In numerical implementations, $+\infty$ can be represented by $\max(Y^{\text{obs}}) - \min(Y^{\text{obs}})$, which weakly exceeds every finite one-sided or absolute difference in means; for the upper-tail comparisons below, this replacement can only weakly increase the computed p-value relative to the formal $+\infty$ convention and is therefore conservative. In the empirical application, this fallback rule was not invoked.

Running Example Continued. Consider the test statistic

$$T(Y(D^{\text{obs}}), D, D^{\text{obs}}) = \bar{Y}_{\mathbb{I}(D^{\text{obs}})}(D^{\text{obs}})_{\{i:D\in\mathcal{D}_i(0)\setminus\mathcal{D}_i(1)\}} - \bar{Y}_{\mathbb{I}(D^{\text{obs}})}(D^{\text{obs}})_{\{i:D\in\mathcal{D}_i(1)\}}.$$

Under the observed assignment $D^{\text{obs}} = (1,0,0,0)$, the imputable set is fixed at $\mathbb{I}(D^{\text{obs}}) = \{2,3,4\}$. What changes with the counterfactual assignment D is which of these imputable units fall into the interval $(0,1]$ and which fall into $(1,\infty)$.

TABLE IV

CONSTRUCTING A PAIRWISE IMPUTABLE STATISTIC

Assignment D	Potential Outcome Y_i				$\bar{Y}_{\mathbb{I}(D^{\text{obs}})}(D^{\text{obs}})$		$T(Y(D^{\text{obs}}), D, D^{\text{obs}})$
	1	2	3	4	$\{i : D \in \mathcal{D}_i(0) \setminus \mathcal{D}_i(1)\}$	$\{i : D \in \mathcal{D}_i(1)\}$	
$(1,0,0,0)$	2	4	3	2	4	2.5	1.5
$(0,1,0,0)$	?	?	3	2	?	2.5	2
$(0,0,1,0)$	?	4	?	2	2	4	-2
$(0,0,0,1)$	?	4	3	?	3	4	-1

Note: The table lists all assignments in the design support, with the first row corresponding to the observed assignment D^{obs}. Entries with ? are not imputable under H_0^0 from the realized dataset. Blue cells denote jointly imputable units in the interval $(0,1]$ under the comparison assignment D, and brown cells denote jointly imputable units in $(1,\infty)$. The red test statistic indicates that a required comparison cell is empty. When this occurs, the formal statistic equals $+\infty$; following the finite implementation convention described in the text, the table reports the placeholder $2 = \max(Y^{\text{obs}}) - \min(Y^{\text{obs}})$.

Table IV records the resulting groupings and statistic values for each assignment. Under the observed assignment $D = D^{\text{obs}}$, unit 2 belongs to $(0,1]$, while units 3 and 4 belong to $(1,\infty)$. Under the counterfactual assignment that treats unit 3 instead of unit 1, however, $\mathbb{I}(D^{\text{obs}}) = \{2,3,4\}$ and $\mathbb{I}(D) = \{1,2,4\}$, so the jointly imputable set is only $\{2,4\}$. As shown in the third row of the table, the statistic is therefore computed using only the outcomes of units 2 and 4, but these units are now assigned to different comparison groups: unit 4 enters $(0,1]$, while unit 2 enters $(1,\infty)$. The table also illustrates the conservative fallback used when a comparison cell is empty.

For further discussion of test-statistic choice in randomization tests and network settings, see Imbens and Rubin (2015, Section 5), Athey et al. (2018), and Hoshino and Yanagi (2026). All test statistics can also be adapted to their absolute-value forms for two-sided testing. While the method remains valid without covariate adjustments, incorporating them

may improve power in practice. See Appendix C of the Supplemental Material for a discussion of covariates and rank-based statistics.

Following Definition 5, I can derive the key property that allows the statistic to be evaluated from observed information alone:

PROPOSITION 1: *Suppose the partially sharp null hypothesis H_0^ϵ is true. Suppose $T(Y(d), d, d')$ is a pairwise imputable statistic. Then,*

$$T(Y(d), d, d') = T(Y(d'), d, d')$$

for any $d, d' \in \{0,1\}^N$.

The proof is provided in Appendix D. By Proposition 1, letting $d = D$ and $d' = D^{\text{obs}}$, we have $T(Y(D), D, D^{\text{obs}}) = T(Y(D^{\text{obs}}), D, D^{\text{obs}})$ under the null H_0^ϵ, which makes the counterfactual statistic computable from observed outcomes.

I focus on the unconditional randomization test framework, which is defined as follows:

DEFINITION 6—Unconditional randomization test: An unconditional randomization test $\phi : \{0,1\}^N \to [0,1]$ is defined such that for any $D^{\text{obs}} \in \{0,1\}^N$,

$$\phi(D^{\text{obs}}) = Q(\tilde{p}(D^{\text{obs}}), \alpha),$$

where $Q : [0,1] \times [0,1] \to [0,1]$ is a measurable function, α is the nominal significance level, and $\tilde{p}(D^{\text{obs}})$ is given by

$$\tilde{p}(D^{\text{obs}}) = \sum_{d \in \{0,1\}^N} g(D^{\text{obs}}, d)\, P(D = d),$$

with P denoting the pre-specified probability distribution over treatment assignments, and $g : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}$ a measurable function.

The key feature of the unconditional randomization test is that the rejection probability is computed by randomizing the treatment assignment according to the same probability distribution P that governs the original assignment. This stands in contrast to methods such as Athey et al. (2018), where the rejection rule is constructed within a conditional probability space. One natural first attempt is the following unconditional analogue of the Fisher test.

DEFINITION 7—Simple randomization test: A *simple randomization test* is an unconditional randomization test defined by $\phi(D^{\text{obs}}) = 1\{pval(D^{\text{obs}}) \leq \alpha\}$, where $pval(D^{\text{obs}}) : \{0,1\}^N \to [0,1]$ is the *p*-value function given by

$$pval(D^{\text{obs}}) = P(T(Y(D^{\text{obs}}), D, D^{\text{obs}}) \geq T(Y(D^{\text{obs}}), D^{\text{obs}}, D^{\text{obs}})) \text{ for } D \sim P,$$

and $T(Y, d, d')$ denotes a pairwise imputable statistic.

