Randomization tests for peer effects in group formation experiments

Guillaume Basse, Peng Ding, Avi Feller, Panos Toulis*

September 5, 2023

Abstract

Measuring the effect of peers on individuals' outcomes is a challenging problem, in part because individuals often select peers who are similar in both observable and unobservable ways. Group formation experiments avoid this problem by randomly assigning individuals to groups and observing their responses; for example, do first-year students have better grades when they are randomly assigned roommates who have stronger academic backgrounds? In this paper, we propose randomization-based permutation tests for group formation experiments, extending classical Fisher Randomization Tests to this setting. The proposed tests are justified by the randomization itself, require relatively few assumptions, and are exact in finite samples. This approach can also complement existing strategies, such as linear-in-means models, by using a regression coefficient as the test statistic. We apply the proposed tests to two recent group formation experiments.

Keywords: Causal inference; Conditional randomization test; Equivariance; Exact p-value; Non-sharp null hypothesis

*We thank Alex Franks, Xinran Li, Sam Pimentel, and Fredrik Sävje as well as seminar participants at UCLA, UC Berkeley, and UCSB for helpful comments. GB acknowledges support from the U.S. National Science Foundation (grant # 1713152). PD acknowledges support from the U.S. National Science Foundation (grants # 1713152 and # 1945136). AF gratefully acknowledges support from a National Academy of Education/Spencer Foundation postdoctoral fellowship. PT is grateful for the John E. Jeuck Fellowship at Booth.

1 Introduction

Peers influence a broad range of individual outcomes, from health to education to co-authoring papers.[1] However, studying these peer effects in practice is challenging in part because individuals typically select peers who are similar in both observed and unobserved ways (Sacerdote, 2014). *Randomized group formation*, also known as exogenous link formation, avoids this problem by randomly assigning individuals to groups and observing their responses. Among its many applications, this approach has been used to assess the effect of dorm-room composition on student grade point average (GPA; Sacerdote, 2001; Bhattacharya, 2009; Li et al., 2019), the effect of squadron composition on individual performance at military academies (Lyle, 2009; Carrell et al., 2013), the effect of business groups on the diffusion of management practices (Fafchamps and Quinn, 2018; Cai and Szeidl, 2017), the effect of group or team assignments on the performance of professional athletes (Guryan et al., 2009), and the effect of co-workers on productivity (Herbst and Mas, 2015; Cornelissen et al., 2017). A typical substantive question is then, for example: what is the effect of randomly assigning an incoming first-year student to a roommate with high academic preparation (the "exposure") on the student's own end-of-year GPA?

In this paper, we propose analyzing randomized group formation designs from the perspective of "randomization inference," in the spirit of Fisher (1935). Like the classic Fisher Randomization Test (FRT), our ultimate proposal is a straightforward permutation test that (conditionally) permutes each individual's exposure. This test is exact in finite-samples, requires relatively few assumptions, and is justified by the randomization itself. Thus, we argue that our approach is a natural benchmark for analyzing randomized group formation designs, building on a growing literature within economics and econometrics (see Lehmann and Romano, 2005; Imbens and Rubin, 2015; Canay et al., 2017; Young, 2019) that seeks to use the randomization itself as the source of uncertainty when analyzing randomized trials. Moreover, we can combine this approach with popular model-based frameworks, such as the linear-in-means model (Manski, 1993), by using a model to generate the test statistics for subsequent randomization tests. When such models are correctly specified, the corresponding randomization tests are likely to have higher power. Even when the models are incorrectly specified, our proposed randomization tests can still ensure that the p-values are finite-sample valid.

To develop this procedure, we have to overcome several technical and computational hurdles.

[1] All of the co-authors entered the same graduate program in the same year.

First, a key challenge for randomization tests under interference is that the null hypotheses of interest are not typically "sharp," in the sense of specifying all potential outcomes for all units (Rosenbaum, 2007; Hudgens and Halloran, 2008). For example, the null hypothesis of no difference between having 0 or 1 students with high academic preparation in a dorm room does not have any information about dorm rooms that have 2 students of that type. An important innovation for causal inference under interference is to restrict the randomization test to a subset of units, known as *focal units*, which "makes the null hypothesis sharp" and allows for otherwise standard conditional randomization tests (Aronow, 2012; Athey et al., 2018; Basse et al., 2019). Our first contribution is to extend these results to randomized group formation designs, and show that restricting our attention to focal units indeed enables valid randomization-based tests, at least in principle.

In practice, however, it is difficult to obtain draws from the appropriate null distribution in group formation designs. The computationally straightforward approach of naively permuting the exposure of interest (e.g., permuting the number of students in a room of a specific type) is not typically valid, since permuted exposures can be incompatible with the original group formation design. Conversely, the conceptually valid approach of repeatedly assigning groups can be computationally prohibitive for testing non-sharp null hypotheses that require conditioning on a specific set of focal units.

Our second main contribution is therefore to develop computationally efficient randomization tests that can be implemented easily via permutations. For a broad class of designs, we show that permuting exposures separately for each level of individuals' own attributes (e.g., high academic preparation) leads to valid randomization tests. Using algebraic group theory, we prove that a key property in all these designs is *equivariance*, which, roughly speaking, ensures that an invariance in the design translates into an invariance on peer exposure. Our paper thus provides one of the first, general theoretical results on efficient implementation of randomization tests of peer effects via permutations.

We apply our results to two studies based on randomized group formation designs: first-year students randomly assigned to dorms (Li et al., 2019) and chief executive officers (CEOs) randomly assigned to group meetings (Cai and Szeidl, 2017). We describe stylized versions of these examples in the next section and discuss the applications in more detail in Section 6. In the appendix, we also include extensive simulation studies showing the validity of the method under a range of scenarios.

Our approach combines two recent threads in the literature on causal inference under in-

terference. In the first thread, Aronow (2012), Athey et al. (2018), and Basse et al. (2019) develop conditional randomization tests that are valid under interference; we discuss this further in Section 3.2. In that setup, the groups are fixed and the intervention itself is randomized. In the second thread, Li et al. (2019) explicitly consider group formation designs and define peer effects using the potential outcomes framework. Their paper mainly considers the *Neymanian* perspective that focuses on randomization-based point and interval estimation based on normal approximations (Imbens and Rubin, 2015; Abadie et al., 2020). By contrast, our paper chiefly considers the *Fisherian* perspective that instead focuses on finite-sample exact p-values via randomization-based testing. This allows us to examine hypotheses for smaller subpopulations, including those in our motivating examples. Moreover, our approach is valid for arbitrary outcome distributions, including possibly heavy-tailed sales revenue in the second example (Rosenbaum, 2002; Lehmann and Romano, 2005).

2 Setup and framework

2.1 From regression to randomization inference for peer effects

To illustrate the notation and the key concepts, we introduce two running examples. Example 1 presents an idealized version of Sacerdote (2001) and Li et al. (2019), in which incoming college first-year students are randomly assigned to dorm rooms. Example 2 presents an idealized version of Cai and Szeidl (2017), in which CEOs of Chinese firms are randomly assigned to attend monthly group meetings. Both examples have a common structure in which individuals are randomly assigned to groups. We observe attribute A and outcome Y for each individual, and the attributes of peer individuals in the group, W. The goal is to estimate the "effect" of W on Y. We make these statements more precise in the next section and analyze the original data from both examples in Section 6.

Example 1. *Suppose that N incoming first-year students are paired into $N/2$ dorm rooms of size 2. We classify incoming first-year students as having high ($A = 1$) or low ($A = 0$) incoming level of academic preparation (e.g., based on standardized test scores and high school grades). We want to understand whether a first-year student's end-of-year GPA varies based on the academic preparation of their roommate (W). Specifically, is there an effect on end-of-year GPA (Y) of being assigned a roommate with 'high' incoming preparation ($W = 1$) relative to being assigned to a roommate with 'low' incoming preparation ($W = 0$)?*

Example 2. *Suppose that N firm CEOs are assigned to $N/3$ monthly meeting groups of size*

3 *where they discuss business and management practices. Each CEO is classified as leading a 'large firm' ($A = 1$) or 'small firm' ($A = 0$). We want to assess whether the revenue of a CEO's company (Y) is affected by the composition of the meeting group (W). Specifically, is there an impact on the firm's revenue of assigning that firm's CEO to a group with two CEOs from large firms ($W = 2$) relative to assigning that firm's CEO to a group with one ($W = 1$) or no CEOs ($W = 0$) from large firms?*

These examples capture the notion of a peer effect as the idea that a given unit's outcome may be affected by their peers' attributes. A vast literature in economics formalizes these ideas; see, among others, Manski (1993), Brock and Durlauf (2001), Sacerdote (2011), Goldsmith-Pinkham and Imbens (2013), and Angrist (2014). We now briefly review common existing approaches and discuss recent work that motivates the use of linear regression from the randomization perspective (Li et al., 2019). Since our eventual goal is a fully randomization-based framework for analyzing randomized group formation designs, our discussion here necessarily focuses on reduced-form approaches, setting aside a vibrant literature on more structural models of peer effects and social interactions (see Bramoullé et al., 2020).

Linear-in-means model. We begin with the workhorse *linear-in-means model*, described in detail in a seminal paper from Manski (1993), which regresses Y on $\bar{A}$, the average attribute in the group. Following a long literature (see Sacerdote, 2011), we initially consider the leave-one-out form of this model, which separates out A, a unit's own attribute, and W (a transformation of) the leave-own-unit-out average attribute:

$$Y_i^{\text{obs}} = \alpha + \beta A_i + \tau W_i + \varepsilon_i,$$

where Y_i^{obs} is the observed outcome for unit i. For Example 1, both A and W are binary; for Example 2, A is binary and W takes on three values, $\{0, 1, 2\}$. The coefficient τ is referred to as the *exogenous peer effect* (Manski, 1993) or the *social return* (Angrist, 2014). Standard errors are typically clustered at the group level. Importantly, we do not include specifications with Y on the right-hand side and therefore do not consider so-called *endogenous peer effects*. While this avoids a range of thorny econometric questions (see Manski, 1993; Angrist, 2014), this choice necessarily restricts the type of substantive questions we can address. Similarly, since we focus on experiments in which individuals are randomly assigned to groups, we also exclude *correlated effects*, which could arise if individuals self-select into groups.

