Identification and Estimation of Spillover Effects in Randomized Experiments*

Gonzalo Vazquez-Bare†

November 5, 2020

Abstract

I study identification, estimation and inference for spillover effects in experiments where units' outcomes may depend on the treatment assignments of other units within a group. I show that the commonly-used linear-in-means (LIM) regression identifies a weighted sum of spillover effects with some negative weights, and that the difference in means between treated and controls identifies a combination of direct and spillover effects entering with different signs. I propose nonparametric estimators for average direct and spillover effects that overcome these issues and are consistent and asymptotically normal under a precise relationship between the number of parameters of interest, the total sample size and the treatment assignment mechanism. These findings are illustrated using data from a conditional cash transfer program and with simulations. The empirical results reveal the potential pitfalls of failing to flexibly account for spillover effects in policy evaluation: the estimated difference in means and the LIM coefficients are all close to zero and statistically insignificant, whereas the nonparametric estimators I propose reveal large, nonlinear and significant spillover effects. (*JEL* C10, C13, C14, C90)

*I am deeply grateful to Matias Cattaneo for continued advice and support. I am indebted to Lutz Kilian, Mel Stephens and Rocío Titiunik for thoughtful feedback and discussions that greatly improved the paper. I thank Clément de Chaisemartin, Catalina Franco, Amelia Hawkins, Nicolás Idrobo, Xinwei Ma, Nicolás Morales, Kenichi Nagasawa, Olga Namen and Doug Steigerwald for valuable discussions and suggestions. I also thank seminar participants at UChicago, UCLA, UC San Diego, University of Michigan, UC Santa Barbara, UChicago Harris School of Public Policy, Cornell University, UChicago Booth School of Business, UT Austin, Stanford University and UC Berkeley for helpful comments.

†Department of Economics, University of California, Santa Barbara. gvazquez@econ.ucsb.edu.

1 Introduction

Spillover effects, which occur when an agent's actions or behaviors indirectly affect other agents' outcomes through peer effects, social interactions or externalities, are ubiquitous in economics and social sciences. A thorough account of spillover effects is crucial to assess the causal impact of policies and programs (Abadie and Cattaneo, 2018; Athey and Imbens, 2017). However, the literature is still evolving in this area, and most of the available methods for analyzing treatment effects either assume no spillovers or allow for them in restrictive ways, sometimes without a precise definition of the parameters of interest or the conditions required to recover them.

This paper studies identification, estimation and inference for average direct and spillover effects in randomized controlled trials, and offers three main contributions. First, I provide conditions for nonparametric identification of causal parameters when the true spillovers structure is possibly unknown. To this end, Section 2 sets up a potential-outcomes framework that nests several models commonly used to analyze spillovers. Under the assumption that interference can occur within (but not between) the groups in which units are clustered, I define a rich set of direct and spillover treatment effects based on a function, the *treatment rule*, that maps peers' treatment assignments and outcomes. Lemma 1 in Section 3 links average potential outcomes to averages of observed variables when the posited treatment rule is possibly misspecified.

The second contribution of this paper is to characterize the difference in means between treated and controls and the coefficients from a linear-in-means (LIM) model, two of the most commonly analyzed estimands when analyzing RCTs and spillover effects in general. Theorem 1 shows that, in the presence of spillovers, the difference in means between treated and controls combines the direct effect of the treatment and the difference in spillover effects for treated and untreated units, and thus the sign of the difference in means is undetermined even when the signs of all direct and spillover effects are known. On the other hand, Theorem 2 shows that a LIM regression recovers a linear combination of spillover effects for different numbers of treated peers where the weights sum to zero, and hence some weights are necessarily negative. As a result, the coefficients from a LIM regression can be zero even when all the spillover effects are non-zero. I then provide sufficient conditions under which the difference in means and the LIM coefficients have a causal interpretation. I also propose a regression-based pooling strategy that is robust to nonlinearities and heterogeneity in spillover effects.

Finally, the third contribution is to analyze nonparametric estimation and inference for spillover effects. In the presence of spillovers, estimation faces two main challenges: the number of treatment effects to estimate can be large, and the probability of observing units under different treatment assignments can be small. Section 4 provides general conditions that ensure uniform consistency and asymptotic normality of the estimators of interest with special focus on the role of group size on estimation and inference. This approach formalizes

the requirement of "many small groups" that is commonly invoked in the literature, and specifies the role that the number of parameters and the assignment mechanism play on the asymptotic properties of nonparametric estimators. More precisely, consistency and asymptotic normality are shown under two main conditions that are formalized in the paper: (i) the number of treatment effects should not be "too large" with respect to the sample size, and (ii) the probability of each treatment assignment should not be "too small". These two requirements are directly linked to modeling assumptions on the potential outcomes, the choice of the set of parameters of interest and the treatment assignment mechanism. As an alternative approach to inference based on the normal approximation, the wild bootstrap is shown to be consistent, and simulation evidence suggests that it can yield better performance compared to the Gaussian approximation for moderately large groups.

The results in this paper are illustrated in a simulation study and using data from a randomized conditional cash transfer. The empirical results clearly highlight the pitfalls of failing to flexibly account for spillovers in policy evaluation: the estimated difference in means and the estimated LIM coefficients are all close to zero and statistically insignificant, whereas the nonparametric estimators I propose reveal large, nonlinear and significant and spillover effects.

This paper is related to the longstanding literature on peer effects and social interactions (see e.g. Moffit, 2001; Duflo and Saez, 2003; Miguel and Kremer, 2004; Blume, Brock, Durlauf, and Jayaraman, 2015; Kline and Tamer, 2019). In a large fraction of this literature, identification and estimation rely on parametric methods or specific experimental designs. My paper contributes to this literature by offering a nonparametric framework that is independent of the experimental design, and can therefore be employed to better understand the estimands commonly analyzed in the literature from a potential-outcomes perspective.

In particular, linear-in-means models, in which the outcome of interest is regressed against own and peers' average characteristics, have been a workhorse model for estimating peer effects in many areas of economics (see e.g. Manski, 1993; Bramoullé, Djebbari, and Fortin, 2009; Davezies, D'Haultfoeuille, and Fougère, 2009). Even in randomized contexts, identification in LIM models relies on the linearity assumption imposed on the structure of spillover effects. The parametric assumptions in the LIM models have been criticized for the unrealistic restrictions that they impose on the structure of peer effects (see Sacerdote, 2014, for a review). While some empirical specifications have attempted to relax parametric assumptions, these models have only been analyzed from a linear regression perspective. As such, the identified parameters can be interpreted as best linear predictors, but their causal interpretation remains unclear, and Angrist (2014) has criticized the usefulness of LIM models to recover causal effects. My paper contributes to this strand of the literature by shedding light on the link between LIM coefficients and potential outcomes, and providing sufficient conditions for the LIM parameters to have a causal interpretation.

Manski (2013) studies partial identification under different restrictions on the structural model, the response functions and the structure of social interactions. The relationship

between reduced-form and structural response functions is discussed in Section A1 of the supplemental appendix. My paper complements this important strand of the literature by offering point identification, estimation and inference results for well-defined, reduced-form causal treatment effects in the presence of spillovers.

Finally, Duflo and Saez (2003) and Hirano and Hahn (2010) analyze the design of experiments to measure social interactions. Section A2 in the supplemental appendix discusses some implications of my inference results for experimental design.

The remainder of the paper is organized as follows. Section 2 describes the setup and defines the parameters of interest. Section 3 provides the main identification results. Section 4 analyzes estimation and inference. Section 5 provides a simulation study, and Section 6 contains the empirical application. Section 7 concludes. The proofs, together with additional results and discussions, are provided in the supplemental appendix.

2 Setup

As a motivating example, consider a program in which parents in low-income households receive a cash transfer from the government as long as their children are enrolled in school and reach a required level of attendance. Suppose that this conditional cash transfer program is evaluated using a randomized pilot in which children are randomly selected to participate. There are several reasons to expect within-household spillovers from this program. On the one hand, the cash transfer may alleviate a financial constraint that was preventing the parents from sending their children to school on a regular basis. The program could also help raise awareness on the importance of school attendance, encouraging parents to worry more about sending their children to school. In both these cases, untreated children may indirectly benefit from the program when they have a treated sibling. On the other hand, the program could create incentives for the parents to reallocate resources towards their treated children and away from their untreated siblings, decreasing school attendance for the latter. In all cases, ignoring spillover effects can severely underestimate the costs and benefits of this policy.

Moreover, these alternative scenarios have drastically different implications on how to assign the program when scaling it up. In the first two situations, treating one child per household can be a cost-effective way to assign the treatment, whereas in the second case, treating all the children in a household can be more beneficial.

2.1 Notation and parameters of interest

In light of the motivating example, consider a random sample of groups indexed by $g = 1, \ldots, G$, each with $n_g + 1$ units, so that each unit i in group g has n_g neighbors or peers. I assume group membership is observable. Units in each group are assigned a binary treatment, and a unit's potential outcomes, defined below, can depend on the assignment of all other

units in the same group. I refer to this phenomenon as *interference*, and to the effect of a neighbor's treatment assignment on unit i's potential outcome as *spillover effect*. Interference is assumed to occur between units in the same group, but not between units in different groups.

The individual treatment assignment of unit i in group g is denoted by D_{ig}, taking values $d \in \{0, 1\}$, and the vector of treatment assignments in each group is given by $\mathbf{D}_g = (D_{1g}, \ldots, D_{n_g+1g})$. For each unit i, D_{jig} is the treatment indicator corresponding to unit i's j-th neighbor, collected in the vector $\mathbf{D}_{(i)g} = (D_{1ig}, D_{2ig}, \ldots, D_{n_gig})$. This vector takes values $\mathbf{d}_g = (d_1, d_2, \ldots, d_{n_g}) \in \mathcal{D} \subseteq \{0, 1\}^{n_g}$.

A key element in this setup will be a function $h_0(\cdot)$ that summarizes how the vector $\mathbf{d}_g$ enters the potential outcome. More precisely, define a function or *treatment rule*:

$$h_0 : \mathcal{D} \to \mathcal{H}_0$$

that maps $\mathbf{d}_g$ into some value $h_0(\mathbf{d}_g)$ of the same or smaller dimension, so that $\dim(\mathcal{H}_0) \leq \dim(\mathcal{D})$. Following Manski (2013)'s terminology, for $h_0(\mathbf{d}_g) = \mathbf{h}_0$, I will refer to the tuple $(d, \mathbf{h}_0)$ as the *effective treatment assignment*, an element in the set $\{0, 1\} \times \mathcal{H}_0$. The potential outcome for unit i in group g depends on the treatment assignment $(d, \mathbf{d}_g)$ through this function $h_0(\cdot)$ and is denoted by the random variable $Y_{ig}(d, \mathbf{h}_0)$ where $\mathbf{h}_0 = h_0(\mathbf{d}_g) \in \mathcal{H}^0$.