However, this simple test is generally invalid because the effective sample $\mathbb{I}(D) \cap \mathbb{I}(D^{\text{obs}})$ varies with the assignment D and is often largest when $D = D^{\text{obs}}$. For sum-type statistics, this imbalance can make the observed statistic mechanically large relative to its counterfactual values, yielding an artificially small p-value. See Zhong (2024) for a fuller discussion of this problem for general test statistics, as well as an illustration using the running example.

The construction of PIRT resolves this problem by replacing the naive comparison above with a comparison between two pairwise imputable statistics. I now turn to that construction.

3.2. *The Pairwise Comparison-Based p-values*

Formally, I define PIRT as follows:

DEFINITION 8—PIRT: The PIRT is an unconditional randomization test defined by

$$\phi^{pair}(D^{\text{obs}}) = 1\left\{pval^{pair}(D^{\text{obs}}) \leq \alpha/2\right\},$$

where $pval^{pair} : \{0,1\}^N \to [0,1]$ is the *p*-value function given by

$$pval^{pair}(D^{\text{obs}}) = P\left(T(Y(D^{\text{obs}}), D, D^{\text{obs}}) \geq T(Y(D^{\text{obs}}), D^{\text{obs}}, D)\right), \quad \text{for } D \sim P,$$

and $T(Y, d, d')$ denotes a pairwise imputable statistic.

The key intuition is the pairwise symmetry induced by Proposition 1. For a realized assignment D^{obs} and a reassignment D, the indicator contributing to $pval^{pair}(D^{\text{obs}})$ compares $T(Y(D^{\text{obs}}), D, D^{\text{obs}})$ with $T(Y(D^{\text{obs}}), D^{\text{obs}}, D)$. Finite-sample size control concerns the distribution of PIRT p-values over all possible realized assignments. If D were realized and D^{obs} served as the reassignment, the corresponding indicator contributing to $pval^{pair}(D)$ would compare $T(Y(D), D^{\text{obs}}, D)$ with $T(Y(D), D, D^{\text{obs}})$. Under the null, Proposition 1 shows that the two indicators compare the same two numerical quantities in

reverse order. Hence, at least one of the two indicators equals one for every assignment pair. This pairwise constraint limits the design probability of assignments with very small PIRT p-values. The theorem below formalizes this argument.

THEOREM 1: *Suppose the partially sharp null hypothesis H_0^ϵ holds. Then, the PIRT, as defined in Definition 8, satisfies $\mathbb{E}_P[\phi^{pair}(D^{\text{obs}})] < \alpha$ for any $\alpha \in (0,1)$, where the expectation is taken with respect to $D^{\text{obs}} \sim P$.*

See Appendix D for the proof, which is inspired by the proof in Guan (2024). As in the conformal prediction literature, Theorem 1 is a worst-case guarantee (Vovk, 2015, Barber et al., 2021). The $\alpha/2$ threshold is the price of this unconditional finite-sample guarantee. Section 4 characterizes the corresponding power cost relative to conditioning-based alternatives. In the simulations below, using the heuristic cutoff α often yields empirical size close to the nominal level, although pathological examples prevent replacing the $\alpha/2$ threshold by α in general. I therefore also report the α cutoff as a sensitivity benchmark, to illustrate how conservative the $\alpha/2$ bound can be.

In large assignment spaces, the exact value of $pval^{pair}(D^{\text{obs}})$ is generally infeasible to compute. Algorithm 1 instead reports a plus-one Monte Carlo PIRT p-value.[7] The first comparison is fixed at $D = D^{\text{obs}}$, and the remaining R assignments are drawn independently of D^{obs} from the design P. This plus-one construction is used as a finite-R randomization test rather than merely as an estimator of the exact p-value: Appendix D shows that, for every finite R, rejecting when $\widehat{pval}_R^{pair}(D^{\text{obs}}) \leq \alpha/2$ controls size at level α, provided the auxiliary assignments are independent draws from P.

Running Example Continued. Using the same difference-in-means statistic as before,

$$T(Y(D^{\text{obs}}), D^{\text{obs}}, D) = \bar{Y}_{\mathbb{I}(D)}(D^{\text{obs}})_{\{i:D^{\text{obs}}\in\mathcal{D}_i(0)\setminus\mathcal{D}_i(1)\}} - \bar{Y}_{\mathbb{I}(D)}(D^{\text{obs}})_{\{i:D^{\text{obs}}\in\mathcal{D}_i(1)\}}.$$

Under the observed assignment $D^{\text{obs}} = (1,0,0,0)$, the comparison groups induced by D^{obs} are fixed: unit 2 lies in the interval $(0,1]$, while units 3 and 4 lie in $(1,\infty)$. For each reassignment D, the statistic $T(Y(D^{\text{obs}}), D^{\text{obs}}, D)$ retains these D^{obs}-defined groups but excludes any unit that is not jointly imputable under D. Table V reports this decom-

[7] I thank Lihua Lei for suggesting the plus-one construction.

Algorithm 1 Plus-one Monte Carlo PIRT p-value

Input : Test statistic T; observed assignment D^{obs}; observed outcome Y^{obs}; treatment assignment mechanism P.

for $r = 1$ **to** R **do**

  Randomly sample $d^r \sim P$ and compute $T_r \equiv T(Y(D^{\text{obs}}), d^r, D^{\text{obs}})$.

  Compute $T_r^{\text{obs}} \equiv T(Y(D^{\text{obs}}), D^{\text{obs}}, d^r)$.

end

Output: p-value: $\widehat{pval}_R^{pair}(D^{\text{obs}}) = \frac{1+\sum_{r=1}^{R} 1\{T_r \geq T_r^{\text{obs}}\}}{1+R}$.

position together with the two statistics that enter the PIRT comparison. The column for $T(Y(D^{\text{obs}}), D, D^{\text{obs}})$ matches the corresponding values already obtained in Table IV.