Heterogeneous treatment effect model. Even when we focus exclusively on exogenous peer effects, there are many challenges with the linear-in-means model. Most immediately, as Sacerdote (2011) notes: "from an empirical point of view, researchers have found that peer effects are not in fact linear-in-means." This has led researchers to instead consider interacted specifications that allow for possible nonlinearities (Sacerdote, 2001; Duncan et al., 2005; Cai and Szeidl, 2017). In the context of our examples these are specifications of the form:

$$Y_i^{\text{obs}} = \alpha + \beta A_i + \tau W_i + \gamma A_i \cdot W_i + \varepsilon_i. \tag{1}$$

Here the relevant effects are appropriate combinations of the coefficients τ and γ, and, as above, the standard errors are typically clustered at the group level. Again, this interacted model is typically motivated by the desire to estimate a more flexible specification for the (sometimes implicit) underlying model of social interactions.

Motivating regression from randomization. Somewhat surprisingly, Li et al. (2019) show that randomization fully justifies the interacted specification (1) above for a broad class of randomized group formation designs. Moreover, Li et al. (2019) argue that the randomization-based perspective justifies the use of *non-clustered* robust standard errors, suggesting that the common practice of clustering standard errors is overly conservative for such designs, analogous to arguments from Abadie et al. (2023). In this case, failing to include the interaction (i.e., simply running the regression of Y on A and W) leads to a precision-weighted average of the subgroup effects.

From regression to randomization-based testing. As we show below, the regression-based approach from Li et al. (2019), while conceptually elegant, can have poor finite-sample performance. In particular, the asymptotic theory in that paper assumes that both A and W have very few levels, and that the number of individuals within each $A \times W$ group is large. This is not a reasonable approximation in our applications, however; for instance, in the roommates application we analyze in Section 6, the size of an $A \times W$ subgroup can be as small as 4 students.

Our main contribution is to justify and implement randomization-based tests for exogenous peer effects, building on recent proposals for randomization tests under interference (Aronow, 2012; Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022). At a high level, we propose the permutation-based analog of the fully interacted regression model discussed above. The primary technical obstacle is justifying this approach from the randomized group formation

design itself. As we will see, this requires substantial technical overhead, even if the final procedure is itself straightforward. To demonstrate this, we also develop theory for general randomization-based tests for non-sharp nulls.

2.2 Notation and setup

We now formalize the problem setup outlined above. Consider N units to be assigned to K different groups; both numbers are fixed. Let $\mathbb{U} = \{1, \ldots, N\}$ denote the set of units. Let $L_i \in \mathbb{L} = \{1, \ldots, K\}$ denote the labeled group to which unit i is assigned, and define $L = (L_i)_{i=1}^N$ as the full group-label assignment vector. Also, let $P(L) \in [0, 1]$ denote the probability distribution of L, which is known from the experimental design. In a group formation design, the individual i's treatment assignment can be defined as

$$Z_i = \big\{j \in \mathbb{U} : j \neq i \text{ and } L_j = L_i\big\}. \tag{2}$$

Assignment Z_i is therefore the set of individuals assigned to the same group as individual i. Let $Z = (Z_i)_{i=1}^N$ be the full assignment vector.

As we discuss above, a key feature of our setting is that each individual i exhibits a salient *attribute*, A_i; for example, $A_i = 1$ if individual i has high academic preparation entering college. This attribute often plays a special role in group formation designs; for example, in the *stratified group formation design* we consider in Section 5.1, a room must have a fixed, pre-defined number of students with $A_i = 1$. Formally, attribute A_i takes values in a set $\mathbb{A}$, which could be a transformation (e.g., coarsened version) of covariates X_i. We let $A = (A_i)_{i=1}^N$ and $X = (X_i)_{i=1}^N$ be the full vector of attributes and matrix of covariates, respectively.

The goal of this paper is to understand how peers' attributes affect unit outcomes, and so we define the *exposure* for each unit i as:

$$W_i = w_i(Z) = \{A_j : j \in Z_i\}, \tag{3}$$

that is, the exposure of unit i is the multiset of attributes of its neighbors, where a multiset is a set with possibly repeated values. Define $W = w(Z) = (w_i(Z))_{i=1}^N$ as the full vector of exposures, and denote by $\mathbb{W} = \{\mathsf{w}_1, \ldots, \mathsf{w}_m\}$ the finite set of possible exposure values in the experiment. Finally, we let $Y_i(Z)$ denote the real-valued potential outcome of unit i under assignment Z.

While this formulation is general, it is often useful to define exposures as simple functions of the attribute vector A. For example, when A is binary, a natural choice is to define

$$W_i = w_i(Z) = \sum_{j \in Z_i} A_j, \tag{4}$$

the number of "neighbors" of unit i with attribute $A = 1$. All results in the paper hold for general exposure mappings as in (3); we use the simpler formulation in (4) in the running examples for simplicity.

Notation. These definitions are nested, so that L determines Z, and Z determines W. As such, any function on one domain is also a function on a 'finer' domain. To ease notation, we will use '$f(Z)$' to denote a function defined on the domain of Z that is implied by $f^\omega(W)$, and, similarly, use '$f^\ell(L)$' to denote the function on the domain of L that is implied by either $f^\omega(W)$ or $f(Z)$, noting that these all map to the same value: $f^\omega(W) = f(Z) = f^\ell(L)$. For instance, we write $W = w^\ell(L)$ to express the exposures in (3) as a function of L.

2.3 Assumptions and exclusion restrictions

The primary goal of our analysis is to estimate the causal effect of exposing a unit to a mix of peers with one set of attributes versus another, known as the *exogenous peer effect* (Manski, 1993) or the *social return* (Angrist, 2014). Formalizing such effects is non-trivial, however, with a substantial literature defining estimands in terms of coefficients in a linear model. Following a more recent set of papers, we instead formalize these effects via exposure mappings based on potential outcomes (Toulis and Kao, 2013; Manski, 2013; Aronow et al., 2017; Li et al., 2019), which capture the summary of Z that is sufficient to define potential outcomes on the unit level.

To do so, we make the critical assumption that the exposure is *properly specified* in the sense defined below (Aronow et al., 2017):

Assumption 1. *For all $i \in \mathbb{U}$ and for all Z, Z', we have*

$$w_i(Z) = w_i(Z') \Rightarrow Y_i(Z) = Y_i(Z').$$

Under Assumption 1, each unit i has $|\mathbb{W}| = m$ potential outcomes, one for each level of

exposure, and we may write

$$Y_i(Z) = Y_i^{\omega}(w_i(Z)) = Y_i^{\omega}(W_i)$$

to indicate that potential outcomes depend only on the exposure level and not the particular group assignment.

Example 1 (continued)**.** *With dorm rooms of size* 2*, the exposure* W_i *of student* i *is then the attribute* A_j *of student* i*'s roommate. More generally, under the exposure mapping in* (4)*, each unit has only two possible exposures, since* $W_i \in \mathbb{W} = \{0, 1\}$*, and thus each unit has two potential outcomes* $\{Y_i^{\omega}(0), Y_i^{\omega}(1)\}$*.*

Example 2 (continued)**.** *Here, each group has size* 3 *and the assignment* Z_i *of unit* i *is the unordered pair of indices of the other two CEOs in the group. CEO* i*'s exposure is then the number of the other CEOs from large firms. In this case, each unit has three possible exposures, since* $W_i \in \mathbb{W} = \{0, 1, 2\}$ *under* (4)*, and thus each unit has three potential outcomes* $\{Y_i^{\omega}(0), Y_i^{\omega}(1), Y_i^{\omega}(2)\}$*.*

Discussion of Assumption 1. Assumption 1, which is *not* justified by the randomization, is the key substantive assumption in our setup and merits further discussion. At its core, this assumption is an *exclusion restriction*: the only impact of the randomization on an individual's outcome is by changing the salient attributes A — and only the salient attributes — of the other individuals in the group. For instance in Example 1, Assumption 1 implies that room assignment affects unit i's GPA only by changing i's roommate's academic ability, excluding other possible channels of peer influence. This necessarily reduces otherwise complex individual and social interactions to a scalar quantity; for discussion, see Sacerdote (2011).[2] Assumption 1 also plays a role analogous to the stable unit treatment value assumption (SUTVA) by ruling out effects from changing *other* groups. Thus, when combined with the exposure mapping of (3), this assumption implies both a form of *partial interference* and a form of *stratified interference* (Hudgens and Halloran, 2008). Finally, beyond assuming that attribute A is the relevant quantity, Assumption 1 also assumes that the functional form is correctly specified, though we typically allow W to be fully flexible with respect to A.

As we discuss in Appendix A.1, the procedure we outline below will still lead to a valid

[2]Similar challenges arise in other econometric applications, such as 'judge fixed effects', where the choice of attribute (e.g., conviction rate) is important in the overall analysis (e.g., Frandsen et al., 2023).

test without imposing Assumption 1 — though interpreting that rejection is challenging. In particular, the test might reject even if the null hypothesis is indeed correct but Assumption 1 does not hold, for instance if an individual's outcome depends on attributes other than A. At present, there is limited guidance for applied researchers on specifying exposure mappings, in part because these mappings can be highly context-dependent. For point estimation, violating Assumption 1 complicates the implied estimand, which will typically correspond to a particular weighted average of treatment effects. See Li et al. (2019, Section 7) for a discussion in the context of peer effects, Sävje (2023) for more a general discussion of inference with misspecified exposure mappings, and Leung (2022) for an alternative approach that considers approximate exposures. For testing, the situation is more complicated, since it is difficult to interpret a rejection in the absence of Assumption 1. This remains an open research area.