Example 1 (SUTVA) If $h_0(\cdot)$ is a constant function, the vector of peers' assignments is ignored and the set of effective treatment assignments becomes $\{0, 1\}$, so the potential outcomes do not depend on peers' assignments. In this case the only effective treatment assignments are $D_{ig} = 1$ and $D_{ig} = 0$ (treated and control). This assumption is often known as the stable unit treatment value assumption or SUTVA (Imbens and Rubin, 2015). □

Example 2 (Exchangeability) When potential outcomes depend on how many peers, but not which ones, are treated, peers are said to be *exchangeable*. Exchangeability can be modeled by setting $h_0(\mathbf{d}_g) = \mathbf{1}_g'\mathbf{d}_g$ so $h_0(\cdot)$ summarizes $\mathbf{d}_g$ through the sum of its elements (i.e. the number of peers assigned to treatment). The set of effective treatment assignments in this case is given by $\{(d, s) : d = 0, 1, s = 0, 1, \ldots, n_g\}$. Exchangeability may be a natural starting point when there is no clear way (or not enough information) to assign identities to peers. □

Example 3 (Stratified exchangeability) Exchangeability may also be imposed by subgroups. For instance, the vector of assignments may be summarized by the number of male and female treated peers separately. In settings where units are geographically located, peers are commonly assumed to be exchangeable within groups defined by distance such as within one block, between one and two blocks, etc, or by different distance radiuses (e.g. within 100 meters, between 100 and 200 meters and so on). □

Example 4 (Reference groups) When each unit interacts only with a strict subset of her peers, we can define for example $h_0(\cdot) : \{0,1\}^{n_g} \to \{0,1\}^{k_g}$ where $k_g < n_g$. For instance, under the assumption that each unit interacts with her two closest neighbors, $h_0(\mathbf{d}_g) = (d_1, d_2)$ so that $k_g = 2$. The subset of peers with which each unit interacts is known as the *reference group* (Manski, 2013). □

Example 5 (Non-exchangeable peers) The case in which $h_0(\mathbf{d}_g) = \mathbf{d}_g$ does not imply any restrictions on the structure of the potential outcomes, and therefore does not provide any dimensionality reduction. This level of generality requires the researcher to be able to determine an ordering between peers to determine who is unit i's nearest neighbor, second nearest neighbor and so on. This ordering can be based for example on geographic distance, a spatial weights matrix as used in spatial econometrics, frequency of interaction in a social network, etc. □

In what follows, $\mathbf{0}_g$ and $\mathbf{1}_g$ will denote n_g-dimensional vectors of zeros and ones, respectively. Throughout the paper, I will assume that all the required moments of the potential outcomes are bounded. Unit-level direct effects are defined as differences in potential outcomes switching own treatment assignment for a fixed peer assignment $\mathbf{h}_0$, $Y_{ig}(1, \mathbf{h}_0) - Y_{ig}(0, \mathbf{h}_0)$. Unit-level spillover effects are defined as differences in potential outcomes switching peer assignments for a fixed own assignment d, $Y_{ig}(d, \mathbf{h}_0) - Y_{ig}(d, \tilde{\mathbf{h}}_0)$.

Given a vector of observed assignments $(D_{ig}, \mathbf{D}_{(i)g})$, the observed outcome is given by $Y_{ig}(D_{ig}, h_0(\mathbf{D}_{(i)g}))$ and can be rewritten as:

$$Y_{ig} = \sum_{d \in \{0,1\}} \sum_{\mathbf{h}_0 \in \mathcal{H}^0} Y_{ig}(d, \mathbf{h}_0) \mathbb{1}(D_{ig} = d) \mathbb{1}(h_0(\mathbf{D}_{(i)g}) = \mathbf{h}_0).$$

To fix ideas, consider a household with three children, $n_g + 1 = 3$. In this household, each kid has two siblings, with assignments d_1 and d_2, so $\mathbf{d}_g = (d_1, d_2)$. If the true treatment rule $h_0(\cdot)$ is the identity function, the potential outcome has the form $Y_{ig}(d, d_1, d_2)$. This treatment rule does not restrict the potential outcomes, and hence each unit can have up to $2^{(n_g+1)} = 8$ different potential outcomes. In this case, $Y_{ig}(1,0,0) - Y_{ig}(0,0,0)$ is the effect of the treatment when both of unit i's siblings are untreated, $Y_{ig}(0,1,0) - Y_{ig}(0,0,0)$ is the spillover effect on unit i of treating unit i's first sibling, and so on. The average effect of assignment (d, d_1, d_2) compared to $(\tilde{d}, \tilde{d}_1, \tilde{d}_2)$ is thus given by $\mathbb{E}[Y_{ig}(d, d_1, d_2)] - \mathbb{E}[Y_{ig}(\tilde{d}, \tilde{d}_1, \tilde{d}_2)]$. On the other hand, under an exchangeable treatment rule, the potential outcome can be written as $Y_{ig}(d, s)$ where $s = 0, 1, 2$ is the number of treated siblings. The total number of different potential outcomes is $2(n_g + 1) = 6$, so exchangeability reduces the dimensionality of the effective assignments set from exponential to linear in group size.

I will assume perfect compliance, which means that all units receive the treatment they are assigned to. I analyze the case of imperfect compliance in Vazquez-Bare (2020). The following assumption establishes the relationship between the treatment and the potential outcomes.

Assumption 1 (Random assignment) *The potential outcomes* $\{Y_{ig}(d, \mathbf{h}_0), d \in \{0,1\}, \mathbf{h}_0 \in \mathcal{H}_0\}_{i=1}^{n_g}$ *and the treatment assignments* D_{ig} *are identically distributed across* i *and* g*, and for each* g*,*

$$\{Y_{ig}(d, \mathbf{h}_0), d \in \{0,1\}, \mathbf{h}_0 \in \mathcal{H}_0\}_{i=1}^{n_g} \perp\!\!\!\perp \mathbf{D}_g.$$

The first part of Assumption 1 states that potential outcomes and the treatment assignments are identically distributed, and thus moments are not indexed by either i or g. The second part states that the treatment is randomly assigned within each group.

In practice, the true $h_0(\cdot)$ is usually unknown, and the researcher needs to posit a candidate $h(\cdot)$ that may or may not coincide with $h_0(\cdot)$. Given the lack of knowledge on the true assignment, a function $h(\cdot)$ that imposes fewer restrictions on the potential outcomes has a lower risk of misspecification. To formalize this idea, I introduce the following definition.

Definition 1 (Coarseness) *Given two treatment rules* $h(\cdot) : \mathcal{D} \to \mathcal{H}$ *and* $\tilde{h}(\cdot) : \mathcal{D} \to \tilde{\mathcal{H}}_g$*, we say* $h(\cdot)$ *is coarser than* $\tilde{h}(\cdot)$ *if there exists another function* $f(\cdot) : \tilde{\mathcal{H}}_g \to \mathcal{H}$ *such that* $h(\mathbf{d}_g) = f \circ \tilde{h}(\mathbf{d}_g)$ *for all* $\mathbf{d}_g \in \mathcal{D}$*.*

Intuitively, this means that $h(\cdot)$ gives a "cruder" summary of $\mathbf{d}_g$ (i.e. discards more information) than $\tilde{h}(\cdot)$. In other words, a coarser function imposes more restrictions on the potential outcomes. For example, the exchangeable assignment $h(\mathbf{d}_g) = \mathbf{1}_g'\mathbf{d}_g$ is coarser than the identity function $\tilde{h}(\mathbf{d}_g) = \mathbf{d}_g$, and the reference group assignment $h(\mathbf{d}_g) = (d_1, d_2)$ is coarser than $\tilde{h}(\mathbf{d}_g) = (d_1, d_2, d_3, d_4)$.

The next section addresses identification of average potential outcomes when the true treatment rule $h_0(\cdot)$ is possibly unknown.

3 Identification

In what follows, let $\mathbf{H}_{ig} = h(\mathbf{D}_{(i)g})$ be the observed value of the chosen treatment rule, and let $\mathbf{H}_{ig}^0 = h_0(\mathbf{D}_{(i)g})$. The following result links observed outcomes, potential outcomes and effective treatment assignments, and will be used for the upcoming theorems.

Lemma 1 (Nonparametric Identification) *Suppose Assumption 1 holds and let* $h_0(\cdot) : \mathcal{D} \to \mathcal{H}_0$ *be the true treatment rule. Given a treatment rule* $h(\cdot) : \mathcal{D} \to \mathcal{H}$*, for any pair* $(d, \mathbf{h}) \in \{0,1\} \times \mathcal{H}$ *such that* $\mathbb{P}[D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h}] > 0$ *and for any measurable function* $m(\cdot)$*,*

$$\mathbb{E}[m(Y_{ig})|D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h}] = \sum_{\mathbf{h}_0 \in \mathcal{H}_0} \mathbb{E}[m(Y_{ig}(d, \mathbf{h}_0))]\mathbb{P}[\mathbf{H}_{ig}^0 = \mathbf{h}_0|D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h}].$$

In particular, if $h_0(\cdot)$ *is coarser than* $h(\cdot)$*, then*

$$\mathbb{E}[m(Y_{ig})|D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h}] = \mathbb{E}[m(Y_{ig}(d, h_0(\mathbf{h})))].$$

In what follows, I will let $m(\cdot)$ be the identity function to reduce notation. More generally, different choices of the function $m(\cdot)$ lead to different estimands of interest. For example, setting $m(\cdot) = \mathbb{1}(\cdot \leq y)$ for some $y \in \mathbb{R}$ yields $\mathbb{E}[m(Y_{ig})|D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h}] = F_Y(y|D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h})$ where $F_Z(\cdot|\cdot)$ is the conditional cdf of a random variable Z. This choice of $m(\cdot)$ can be used to identify the distribution of potential outcomes.

Lemma 1 shows that the average observed outcome among units facing $D_{ig} = d$ and $\mathbf{H}_{ig} = \mathbf{h}$ averages the potential outcomes over all the assignments $\mathbf{h}_0$ that are consistent with $(D_{ig}, \mathbf{H}_{ig}) = (d, \mathbf{h})$, as long as the probability of $(d, \mathbf{h})$ is not zero.

To illustrate Lemma 1, consider the previous example with three units and where the $h_0(\cdot)$ is the identity function so the potential outcome has the form $Y_{ig}(d, d_1, d_2)$. Suppose we posit an exchangeable treatment rule $h(\mathbf{d}_g) = \mathbf{1}_g'\mathbf{d}_g$ and thus $\mathbf{H}_{ig} = S_{ig} = \sum_{j \neq i} D_{jg}$ which is a scalar counting how many of unit i's peers are treated. By Lemma 1, if $\mathbb{P}[D_{ig} = 0, S_{ig} = 1] > 0$, $\mathbb{E}[Y_{ig}|D_{ig} = 0, S_{ig} = 1]$ equals a weighted average of $\mathbb{E}[Y_{ig}(0, 1, 0)]$ and $\mathbb{E}[Y_{ig}(0, 0, 1)]$, with weights given by the conditional probabilities of these different assignments.