TABLE V

PIRT IN THE RUNNING EXAMPLE

Assignment D	Potential Outcome Y_i				Means		$T_r(D)$	$T_o(D)$
	1	2	3	4	$\mu_N(D)$	$\mu_C(D)$		
$(1,0,0,0)$	2	4	3	2	4	2.5	1.5	1.5
$(0,1,0,0)$	?	?	3	2	?	2.5	2	2
$(0,0,1,0)$	?	4	?	2	4	2	-2	2
$(0,0,0,1)$	?	4	3	?	4	3	-1	1

Note: $T_r(D) = T(Y(D^{\text{obs}}), D, D^{\text{obs}})$ and $T_o(D) = T(Y(D^{\text{obs}}), D^{\text{obs}}, D)$. Also, $\mu_N(D) = \bar{Y}_{\mathbb{I}(D)}(D^{\text{obs}})_{\{i:D^{\text{obs}} \in \mathcal{D}_i(0) \setminus \mathcal{D}_i(1)\}}$ and $\mu_C(D) = \bar{Y}_{\mathbb{I}(D)}(D^{\text{obs}})_{\{i:D^{\text{obs}} \in \mathcal{D}_i(1)\}}$. The first row corresponds to the observed assignment D^{obs}. Entries marked ? are not imputable under H_0^0. Blue cells denote units in the interval $(0, 1]$ under the observed assignment D^{obs}, and brown cells denote units in $(1, \infty)$ under D^{obs}. The red test statistic indicates that a required comparison cell is empty. When this occurs, the formal statistic equals $+\infty$; following the finite implementation convention described in the text, the table reports the placeholder $2 = \max(Y^{\text{obs}}) - \min(Y^{\text{obs}})$.

Only when D treats unit 1 or unit 2 does $T(Y(D^{\text{obs}}), D, D^{\text{obs}}) \geq T(Y(D^{\text{obs}}), D^{\text{obs}}, D)$. Hence $pval^{pair}(D^{\text{obs}}) = 2/4$. As in Guan (2024), ties can also be handled by using a factor of $1/2$ without affecting validity. For a directional alternative, orient the statistic so that larger values favor the alternative. For a two-sided alternative, apply the same construction to the absolute value of the statistic.

4. COMPARISON TO CONDITIONAL RANDOMIZATION TESTS UNDER CLUSTERED INTERFERENCE

This section compares PIRT with conditional randomization tests (CRTs) for testing the null of no spillover effect on untreated units under the two-stage clustered design. Throughout, I use the same one-sided difference-in-means statistic, comparing untreated units in spillover exposure with untreated units in control exposure. Rather than seeking a uniform ranking, the goal is to characterize when each procedure is more favorable and which features of the design drive the ranking.

Three forces are central. First, *effective overlap* controls the precision of the observed spillover-control contrast: larger and better-balanced jointly imputable sets provide more information for estimation. Second, *leave-one-out variation* affects the spread of the randomization reference statistic when reassignment changes which untreated units enter the pairwise comparison. Third, the cutoff matters: the conditional-focals CRT uses cutoff α, whereas the finite-sample-valid PIRT rule uses cutoff $\alpha/2$. I first isolate these forces by comparing PIRT with the conditional-focals CRT of Basse et al. (2019) in a stylized two-stage benchmark, and then use the clustered-interference simulation design of Puelz et al. (2022) to examine the same forces in finite samples.

4.1. *A two-stage benchmark: effective overlap and leave-one-out variation*

There are $K = 2n$ clusters, indexed by c, and $m \geq 2$ units per cluster, indexed by u. The design assigns exactly n clusters to treatment and then assigns exactly one unit to treatment in each treated cluster. Let $W_c(D) \in \{0, 1\}$ denote the cluster-treatment indicator under assignment D, and let $J_c(D)$ denote the treated unit in cluster c when $W_c(D) = 1$. Thus, $D_{c,u} = 1$ if and only if $W_c(D) = 1$ and $u = J_c(D)$.

Under the null hypothesis of no spillovers on untreated units, unit (c, u) is jointly imputable for the pair (D, D^{obs}) if and only if it remains untreated under both assignments: $D_{c,u} = D^{obs}_{c,u} = 0$. Under an assignment d, I refer to untreated units in treated clusters as *spillover-exposure* units and units in control clusters as *control-exposure* units. The conditional-focals CRT of Basse et al. (2019) fixes one focal unit per cluster: one untreated unit in each observed treated cluster and one unit in each observed control cluster. It then randomizes conditionally on this focal set. In contrast, PIRT fixes no focal set. For each comparison assignment D, it utilizes the jointly imputable units in the pair (D, D^{obs}); the

PIRT statistics compare the exposure groups induced by D and by D^{obs} on this common set.

To define the power calculation, suppose outcomes satisfy the potential outcomes model

$$Y_{c,u}(D) = \mu_c + \varepsilon_{c,u} + \delta D_{c,u} + \tau\, \mathbf{1}\{W_c(D) = 1,\ D_{c,u} = 0\}, \tag{1}$$

where $\mu_c \overset{iid}{\sim} N(0, \sigma_\mu^2)$ and $\varepsilon_{c,u} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$, independently across clusters and units. The parameter τ represents the spillover effect for untreated units in treated clusters, while δ captures an arbitrary direct-treatment effect. Because the spillover-control comparison relies exclusively on untreated units, the term $\delta D_{c,u}$ drops out of all test statistics below.

Let P_n denote the law induced by the random draws of $\{\mu_c, \varepsilon_{c,u}\}$, the observed assignment, and, for the conditional-focals CRT, the auxiliary focal draw. For a method $M \in \{PIRT, CRT\}$ with an exact one-sided p-value $p_{M,m,n}(D^{obs}, Y^{obs})$ and significance cutoff a, we define the power as

$$\beta_{M,m,n}(\tau; a) := P_n\Big(p_{M,m,n}(D^{obs}, Y^{obs}) \le a\Big). \tag{2}$$

Thus, $\beta_{M,m,n}$ represents an average rejection probability over random finite populations and assignments, rather than the conditional power for a single realized finite population.