2.4 Sharp and non-sharp null hypotheses

Following the literature on FRTs, we focus on hypotheses defined at the unit level, unlike the regression-based approaches in Section 2.1, which focus on so-called *weak null* hypotheses that average over units. A key technical challenge is that many unit-level null hypotheses of interest are *non-sharp*; a primary goal in this paper is to develop procedures that are both theoretically valid (Section 3) and computationally tractable (Section 4) for such hypotheses.

To illustrate the distinction between sharp and non-sharp null hypotheses, let Z^{obs}, $W^{\text{obs}} = w(Z^{\text{obs}})$, and $Y^{\text{obs}} = Y(Z^{\text{obs}})$ be, respectively, the observed assignment, exposure, and outcome vectors. We say a null hypothesis is *sharp* if, given the null and the observed data, the potential outcomes $\{Y_i^\omega(\mathsf{w}_1), Y_i^\omega(\mathsf{w}_2), \ldots, Y_i^\omega(\mathsf{w}_m)\}$ are imputable for all units $i \in \mathbb{U}$.

First, consider the global null hypothesis:

$$H_0 : Y_i^\omega(\mathsf{w}_1) = Y_i^\omega(\mathsf{w}_2) = \cdots = Y_i^\omega(\mathsf{w}_m) \text{ for all } i \in \mathbb{U}. \tag{5}$$

The null hypothesis in (5) is sharp. As we show in Section 3.1, we can test this hypothesis using a standard FRT; Li et al. (2019, Section 7.1) briefly consider this approach as well. This global sharp null is analogous to the omnibus null hypothesis in a classical analysis of variance (Ding and Dasgupta, 2018) and is a useful starting point for analyses: if there is no evidence of any effect at all, then further analyses are likely less interesting. See Lehmann and Romano (2005, Ch. 15).

At the same time, many substantively interesting causal hypotheses for peer effects are not sharp. One important example is the pairwise null hypothesis of the type:

$$H_0^{\mathsf{w}_1,\mathsf{w}_2} : Y_i^\omega(\mathsf{w}_1) = Y_i^\omega(\mathsf{w}_2) \text{ for all } i \in \mathbb{U}, \tag{6}$$

where $\mathsf{w}_1, \mathsf{w}_2 \in \mathbb{W}$. To illustrate, Example 2 has three possible exposures $\mathbb{W} = \{0, 1, 2\}$, and the sharp null hypothesis of (5) can be written as: $H_0 : Y_i^\omega(0) = Y_i^\omega(1) = Y_i^\omega(2)$ for all $i \in \mathbb{U}$. This contains strictly more information about the missing potential outcomes than a pairwise null hypothesis (6), such as $H_0^{1,2} : Y_i^\omega(1) = Y_i^\omega(2)$ for all $i \in \mathbb{U}$. Substantively, the global sharp null hypothesis assumes that changing the number of peer CEOs from large firms has no effect whatsoever on a firm's revenue. By contrast, the pairwise non-sharp null hypothesis instead imposes that there is no impact on firm revenue of having one versus two peer CEOs from large firms, without imposing any restrictions on revenue in the absence of any peer CEOs from large firms. Thus, the ability to test pairwise null hypotheses is critical for learning more from the experiment than the initial conclusion that the experiment indeed had some effect somewhere.

Finally, we are often interested in null hypotheses for the subset of units with a given attribute $A_i = a$. As we discuss in our applications below, we frequently believe that the exposure will have differential effects depending on an individual's own attribute. Specifically, we can modify both (5) and (6) to only consider units with $A_i = a$:

$$H_0(a) : Y_i^\omega(\mathsf{w}_1) = Y_i^\omega(\mathsf{w}_2) = \cdots = Y_i^\omega(\mathsf{w}_m) \text{ for all } i \in \mathbb{U} \text{ such that } A_i = a \tag{7}$$

and

$$H_0^{\mathsf{w}_1,\mathsf{w}_2}(a) : Y_i^\omega(\mathsf{w}_1) = Y_i^\omega(\mathsf{w}_2) \text{ for all } i \in \mathbb{U} \text{ such that } A_i = a. \tag{8}$$

The results below immediately carry over to these subgroup null hypotheses by conditioning on the set of units with $A_i = a$. We therefore focus on the simpler null hypotheses of (5) and (6), returning to subgroup null hypotheses in Section 6.

We note that this framework does not require formally specifying an alternative hypothesis; see Athey et al. (2018) for a discussion in the context of randomization tests under network interference. In our applications, the choice of the test statistic is motivated by having power against two-sided alternative hypotheses on coefficients from a linear regression model, such as the coefficient on W in the regression of Y on A and W.

2.5 Toy example and sketch of key ideas

Before turning to the theoretical results, we first illustrate the key challenges through a toy example, shown in Figure 1. For this example, individuals possess a binary attribute, represented by squares ($A_i = 1$) and circles ($A_i = 0$), and are assigned to one of three dorm rooms, one with size 3 (Room I, a "triple") and two with size 2 (Rooms II and III, "doubles"), shown as large rectangles.[3] Rooms are assigned via a *completely randomized group formation design* (see Section 5.2), which means that the sizes of the three rooms are fixed, but that the number of square roommates in each room can vary. Here the exposure mapping is the number of roommates with $A_j = 1$ as defined in (4), so that $\mathbb{W} = \{0, 1, 2\}$. Figure 1 shows the realized assignment Z^{obs} and induced exposure W^{obs}.

In this toy example, we are interested in testing two null hypotheses. First, the global sharp null hypothesis is that individuals' outcomes are the same regardless of the number of "square" roommates. Written in terms of unit-level outcomes, this is $H_0 : Y_i^\omega(0) = Y_i^\omega(1) = Y_i^\omega(2)$ for all $i \in \mathbb{U}$. Second, a non-sharp, pairwise null hypothesis is whether there is an effect of having zero versus one "square" roommate, $H_0^{0,1} : Y_i^\omega(0) = Y_i^\omega(1)$ for all $i \in \mathbb{U}$.

Our starting place for testing these null hypotheses is a permutation test based on permuting the exposure vector, W^{obs}. The right-hand columns of Figure 1 show three possible permutations, swapping the observed exposure for unit 5, W_5^{obs}, with, respectively, the exposures for units 4, 3, and 2 (W_4^{obs}, W_3^{obs}, W_2^{obs}).

Naive permutation tests can fail. While seemingly natural, the first two permutations in Figure 1, W' and W'', are invalid. The first permutation W', which swaps the exposures of units 4 and 5, leads to invalid tests for both H_0 and $H_0^{0,1}$ because it is incompatible with the group formation design; that is, there are no assignments Z' such that, $w(Z') = W'$. To see this, note that under W', units 1, 2, and 5 — the only "square" units in the set — would each need to have exactly one other "square" roommate. But this configuration is impossible as it requires an even number of "square" units.

The second permutation W'', which swaps the exposures of units 3 and 5, leads to an invalid test for $H_0^{0,1}$. In particular, we observe $Y_3^{\text{obs}} = Y_3^\omega(2)$ for unit 3; under $H_0^{0,1}$ we have no information about either $Y_3^\omega(0)$ or $Y_3^\omega(1)$, since $H_0^{0,1}$ is only about treatment exposures 0 and 1. And since $W_3'' = 0$, we cannot construct a valid test statistic under W''.

[3] The sizes of the rooms themselves are not central here, and merely restrict the set of possible exposures. We also mean no disrespect to any of our former roommates, several of whom could be described as "squares."

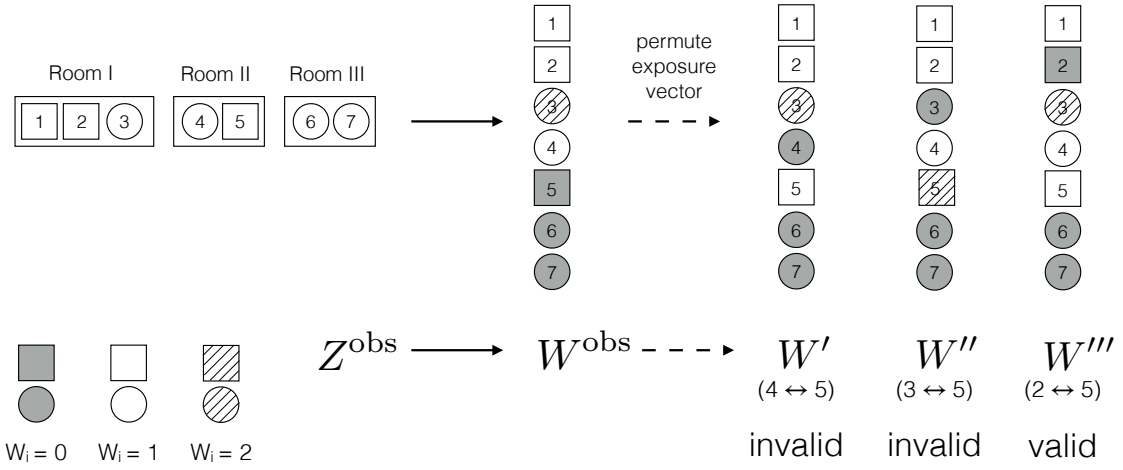

Figure 1: Example of a group formation design. Squares represent units with attribute $A_i = 1$ and circles units with attribute $A_i = 0$. Units with exposure $W_i = 0$ (i.e., zero "square" roommates) are shaded grey; units with exposure $W_i = 1$ have no color; and units with exposure $W_i = 2$ (only unit 3) have a patterned background.