In general, $(D_{ig}, \mathbf{H}_{ig}) = (d, \mathbf{h})$ may be consistent with many different effective assignments $\mathbf{h}_0$. When $h_0(\cdot)$ is coarser than $h(\cdot)$, however, the value of $\mathbf{h_0}$ is uniquely determined. In such cases, the second part of Lemma 1 shows that $\mathbb{E}[Y_{ig}|D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h}]$ identifies the value of the average potential outcome consistent with that assignment. For example, suppose that the true $h_0(\cdot)$ is exchangeable, so that outcomes have the form $Y_{ig}(d, s)$ with $s = 0, 1, 2$, and suppose we posit $h(\mathbf{d}_g) = \mathbf{d}_g = (d, d_1, d_2)$. Setting $\mathbf{H}_{ig} = (0, 1, 1)$ implies that the sum of treated peers is equal to 2, and therefore $\mathbb{E}[Y_{ig}|D_{ig} = 0, \mathbf{H}_{ig} = (1, 1)] = \mathbb{E}[Y_{ig}(0, 2)]$. In particular, this result implies that if the function $h(\cdot)$ is correctly specified, $\mathbb{E}[Y_{ig}|D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h}] = \mathbb{E}[Y_{ig}(d, \mathbf{h})]$.

Remark 1 (Overidentification) When $h_0(\cdot)$ is coarser than $h(\cdot)$, there may be multiple values of $\mathbb{E}[Y_{ig}|D_{ig} = d, \mathbf{H}_{ig} = \mathbf{h}]$ that equal the same value of the average potential outcome $\mathbb{E}[Y_{ig}(d, h_0(\mathbf{h}))]$. In the example above, if $h(\cdot)$ is the identity treatment rule but the true treatment rule is exchangeable, we have that $\mathbb{E}[Y_{ig}|D_{ig} = 0, \mathbf{H}_{ig} = (1, 0)] = \mathbb{E}[Y_{ig}|D_{ig} = 0, \mathbf{H}_{ig} = (0, 1)] = \mathbb{E}[Y_{ig}(0, 1)]$. This fact can be used to test, to some extent, hypotheses about the true unknown treatment rule $h_0(\cdot)$. For instance, $\mathbb{E}[Y_{ig}|D_{ig} = 0, \mathbf{H}_{ig} = (1, 0)] \neq \mathbb{E}[Y_{ig}|D_{ig} = 0, \mathbf{H}_{ig} = (0, 1)]$ would suggest that the true treatment rule is not exchangeable. Note, however, that failing to reject this hypothesis does not imply that exchangeability holds, as exchangeability could hold for $\mathbb{E}[Y_{ig}(d, \mathbf{d}_g)]$ but not for other moments. □

Remark 2 (Pooled estimands) Coarse treatment rules can be used not only as a modeling assumption on potential outcomes but also as summary measures of average potential outcomes and treatment effects. This may be particularly useful when $h_0(\cdot)$ is suspected to be high-dimensional. For instance, setting $h(\cdot)$ equal to a constant function, which ignores $\mathbf{d}_g$, averages over all possible peers' assignments: $\mathbb{E}[Y_{ig}|D_{ig} = d] = \sum_{\mathbf{h}_0} \mathbb{E}[Y_{ig}(d, \mathbf{h}_0)]\mathbb{P}[\mathbf{H}_{ig}^0 = \mathbf{h}_0|D_{ig} = d]$. As a less extreme example, let $s = \mathbf{1}_g'\mathbf{d}_g$, and define $h(\mathbf{d}_g) = \mathbb{1}(s > 0)$ which equals one if there is at least one treated peer. Let $S_{ig} = \sum_{j \neq i} D_{jg}$ be the observed number of treated peers for unit i. Then, by Lemma 1, $\mathbb{E}[Y_{ig}|D_{ig} = d, S_{ig} > 0] =$

$\sum_{\mathbf{h}_0} \mathbb{E}[Y_{ig}(d,\mathbf{h}_0)]\mathbb{P}[\mathbf{H}_{ig}^0 = \mathbf{h}_0 | D_{ig} = d, S_{ig} > 0]$. Consider the difference between untreated units with at least one treated peer and untreated units with no treated peers:

$$\Delta = \mathbb{E}[Y_{ig}|D_{ig} = 0, S_{ig} > 0] - \mathbb{E}[Y_{ig}|D_{ig} = 0, S_{ig} = 0].$$

Then, given that $S_{ig} = 0$ implies that $\mathbf{D}_{(i)g} = \mathbf{0}_g$, we have that:

$$\Delta = \sum_{\mathbf{h}_0} \mathbb{E}[Y_{ig}(0,\mathbf{h}_0) - Y_{ig}(0,\mathbf{0})]\mathbb{P}[\mathbf{H}_{ig}^0 = \mathbf{h}_0 | D_{ig} = 0, S_{ig} > 0]$$

where $\mathbf{0} = h_0(\mathbf{0}_g)$. Thus, Δ recovers a weighted average of spillover effects on untreated units weighted by the probabilities of the different assignments, a well-defined causal parameter. A natural generalization of this idea is to split S_{ig} into categories such as $S_{ig} = 0$, $1 \leq S_{ig} \leq m$, $m+1 \leq S_{ig} \leq n_g$ and so on. Section 6 illustrates how to estimate these pooled parameters using linear regressions. □

Remark 3 (Partial population experiments) A popular design when analyzing spillover effects is the partial population design (Moffit, 2001; Duflo and Saez, 2003; Hirano and Hahn, 2010). In its simplest form, groups are randomly divided into treated and controls based on binary indicator T_g. Then, within the groups with $T_g = 1$, treatment D_{ig} is randomly assigned at the individual level. In these type of experiments, spillover effects are often estimated as the average difference between control units in treated groups and control units in pure control groups,

$$\Delta_{\mathsf{PP}} = \mathbb{E}[Y_{ig}|D_{ig} = 0, T_g = 1] - \mathbb{E}[Y_{ig}|T_g = 0].$$

For recent examples of this or similar strategies, see Beuermann, Cristia, Cueto, Malamud, and Cruz-Aguayo (2015), Beshears, Choi, Laibson, Madrian, and Milkman (2015) and Giné and Mansuri (2018), among others. Redefining the vector of treatment assignments as $(D_{ig}, \mathbf{D}_{(i)g}, T_g) = (d, \mathbf{d}_g, t)$ and setting $h(\mathbf{d}_g, t) = t$, if $(\mathbf{D}_g, T_g)$ is independent of potential outcomes, then Lemma 1 implies that:

$$\Delta_{\mathsf{PP}} = \sum_{\mathbf{h}_0} \mathbb{E}[Y_{ig}(0,\mathbf{h}_0) - Y_{ig}(0,\mathbf{0})]\mathbb{P}[\mathbf{H}_{ig}^0 = \mathbf{h}_0 | D_{ig} = 0, T_g = 1]$$

which averages over all the possible number of treated peers that an untreated unit can have in a treated group. The generalization to experiments with more than two categories (see e.g. Crépon, Duflo, Gurgand, Rathelot, and Zamora, 2013) is straightforward. □

3.1 Difference in means

The difference in means estimand $\beta_{\mathsf{D}} = \mathbb{E}[Y_{ig}|D_{ig} = 1] - \mathbb{E}[Y_{ig}|D_{ig} = 0]$, which compares the average observed outcomes between treated and controls, is arguably the most common estimand when analyzing randomized experiments. It is well known that, in the absence of

spillovers, β_{D} equals the average treatment effect (ATE) when the treatment is randomly assigned. An estimate for the ATE can be calculated by estimating the model:

$$Y_{ig} = \alpha_{\mathsf{D}} + \beta_{\mathsf{D}} D_{ig} + u_{ig}. \tag{1}$$

The following results characterizes the difference in means β_{D} in the presence of spillovers. In what follows, let $\mathbf{0} = h_0(\mathbf{0}_g)$.

Theorem 1 (Difference in means) *Under Assumption 1, the coefficient β_{D} from Equation (1) can be written as:*

$$\begin{aligned}\beta_{\mathsf{D}} = \mathbb{E}[Y_{ig}(1,\mathbf{0}) - Y_{ig}(0,\mathbf{0})] &+ \sum_{\mathbf{h}_0 \in \mathcal{H}_0} \mathbb{E}[Y_{ig}(1,\mathbf{h}_0) - Y_{ig}(1,\mathbf{0})]\mathbb{P}[\mathbf{H}_{ig}^0 = \mathbf{h}_0 | D_{ig} = 1] \\ &- \sum_{\mathbf{h}_0 \in \mathcal{H}_0} \mathbb{E}[Y_{ig}(0,\mathbf{h}_0) - Y_{ig}(0,\mathbf{0})]\mathbb{P}[\mathbf{H}_{ig}^0 = \mathbf{h}_0 | D_{ig} = 0].\end{aligned}$$

Hence, the difference-in-means estimand equals the direct effect without treated peers $\mathbb{E}[Y_{ig}(1,\mathbf{0}) - Y_{ig}(0,\mathbf{0})]$ plus the difference in weighted averages of spillover effects under treatment and under control. In general, the sign of this difference is undetermined, as it depends on the relative magnitudes of the average spillover effects on treated and controls. Hence, the difference in means can be larger, smaller or equal than the average direct effect $\mathbb{E}[Y_{ig}(1,\mathbf{0}) - Y_{ig}(0,\mathbf{0})]$. In particular, if the spillover effects on treated units are equal to zero and the spillover effects on controls are of the same sign that the direct effects, the difference in means will underestimate the average direct effect without treated peers. In this case, the presence of spillovers pushes the difference in means towards zero, which captures the idea of "contamination" of the control group.

3.2 Linear-in-means models

Equation (1) may give an incomplete assessment of the effect of a program because it completely ignores the presence of spillovers. When trying to explicitly estimate spillover effects, a common strategy is to estimate a reduced-form linear-in-means (LIM) model, which is given by:

$$Y_{ig} = \alpha_\ell + \beta_\ell D_{ig} + \gamma_\ell \bar{D}_g^{(i)} + \eta_{ig}, \qquad \bar{D}_g^{(i)} = \frac{1}{n_g} \sum_{j \neq i} D_{jg} \tag{2}$$

that is, a regression of the outcome on a treatment indicator and the proportion of treated neighbors. In this specification, β_ℓ intends to capture a direct effect whereas γ_ℓ is seen as a measure of spillover effects, since it captures the average change in outcomes in response to a change in the proportion of treated neighbors. While the parameters $(\beta_\ell, \gamma_\ell)$ can be interpreted as coefficients from a linear projection, the following result shows that they do not have a causal interpretation in general.