An important determinant of power is the variance of the observed contrast. Let $T^{obs}_{M,m}$ denote the observed contrast used by method M. For the conditional-focals CRT, $T^{obs}_{\mathrm{CRT},m}$ is the difference in means over the fixed focal units. For PIRT, the observed contrast is the realized value $T(Y(D^{obs}), D^{obs}, D^{obs})$ of the pairwise imputable statistic:

$$T^{obs}_{\mathrm{PIRT},m} = \frac{1}{n(m-1)} \sum_{c:W_c(D^{obs})=1} \sum_{u:D^{obs}_{c,u}=0} Y^{obs}_{c,u} - \frac{1}{nm} \sum_{c:W_c(D^{obs})=0} \sum_{u=1}^{m} Y^{obs}_{c,u}.$$

I define $V^{obs}_{M,m} := n \operatorname{Var}_{P_n}(T^{obs}_{M,m})$. A smaller $V^{obs}_{M,m}$ implies that a given local spillover signal is easier to detect.

The first design diagnostic is *effective overlap*. For a difference in means with N_1 jointly imputable untreated units in spillover exposure and N_0 jointly imputable untreated units in control exposure, define

$$\Omega^{\mathrm{eff}} := \left(\frac{1}{N_1} + \frac{1}{N_0}\right)^{-1}. \tag{3}$$

This is the harmonic effective size of the spillover-control comparison: it is large only when both comparison cells contain many jointly imputable units, and it is small when either cell

is thin. Appendix D shows that, in this benchmark,

$$V_{M,m}^{obs} = 2\sigma_\mu^2 + \frac{n}{\Omega_{M,m}^{\text{eff}}}\sigma_\varepsilon^2. \tag{4}$$

Thus $\Omega_{M,m}^{\text{eff}}$ governs the unit-level-error contribution to the observed-statistic variance, while $2\sigma_\mu^2$ is the common cluster-level contribution.

For the conditional-focals CRT, $N_1 = N_0 = n$. For PIRT, $N_1 = n(m-1)$ and $N_0 = nm$. Hence,

$$\Omega_{\text{CRT},m}^{\text{eff}} = \frac{n}{2}, \qquad \Omega_{\text{PIRT},m}^{\text{eff}} = \frac{nm(m-1)}{2m-1}, \tag{5}$$

which yields the respective variances

$$V_{\text{CRT},m}^{obs} = 2\sigma_\mu^2 + 2\sigma_\varepsilon^2, \qquad V_{\text{PIRT},m}^{obs} = 2\sigma_\mu^2 + \frac{2m-1}{m(m-1)}\sigma_\varepsilon^2.$$

This highlights the overlap gain achieved by PIRT: relative to the conditional-focals CRT, PIRT averages over a larger number of jointly imputable untreated units in the observed contrast, thereby reducing the observed-statistic variance whenever $\sigma_\varepsilon^2 > 0$.

Precision of the observed statistic is not the sole determinant of power. The second diagnostic is *leave-one-out variation*, which dictates the spread of the randomization reference distribution rather than the observed contrast. Consider an observed control cluster that becomes treated under a comparison assignment D. Under D^{obs}, all m units in this cluster contribute to the contrast through the cluster mean

$$\bar{\varepsilon}_c := \frac{1}{m}\sum_{u=1}^{m} \varepsilon_{c,u}.$$

Under D, however, the directly treated unit $J_c(D)$ is excluded from the spillover average, meaning PIRT utilizes

$$\tilde{\varepsilon}_c(D) := \frac{1}{m-1}\sum_{u \neq J_c(D)} \varepsilon_{c,u}.$$

It follows that

$$\tilde{\varepsilon}_c(D) - \bar{\varepsilon}_c = -\frac{\varepsilon_{c,J_c(D)} - \bar{\varepsilon}_c}{m-1}, \qquad \text{Var}\{\tilde{\varepsilon}_c(D) - \bar{\varepsilon}_c\} = \frac{\sigma_\varepsilon^2}{m(m-1)}.$$

This motivates the leave-one-out variation index

$$\mathcal{L}_{\text{PIRT},m} := \frac{1}{m(m-1)}, \qquad \mathcal{L}_{\text{CRT},m} := 0. \tag{6}$$

Fix $h > 0$ and consider local alternatives $\tau_n = h/\sqrt{n}$, writing $\beta_{M,m,n}(h;\cdot)$ for $\beta_{M,m,n}(\tau_n;\cdot)$. Appendix D derives the limiting power curves. Since these curves are monotone in h and cross once, one may equivalently compare methods by fixing h and comparing power, or by fixing a target power π and comparing the local effect required to attain it. I use the latter representation because π is unit-free.

PROPOSITION 2—Stylized local-power comparison under equal cluster size: *Suppose the assignment mechanism is the balanced two-stage clustered design described above and outcomes satisfy* (1)*. Then, for* $M \in \{\mathrm{CRT}, \mathrm{PIRT}\}$ *and any cutoff* $a \in (0,1)$*,*

$$\beta_{M,m,n}(h;a) = \Phi\left(\frac{h - z_{1-a}\sqrt{V^{obs}_{M,m} + 2\mathcal{L}_{M,m}\sigma^2_\varepsilon}}{\sqrt{V^{obs}_{M,m}}}\right) + o(1).$$

Furthermore, setting $a = \alpha$ *for the CRT and* $a = \alpha/2$ *for PIRT, assume* $\sigma^2_\varepsilon > 0$*. There exists a unique crossover power* $\pi^\times_m = \pi^\times_m(\alpha, V^{obs}_{\mathrm{CRT},m}, V^{obs}_{\mathrm{PIRT},m}, \mathcal{L}_{\mathrm{PIRT},m}, \sigma^2_\varepsilon) \in (0,1)$*, equal to the common limiting power at the unique crossing of the two power curves. For any target limiting power* $\pi \in (0,1)$*, PIRT requires a smaller local effect than the CRT if and only if* $\pi > \pi^\times_m$*. If* $\sigma^2_\varepsilon = 0$*, the observed-variance gain vanishes and the level-*α *CRT dominates PIRT at the* $\alpha/2$ *cutoff.*