Randomization tests based on draws from the assignment distribution are valid but computationally prohibitive. An alternative to naively permuting W^{obs} is to instead re-draw room assignments directly, $Z' \sim P(Z')$, and compute the induced exposures for each assignment, $W' = w(Z')$. This will always lead to valid, direct randomization-based tests for the *sharp* global null hypothesis, H_0, though these are not always permutation tests.

However, extending this to non-sharp null hypotheses like $H_0^{0,1}$ is non-trivial. In Figure 1, for instance, we need to sample from all room assignments such that unit 3 has exposure $W_i = 2$. Enumerating all such room assignments becomes exponentially hard (increasing in the sample size), and is especially challenging when $w(\cdot)$ is complex. As such, exact or approximate sampling (e.g., rejection sampling) from the conditional treatment assignment distribution is prohibitive.[4]

Permutation tests stratified by attribute are valid and tractable for both sharp and non-sharp null hypotheses. Remarkably, we can generate valid, computationally

[4]To illustrate, consider Example 1 with $N = 32$ units in $K = 8$ rooms of 4 students. Drawing 1,000 samples from the conditional exposure distribution via rejection sampling requires over 400 hours on a conventional laptop, though this can be parallelized. By contrast, the actual roommates application in Section 6 has $N = 156$ units in $K = 39$ groups. Since computation time increases exponentially in the number of groups, a practical test based on rejection sampling is infeasible.

tractable randomization tests for both null hypotheses by simply stratifying the permutations based on attribute A. In Figure 1, this is the set of permutations that separately permute the exposures for circles and squares. The rightmost column of Figure 1 shows one such permutation W''', which swaps the exposures for units 2 and 5; this is valid because both units 2 and 5 are "squares."

While the final procedure is straightforward, the mathematical justification is intricate and stems from a key property called *equivariance*. As we formalize in Theorem 1, equivariance guarantees that, under some technical restrictions on the design and the exposure function, permuting room assignment *stratified by attribute* is equivalent to permuting the exposure directly, and these permutations lead to a valid test. These technical conditions are satisfied for many common designs, including those in the applications we re-analyze in Section 6. The final permutation in Figure 1 illustrates this idea: due to equivariance, swapping the room assignments Z for units 2 and 5 is equivalent to swapping their implied exposures. Thus, W''' could form the basis of a valid permutation test for $H_0^{0,1}$.

3 Valid tests in arbitrary group formation designs

In this section, we introduce conceptually general — albeit possibly infeasible — procedures for constructing valid tests for sharp and non-sharp null hypotheses for arbitrary group formation designs. For sharp null hypotheses, the procedure is a straightforward application of the standard FRT to our setting. For non-sharp null hypotheses, however, the procedure requires greater care to ensure validity. We turn to constructing feasible randomization tests in the next section.

3.1 Randomization test for the sharp null

We start with a brief review of the classical FRT for sharp null hypotheses (Fisher, 1935; Lehmann and Romano, 2005; Imbens and Rubin, 2015), as a stepping stone to the more challenging non-sharp null hypotheses discussed in Section 3.2. Consider a test statistic $T(z; Y)$ as a function of the observed treatment and outcome vectors; any choice will lead to a valid test, but certain statistics will lead to more power. One reasonable choice, for example, is the coefficient of W in the regression of Y on (W, A) and other covariates; see also Section 6.2 for an applied example. We can test the sharp null hypothesis H_0 with Procedure 1 below.

Procedure 1. *Consider observed assignment* $Z^{\text{obs}} \sim P(Z^{\text{obs}})$.

1. *Observe outcomes,* $Y^{\text{obs}} = Y(Z^{\text{obs}})$.
2. *Compute test statistic* $T^{\text{obs}} = T(Z^{\text{obs}}; Y^{\text{obs}})$.
3. *For* $Z' \sim P(Z')$, *let* $T' = T(Z'; Y^{\text{obs}})$ *and define* $\text{pval}(Z^{\text{obs}}) = P(T' \geq T^{\text{obs}})$, *where* T^{obs} *is fixed and the randomization distribution is with respect to* $P(Z')$.

This procedure is computationally straightforward if the analyst has access to the assignment mechanism $P(Z)$, which is necessary for Step 3.

Proposition 1. *The p-value obtained in Procedure 1 is valid, in the sense that if* H_0 *is true, then* $P\{\text{pval}(Z^{\text{obs}}) \leq \alpha\} \leq \alpha$ *for any* $\alpha \in [0, 1]$.

In general, it is difficult to compute $\text{pval}(Z^{\text{obs}})$ exactly, and we must rely on Monte Carlo approximation. This can be done by replacing the third step above by:

3. For $r = 1, \ldots R$, draw $Z^{(r)} \sim P(Z^{(r)})$ and compute $T^{(r)} = T(Z^{(r)}; Y^{\text{obs}})$. Then compute the approximation $\text{pval}(Z^{\text{obs}}) \approx R^{-1} \sum_{r=1}^{R} \mathbb{1}(T^{(r)} \geq T^{\text{obs}})$.

In practice, the test statistic T used in Procedure 1 is chosen to depend on Z only through the exposures $W = w(Z)$. Following our convention in Section 2.2, we can re-write this test statistic as $T(Z; Y^{\text{obs}}) = T^{\omega}(W; Y^{\text{obs}})$. Procedure 1 can then be reformulated as:

Procedure 1b (special case)**.** *Consider observed assignment* $Z^{\text{obs}} \sim P(Z^{\text{obs}})$.

1. *Observe outcomes,* $Y^{\text{obs}} = Y^{\omega}(W^{\text{obs}})$.
2. *Compute test statistic* $T^{\text{obs}} = T^{\omega}(W^{\text{obs}}; Y^{\text{obs}})$.
3. *For* $W' \sim P(W')$, *let* $T' = T^{\omega}(W'; Y^{\text{obs}})$ *and define* $\text{pval}(Z^{\text{obs}}) = P(T' \geq T^{\text{obs}})$, *where* T^{obs} *is fixed and the randomization distribution is with respect to* $P(W')$.

The distribution $P(W')$ used above is directly induced by $P(Z')$, as $P(W') = P\{w(Z')\}$, and the validity of Procedure 1b follows from that of Procedure 1, as established by Proposition 1.

3.2 Randomization tests for non-sharp nulls

We now turn to the more challenging problem of testing non-sharp pairwise hypotheses such as $H_0^{\mathsf{w}_1, \mathsf{w}_2}$. In general, Procedure 1 can only be valid if the test statistic is imputable under H_0 (Basse et al., 2019); that is, $T(Z; Y(Z)) = T(Z; Y^{\text{obs}})$ under H_0, for all Z for which $P(Z) > 0$. This property holds because H_0 is sharp, which implies that $Y(Z) = Y^{\text{obs}}$

under H_0. In contrast, pairwise null hypotheses like $H_0^{\mathsf{w}_1,\mathsf{w}_2}$ are not sharp, and the FRT methodology does not apply directly.

3.2.1 Focal units

One popular technical tool for randomization tests under interference is to restrict the test statistic to use only outcomes from a subpopulation of units, known as *focal units* (Aronow, 2012; Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022). Here, we use this device to construct valid tests: we effectively "make the null hypothesis sharp" by restricting the test to the set of focal units. We formalize this next.

Let a binary variable U_i to indicate whether unit i is selected as a focal unit. To test $H_0^{\mathsf{w}_1,\mathsf{w}_2}$ we can define U as follows:

$$U = u(Z) = (U_1, \ldots, U_N) \in \{0,1\}^N, \text{ with } U_i = 1 \text{ if and only if } w_i(Z) \in \{\mathsf{w}_1, \mathsf{w}_2\}. \tag{9}$$

That is, we select as focal units the set of units that receive either exposure w_1 or exposure w_2 under assignment Z. The realized set of focal units, $U^{\text{obs}} = u(Z^{\text{obs}})$, therefore denotes the set of all units with *observed* exposure w_1 or w_2, the null exposures of interest. To illustrate, for testing the pairwise null hypothesis $H_0^{1,2}$ in Example 2, the focal units are all CEOs who have $W_i^{\text{obs}} = 1$ or $W_i^{\text{obs}} = 2$ peer CEOs from large firms. So long as we restrict testing to this subset of units — and under some restrictions on the possible assignment vectors — the null hypothesis $H_0^{\mathsf{w}_1,\mathsf{w}_2}$ behaves like a sharp null hypothesis. Basse et al. (2019) build on this intuition and develop a valid conditional testing procedure.

Adapting the formulation from Basse et al. (2019) to the peer effects setting requires two changes to Procedure 1. First, we need to resample assignments (Step 3 of Procedure 1) with respect to the conditional distribution of treatment assignment,

$$P\{Z' \mid u(Z') = U^{\text{obs}}\} \propto \mathbb{1}\{u(Z') = U^{\text{obs}}\} P(Z'), \tag{10}$$

rather than with respect to the unconditional distribution. In the terminology of Basse et al. (2019), U^{obs} is the conditioning event of the test, and its (degenerate) conditional distribution $P(U \mid Z) = \mathbb{1}\{u(Z) = U\}$ is the conditioning mechanism. Second, to ensure that the potential outcomes used by the test are imputable, we need to restrict the test statistic to the units in the focal set; we denote this new test statistic as $T(z; Y, U)$.

3.2.2 Valid tests

The following procedure leads to a valid test of the pairwise non-sharp hypothesis $H_0^{\mathsf{w}_1,\mathsf{w}_2}$.

Procedure 2. *Consider observed assignment* $Z^{\text{obs}} \sim P(Z^{\text{obs}})$.