Theorem 2 (LIM regression) *Under Assumption 1, the coefficients* $(\beta_\ell, \gamma_\ell)$ *from Equation* (2) *can be written as:*

$$\begin{aligned}
\beta_\ell &= \mathbb{E}[Y_{ig}|D_{ig}=1] - \mathbb{E}[Y_{ig}|D_{ig}=0] - \frac{\gamma_\ell}{n_g}(\mathbb{E}[S_{ig}|D_{ig}=1] - \mathbb{E}[S_{ig}|D_{ig}=0]) \\
\gamma_\ell &= \sum_{s=1}^{n_g} \phi_0(s)(\mathbb{E}[Y_{ig}|D_{ig}=0, S_{ig}=s] - \mathbb{E}[Y_{ig}|D_{ig}=0, S_{ig}=0]) \\
&\quad + \sum_{s=1}^{n_g} \phi_1(s)(\mathbb{E}[Y_{ig}|D_{ig}=1, S_{ig}=s] - \mathbb{E}[Y_{ig}|D_{ig}=1, S_{ig}=s])
\end{aligned}$$

where for $d = 0, 1$,

$$\phi_d(s) = \frac{n_g \mathbb{P}[D_{ig}=d]\mathbb{P}[S_{ig}=s|D_{ig}=d]}{\mathbb{P}[D_{ig}=0]\mathbb{V}[S_{ig}|D_{ig}=0] + \mathbb{P}[D_{ig}=1]\mathbb{V}[S_{ig}|D_{ig}=1]} \cdot (s - \mathbb{E}[S_{ig}|D_{ig}=d]).$$

Theorem 2 characterizes the coefficients from the linear projection of Y_{ig} into D_{ig} and $\bar{D}_g^{(i)}$. The coefficient β_ℓ equals the difference in means minus and adjustment factor that depends on γ_ℓ and the relationship between treatment assignments within group. This parameter is similar to the one analyzed in Theorem 1 and hence does not have a direct causal interpretation in general.

The coefficient γ_ℓ equals a linear combination of differences $\mathbb{E}[Y_{ig}|D_{ig}=d, S_{ig}=s] - \mathbb{E}[Y_{ig}|D_{ig}=d, S_{ig}=0]$ across all values of s with weights $(\phi_0(s), \phi_1(s))_s$. Two factors obscure the causal interpretation of this coefficient. On the one hand, the magnitudes $\mathbb{E}[Y_{ig}|D_{ig}=d, S_{ig}=s]$ implicitly impose an exchangeable treatment rule that may be misspecified in general. The interpretation of these expectations is given in Lemma 1. On the other hand, even if this treatment rule was correctly specified, so that $\mathbb{E}[Y_{ig}|D_{ig}=d, S_{ig}=s] - \mathbb{E}[Y_{ig}|D_{ig}=d, S_{ig}=0] = \mathbb{E}[Y_{ig}(d,s) - Y_{ig}(d,0)]$, these effects are combined using weights that can be negative. Indeed, it can be seen that $\sum_{s=0}^{n_g} \phi_d(s) = 0$ for $d = 0, 1$ and hence some of the weights are necessarily negative (specifically, the ones corresponding to low values of s). Thus, γ_ℓ may be zero or negative when all the average spillover effects are positive or vice versa.

To illustrate the importance of these issues in an empirical setting, Section 6 shows a case in which the estimates of β_ℓ and γ_ℓ are close to zero and not statistically significant, even when the estimated average spillover effects are all large and statistically significant when estimated nonparametrically.

The following result provides sufficient conditions for the coefficients from a LIM regression to have a causal interpretation.

Corollary 1 (Correctly-specified LIM regression) *Suppose that, in addition to Assumption 1, the following conditions hold:*

(i) Exchangeability: $Y_{ig}(d, h_0(\mathbf{d}_g)) = Y_{ig}(d, s)$ *where* $s = \mathbf{1}_g'\mathbf{d}_g$,

(ii) Linearity: for each $d = 0, 1$ *there is a constant* κ_d *such that* $\mathbb{E}[Y_{ig}(d,s) - Y_{ig}(d,s-1)] =$

κ_d *for all* $s \geq 1$.

Then, the coefficients $(\beta_\ell, \gamma_\ell)$ *from Equation* (2) *are:*

$$\beta_\ell = \mathbb{E}[Y_{ig}(1,0) - Y_{ig}(0,0)] + (\kappa_1 - \kappa_0)\{(1-\lambda)\mathbb{E}[S_{ig}|D_{ig}=1] + \lambda\mathbb{E}[S_{ig}|D_{ig}=0]\}$$
$$\gamma_\ell = \lambda\mathbb{E}[Y_{ig}(1,n_g) - Y_{ig}(1,0)] + (1-\lambda)\mathbb{E}[Y_{ig}(0,n_g) - Y_{ig}(0,0)]$$

where

$$\lambda = \frac{\mathbb{P}[D_{ig}=1]\mathbb{V}[S_{ig}|D_{ig}=1]}{\mathbb{P}[D_{ig}=1]\mathbb{V}[S_{ig}|D_{ig}=1] + \mathbb{P}[D_{ig}=0]\mathbb{V}[S_{ig}|D_{ig}=0]} \in (0,1).$$

The above result highlights two restrictions that the LIM model implicitly imposes on potential outcomes: (i) peers are exchangeable, so potential outcomes only depend on own treatment and the number of treated peers, and (ii) average spillover effects are linear in s so that, for instance, the effect of having two treated peers is twice as large as the effect of having one treated peer. If these conditions hold, γ_ℓ recovers a weighted average of the effects of having all treated peers for treated and untreated units, where the weights are positive and sum to one. Hence, the LIM regression is robust to some heterogeneity in spillover effects $\mathbb{E}[Y_{ig}(d,s) - Y_{ig}(d,0)]$ both over s and over d, but suffers from potentially severe misspecification when spillover effects are nonlinear.

On the other hand, β_ℓ does not recover a causal effect in general. In the particular case in which $\kappa_1 = \kappa_0$, that is, the average spillover effects are equal for treated and untreated units, β_ℓ equals the average direct effect with no treated peers $\mathbb{E}[Y_{ig}(1,0) - Y_{ig}(0,0)]$.

A straightforward way to make Equation (2) more flexible is to include an interaction term between own treatment indicator and the proportion of treated peers:

$$Y_{ig} = \tilde{\alpha}_\ell + \tilde{\beta}_\ell D_{ig} + \gamma_\ell^0 \bar{D}_g^{(i)}(1 - D_{ig}) + \gamma_\ell^1 \bar{D}_g^{(i)} D_{ig} + \xi_{ig} \tag{3}$$

The following result characterizes the coefficients from this specification.

Theorem 3 (Interacted LIM regression) *Under Assumption 1, the coefficients* $(\tilde{\beta}_\ell, \gamma_\ell^0, \gamma_\ell^1)$ *from Equation* (3) *can be written as:*

$$\tilde{\beta}_\ell = \mathbb{E}[Y_{ig}|D_{ig}=1] - \mathbb{E}[Y_{ig}|D_{ig}=0] - \left(\frac{\gamma_\ell^1}{n_g}\mathbb{E}[S_{ig}|D_{ig}=1] - \frac{\gamma_\ell^0}{n_g}\mathbb{E}[S_{ig}|D_{ig}=0]\right)$$
$$\gamma_\ell^0 = \sum_{s=1}^{n_g} \omega_0(s)(\mathbb{E}[Y_{ig}|D_{ig}=0, S_{ig}=s] - \mathbb{E}[Y_{ig}|D_{ig}=0, S_{ig}=0])$$
$$\gamma_\ell^1 = \sum_{s=1}^{n_g} \omega_1(s)(\mathbb{E}[Y_{ig}|D_{ig}=1, S_{ig}=s] - \mathbb{E}[Y_{ig}|D_{ig}=1, S_{ig}=s])$$

where for $d = 0, 1$,

$$\omega_d(s) = \frac{n_g \mathbb{P}[S_{ig}=s|D_{ig}=d]}{\mathbb{V}[S_{ig}|D_{ig}=d]} \cdot (s - \mathbb{E}[S_{ig}|D_{ig}=d]).$$

According to this theorem, an interacted LIM regression separates the spillover components $\mathbb{E}[Y_{ig}|D_{ig}=0,S_{ig}=s]-\mathbb{E}[Y_{ig}|D_{ig}=0,S_{ig}=0]$ and $\mathbb{E}[Y_{ig}|D_{ig}=1,S_{ig}=s]-\mathbb{E}[Y_{ig}|D_{ig}=1,S_{ig}=0]$. However, the issue of negative weights remains for each component, since $\sum_{s=0}^{n_g}\omega_d(s)=0$ for $d=0,1$ and hence some of the weights are necessarily negative.

Finally, the following result shows that when peers are exchangeable and spillover effects are linear, the interacted LIM model can recover all the causal parameters of interest.

Corollary 2 (Correctly-specified interacted LIM regression) *Suppose that, in addition to Assumption 1, the following conditions hold:*

(i) Exchangeability: $Y_{ig}(d,h_0(\mathbf{d}_g))=Y_{ig}(d,s)$ *where* $s=\mathbf{1}_g'\mathbf{d}_g$,

(ii) Linearity: for each $d=0,1$ *there is a constant* κ_d *such that* $\mathbb{E}[Y_{ig}(d,s)-Y_{ig}(d,s-1)]=\kappa_d$ *for all* $s\geq 1$.

Then, the coefficients $(\tilde{\beta}_\ell,\gamma_\ell^0,\gamma_\ell^1)$ *from Equation (3) can be written as:*

$$\begin{aligned}\tilde{\beta}_\ell &= \mathbb{E}[Y_{ig}(1,0)-Y_{ig}(0,0)]\\ \gamma_\ell^0 &= \mathbb{E}[Y_{ig}(0,n_g)-Y_{ig}(0,0)]\\ \gamma_\ell^1 &= \mathbb{E}[Y_{ig}(1,n_g)-Y_{ig}(1,0)].\end{aligned}$$

According to this result, the coefficients from a correctly-specified interacted LIM regression recover the average direct effect without treated peers and the spillover effects of having all peers treated, for treated and untreated units separately. Because of linearity, all the remaining spillover effects can be recovered by appropriately rescaling γ_ℓ^d. For example, the average spillover effect from having one treated peer on an untreated unit is $\mathbb{E}[Y_{ig}(0,1)-Y_{ig}(0,0)]=\gamma_\ell^0/n_g$.

4 Estimation and inference

The previous sections provide conditions under which average direct and spillover effects can be nonparametrically identified by exploiting variation over own and peers' assignments. Because these magnitudes can be written as population averages, it is straightforward to construct their corresponding estimators based on sample cell means. The main challenge for estimation arises when groups are large. A large number of units per group requires estimating a large number of means in each of the cells defined by the treatment assignments. When groups have many units (as in households with many siblings or classrooms with a large number of students), the probability of observing some assignments can be close to zero and the number of observations in each cell can be too small to estimate the average potential outcomes.