The explicit algebraic formula for the *crossover power* is provided in Appendix D. The framework cleanly illuminates the central design tradeoff: PIRT steepens its power curve by reducing $V^{obs}_{\mathrm{PIRT},m}$ through superior effective overlap, but pays a penalty via a stricter critical value $z_{1-\alpha/2}$ and the leave-one-out reference variance $\mathcal{L}_{\mathrm{PIRT},m}\sigma^2_\varepsilon$. This variance is fundamentally the price of not conditioning on a fixed focal set. Whereas the conditional-focals CRT freezes the identities of the specific units entering the contrast, PIRT's reliance on the full two-stage assignment law means that the identity of the unit excluded from the spillover average varies across reassignments. This variation injects within-cluster leave-one-out noise into the randomization reference distribution.

The crossover power summarizes these competing forces as a target-power threshold: below it, the CRT requires a smaller local effect; above it, PIRT requires a smaller local effect. Because the leave-one-out penalty is largest at $m = 2$ and declines rapidly with cluster size, Proposition 2 implies that PIRT's overlap advantage becomes more valuable as m increases. For example, when $\alpha = 0.05$ and $\sigma^2_\mu = \sigma^2_\varepsilon$, the crossover power is essentially

1.00 for $m = 2$, about 0.73 for $m = 3$, 0.46 for $m = 4$, 0.31 for $m = 6$, and 0.24 for $m = 10$. Thus, PIRT does not overtake the CRT at conventional targets such as 80% power when $m = 2$, but it does so at such targets once clusters are moderately larger. Larger m helps twice: it increases effective overlap and reduces leave-one-out variation. The simulations in Section 4.2 evaluate these asymptotic predictions in finite samples.

4.2. *Simulation evidence under clustered interference*

I replicate the balanced two-stage clustered-interference design studied by Puelz et al. (2022, Section 6.2), matching their outcome model, calibration, biclique-construction procedure, and number of sampled assignments. The only substantive extension is a finer grid of cluster counts $K \in \{20, 30, 50, 75, 100, 150\}$, with $N = 300$ fixed and corresponding cluster sizes $m = N/K \in \{15, 10, 6, 4, 3, 2\}$.

I compare four procedures: the biclique CRT of Puelz et al. (2022), which uses a biclique decomposition of the null-exposure graph to select both focal units and focal assignments; the conditional-focals CRT of Basse et al. (2019), which is the CRT used in the benchmark above; PIRT with cutoff $\alpha/2$; and PIRT with cutoff α. The last version is reported only as a heuristic benchmark for the conservativeness of the finite-sample-valid $\alpha/2$ rule.

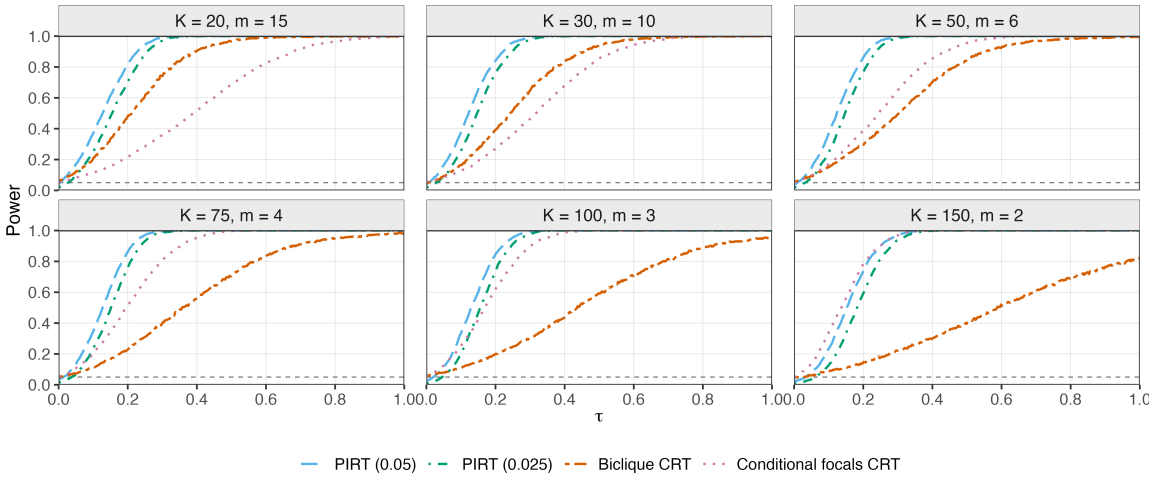


FIGURE 2.—Monte Carlo rejection probabilities under the two-stage clustered-interference design, computed using a one-sided difference-in-means statistic. Panel headers report the number of clusters K and cluster size $m = N/K$, with $N = 300$. The horizontal axis is the mean spillover effect τ, and the dashed line marks the nominal level $\alpha = 0.05$. Rejection rates are computed from 2,000 Monte Carlo repetitions. For the biclique CRT, the conditioning sets are constructed from 5,000 sampled assignments, following Puelz et al. (2022).

For the comparison between PIRT and the conditional-focals CRT, Figure 2 confirms the qualitative crossover pattern predicted by Proposition 2. Size control is satisfactory throughout: both CRT procedures begin near the nominal level, while PIRT with cutoff $\alpha/2$ is conservative, as expected. When $m = 15, 10, 6$, PIRT combines large effective overlap with little leave-one-out variation and is correspondingly more powerful over much of the simulated range. As m shrinks, the overlap advantage weakens while leave-one-out variation becomes more important. At $m = 2$, Proposition 2 predicts a very high crossover power under the valid PIRT cutoff, and the conditional-focals CRT is indeed more powerful throughout the simulated range.