1. *Observe outcomes,* $Y^{\text{obs}} = Y(Z^{\text{obs}})$.
2. *Let* $U^{\text{obs}} = u(Z^{\text{obs}})$ *and compute* $T^{\text{obs}} = T(Z^{\text{obs}}; Y^{\text{obs}}, U^{\text{obs}})$.
3. *For* $Z' \sim P(Z' \mid U^{\text{obs}})$, *let* $T' = T(Z'; Y^{\text{obs}}, U^{\text{obs}})$ *and define the p-value as* $\text{pval}(Z^{\text{obs}}) = P(T' \geq T^{\text{obs}} \mid U^{\text{obs}})$, *where* T^{obs} *is fixed and the randomization distribution is with respect to* $P(Z' \mid U^{\text{obs}})$ *as defined in* (10).

As in Section 3.1, we generally consider test statistics that depend on Z only through the exposure vector $W = w(Z)$. In addition, notice that the focal indicator $U = u(Z)$ in (9) also depends on Z only through W. Following our convention in Section 2.2, this allows us to redefine the focal indicator as $U = u(Z) = u^\omega(W)$, and rewrite Procedure 2 as follows:

Procedure 2b (special case)**.** *Consider observed assignment* $Z^{\text{obs}} \sim P(Z^{\text{obs}})$.

1. *Observe outcomes,* $Y^{\text{obs}} = Y^\omega(W^{\text{obs}})$.
2. *Compute* $U^{\text{obs}} = u^\omega(W^{\text{obs}})$ *and* $T^{\text{obs}} = T^\omega(W^{\text{obs}}; Y^{\text{obs}}, U^{\text{obs}})$.
3. *For* $W' \sim P(W' \mid U^{\text{obs}})$, *let* $T' = T^\omega(W'; Y^{\text{obs}}, U^{\text{obs}})$ *and define the p-value as* $\text{pval}(Z^{\text{obs}}) = P(T' \geq T^{\text{obs}})$, *where* T^{obs} *is fixed and the randomization distribution is with respect to* $P(W' \mid U^{\text{obs}})$. *Note again that the distribution* $P(W' \mid U^{\text{obs}})$ *is induced by that of* $P(Z' \mid u(Z') = U^{\text{obs}})$.

Proposition 2. *Procedure 2 and its special case, Procedure 2b, lead to valid p-values conditionally and marginally for* $H_0^{\mathsf{w}_1,\mathsf{w}_2}$. *That is, if* $H_0^{\mathsf{w}_1,\mathsf{w}_2}$ *is true then* $P\{\text{pval}(Z^{\text{obs}}) \leq \alpha \mid U^{\text{obs}}\} \leq \alpha$ *for any* U^{obs} *and any* $\alpha \in [0, 1]$, *and thus* $P\{\text{pval}(Z^{\text{obs}}) \leq \alpha\} \leq \alpha$ *as well.*

The proof for Proposition 2 uses Theorem 1 of Basse et al. (2019). For the rest of this paper, we only consider test statistics that depend on Z through $W = w(Z)$ alone. All statements in subsequent sections will thus be in terms of Procedures 1b and 2b instead of 1 and 2.

The conditional randomization tests described in this section differ from standard conditional tests in several important ways. First, the goal of standard conditional tests is typically to make the test more powerful (Lehmann and Romano, 2005; Hennessy et al., 2016), rather than to ensure validity. The conditioning in Procedures 2 and 2b, by contrast, is necessary to

ensure that the test is valid. Second, the procedure depends strongly on the non-sharp null hypothesis being tested. Indeed, conditional randomization tests can only test certain non-sharp null hypotheses, such as $H_0^{\mathsf{w}_1,\mathsf{w}_2}$, which typically dictate the conditioning mechanism.

Computational challenges with testing non-sharp nulls. As discussed in Section 2.5, testing non-sharp null hypotheses is computationally intractable in realistic settings. Indeed, while we can easily draw samples from the unconditional distribution $P(W)$ through $w(Z)$, where $Z \sim P(Z)$, Step 3 of Procedure 2b requires draws from the unwieldy conditional distribution $P(W \mid U^{\text{obs}})$. Our main proposal in the next section directly addresses this computational issue.

4 Using design symmetry to construct computationally tractable permutation tests

In this section, we show that certain designs can lead to computationally tractable conditional distributions $P(W \mid U)$, which are crucial in the randomization tests discussed above. Our analysis relies on results from algebraic group theory; readers interested in the concrete consequences of these results on the design of randomization tests in our setting may skip ahead to Section 5.

4.1 Equivariant maps and stabilizers

This subsection introduces three key algebraic concepts for our main theoretical result. Let $\mathbb{S}_N$ be the symmetric group containing all permutations of N elements; i.e., bijections of $\{1, \ldots, N\}$ onto itself. For any permutation $\pi \in \mathbb{S}_N$ and a real-valued N-length vector $X \in \mathbb{X} \subseteq \mathbb{R}^N$, let $\pi X = (X_{\pi^{-1}(i)})_{i=1}^N$ be the vector obtained by permuting the indices of X according to π.

Definition 1 (Stabilizer)**.** *$\mathbb{X}$ is closed under $\mathbb{S}_N$ in the sense that $\pi X \in \mathbb{X}$ for all $\pi \in \mathbb{S}_N$ and $X \in \mathbb{X}$. Fix $X \in \mathbb{X}$. The set $\mathbb{S}_N(X) = \{\pi \in \mathbb{S}_N : \pi X = X\}$ also forms a group and is called the stabilizer of X in $\mathbb{S}_N$.*

A stabilizer $\mathbb{S}_N(X)$ captures all possible ways of permuting X without changing X. For instance, if X is a binary vector, then a permutation $\pi \in \mathbb{S}_N(X)$ separately permutes elements with $X_i = 0$ and $X_i = 1$, respectively. This formalizes the argument we sketched out in

Section 2.5: the operations that "permute units with the same attribute" are precisely the elements of $\mathbb{S}_N(A)$, the stabilizer of the attribute vector $A = (A_i)_{i=1}^N$ in the symmetric group.

Definition 2 (Orbits and Partitions)**.** *Fix a subgroup of the symmetric group $\Pi \subseteq \mathbb{S}_N$. Fix $X \in \mathbb{X}$, where $\mathbb{X}$ is closed under Π. Then, the set $\{\pi X : \pi \in \Pi\}$ is called the orbit of X with respect to Π. These orbits define a unique partition of $\mathbb{X}$, denoted by $\mathcal{O}(\mathbb{X}; \Pi)$.*

An orbit is a collection of vectors that are permuted versions of one another. A key property of orbits is that they partition the set that the permutations act upon. This is important in our application because our permutation test on W essentially conditions on an orbit, and we would like the symmetries of our design $P(L)$ to be propagated to the conditional distribution of L given an orbit. The final property that guarantees such symmetry propagation is equivariance.

Definition 3 (Equivariant maps)**.** *Fix a subgroup of the symmetric group $\Pi \subseteq \mathbb{S}_N$. Sets $\mathbb{X}$ and $\mathbb{X}'$ are closed under Π in the sense that $\pi X \in \mathbb{X}$ and $\pi X' \in \mathbb{X}'$ for all $X \in \mathbb{X}, X' \in \mathbb{X}'$ and $\pi \in \Pi$. A function $f : \mathbb{X} \to \mathbb{X}'$ is equivariant with respect to Π if*

$$f(\pi X) = \pi f(X), \text{ for all } X \in \mathbb{X}, \ \pi \in \Pi.$$

By definition, equivariant maps preserve a symmetry from their domain to their target set. This concept is crucial for our main theoretical result, which we turn to next.

4.2 Main result: Sufficient conditions for valid permutation tests on exposures

We now state our main theoretical result, which establishes that if the exposure function, $w^\ell(\cdot)$, and the focal unit selection function, $u^\ell(\cdot)$, are equivariant with respect to a particular permutation subgroup, then the treatment exposure W is uniformly distributed within an orbit defined by that subgroup.

Theorem 1. *Let $P(L)$ denote a distribution of the group labels with support $\mathbb{L} = \{1, \ldots, K\}^N$. Let $W = w^\ell(L) \in \mathbb{W}^N$ be the corresponding exposures, and let $U = u^\ell(L) \in \{0, 1\}^N$ be the focal indicator vector, for some $w^\ell(\cdot), u^\ell(\cdot)$ defined by the analyst. Define $\mathbb{S}_{A,U} = \mathbb{S}_N(A) \cap \mathbb{S}_N(U)$, which is the permutation subgroup of $\mathbb{S}_N$ that leaves A (the attribute vector) and U (the focal unit vector) unchanged. Suppose that the following conditions hold.*

(a) $P(L) = P(\pi L)$, *for all* $\pi \in \mathbb{S}_{A,U}$ *and* $L \in \mathbb{L}$.

(b) $w^\ell(\cdot)$ *is equivariant with respect to* $\mathbb{S}_{A,U}$.

(c) $u^\ell(\cdot)$ *is equivariant with respect to* $\mathbb{S}_{A,U}$.

Then, W *is uniformly distributed conditional on the event* $\{W \in \mathcal{B}\}$, *where* $\mathcal{B} \in \mathcal{O}(\mathbb{W}^N; \mathbb{S}_{A,U})$.

Theorem 1 formalizes the intuition behind the example in Section 2.5: under the conditions of the theorem, we can implement Procedure 2b by directly permuting the exposures of *only* the focal units, and making sure that these permutations are stratified with respect to the attribute value; the space of these permutations is exactly $\mathbb{S}_{A,U}$. The sharp null of Procedure 1b is a special case of this result by defining $u^\ell(L) = \mathbf{1}_N$, i.e., by selecting all units to be focals. In this special case, $\mathbb{S}_{A,U} = \mathbb{S}_N(A)$ and so we can directly permute the entire exposure vector, W, across units with the same attribute value.