For example, suppose the treatment is assigned as an independent coin flip with probability $p=1/2$. Under this assignment we would expect most groups to have about half its units treated, so when groups have, say, 10 units, 5 of them would be treated on average. The

probability of observing groups with zero or all treated units, on the other hand, will be close to zero, and thus the average potential outcomes corresponding to these "tail assignments" will be very hard to estimate.

So far, the analysis has been done taking group size as fixed. When group size is fixed, small cells are a finite sample problem that disappears as the sample grows. To account for this phenomenon asymptotically, in this section I will generalize this setting to allow group size to grow with the sample size. The goal is to answer the question of how large groups can be relative to the total sample size to allow for valid estimation and inference. The key issue to obtain consistency and asymptotic normality will be to ensure that the number of observations in all cells grows to infinity as the sample size increases. This setup is not intended to model a population in which groups are effectively infinitely large, but as a statistical device to approximate the distribution of estimators in finite samples when the number of parameters can be "moderately" large, in a sense that will be made more precise in this section. The case with fixed group size is a particular case in this setting.

In this section I will assume that groups are equally sized, so that $n_g = n$. Recall that given a candidate treatment rule $h(\cdot)$ and $\mathbf{h} = h(\mathbf{d}_g)$, the effective treatment assignments are given by $(d, \mathbf{h}_g) \in \{0,1\} \times \mathcal{H}$. As formalized in Assumption 2 below, $h(\cdot)$ is not assumed to equal the true treatment rule, but the true treatment rule $h_0(\cdot)$ has to be coarser than $h(\cdot)$ as specified in Definition 1. To make the notation more compact, I will let $\mathcal{A}_n = \{0,1\} \times \mathcal{H}$ where the notation makes the dependence of this set on the group size explicit. The cardinality of this set is denoted by $|\mathcal{A}_n|$, which indicates the total number of parameters to be estimated. The observed effective treatment assignments will be $(D_{ig}, \mathbf{H}_{ig}) = \mathbf{A}_{ig}$, taking values $\mathbf{a} \in \mathcal{A}_n$, and $\mu(\mathbf{a}) = \mathbb{E}[Y_{ig} | \mathbf{A}_{ig} = \mathbf{a}]$.

Each treatment assignment mechanism determines a distribution $\pi(\cdot)$ over $\mathcal{A}_n$ where $\pi(\mathbf{a}) = \mathbb{P}[\mathbf{A}_{ig} = \mathbf{a}]$ for $\mathbf{a} \in \mathcal{A}_n$. For example, when $\mathcal{A}_n = \{0,1\}$, if the treatment is assigned independently as a coin flip, $\pi(1) = \mathbb{P}[D_{ig} = 1] = p$ and $\pi(0) = 1 - p$. Under the same assignment, with an exchangeable treatment rule, $\pi(d,s) = \mathbb{P}[D_{ig} = d, S_{ig} = s] = \binom{n}{s} p^{s+d}(1-p)^{n+1-s-d}$. Importantly, since the size of the set $\mathcal{A}_n$ can increase with group size, the probabilities $\pi(\mathbf{a})$ can shrink towards zero for some (or all) assignments $\mathbf{a} \in \mathcal{A}_n$, as illustrated in the previous example. Finally, define:

$$\underline{\pi}_n = \min_{\mathbf{a} \in \mathcal{A}_n} \pi(\mathbf{a})$$

which is the probability of the least likely treatment assignment. This probability, together with the total sample size, will determine the number of observations in the smallest assignment cell, that is, the number of observations available to estimate the "hardest" average potential outcome.

Let $\mathbf{A}_g = (\mathbf{A}_{1g}, \ldots, \mathbf{A}_{n_g+1,g})$, $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_G)$, and $\mathbf{Y}_g = (Y_{1g}, Y_{2g}, \ldots Y_{n_g+1,g})'$. I will assume the following.

Assumption 2 (Sampling and design)

(i) *For* $g = 1, \ldots, G$, $(\mathbf{Y}_g', \mathbf{A}_g')$ *are iid, and* $n_g = n$.

(ii) *The true treatment rule* $h_0(\cdot)$ *is coarser than* $h(\cdot)$.

(iii) *The potential outcomes are independent across* i *within groups.*

(iv) $|\mathcal{A}_n| = O(G(n+1)\underline{\pi}_n)$, *as* $G \to \infty$ *and* $n \to \infty$.

Part (i) in Assumption 2 states that the researcher has access to a sample of G independent and identically distributed equally-sized groups. When groups have different sizes (for example, households with 3, 4 or 5 siblings), the analysis can be performed separately for each group size. Section A3 of the supplemental appendix further discusses the case of unequally-sized groups. Part (ii) allows the posited treatment rule $h(\cdot)$ to be different from the true treatment rule, but requires it to be flexible enough to break the dependence between Y_{ig} and $\mathbf{A}_{jg}$ conditional on $\mathbf{A}_{ig}$ for $j \neq i$. Part (iii) assumes that potential outcomes are independent within groups, and hence the only source of dependence between the observed outcomes is the assignment $\mathbf{A}_g$. In the setting with growing group sizes, this assumption can be relaxed at the expense of introducing more restrictions on the within-group network structure (for example, assuming the rate at which links are formed is slow enough). On the other hand, this condition is easily relaxed to arbitrary dependence structures when the group size is fixed. Together, conditions (ii) and (iii) imply that observed outcomes are independent conditional on the assignments. Importantly, note that this condition does not restrict in any way the correlation between treatment assignments $\mathbf{A}_{ig}$ and $\mathbf{A}_{jg}$. In fact, it can be seen that effective treatment assignments are always correlated by construction, since $\mathbf{A}_{ig}$ depends on D_{jg} and vice versa. Finally, part (iv) requires that the total number of parameters not to grow faster than the effective sample size, that is, the expected sample size in the smallest cell.

Given a sample of G groups with $n+1$ units each, let $\mathbb{1}_{ig}(\mathbf{a}) = \mathbb{1}(\mathbf{A}_{ig} = \mathbf{a})$, $N_g(\mathbf{a}) = \sum_{i=1}^{n+1} \mathbb{1}_{ig}(\mathbf{a})$ and $N(\mathbf{a}) = \sum_{g=1}^{G} N_g(\mathbf{a})$, so that $N_g(\mathbf{a})$ is the total number of observations receiving effective assignment $\mathbf{a}$ in group g and $N(\mathbf{a})$ is the total number of observations receiving effective assignment $\mathbf{a}$ in the sample. The estimator for $\mu(\mathbf{a})$ is defined as:

$$\hat{\mu}(\mathbf{a}) = \begin{cases} \frac{\sum_{g=1}^{G} \sum_{i=1}^{n+1} Y_{ig} \mathbb{1}_{ig}(\mathbf{a})}{N(\mathbf{a})} & \text{if } N(\mathbf{a}) > 0 \\ \nexists & \text{if } N(\mathbf{a}) = 0 \end{cases}$$

Thus, the estimator for $\mu(\mathbf{a})$ is simply the sample average of the outcome for observations receiving assignment $\mathbf{a}$, whenever there is at least one observation receiving this assignment.

The following assumption imposes some regularity conditions that are required for upcoming theorems. Let $\sigma^2(\mathbf{a}) = \mathbb{V}[Y_{ig} | \mathbf{A}_{ig} = \mathbf{a}]$.

Assumption 3 (Moments) *There are constants* $\underline{\sigma}$ *and* b *such that:*

$$(i) \quad \inf_n \min_{\mathbf{a} \in \mathcal{A}_n} \sigma^2(\mathbf{a}) \geq \underline{\sigma}^2 > 0, \qquad (ii) \quad \sup_n \max_{\mathbf{a} \in \mathcal{A}_n} \mathbb{E}[Y_{ig}^6 | \mathbf{A}_{ig} = \mathbf{a}] \leq b < \infty$$

Then we have the following result.

Lemma 2 (Effective sample size) *Suppose Assumptions 2 and 3 hold, and consider an assignment mechanism $\pi(\cdot)$ such that $\pi(\mathbf{a}) > 0$ for all $\mathbf{a} \in \mathcal{A}_n$. If*

$$\frac{\log |\mathcal{A}_n|}{G\underline{\pi}_n} \to 0 \tag{4}$$

then for any $c \in \mathbb{R}$

$$\mathbb{P}\left[\min_{\mathbf{a}\in\mathcal{A}_n} N(\mathbf{a}) > c\right] \to 1.$$

Lemma 2 says that, under condition (4), the number of observations in the smallest cell will go to infinity, which implies that all the estimators are well defined asymptotically. Hence, condition (4) formalizes the meaning of "large sample" in this context, and states that the number of groups has to be large relative to the total number of parameters and the probability of the least likely assignment. This expression can be interpreted as an invertibility condition for the design matrix of a linear regression model, in the specific case in which the regressors are mutually exclusive indicator variables. This requirement can be seen as a low-level condition that justifies the assumption of invertibility of the design matrix (see e.g. Assumption 2 in Cattaneo, Jansson, and Newey, 2018). When this condition does not hold, small cells may result in estimators with poor finite sample behavior and whose asymptotic distribution, if it exists, may be non-Gaussian. See Ma and Wang (forthcoming) for an example in the context of inverse-probability weighting estimators, which includes randomized experiments as a special case.

Next, let

$$\hat{\sigma}^2(\mathbf{a}) = \frac{\sum_{g=1}^{G}\sum_{i=1}^{n+1}(Y_{ig} - \hat{\mu}(\mathbf{a}))^2 \mathbb{1}_{ig}(\mathbf{a})}{N(\mathbf{a})} \mathbb{1}(N(\mathbf{a}) > 0)$$

be the standard error estimators. Then we have the following result.

Theorem 4 (Consistency and asymptotic normality) *Suppose Assumptions 1, 2 and 3 hold. Under the conditions for Lemma 2,*

$$\begin{aligned} \max_{\mathbf{a}\in\mathcal{A}_n} |\hat{\mu}(\mathbf{a}) - \mu(\mathbf{a})| &= O_{\mathbb{P}}\left(\sqrt{\frac{\log |\mathcal{A}_n|}{G(n+1)\underline{\pi}_n}}\right), \\ \max_{\mathbf{a}\in\mathcal{A}_n} \left|\hat{\sigma}^2(\mathbf{a}) - \sigma^2(\mathbf{a})\right| &= O_{\mathbb{P}}\left(\sqrt{\frac{\log |\mathcal{A}_n|}{G(n+1)\underline{\pi}_n}}\right), \end{aligned} \tag{5}$$

and

$$\max_{\mathbf{a}\in\mathcal{A}_n} \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left[\frac{\hat{\mu}(\mathbf{a}) - \mu(\mathbf{a})}{\sqrt{\mathbb{V}[\hat{\mu}(\mathbf{a})|\mathbf{A}]}} \le x\right] - \Phi(x)\right| = O\left(\frac{1}{\sqrt{G(n+1)\underline{\pi}_n}}\right) \tag{6}$$

where $\Phi(x)$ is the cdf of a standard Gaussian random variable.