The ranking is already non-universal within CRT. The conditional-focals CRT fixes one focal unit per cluster, giving stable but thin comparison exposures. The biclique CRT can retain more focal units when the null-exposure graph is dense; for example, in the $N = 300, K = 20$ calibration, Puelz et al. (2022, Section 6.2) report that the biclique method averages 5.24 focal units per cluster. But biclique power also depends on the number of focal assignments, and sparse null-exposure graphs can force the biclique construction to discard units or assignments. This focal-unit and focal-assignment tradeoff explains why neither CRT uniformly dominates the other.

The practical implication is that method choice should be evaluated at the design stage conditional on a prespecified exposure contrast. In the balanced two-stage benchmark, Proposition 2 characterizes the target-power range in which PIRT is expected to outperform the conditional-focals CRT. More generally, PIRT is relatively favorable when jointly imputable units remain numerous and well balanced across the spillover and control exposures. For each substantively specified contrast, application-calibrated design-stage simulations can therefore be used to compare the operating characteristics of PIRT and feasible CRTs under plausible outcome heterogeneity.

In spatial applications, the contrast is determined by the null threshold ϵ and the control threshold ϵ_c. These thresholds should be prespecified to represent substantively meaningful distance ranges. A larger ϵ_c places controls farther from treated units but can thin the control cell, whereas a smaller ϵ_c expands the control region but may weaken the contrast if spillovers extend into it. Appendix B discusses this threshold choice in more detail. The spatial simulation below follows this sequence: the distance thresholds are fixed first, and the procedures are then compared for the resulting tests.

5. EMPIRICAL APPLICATION: SPATIAL INTERFERENCE

I illustrate PIRT using the hotspot-policing experiment of Blattman et al. (2021) in Bogotá, Colombia. The study covers 136,984 street segments, of which 1,919 were identified as crime hotspots. Among these hotspots, 756 were randomly assigned to increased daily police patrolling, from 92 to 169 minutes per day over eight months. The main outcomes are the number of violent and property crimes recorded on each street segment during the intervention period.

I test whether concentrating police resources at treated hotspots displaced or deterred crime in nearby areas, and over what distance, by asking whether the data reject partially sharp nulls that rule out spillovers beyond specified thresholds. These tests target the spatial extent of interference while leaving direct effects and nearby spillovers unrestricted.[8]

This is a natural setting for PIRT. The spatial network is dense, and dependence across nearby street segments is difficult to model directly. The original study conducted Fisher randomization tests under the sharp null and reported no significant displacement effect for violent crime and a marginally significant effect for property crime. The question here is whether those conclusions change when inference is based on a finite-sample-valid approach to partially sharp spillover nulls.

5.1. *Power comparison under spatial interference*

Before turning to the data, I compare methods in a simulation calibrated to this setting. The simulation contains 1,000 units, including 20 hotspots and 7 randomly treated units, approximately matching the treatment intensity in the experiment. I use thresholds $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 0.1, 0.2)$. Potential outcomes are calibrated from gamma distributions chosen to match the observed mean and variance of total crime counts. I impose a direct treatment effect of -1, while preserving nonnegative outcomes, and introduce a decreasing displacement effect governed by a positive parameter τ, so that spillovers decay with distance from treated units.[9]

[8] Estimating effect magnitudes or welfare consequences would require an additional effect model.

[9] Additional simulation-design details and descriptive evidence for the Bogotá calibration, including the experimental-sample map and crime-count summaries, are reported in Zhong (2024).

For $k = 0, 1$, I test the partially sharp null

$$H_0^{\epsilon_k} : Y_i(d) = Y_i(d') \quad \text{for all } i \in \{1, \ldots, N\}, \text{ and any } d, d' \in \mathcal{D}_i(\epsilon_k).$$

I compare the sharp-null FRT of Blattman et al. (2021), the biclique CRT of Puelz et al. (2022), PIRT with cutoff $\alpha/2$, and PIRT with heuristic benchmark cutoff α to evaluate conservativeness. For each method, I study the size at $\tau = 0$ and power over 50 equally spaced values of $\tau \in [0, 1]$, using 2,000 Monte Carlo repetitions at each value. For the biclique CRT, the conditioning set is built from 5,000 sampled assignments.[10]

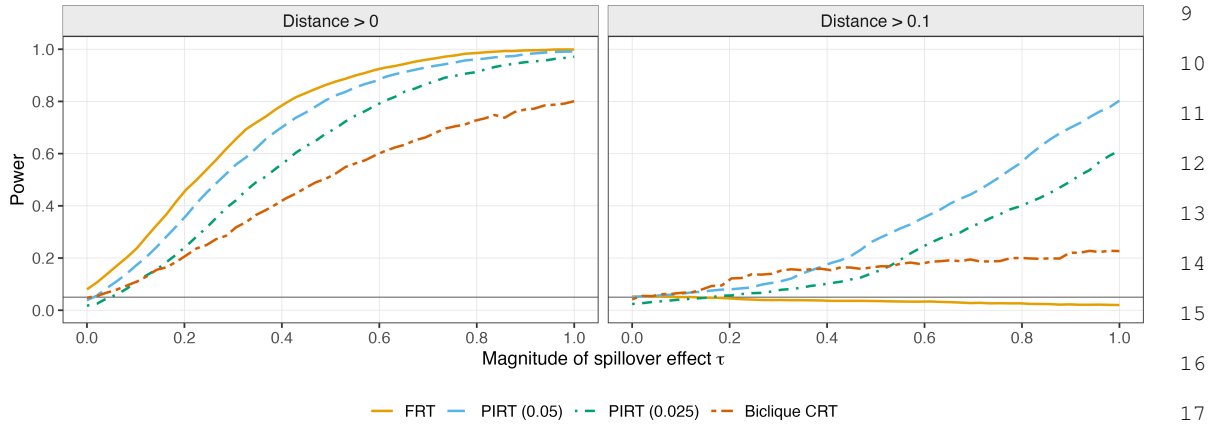


FIGURE 3.—Monte Carlo rejection probabilities under the spatial simulation calibrated to the Bogotá application. The simulation has $N = 1{,}000$ units, including 20 hotspots and 7 randomly treated hotspots, and uses a one-sided difference-in-means statistic. The left panel reports results for testing H_0^0, and the right panel reports results for testing $H_0^{0.1}$. Rejection rates are computed from 2,000 Monte Carlo repetitions. For the biclique CRT, the conditioning set is built from 5,000 sampled assignments. The red line marks the nominal level $\alpha = 0.05$. "PIRT (0.025)" denotes PIRT with rejection at $\alpha/2$, and "PIRT (0.05)" denotes PIRT with rejection at α.