All three conditions in Theorem 1 are intuitive and testable in practice. Condition (a) expresses a *design symmetry* condition. This depends on the experimental design, and will generally be satisfied for a permutation group that is larger than $\mathbb{S}_{A,U}$, such as in the stratified and completely randomized designs we consider in the next section. In particular, the design symmetry condition holds for both our applications. For instance, in Cai and Szeidl (2017), the design is invariant to permutations between firms of the same size and industry in the same subregion (i.e., the attribute A is a vector of length 3); we discuss this condition more in Section 6.

Condition (b) depends on the definition of the exposures, and is part of the analysis rather than the design. This condition posits that, for two units with the same attribute A and focal status U, swapping the group label assignments also swaps their exposures; Condition (b) does not require the exclusion restriction in Assumption 1. Finally, Condition (c) is also under the analyst's control and requires that swapping group label assignments for two units also swaps their selection as focal units.

We note that Theorem 1 is more general than the specific group formation design settings we consider in this paper. In particular, our definition of the exposure function $w^\ell(\cdot)$ in Eq. (3) satisfies Condition (b), and our definition of the focal selection function $u^\ell(\cdot)$ in Eq. (9) satisfies Condition (c). In fact, Condition (c) holds more generally whenever focal selection depends on whether the observed exposure belongs to a predefined set. We summarize these results in the following lemma.

Lemma 1. *Conditions (b) and (c) of Theorem 1 hold under definitions in Eqs.* (3) *and* (9).

Since Conditions (b) and (c) hold in our setting, we will only check the design symmetry in Condition (a) going forward.

As a technical note, Theorem 1 contributes to the existing theory of randomization tests by providing sufficient conditions under which symmetry in distribution of a random variable implies symmetry in distribution to a *function* of that variable. In our context, while the standard theory of randomization tests (Lehmann and Romano, 2005) could be applied on hypotheses in the space of labels (L), it is not directly applicable in the exposure space, $W = w^{\ell}(L)$ because W is not generally invariant to permutations even when L is. The toy example in Section 2.5 illustrated this point through permutations of the exposure vector that were inconsistent with the experimental design. Theorem 1 delivers conditions under which W maintains a permutation symmetry like L. Crucially, the theorem also characterizes the permutation subgroup ($\mathbb{S}_{A,U}$) for which such symmetry propagation is possible.

5 Permutation tests in two group formation designs

We now apply the theory of the previous section in practice. We consider two designs, the stratified randomized design and completely randomized design, and show that these designs have the required symmetries for permutation tests on exposures.

5.1 Stratified randomized design

The stratified randomized design is an important special case of group formation design that satisfies the design symmetry condition in Theorem 1(a). Specifically, we consider designs that, separately for each level of attribute A, assign K group-labels to N units completely at random. In a simplified setting with a binary attribute and two individuals per group, this design randomly assigns one individual of each type to each group.

Definition 4 (Stratified randomized design)**.** *Consider a distribution of group labels, $P(L)$, that assigns equal probability to all vectors L such that for every attribute $a \in \mathbb{A}$ and every group-label $k \in \{1, \ldots, K\}$, the number of units with attribute $A_i = a$ assigned to group-label k is equal to a fixed integer $n_{a,k}$. The design $P(Z)$ induced by such $P(L)$ is called a stratified randomized group formation design, denoted by* $\mathrm{SR}(\mathbf{n}_A)$*, where $\mathbf{n}_A = (n_{a,k})$ satisfies the constraint that $\sum_{k=1}^{K} n_{a,k} = |\{i \in \mathbb{U} : A_i = a\}|$.*

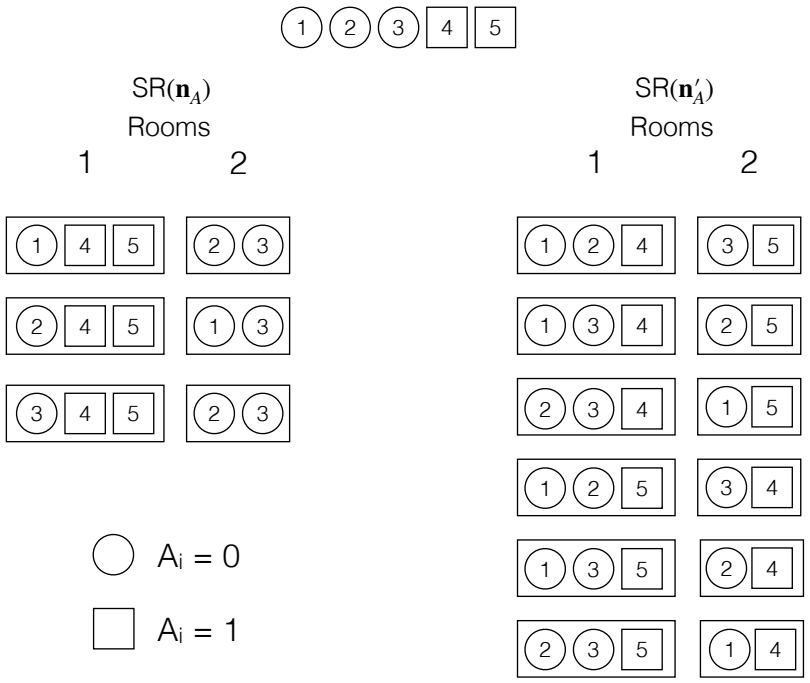

Figure 2: Example of supports for two latent distributions $P(L)$ inducing two stratified randomized experiments. Both examples have $N = 5$ units, $K = 2$ rooms labelled 1 and 2, and a binary attribute. Left: $(n_{0,1}, n_{0,2}) = (1, 2)$ and $(n_{1,1}, n_{1,2}) = (2, 0)$. Right: $(n'_{0,1}, n'_{0,2}) = (2, 1)$ and $(n'_{1,1}, n'_{1,2}) = (1, 1)$.

The stratified randomized design generalizes the design in Li et al. (2019, Section 2.4.2) by allowing the group sizes to vary. As an illustration, Figure 2 shows all possible assignments for two stratified randomized designs in a setting in which we allocate students with a binary attribute to their dorm rooms. The design on the left is SR($\mathbf{n}_A$) with $(n_{0,1}, n_{0,2}) = (1, 2)$, meaning that there is one unit with attribute $A_i = 0$ assigned to room 1, and two to room 2; and $(n_{1,1}, n_{1,2}) = (2, 0)$, meaning that that there are two units with attribute $A_i = 1$ assigned to room 1, and no unit assigned to room 2. The design on the right is SR($\mathbf{n}'_A$) with $(n'_{0,1}, n'_{0,2}) = (2, 1)$ and $(n'_{1,1}, n'_{1,2}) = (1, 1)$.

As stated in Lemma 2 below, the stratified randomized design satisfies the design symmetry condition in Theorem 1 since the number of units assigned to any attribute-label pair remains fixed under any permutation of the labels that stratifies on A.

Lemma 2. *Definition 4 satisfies Condition (a) in Theorem 1.*

Our recommended procedure for testing the sharp null under a stratified design is as follows:

Procedure 1c (Sharp null under the stratified randomized design)**.** *Consider observed assignment* $Z^{\text{obs}} \sim \text{SR}(\mathbf{n}_A)$ *and corresponding exposure* W^{obs}*.*

1. *Observe outcomes,* $Y^{\text{obs}} = Y^{\omega}(W^{\text{obs}})$*.*
2. *Compute* $T^{\text{obs}} = T^{\omega}(W^{\text{obs}}; Y^{\text{obs}})$*.*
3. *For* $r = 1, \ldots, R$*, obtain* $W^{(r)}$ *via a random permutation of* W^{obs}*, stratifying on the attribute* A*, and then compute* $T^{(r)} = T^{\omega}(W^{(r)}; Y^{\text{obs}})$*.*
4. *Compute the approximate p-value* $\text{pval}(W^{\text{obs}}) = R^{-1} \sum_{r=1}^{R} \mathbb{1}(T^{(r)} \geq T^{\text{obs}})$*.*

In Step 3 above, we randomly permute W^{obs} stratifying on attribute A, that is, we randomly permute within each subvector of W^{obs} corresponding to a given value of A. This procedure is identical to how one would analyze a stratified completely randomized multi-arm trial in the non-interference setting — with the exposure vector W^{obs} being the analog to the treatment vector in that case (Imbens and Rubin, 2015, Chapter 9). That is, given the data $(Y_i, W_i, A_i)_{i=1}^{N}$, the analyst simply perform a complete randomization test stratified on A.

The analogy with the traditional setting extends to testing the non-sharp nulls introduced in Section 3.2, with only minor modifications. Recall that for Procedure 2c, the test statistics are restricted to focal units, i.e., $T(z; Y, U)$. Our recommended procedure for testing non-sharp nulls under a stratified design is then:

Procedure 2c (Non-sharp nulls under the stratified randomized design)**.** *Consider observed assignment* $Z^{\text{obs}} \sim \text{SR}(\mathbf{n}_A)$ *and corresponding exposure* W^{obs}*.*

1. *Observe outcomes,* $Y^{\text{obs}} = Y^{\omega}(W^{\text{obs}})$*.*
2. *Let* $U^{\text{obs}} = u(Z^{\text{obs}})$ *be the focal unit selection as in* (9)*.*
3. *Compute* $T^{\text{obs}} = T^{\omega}(W^{\text{obs}}; Y^{\text{obs}}, U^{\text{obs}})$*.*
4. *For* $r = 1, \ldots, R$*, obtain* $W^{(r)}$ *via a random permutation of* W^{obs}*, restricted only to focal units* ($U_i^{\text{obs}} = 1$) *and stratifying on the attribute* A*. Compute* $T^{(r)} = T^{\omega}(W^{(r)}; Y^{\text{obs}}, U^{\text{obs}})$*.*
5. *Compute the approximate p-value* $\text{pval}(W^{\text{obs}}) = R^{-1} \sum_{r=1}^{R} \mathbb{1}(T^{(r)} \geq T^{\text{obs}})$*.*

Although less obvious than in the case of Procedure 1c, Procedure 2c also connects to traditional randomization tests. Given the data $(Y_i, W_i, A_i)_{i=1}^{N}$, the analyst first subsets the array to contain only focal units ($U_i^{\text{obs}} = 1$), and then simply performs a stratified complete randomization test on this reduced data, stratifying on A. Interestingly, there is a gap

in the literature for randomization tests for non-sharp null hypotheses, even in traditional stratified randomized experiments without peer effects. Our permutation test applies to the traditional setting as well. Finally, we note that both Procedures 1c and 2c are finite-sample exact with a direct application of Theorem 1. See Appendix D.3 for details.