Equation (5) shows that both the average potential outcome and standard error estimators converge in probability to their true values, uniformly over treatment assignments, at the rate $\sqrt{\log|\mathcal{A}_n|/(G(n+1)\underline{\pi}_n)}$. The denominator in this rate can be seen as the minimum expected cell size, whereas the numerator is a penalty for having an increasing number of parameters. Equation (6) bounds the difference between the distributions of the standardized potential outcomes estimators and the standard normal distribution, uniformly over the treatment assignments. Under condition (4), $G(n+1)\underline{\pi}_n \to \infty$, which gives asymptotic normality. Furthermore, this bound also reveals the rate at which the distribution of the standardized estimator approaches the standard normal, namely, $\sqrt{G(n+1)\underline{\pi}_n}$.

Remark 4 (Inference with many small groups) When the number of units per group is small compared to the total sample size, the effect of group size disappears asymptotically and inference can be based on a large G small n approximation as commonly done, for example, in panel data models. In this context, n, $|\mathcal{A}_n|$ and $\underline{\pi}_n$ are fixed so condition (4) holds automatically as long as the number of groups goes to infinity. Consistency and asymptotic normality of the estimators can be achieved under the usual regularity conditions as $G \to \infty$, and the variance estimator can easily account for both heteroskedasticity and intragroup correlation using standard techniques. □

Remark 5 (Inference for pooled parameters) When focusing on pooled estimands as described in Section 3 (see Remarks 2 and 3), the set $\mathcal{A}_n$ does not change with n. For example, for the parameter Δ_{pp} in Remark 3, the set of effective treatment assignments can be defined as $\mathcal{A}_n = \{(t,d) = (0,0),(1,0),(1,1)\}$ corresponding to the average outcomes for units in pure control groups ($T_g = 0$), control units in treated groups ($T_g = 1, D_{ig} = 0$) and treated units in treated groups ($T_g = 1, D_{ig} = 1$), respectively. In this case, $|\mathcal{A}_n| = 3$ and hence the number of parameters and the probabilities of each assignment do not change with group size. Hence, inference can be conducted using standard tools when when $G \to \infty$. Inference with a small number of large groups, as in the case where groups are villages or large geographical units (see e.g. Giné and Mansuri, 2018; Crépon, Duflo, Gurgand, Rathelot, and Zamora, 2013) may require other strategies such as the wild bootstrap; see MacKinnon and Webb (2020) for a recent review. □

Remark 6 (Connection to multi-valued treatments) This asymptotic framework can also be applied to multi-valued treatments setting (Imbens, 2000) where $\mathbf{A}_{ig}$ corresponds to the treatment of unit i in group g. Estimation and inference for multi-valued treatments taking on a finite number of values have been extensively analyzed in a variety of contexts (see e.g. Cattaneo, 2010; Farrell, 2015; Ao, Calonico, and Lee, forthcoming, and references therein). The results in this section complement this literature in two ways. First, I consider double-array asymptotics in which the number of treatment values is allowed to grow with the sample size. Second, the results account for the dependence between treatment assignments $\mathbf{A}_{ig}$ and $\mathbf{A}_{jg}$ of units in the same group, a feature that is specific of the spillovers setting. □

Remark 7 (Other estimands and uniform inference) Using standard empirical process theory, the estimation and inference results in this section can be extended to hold uniformly on $m(\cdot) \in \mathcal{M}$ for appropriate classes $\mathcal{M}$. See van der Vaart and Wellner (1996) for a classical reference. For example, letting $\mathcal{M} = \{\mathbb{1}(\cdot \leq y) : y \in \mathbb{R}\}$ leads to the distribution of potential outcomes as discussed in Lemma 1, which can be used to conduct uniform inference on quantile treatment effects and related estimands. □

4.1 Bootstrap approximation

An alternative approach to perform inference in this setting is the bootstrap. Since the challenge for inference is that cells can have too few observations for the Gaussian distribution to provide a good approximation, the wild bootstrap (Shao and Tu, 1995) can offer a more accurate approximation when groups are relatively large. This type of bootstrap can be performed by defining weights $w_{ig} \in \{-1, 1\}$ with probability 1/2 independently of the sample. The bootstrap estimator for $\mu(\mathbf{a})$ is given by:

$$\hat{\mu}^*(\mathbf{a}) = \frac{\sum_g \sum_i Y_{ig}^* \mathbb{1}_{ig}(\mathbf{a})}{N(\mathbf{a})}$$

whenever the denominator is non-zero, where

$$Y_{ig}^* \mathbb{1}_{ig}(\mathbf{a}) = (\bar{Y}(\mathbf{a}) + (Y_{ig} - \bar{Y}(\mathbf{a}))w_{ig})\mathbb{1}_{ig}(\mathbf{a}) = (\bar{Y}(\mathbf{a}) + \hat{\varepsilon}_{ig} w_{ig})\mathbb{1}_{ig}(\mathbf{a})$$

In what follows, $\mathbb{P}^*[\cdot]$ denotes a probability calculated over the distribution of w_{ig}, conditional on the sample, and $\mathbb{E}^*[\cdot]$ and $\mathbb{V}^*[\cdot]$ the expectation and variance calculated over $\mathbb{P}^*[\cdot]$. The validity of the wild bootstrap is established in the following theorem.

Theorem 5 (Wild bootstrap) *Under Assumptions 1, 2 and 3,*

$$\max_{\mathbf{a} \in \mathcal{A}_n} \sup_{x \in \mathbb{R}} \left| \mathbb{P}^* \left[\frac{\hat{\mu}^*(\mathbf{a}) - \hat{\mu}(\mathbf{a})}{\sqrt{\mathbb{V}^*[\hat{\mu}^*(\mathbf{a})]}} \leq x \right] - \mathbb{P}\left[\frac{\hat{\mu}(\mathbf{a}) - \mu(\mathbf{a})}{\sqrt{\mathbb{V}[\hat{\mu}(\mathbf{a})|\mathbf{A}]}} \leq x \right] \right| \to_{\mathbb{P}} 0.$$

This theorem shows that the wild bootstrap can be used to approximate the distribution of the estimator as an alternative to the standard normal, which may not be accurate when cells have few observations. The performance of the wild bootstrap will be illustrated in Section 5 using simulation data.

5 Simulations

This section illustrates the above results in a simulation setting. The outcome will be binary and generated by the following DGP:

$$\mathbb{P}[Y_{ig}(d, s) = 1] = \mu(d, s) = 0.75 + 0.13 \times d + 0.12 \times (1 - d)\mathbb{1}(s > 0)$$

which corresponds to the case with $\mu(0,0) = 0.75$, $\tau(0) = 0.13$, $\theta_0(s) = 0.12$ for all s and $\theta_1(s) = 0$ for all s. That is, the spillover effects on an untreated unit is equal to 0.12 whenever at least one neighbor is treated, and zero for treated units. This DGP is based on the empirical application in Section 6.

The simulations consider two assignment mechanisms. First, *simple random assignment* (SR), where treatment is assigned independently with $\mathbb{P}[D_{ig} = 1] = 0.5$. Second, *two-stage randomization with fixed margins* (2SR-FM), where in the first stage groups are assigned the total number of treated units between 0, 1, 2 and 3 with equal probability, and then treated units are chosen according to this assignment within each group. Corollaries A1 and A2 in the supplemental appendix show that when peers are exchangeable, under SR, condition (4) from Theorem 4 holds whenever $(n+1)/\log G \to 0$, whereas under 2SR-FM the condition holds when $\log(n+1)/\log G \to 0$. Because the second condition is weaker, 2SR-FM is expected to perform better for larger groups.

The parameter of interest will be $\theta_0(n) = \mathbb{E}[Y_{ig}(0,n)] - \mathbb{E}[Y_{ig}(0,0)]$, which is the average spillover effect for an untreated unit with all peers treated. In this simulation, $\theta_0(n) = 0.12$. This parameters can be seen as a "worst-case scenario" given that the probability of the assignment $(D_{ig}, S_{ig}) = (0,n)$ is one of the smallest (in fact, the smallest under 2SR-FM). The estimator will be the difference in cell means:

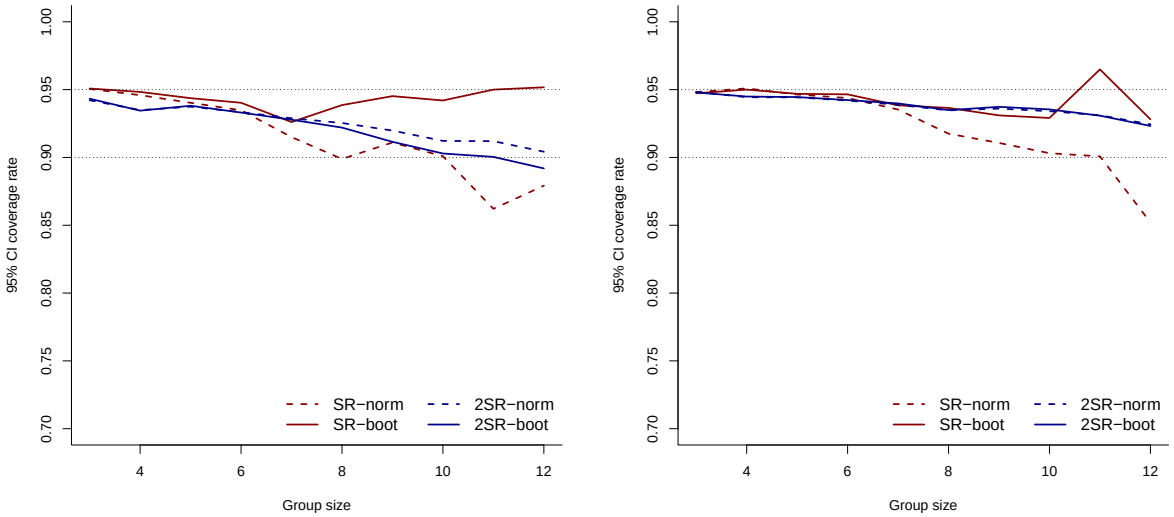
$$\hat{\theta}_0(n) = \frac{\sum_g \sum_i Y_{ig} \mathbb{1}_{ig}(0,n)}{N(0,n)} - \frac{\sum_g \sum_i Y_{ig} \mathbb{1}_{ig}(0,0)}{N(0,0)}$$

whenever $N(0,n) > 1$ and $N(0,0) > 1$, so that both the estimator and its standard error can be calculated. When at least one of the cells has one or zero observations, the estimator is undefined.