Figure 3 illustrates the practical danger of using a sharp-null FRT as a proxy test for partially sharp spillover hypotheses. In the left panel, $\tau = 0$ satisfies H_0^0, because there is no spillover effect on untreated units. However, the sharp null tested by the FRT is false because the simulation includes a direct treatment effect. The approximately 10% rejection rate at $\tau = 0$ therefore reflects the core problem: a sharp-null FRT can confound direct treatment effects with evidence against a spillover null. By contrast, the biclique CRT and PIRT,

[10]The current implementation still requires more than four hours to construct the conditioning sets in this simulation.

at both cutoffs, remain at or below the nominal level for the partially sharp null. Among the methods with theoretical finite-sample guarantees, the biclique CRT is somewhat more powerful for very small τ, but PIRT with cutoff $\alpha/2$ catches up and eventually exceeds it as the signal grows.

The right panel, for testing $H_0^{0.1}$, is harder for all procedures. Here, size is close to nominal throughout, but power is lower because fewer units are relevant to the null, and the spillover effect is smaller. The FRT has almost no power.[11] PIRT nevertheless continues to gain with τ and eventually exceeds the biclique CRT.

These simulations show that the sharp-null FRT can reject for the wrong reason or have little power when applied to partially sharp spillover questions. In this spatial setting, PIRT offers valid inference, good empirical power, and much lower computational burden than the biclique CRT, so I use it in the empirical analysis below.

5.2. *PIRT on the observed data*

I apply PIRT to the publicly available replication materials for Blattman et al. (2021). The materials include the realized assignment, distance intervals defined at 125, 250, and 500 meters, and 1,000 pseudo-randomized assignments used in the original study.[12] The replication materials describe these pseudo-randomized assignments as independent draws from the original assignment mechanism. I therefore use them as the auxiliary assignments in Algorithm 1 and report the corresponding plus-one Monte Carlo PIRT p-values.

Because the crime-count outcomes contain a large mass of zeros, I report results for both an indicator for any crime and the raw crime count. Table VI presents the resulting PIRT p-values across distance thresholds.

The unadjusted p-values show different patterns for violent and property crime. For both violent-crime outcomes, the p-values for H_0^0 are below 0.05, corresponding to rejection at the 10% level. None of the property-crime nulls is rejected at that level. To account for testing multiple distance thresholds, I apply the fixed-sequence procedure described in Ap-

[11] The FRT p-value remains large because units close to treated hotspots under the observed assignment can contribute to the statistic under reassignment, while still being affected by spillovers of size τ, which dampens the sharp-null contrast.

[12] Because the exact street-segment locations are not observed, I cannot generate additional assignments that are guaranteed to satisfy the original design.

TABLE VI

p-VALUES: TESTING THE DISPLACEMENT EFFECT AT DIFFERENT DISTANCES

	Unadjusted p-values		
	$(0m, \infty)$	$(125m, \infty)$	$(250m, \infty)$
Violent crime			
Indicator of > 0	*0.027*	0.812	0.060
# of crimes	*0.047*	0.546	*0.045*
Property crime			
Indicator of > 0	0.390	0.466	0.486
# of crimes	0.325	0.346	0.394

Note: The table reports PIRT p-values for tests of displacement in violent and property crime. Thus, for a 5% test the relevant cutoff is 0.025; for a 10% test the relevant cutoff is 0.05. "Indicator of > 0" refers to an indicator for any crime occurrence, while "# of crimes" refers to the raw number of reported crimes. p-values are constructed using the difference-in-means statistic.

pendix B separately within each outcome. For both violent-crime outcomes, the procedure rejects H_0^0 and then stops after failing to reject $H_0^{125\text{m}}$. Because the statistic for H_0^0 contrasts untreated streets within 125 meters of a treated hotspot with more distant streets, these results are consistent with short-range displacement.

For violent-crime counts, the p-value at the 250-meter threshold is also 0.045. Although this result lies beyond the stopping point of the fixed-sequence procedure and is therefore exploratory, it suggests that displacement may vary nonmonotonically across distance or differ across types of violent crime. Appendix E examines the latter possibility by decomposing violent crime into homicides and sexual assaults on the one hand, and all other violent crimes on the other.

Taken together, these results point to displacement in violent crime but not property crime and illustrate how PIRT can reveal spatial patterns with potential policy relevance for Bogotá. From a policy perspective, they suggest that hotspot policing should be evaluated not only at treated locations, but also in nearby areas that may absorb displaced crime. The contrast between violent and property crime is consistent with the interpretation in Blattman et al. (2021) that different offenses may respond to policing through different mechanisms. Some violent crimes may involve more mobile criminal opportunities, whereas property crimes may be more closely tied to location-specific targets. Policy

design should therefore account for differences in displacement patterns across offenses and locations.

6. CONCLUSION

This paper develops a randomization-based framework for testing partially sharp null hypotheses under interference. The difficulty is that only part of the potential outcome schedule is imputable under the null, and the relevant imputable set varies across assignments. PIRT handles the first problem through pairwise imputable statistics and the second through pairwise comparison-based p-values, delivering a procedure that is finite-sample valid and straightforward to implement.

The comparison with conditional randomization tests shows that relative power is design-specific. PIRT gains from greater effective overlap, whereas conditioning can stabilize the randomization comparison, and PIRT additionally pays the stricter $\alpha/2$ cutoff required for its unconditional finite-sample guarantee. This suggests a natural direction for future work: combining pairwise imputability with restricted or selectively conditioned assignment spaces to improve power while preserving finite-sample validity.