5.2 Completely randomized design

Another common design is the completely randomized design, which fixes the *overall* number of units that receive each group-label, without stratifying on the attribute. Despite this difference, we will show that the completely randomized design can be analyzed exactly like a stratified randomized design by conditioning on the observed attribute-group assignments.

Definition 5 (Completely randomized design)**.** *Consider a distribution of group labels, $P(L)$, that assigns equal probability to all vectors L such that for every group-label $k \in \{1, \ldots, K\}$, the number of units assigned to group-label k is equal to a fixed integer n_k. The design $P(Z)$ induced by such $P(L)$ is a completely randomized group formation design, denoted by* $\mathrm{CR}(\mathbf{n})$*, where $\mathbf{n} = (n_1, \ldots, n_K)$ satisfies $\sum_{k=1}^{K} n_k = N$.*

Lemma 3. *Definition 5 satisfies Condition (a) in Theorem 1.*

The completely randomized design generalizes the design in Li et al. (2019, Section 2.4.1) by allowing the size of the groups to vary. Importantly, we can construct a stratified randomized design from a completely randomized design by conditioning on the number of units with each level of the attribute in each group. As a result, conditional on $\mathbf{n}_A$, we can analyze a completely randomized group formation design exactly like a stratified randomized design.

Corollary 1. *Consider $P(Z) \sim \mathrm{CR}(\mathbf{n})$. The null hypotheses H_0 (resp. $H_0^{\mathsf{w}_1,\mathsf{w}_2}$) can be tested with Procedure 1c (resp. Procedure 2c) as if the design were* $\mathrm{SR}(\mathbf{n}_A)$*, where $\mathbf{n}_A$ is the observed number of units with each value of the attribute A assigned to each group.*

This connection is important since many designs are not stratified on the attribute of interest; e.g., the application we analyze in Section 6.1 uses a completely randomized design rather than a stratified randomization design. Importantly, conditioning on $\mathbf{n}_A$ is necessary to ensure the validity of the permutation test even in completely randomized designs. Figure 1 gives an example in which the unconditional permutation test is invalid.

Remark 1 (Incorporating additional covariates)**.** *All our procedures can be extended to incorporate additional covariates in the design and analysis stages. These strategies will*

generally increase the power of the test, so long as covariates are predictive of the potential outcomes (Zhao and Ding, 2021). Most immediately, we could stratify both the permutations and the test statistic by an additional discrete covariate. We could also consider regression-adjusted test statistics, rather than test statistics based on the raw outcomes (Rosenbaum, 2002). We could further tailor these models to a particular interference structure; for instance, Athey et al. (2018) propose a test statistic derived from the linear-in-means model. Importantly, this approach does not *assume that the linear-in-means model is correct, but rather that this parameterization captures departures from the null hypothesis.*

6 Applications

We illustrate our approach by re-analyzing two randomized group formation experiments. The first application is from Li et al. (2019), who assess the impact of randomly assigned roommates on student GPA. Our conditional testing approach yields results that are consistent with their randomization-based estimate. The second application is from Cai and Szeidl (2017), who conduct a randomized experiment to estimate the effect of social connections on firm performance. Our approach complements the results from their regression-based estimates by uncovering interesting heterogeneity in the peer group effect.

6.1 Random roommate assignment

Li et al. (2019) explore the impact of the composition of randomly assigned roommates on student academic performance among students at a top Chinese university. For ease of exposition, we restrict our analysis to the $N = 156$ male students admitted to the Department of Informatics, the largest department in the original study. The attribute of interest is whether students are admitted via a college entrance exam ($A_i = 1$), known as *Gaokao*, or via an external recommendation ($A_i = 0$). Students are assigned to dorm rooms of size four via complete randomization, as described in Section 5.2; that is, the number of students of each background in each room is a random quantity.

The exposure of interest is the number of roommates admitted via the entrance exam $w_i(Z) = \sum_{j \in Z_i} A_j$. We focus on the null hypothesis $H_0^{0,3} : Y_i^\omega(0) = Y_i^\omega(3)$ for all $i = 1, \ldots, N = 156$, that is, a student's end-of-year GPA is the same if he is randomly assigned to have zero *Gaokao* roommates versus three *Gaokao* roommates. Moreover, following Li et al. (2019), we want to test this null hypothesis separately for *Gaokao* and recommendation students,

Table 1: p-values, difference-in-means point estimates and 95% confidence intervals for the application of Li et al. (2019).

	p-value	estimate	confidence interval
$H_0^{0,3}$	0.04	-0.31	$(-0.67, -0.02)$
$H_0^{0,3}(0)$	0.02	-0.37	$(-0.73, -0.05)$
$H_0^{0,3}(1)$	0.23	-0.28	$(-0.81, 0.12)$

which we denote $H_0^{0,3}(1)$ and $H_0^{0,3}(0)$ respectively. Here, Assumption 1 states that group formation only affects end-of-year GPA by changing the number of *Gaokao* roommates for a student. This excludes, for example, the subject area or sociability of roommates as important mechanisms for group peer effects. Among 17 students from *Gaokao*, 13 have observed exposure $W_i^{\text{obs}} = 0$ and 4 have observed exposure $W_i^{\text{obs}} = 3$; among 45 students from recommendation, 40 have observed exposure $W_i^{\text{obs}} = 0$ and 5 have observed exposure $W_i^{\text{obs}} = 3$. Table 1 reports the p-value using a difference-in-means test statistic, and the corresponding inverted confidence intervals for the overall null hypothesis $H_0^{0,3}$ and the subgroup null hypotheses $H_0^{0,3}(1)$ and $H_0^{0,3}(0)$.

Our results are substantively close to those obtained by Li et al. (2019). First, our point estimates are identical to those from Li et al. (2019), since both are based on a difference in means. Our p-values and confidence intervals are also similar, with the exception of $H_0^{0,3}(1)$, the separate null hypothesis on *Gaokao* students. For this, Li et al. (2019) find a p-value ≤ 0.05, while we cannot reject that null hypothesis. One possible explanation for this discrepancy is that, while our p-values are exact, Li et al. (2019) instead use an asymptotic approximation, which may be unwarranted given the small sample size. We investigate this more in Appendix C.1, where we conduct simulation calibrated on this application and show that normal asymptotics can fail severely.

6.2 Meeting groups among firm managers

We now turn to the study from Cai and Szeidl (2017), in which CEOs of Chinese firms were randomly assigned to meetings where they discussed management practices, with ten managers per group. Groups were encouraged to meet monthly for roughly a year; firms assigned to control did not meet. The primary outcome of interest is growth in firm sales, defined as the difference in (log) firm sales from endline to baseline.[5]

[5] Cai and Szeidl (2017) collected survey data at baseline, midline, and endline. While the authors analyzed the experiment using panel data regression, we side-step the panel structure here by defining the outcome

Cai and Szeidl (2017) focused on the impact of assigning CEOs to meeting groups versus a business-as-usual control group. Here we revisit a secondary analysis in their paper that explores the role of peer composition. In particular, among treated firms, the group formation design was stratified across three attributes: firm sector (manufacturing/service), location (26 subregions), and firm size (small/large).[6] Using this design, Cai and Szeidl (2017) "ask whether firms randomized into groups with larger peers grew faster," finding evidence in the affirmative.

We revisit this question using our proposed randomization inference framework, where firm size is the exposure of interest. In particular, we focus on the 1,323 firms with non-missing data (on size and revenue) that were randomly assigned to meetings. We first consider the global sharp null of any effect of peer size on sales, and then highlight a source of peer effect heterogeneity by testing the sharp null within subgroups defined by sector and size. In the Appendix, we also consider alternative exposure definitions and look at pairwise, non-sharp null hypotheses to further explore this source of heterogeneity.

Global sharp null hypothesis. We start with the global sharp null hypothesis that there is no effect whatsoever of peer size on sales. The exposure of interest is $W_i = \frac{1}{|Z_i|}\sum_{j\in Z_i} \text{size}_j$, where Z_i is the set of peer firms for firm i, and size_j is the log-number of employees in firm j at baseline. Let $\mathbb{W} \subset \mathbb{R}$ be the exposure domain, then the global sharp null hypothesis is:

$$H_0 : Y_i^\omega(\mathsf{w}) = Y_i^\omega(\mathsf{w}') \text{ for all } i \in \mathbb{U} \text{ and } \mathsf{w}, \mathsf{w}' \in \mathbb{W}. \tag{11}$$

That is, under H_0, the average employee size of firm i's peer group does not affect the firm's revenue. As we discuss in Section 2.3, Assumption 1 plays a critical role in interpreting a rejection of our null hypothesis. In this application, Assumption 1 states that group formation only affects sales by changing the size of a firm's peer companies. This excludes, for example, the number of other peer firms' *clients* (rather than number of employees) from affecting a firm's own revenue. To check robustness, we explore alternative definitions of the exposure in Appendix B.

To mirror the analysis in Cai and Szeidl (2017), we set the test statistic to be the coefficient

as the difference in log firm sales between endline and baseline. We note, however, that our framework accommodates a wide range of outcomes and test statistics, including those generated by panel regressions.