Table 1 presents the results for a sample with 300 groups, for four group sizes, $n+1 = 3, 6, 9, 12$. The upper panel shows the results under SR while the lower panel corresponds to the 2SR-FM assignment. In each panel, the first row gives the value of the condition to achieve consistency and asymptotic normality; intuitively, the closer this value is to zero, the better the approximation based on the Gaussian distribution should be. The second and third rows show the bias and the variance of $\hat{\theta}_0(n)$, calculated over the values of the simulated estimates conditional on the estimate being well defined (i.e. when the cells have enough observations to calculate the estimator). The third and fourth rows show the coverage rate of a 95% confidence interval based on the Gaussian approximation and a wild bootstrap confidence interval. Finally, the sixth row, labeled "proportion of empty cells", gives the proportion of the simulations in which the estimator or its standard error could not be calculated due to insufficient number of observations.

The simulations reveal that under both assignment mechanisms, the estimators perform well for $n = 2$ and $n = 5$, with biases close to zero and coverage rate close to 95%. In both cases the coverage rate decreases as group size increases reaching 88% under SR and 90% for 2SR-FM. For $n = 11$, the variance under SR is much larger than the one under 2SR-FM.

Figure 1: Coverage rate of the 95% confidence interval.



(a) $\hat{\theta}_n(0)$, $G = 300$ (b) $\hat{\theta}_n(0)$, $G = 600$

Notes: the dashed lines show the coverage rate of the 95% confidence interval for $\theta_n(0)$ based on the normal approximation under simple random assignment (red line) and two-stage randomization (blue line) for a sample with 300 (left) and 600 (right) groups. The solid lines show the coverage rates for the confidence interval constructed using wild bootstrap.

These sharp differences in precision are due to the fact that, under simple randomization, when $n = 11$ the probability of observing observations in the cells $(0, 0)$ and $(1, n)$ is very close to zero; as shown in the fourth row of the upper panel, the estimator is undefined in almost 99% of the simulations, and, when it is defined, it relies on a very small number of observations. In fact, the expected number of observations in these cells is about 1.6, not enough to calculate a standard error. On the other hand, the variance under 2SR-FM is much more stable across group sizes, and the estimator can be defined in 100% of the cases. The difference in coverage rates under the two assignment mechanisms becomes more evident when $G = 600$, as shown in Figure 1. On the other hand, the wild bootstrap-based confidence interval maintains coverage close to 95% for all group sizes under simple randomization, whereas both the normal-based and the bootstrap-based confidence intervals perform similarly under 2SR. These results are also illustrated in Figure 1.

Table 2 shows the same results for a sample with 600 groups. As expected, the estimator and confidence intervals show better performance compare to the case with $G = 300$.

6 Empirical illustration

In this section I reanalyze the data from Barrera-Osorio, Bertrand, Linden, and Perez-Calle (2011). The authors conducted a pilot experiment designed to evaluate the effect of a con-

Table 1: Simulation results, $G = 300$

	$n = 2$	$n = 5$	$n = 8$	$n = 11$
Simple randomization				
$(n+1)/\log(G)$	0.5260	1.0519	1.5779	2.1039
Bias	−0.0006	0.0007	0.0041	−0.0118
Variance	0.0027	0.0128	0.0433	0.0654
95% CI coverage - normal	0.9505	0.9348	0.9110	0.8792
95% CI coverage - bootstrap	0.9508	0.9403	0.9452	0.9517
Prop. empty cells	0.0000	0.0087	0.5730	0.9851
Two-stage randomization				
$\log(n+1)/\log(G)$	0.1926	0.3141	0.3852	0.4357
Bias	−0.0001	0.0000	0.0003	−0.0003
Variance	0.0024	0.0034	0.0046	0.0058
95% CI coverage - normal	0.9422	0.9334	0.9198	0.9042
95% CI coverage - bootstrap	0.9433	0.9331	0.9115	0.8919
Prop. empty cells	0.0000	0.0000	0.0000	0.0000

Notes: simulation results for $G = 300$ groups. The second and third rows in each panel show the bias and variance of $\hat{\theta}_n(0)$. The fourth and fifth rows show the coverage rate of a normal-based and wild-bootstrap-based 95% confidence intervals, respectively. The sixth row shows the proportion of simulations in which $\hat{\theta}_n(0)$ is undefined due to the small number of observations in the corresponding cell. Results from 10,000 simulations with 2,000 bootstrap repetitions.

Table 2: Simulation results, $G = 600$

	$n = 2$	$n = 5$	$n = 8$	$n = 11$
Simple randomization				
$(n+1)/\log(G)$	0.4690	0.9379	1.4069	1.8759
Bias	0.0001	−0.0004	0.0028	0.0022
Variance	0.0013	0.0059	0.0270	0.0613
95% CI coverage - normal	0.9482	0.9438	0.9107	0.8521
95% CI coverage - bootstrap	0.9473	0.9465	0.9310	0.9281
Prop. empty cells	0.0000	0.0000	0.3112	0.9263
Two-stage randomization				
$\log(n+1)/\log(G)$	0.1717	0.2801	0.3435	0.3885
Bias	0.0005	0.0002	−0.0004	0.0005
Variance	0.0012	0.0017	0.0023	0.0028
95% CI coverage - normal	0.9483	0.9419	0.9360	0.9244
95% CI coverage - bootstrap	0.9479	0.9423	0.9373	0.9232
Prop. empty cells	0.0000	0.0000	0.0000	0.0000

Notes: simulation results for $G = 600$ groups. The second and third rows in each panel show the bias and variance of $\hat{\theta}_n(0)$. The fourth and fifth rows show the coverage rate of a normal-based and wild-bootstrap-based 95% confidence intervals, respectively. The sixth row shows the proportion of simulations in which $\hat{\theta}_n(0)$ is undefined due to the small number of observations in the corresponding cell. Results from 10,000 simulations with 2,000 bootstrap repetitions.

Table 3: Distribution of household size

	Frequency
1	5205
2	1410
3	168
4	15
5	1
Total	6799

Table 4: Treated per household

	Frequency
0	2355
1	3782
2	607
3	52
4	3
Total	6799

ditional cash transfer program, *Subsidios Condicionados a la Asistencia Escolar*, in Bogotá, Colombia. The program aimed at increasing student retention and reducing drop-out and child labor. Eligible registrants ranging from grade 6-11 were randomly assigned to treatment and control.[1] The assignment was performed at the student level. In addition to administrative and enrollment data, the authors collected baseline and follow-up data from students in the largest 68 of the 251 schools. This survey contains attendance data and was conducted in the household. As shown in Table 3, 1,594 households have more than one registered child, and since the treatment was assigned at the child level, this gives variation in the number of treated children per household, as can be seen in Table 4.

I analyze direct and spillover effects restricting the sample to households with three registered siblings, which gives a total of 168 households and 504 observations. The outcome of interest will be school attendance. Because groups are very small in this case, inference can be conducted using standard methods. Indeed, the simulations in Section 5 shows that coverage rates are very close to 95% when $n + 1 = 3$ and G is larger than 300.

I will start by estimating the average direct and spillover effects assuming exchangeability, so that potential outcomes are assumed to depend on the vector of peers' assignments only through the total number of treated peers, $Y_{ig}(d, s)$.

In this application, exchangeability may be a reasonable assumption as parents can be expected to make schooling decisions based on how many of their children are treated (for example, to determine whether the cash transfer covers the direct and opportunity costs of sending their kids to school), regardless of which of their children are treated, especially so given that all eligible children are similar in age and often attend the same schools. In addition, Table 5 provides additional results that do not impose this assumption, and also a test of exchangeability.

Letting $\mu(d, s) = \mathbb{E}[Y_{ig}(d, s)]$, the goal is to estimate the direct effects:

$$\tau(s) = \mathbb{E}[Y_{ig}(1, s) - Y_{ig}(0, s)] = \mu(1, s) - \mu(0, s)$$

[1] The experiment had two different treatments that varied the timing of the payments, but, following the authors, I pool the two treatment arms to increase the sample size. See Barrera-Osorio, Bertrand, Linden, and Perez-Calle (2011) for details.

and the spillover effects:

$$\theta_d(s) = \mathbb{E}[Y_{ig}(d,s) - Y_{ig}(d,0)] = \mu(d,s) - \mu(d,0)$$

where $s = 0, 1, 2$.

The parameters of interest are estimated through the following regression:

$$\mathbb{E}[Y_{ig}|D_{ig}, S_{ig}] = \alpha + \tau(0)D_{ig} + \sum_{s=1}^{n_g} \theta_0(s)\mathbb{1}(S_{ig} = s)(1 - D_{ig}) + \sum_{s=1}^{n_g} \theta_1(s)\mathbb{1}(S_{ig} = s)D_{ig} \quad (7)$$

where $\alpha = \mu(0,0)$. Because it is equivalent to estimating averages at each cell separately, Equation (7) does not impose any parametric assumptions. The total number of parameters in this regression is $2(n+1) = 6$, so the number of coefficients equals the number of average potential outcomes to estimate.

The estimates from Equation (7) are shown in the third panel, "**Full**", of Table 6. These estimates suggest a positive direct effect of the treatment of 16.4 percentage points, significant at the 5 percent level, with almost equally large spillover effects on the untreated units. More precisely, the estimated effect on an untreated kid of having one treated sibling is 14.6 percentage points, while the effect of having two treated siblings is 14 percentage points. The fact that we cannot reject the hypothesis that $\theta_0(1) = \theta_0(2)$ suggests some form of crowding-out: given that one sibling is treated, treating one more sibling does not affect attendance. These facts are consistent with the idea that, for example, the conditional cash transfer alleviates some financial constraint that was preventing the parents from sending their children to school regularly, or with the program increasing awareness on the importance of school attendance, since in these cases the effect occurs as soon as at least one kid in the household is treated, and does not amplify with more treated children.

On the other hand, spillover effects on treated children are much smaller in magnitude and negative. The fact that these estimates are negative does not mean that the program hurts treated children, but that treating more siblings reduces the benefits of the program. For example, the effect of being treated with two treated siblings, compared to nobody treated, can be written as $\mu(1,2) - \mu(0,0) = \mu(1,0) - \mu(0,0) + \mu(1,2) - \mu(1,0) = \tau(0) + \theta_1(2)$, so it can be estimated by $\hat{\tau}(0) + \hat{\theta}_1(2) \approx 0.113$. Thus, a treated kid with two treated siblings increases her attendance in 11 percentage points starting from a baseline in which nobody in the household is treated.

In all, the estimates suggest large and positive direct and spillover effects on the untreated, with some evidence of crowding-out between treated siblings. In addition to these results, Table 5 in the supplemental appendix shows the estimates when relaxing exchangeability and defining sibling identities based on difference in ages. The results and the formal test of equality of coefficients suggest that exchangeability cannot be rejected in this case.

6.1 Difference in means

Using the results from the nonparametric specification as a benchmark, I now estimate the effects of the program using the difference in means analyzed in Section 3.1. The left panel, "**Diff. Means**", of Table 6 shows the difference in means, calculated as the OLS estimator for β_{D} in Equation (1). The results show that the difference in means is practically zero and not significant. Hence, by ignoring the presence of spillover effects, a researcher estimating the effect of the program in this way would conclude that the treatment has no effect.