The Bogotá application illustrates how PIRT can characterize the spatial extent of spillovers while allowing direct effects and nearby interference, yielding suggestive evidence of violent-crime displacement but not property-crime displacement. More broadly, the pairwise-imputation principle may be useful in other settings where a justified randomization distribution is available and policy-relevant null hypotheses are only partially sharp, including staggered-adoption designs.

REFERENCES

ABADIE, ALBERTO, SUSAN ATHEY, GUIDO W. IMBENS, AND JEFFREY M. WOOLDRIDGE (2020): "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 88 (1), pp. 265–296. [2]

ABADIE, ALBERTO, SUSAN ATHEY, GUIDO W IMBENS, AND JEFFREY M WOOLDRIDGE (2023): "When should you adjust standard errors for clustering?" *The Quarterly Journal of Economics*, 138 (1), 1–35. [2]

ANGRIST, JOSHUA D. (2014): "The perils of peer effects," *Labour Economics*, 30, 98–108. [2]

ARONOW, PETER M. (2012): "A General Method for Detecting Interference Between Units in Randomized Experiments," *Sociological Methods & Research*, 41 (1), 3–16. [2, 4]

ATHEY, SUSAN, DEAN ECKLES, AND GUIDO W. IMBENS (2018): "Exact p-Values for Network Interference," *Journal of the American Statistical Association*, 113 (521), 230–240. [2, 4, 13, 14]

BAIRD, SARAH, J. AISLINN BOHREN, CRAIG MCINTOSH, AND BERK ÖZLER (2018): "Optimal Design of Experiments in the Presence of Interference," *The Review of Economics and Statistics*, 100 (5), 844–860. [2]

BARBER, RINA FOYGEL, EMMANUEL J CANDES, AADITYA RAMDAS, AND RYAN J TIBSHIRANI (2021): "Predictive inference with the jackknife+," *The Annals of Statistics*, 49 (1), 486–507. [3, 16]

BASSE, GUILLAUME, PENG DING, AVI FELLER, AND PANOS TOULIS (2024): "Randomization Tests for Peer Effects in Group Formation Experiments," *Econometrica*, 92 (2), 567–590. [4, 11]

BASSE, G W, A FELLER, AND P TOULIS (2019): "Randomization tests of causal effects under interference," *Biometrika*, 106 (2), 487–494. [4, 18, 22]

BLATTMAN, CHRISTOPHER, DONALD P GREEN, DANIEL ORTEGA, AND SANTIAGO TOBÓN (2021): "Place-based interventions at scale: The direct and spillover effects of policing and city services on crime," *Journal of the European Economic Association*, 19 (4), 2022–2051. [2, 4, 24, 25, 26, 27]

BOND, ROBERT M., CHRISTOPHER J. FARISS, JASON J. JONES, ADAM D. I. KRAMER, CAMERON MARLOW, JAIME E. SETTLE, AND JAMES H. FOWLER (2012): "A 61-million-person experiment in social influence and political mobilization," *Nature*, 489 (7415), 295–298. [2]

BREZA, EMILY, ARUN G. CHANDRASEKHAR, TYLER H. MCCORMICK, AND MENGJIE PAN (2020): "Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data," *American Economic Review*, 110 (8), 2454–84. [3]

BROLLO, FERNANDA, KATJA MARIA KAUFMANN, AND ELIANA LA FERRARA (2020): "Learning Spillovers in Conditional Welfare Programmes: Evidence from Brazil," *The Economic Journal*, 130 (628), 853–879. [2]

COX, D. R. (1958): *Planning of Experiments*, New York: Wiley. [6]

GRAHAM, BRYAN S (2020): "Network data," in *Handbook of econometrics*, Elsevier, vol. 7, 111–218. [4]

GUAN, LEYING (2024): "A conformal test of linear models via permutation-augmented regressions," *The Annals of Statistics*, 52 (5), 2059–2080. [3, 16, 17]

HOSHINO, TADAO AND TAKAHIDE YANAGI (2026): "Conditional Randomization Tests for the Specification of Interference Structure," *Journal of Applied Econometrics*. [7, 13]

IMBENS, GUIDO W. AND DONALD B. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press. [3, 12, 13]

KELLY, MORGAN (2021): "Persistence, Randomization, and Spatial Noise," Working Papers 202124, School of Economics, University College Dublin. [2]

MANSKI, CHARLES F. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60, 531–542. [4]

OWUSU, JULIUS (2023): "Randomization inference of heterogeneous treatment effects under network interference," *arXiv preprint arXiv:2308.00202*. [11]

POLLMANN, MICHAEL (2020): "Causal inference for spatial treatments," *arXiv preprint arXiv:2011.00373*. [2]

PUELZ, DAVID, GUILLAUME BASSE, AVI FELLER, AND PANOS TOULIS (2022): "A graph-theoretic approach to randomization tests of causal effects under general interference," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84 (1), 174–204. [4, 7, 18, 22, 23, 25]

ROSENBAUM, PAUL (2007): "Interference Between Units in Randomized Experiments," *Journal of the American Statistical Association*, 102 (477), 191–200. [2]

TAYLOR, SEAN J. AND DEAN ECKLES (2018): *Randomized Experiments to Detect and Estimate Social Influence in Networks*, Cham: Springer International Publishing, 289–322. [3]

TELLA, RAFAEL DI AND ERNESTO SCHARGRODSKY (2004): "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack," *The American Economic Review*, 94 (1), 115–133. [2]

VOVK, VLADIMIR (2015): "Cross-conformal predictors," *Annals of Mathematics and Artificial Intelligence*, 74 (1), 9–28. [3, 16]

ZHANG, YAO AND QINGYUAN ZHAO (2023): "What is a Randomization Test?" *Journal of the American Statistical Association*, 118 (544), 2928–2942. [2, 7, 12]

——— (2025): "Multiple conditional randomization tests for lagged and spillover treatment effects," *Biometrika*, 112 (1), asae042. [5]

ZHONG, LIANG (2024): "Unconditional randomization tests for interference," *arXiv preprint arXiv:2409.09243*. [10, 15, 24]