[6] Firm size is dichotomized at median employment of the sample of firms in the corresponding subregion, where the authors use the number of employees at baseline as a proxy for the quality of the firm.

of W in the following linear regression:[7]

$$Y_i^{\text{obs}} = \alpha + \beta A_i^* + \tau W_i + \varepsilon_i, \tag{12}$$

where $A_i^* = \text{sector}_i \times \text{location}_i \times \text{size}_i$ includes all interactions between firm sector, location, and size for unit i. We can now employ Procedure 1b to test H_0, computing a one-sided p-value of $p = 0.02$ over 20,000 replications. Importantly, even if the linear model in Equation (12) is not correctly specified, the randomization test remains finite-sample valid.

Heterogeneity by firm size and type. Since our approach is exact in finite-samples, we can easily restrict our analysis to subsets of firms, here defined by sector and size following Cai and Szeidl (2017). We repeat Procedure 1b separately within each subgroup, using the estimated coefficient τ from Equation (12), except with the levels of A^* restricted to the appropriate subgroup. The results in Figure 3 show substantial heterogeneity in peer group effects. In particular, the signal is concentrated entirely among small service firms ($p = 0.0015$), and is essentially zero for the other three subgroups.

Cai and Szeidl (2017) also explored heterogeneity, albeit only in the "direct effects" from treatment (i.e., meetings versus no meetings) rather than in peer effects; they find larger firms benefited more from the meetings. Our analysis complements this picture by showing that the impact of larger peers was concentrated mainly among small service firms. We emphasize that the regression specification of Cai and Szeidl (2017) in (11) cannot easily capture the heterogeneity we show here. In particular, their regression model needs to include all size-sector-subregion interactions (∼85 in total) dictated by the experimental design in order to identify τ (see Section III.B in Cai and Szeidl, 2017). These interactions, however, essentially "wash out" the size-sector interaction effect we observe here. Thus our randomization-based analysis complements the regression-based analyses and offers new insights. Finally, an additional benefit of our analysis is that our p-values are exact, which is especially important for subgroups. In Appendix C.2, we highlight this through a simulation study showing that regression-based tests can be severely distorted in simple but realistic group formation designs motivated by Cai and Szeidl (2017).

[7] To aid interpretation, we follow the regression specification in Cai and Szeidl (2017). However, the randomization inference theory from Li et al. (2019) shows that a regression specification that includes the interaction of A and W is also justified by the randomization itself.

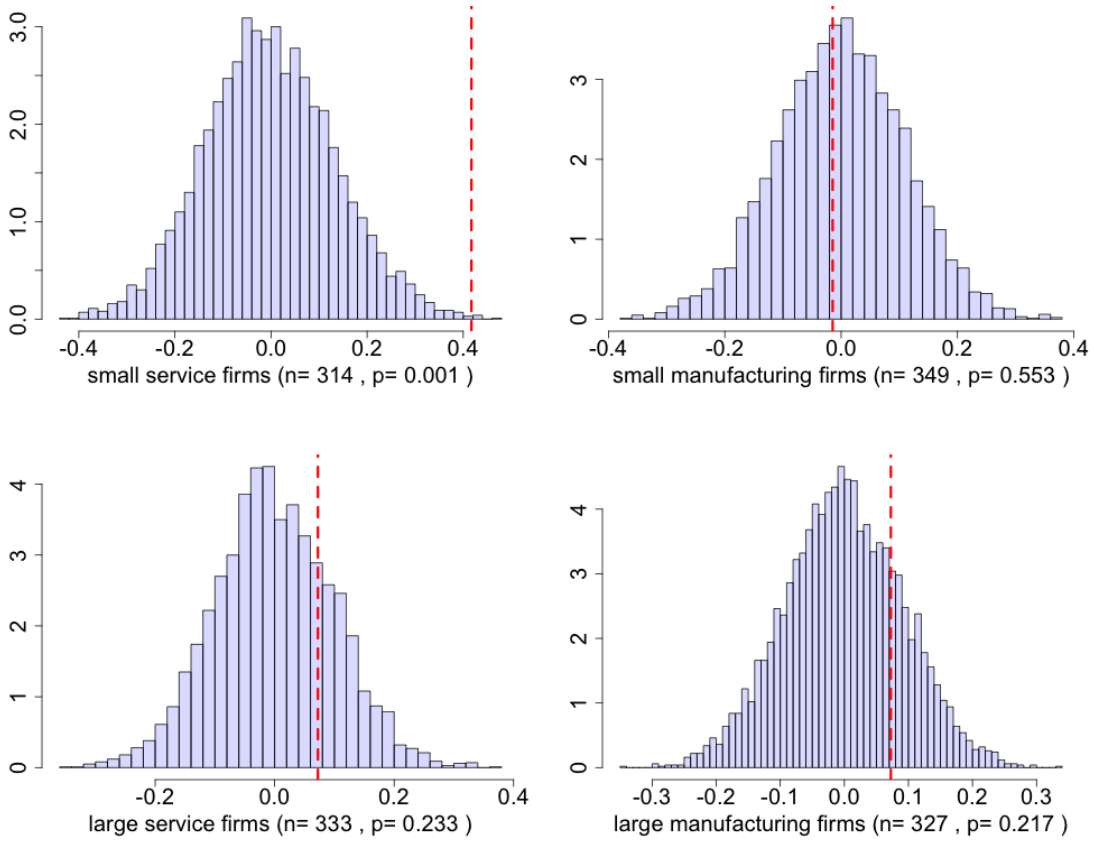

Figure 3: Randomization distributions for H_0 in (11) within firm subpopulations according to size and sector. The dashed lines indicate observed values of the test statistic; 'n' is the subpopulation size, and 'p' is the one-sided p-value calculated from Procedure 1b.

7 Discussion

We have proposed valid randomization tests for testing peer effects in group formation experiments. While a promising first step, there remain several open questions. First, our results motivate new considerations for the design of group formation experiments. In particular, arbitrary designs do not necessarily satisfy the sufficient conditions we propose for valid permutation tests. We therefore recommend using the experimental designs like the stratified and completely randomized designs in Section 5 if researchers want to use our permutation-based tests.

Second, sometimes the group structures may be more elaborate than what we have studied in this paper. For example, we might assign students to classrooms and then separately assign

teachers to those classrooms. Alternatively, there may be multiple, possibly overlapping groups; e.g., students nested within classrooms nested within schools. Finally, randomizing peers may often be infeasible or raise ethical concerns. Thus, extending the ideas in this paper to the observational study setting, especially for sensitivity analysis, is a promising avenue for future work.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica 88*, 265–296.

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics 138*, 1–35.

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics 30*, 98–108.

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research 41*, 3–16.

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

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

Basse, G. W., A. Feller, and P. Toulis (2019). Randomization tests of causal effects under interference. *Biometrika 106*, 487–494.

Bhattacharya, D. (2009). Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association 104*, 486–500.

Bramoullé, Y., H. Djebbari, and B. Fortin (2020). Peer effects in networks: A survey. *Annual Review of Economics 12*, 603–629.

Brock, W. A. and S. N. Durlauf (2001). Interactions-based models. In *Handbook of econometrics*, Volume 5, pp. 3297–3380. Elsevier.

Cai, J. and A. Szeidl (2017). Interfirm relationships and business performance. *The Quarterly Journal of Economics 133*, 1229–1282.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica 85*, 1013–1030.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica 81*, 855–882.

Cornelissen, T., C. Dustmann, and U. Schönberg (2017). Peer effects in the workplace. *American Economic Review 107*, 425–56.

Ding, P. and T. Dasgupta (2018). A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika 105*, 45–56.

Duncan, G. J., J. Boisjoly, M. Kremer, D. M. Levy, and J. Eccles (2005). Peer effects in drug use and sex among college students. *Journal of Abnormal Child Psychology 33*, 375–385.

Fafchamps, M. and S. Quinn (2018). Networks and manufacturing firms in Africa: Results from a randomized field experiment. *The World Bank Economic Review 32*, 656–675.

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

Frandsen, B., L. Lefgren, and E. Leslie (2023). Judging judge fixed effects. *American Economic Review 113*, 253–77.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*, 253–264.

Guryan, J., K. Kroft, and M. J. Notowidigdo (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics 1*, 34–68.

Hennessy, J., T. Dasgupta, L. Miratrix, C. Pattanayak, and P. Sarkar (2016). A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference 4*, 61–80.

Herbst, D. and A. Mas (2015). Peer effects on worker output in the laboratory generalize to the field. *Science 350*, 545–549.

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

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

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. New York: Springer.

Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica 90*(1), 267–293.

Li, X., P. Ding, Q. Lin, D. Yang, and J. S. Liu (2019). Randomization inference for peer effects. *Journal of the American Statistical Association 114*, 1651–1664.

Lyle, D. S. (2009). The effects of peer group heterogeneity on the production of human capital at west point. *American Economic Journal: Applied Economics 1*, 69–84.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies 60*, 531–542.

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

Puelz, D., G. Basse, A. Feller, and P. Toulis (2022). A graph-theoretic approach to randomization tests of causal effects under general interference. *Journal of the Royal Statistical Society Series B 84*, 174–204.

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

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

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics 116*, 681–704.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*, Volume 3, pp. 249–277. Elsevier.

Sacerdote, B. (2014). Experimental and quasi-experimental analysis of peer effects: two steps forward? *Annual Reviews of Economics 6*, 253–272.

Sävje, F. (2023). Causal inference with misspecified exposure mappings. *Biometrika*.

Toulis, P. and E. Kao (2013). Estimation of causal peer influence effects. In *International Conference on Machine Learning*, pp. 1489–1497.

Wu, J. and P. Ding (2020). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association 116*, 1898–1913.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics 134*(2), 557–598.

Zhao, A. and P. Ding (2021). Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics 225*, 278–294.