This finding is due to the fact that the difference in means combines all the effects in the third panel into a single number, as shown in Theorem 1. From Table 6, the estimated spillover effects in this case are larger under control that under treatment, and have different signs, so $\hat{\theta}_1(s) - \hat{\theta}_0(s) < 0$. Therefore, the spillover effects push the difference in means towards zero in this case.

6.2 Linear-in-means models

Next, I estimate the effects using the LIM regression analyzed in Section 3.2. The estimates from Equation (2) are given in the first column of the middle panel, "**Linear-in-Means**", in Table 6. The estimates reveal very small and not significant direct and spillover effects, substantially different from results using Equation (7).

Theorem 2 and Corollary 1 show that a LIM regression implicitly imposes linearity of spillover effects and may suffer from misspecification when spillover effects are nonlinear. The estimates from the full nonparametric specification shows that spillover effects are highly nonlinear in this case, which explains why the LIM regression fails at recovering these effects.

The second column in the LIM panel presents the estimates from the interacted LIM regression shown in Equation (3). The results reveal that separately estimating the spillover effects for treated and controls mitigates the misspecification in this case, and the estimates are closer to the ones from the nonparametric specification, although the fact that 0.169 is not a weighted average of 0.146 and 0.14 suggests that some extrapolation bias remains due to the nonlinearity of spillover effects.

6.3 Pooled effects

I now illustrate how to estimate pooled effects by averaging over the possible number of treated siblings (2 and 3 in this case). For this, I estimate the following regression:

$$Y_{ig} = \alpha_p + \beta_p D_{ig} + \gamma_p^0 \mathbb{1}(S_{ig} > 0)(1 - D_{ig}) + \gamma_p^1 \mathbb{1}(S_{ig} > 0) D_{ig} + \nu_{ig}$$

where

$$\gamma_p^d = \sum_{s=1}^{2} \theta_d(s) \mathbb{P}[S_{ig} = s | D_{ig} = d, S_{ig} > 0]$$

(see Remark 2). From Table 6 we can see that the estimated pooled spillover effects are 0.144 for controls and -0.045 for treated, which as expected lie between the effects found with the saturated regression. These results illustrate how this type of pooling can provide a useful summary of spillover effects, which may be a feasible alternative when the total number of spillover effects is too large or cell sizes are small to estimate them separately.

6.4 Non-exchangeable peers

Finally, I illustrate how to relax the exchangeability assumption. I define an ordering between siblings by looking at differences (in absolute value) in ages, defining sibling 1 as the sibling closest in age and sibling 2 as the sibling farthest in age. Then, estimation is conducted by simply adding indicator variables for the possible different assignments. Table 5 shows the estimates from this specification. The estimates reveal similar results to Table 6, with a direct effect of 0.165, spillover effects on the untreated ranging from 0.133 to 0.14 and spillover effects on the treated ranging from -0.039 to -0.051.

In fact, exchangeability can be tested by assessing whether the spillover effects of siblings 1 and 2 are the same, which in this case amounts to testing equality of coefficients between rows 2 and 3, and between rows 5 and 6. The test statistic and corresponding p-value from this test are given in the last two rows of the table, where it is clear that exchangeability cannot be rejected in this case.

Table 5: Estimation results

	coef	s.e.
D_{ig}	0.165**	0.066
$(1-D_{ig})D_{i1g}(1-D_{i2g})$	0.133**	0.067
$(1-D_{ig})(1-D_{i1g})D_{i2g})$	0.163**	0.07
$(1-D_{ig})D_{i1g}D_{i2g}$	0.14**	0.056
$D_{ig}D_{i1g}(1-D_{i2g})$	-0.039	0.027
$D_{ig}(1-D_{i1g})D_{i2g}$	-0.044*	0.026
$D_{ig}D_{i1g}D_{i2g}$	-0.051**	0.025
Constant	0.706***	0.057
Observations		504
Chi-squared test	0.444	
p-value	0.642	

Notes: Cluster-robust s.e. Regressions include school FE. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

7 Discussion

The findings in this paper offer several takeaways for analyzing spillover effects in randomized experiments. First, commonly analyzed estimands such as the difference in means and LIM

Table 6: Estimation results

	Diff. Means		**Linear-in-Means**				**Full**		**Pooled**	
	coef	s.e.	coef	s.e.	coef	s.e.	coef	s.e.	coef	s.e.
D_{ig}	0.006	0.016	0.007	0.016	0.102**	0.042	0.164**	0.066	0.165**	0.065
$\bar{D}_g^{(i)}$			0.027	0.034						
$\bar{D}_g^{(i)}(1-D_{ig})$					0.169**	0.068				
$\bar{D}_g^{(i)}D_{ig}$					-0.064	0.039				
$\mathbb{1}(S_{ig}=1)(1-D_{ig})$							0.146**	0.066		
$\mathbb{1}(S_{ig}=2)(1-D_{ig})$							0.14**	0.056		
$\mathbb{1}(S_{ig}=1)D_{ig}$							-0.041*	0.023		
$\mathbb{1}(S_{ig}=2)D_{ig}$							-0.051**	0.025		
$\mathbb{1}(S_{ig}>0)(1-D_{ig})$									0.144**	0.06
$\mathbb{1}(S_{ig}>0)D_{ig}$									-0.045**	0.02
Constant	0.822***	0.013	0.811***	0.024	0.756***	0.037	0.706***	0.057	0.705***	0.057
Observations		504		504		504		504		504

Notes: S.e. clustered at the household level. Regressions include school FE. $^{***}p<0.01,^{**}p<0.05,^{*}p<0.1$.

coefficients implicitly impose strong assumptions on the structure of spillover effects, and are therefore not recommended as they generally do not have a causal interpretation. On the other hand, the full vector of spillover effects is identifiable whenever the design generates enough variation in the number of treated units in each group and the researcher assumes a treatment rule that is flexible enough.

Second, while nonparametric estimation of all direct and spillover effects can give a complete picture of the effects of the treatment, it can be difficult to implement in practice when groups are large. As a guideline to determine in which cases spillover effects can be estimated nonparametrically, Theorem 4 formalizes the notion of a "large enough sample" in this context, and provides a way to assess the performance of the different types of treatment effect estimators depending on the number of groups, number of parameters of interest and treatment assignment mechanism. As an alternative, pooled estimands can recover weighted averages of spillover effects with known weights for which inference can be conducted under mild conditions.

The analysis in this paper leaves several questions open for future research, some of which are discussed in the supplemental appendix. Sections A3 and A4 discuss extensions to unequal group sizes and the inclusion of covariates. On the other hand, the results in Section 4 and the simulations in Section 5 highlight two issues that need to be further explored. First, the fact that the rate of convergence of the proposed estimators depend on the design suggests that these results can be used to rank treatment assignment mechanisms, and this has implications for experimental design, as discussed in Section A2. Second, simulation evidence suggests that the wild bootstrap can provide a more accurate approximation to the distribution of estimators in some cases. Future work should also formally address other issues that arise frequently in empirical studies measuring spillovers, such as imperfectly measured groups or treatment missclasification.

References

Abadie, A., and M. D. Cattaneo (2018): "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503.

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

Ao, W., S. Calonico, and Y.-Y. Lee (forthcoming): "Multivalued Treatments and Decomposition Analysis: An Application to the WIA Program," *Journal of Business & Economic Statistics*.

Athey, S., and G. Imbens (2017): "The Econometrics of Randomized Experiments," in *Handbook of Field Experiments*, ed. by A. V. Banerjee, and E. Duflo, vol. 1 of *Handbook of Economic Field Experiments*, pp. 73 – 140. North-Holland.

Barrera-Osorio, F., M. Bertrand, L. L. Linden, and F. Perez-Calle (2011): "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia," *American Economic Journal: Applied Economics*, 3(2), 167–195.

Beshears, J., J. J. Choi, D. Laibson, B. C. Madrian, and K. L. Milkman (2015): "The Effect of Providing Peer Information on Retirement Savings Decisions," *The Journal of Finance*, 70(3), 1161–1201.

Beuermann, D. W., J. Cristia, S. Cueto, O. Malamud, and Y. Cruz-Aguayo (2015): "One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru," *American Economic Journal: Applied Economics*, 7(2), 53–80.

Blume, L. E., W. A. Brock, S. N. Durlauf, and R. Jayaraman (2015): "Linear Social Interactions Models," *Journal of Political Economy*, 123(2), 444–496.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009): "Identification of peer effects through social networks," *Journal of Econometrics*, 150(1), 41 – 55.

Cattaneo, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138 – 154.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018): "Inference in Linear Regression Models with Many Covariates and Heteroskedasticity," *Journal of the American Statistical Association*, 113(523), 1350–1361.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013): "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment," *The Quarterly Journal of Economics*, 128(2), 531–580.

Davezies, L., X. D'Haultfoeuille, and D. Fougère (2009): "Identification of peer effects using group size variation," *Econometrics Journal*, 12(3), 397–413.

Duflo, E., and E. Saez (2003): "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," *The Quarterly Journal of Economics*, 118(3), 815–842.

Farrell, M. H. (2015): "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189(1), 1 – 23.

Giné, X., and G. Mansuri (2018): "Together We Will: Experimental Evidence on Female Voting Behavior in Pakistan," *American Economic Journal: Applied Economics*, 10(1), 207–235.

Hirano, K., and J. Hahn (2010): "Design of Randomized Experiments to Measure Social Interaction Effects," *Economics Letters*, 106(1), 51–53.

Imbens, G. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87(3), 706–710.

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

Kline, B., and E. Tamer (2019): "Econometric Analysis of Models with Social Interactions," in *The Econometric Analysis of Network Data*, ed. by B. S. Graham, and A. De Paula. Elsevier.

Ma, X., and J. Wang (forthcoming): "Robust Inference Using Inverse Probability Weighting," *Journal of the American Statistical Association*.

MacKinnon, J. G., and M. D. Webb (2020): "When and How to Deal with Clustered Errors in Regression Models," Working Paper 1421, Economics Department, Queen's University.

Manski, C. F. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60(3), 531–542.

——— (2013): "Identification of treatment response with social interactions," *The Econometrics Journal*, 16(1), S1–S23.

Miguel, E., and M. Kremer (2004): "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72(1), 159–217.

Moffit, R. (2001): "Policy Interventions, Low-level Equilibria and Social Interactions," in *Social Dynamics*, ed. by S. N. Durlauf, and P. Young, pp. 45–82. MIT Press.

Sacerdote, B. (2014): "Experimental and Quasi-Experimental Analysis of Peer Effects: Two Steps Forward?," *Annual Review of Economics*, 6(1), 253–272.

Shao, J., and D. Tu (1995): *The Jackknife and Bootstrap*. Springer New York, New York, NY.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York, New York, NY.

Vazquez-Bare, G. (2020): "Causal Spillover Effects Using Instrumental Variables," *arXiv:2003.06023*.