Analyzing Two-Stage Experiments in the Presence of Interference

Guillaume Basse & Avi Feller

Taylor & Francis
Taylor & Francis Group

Check for updates

Analyzing Two-Stage Experiments in the Presence of Interference

Guillaume Basse[a] and Avi Feller[b]

[a]Harvard University, Cambridge, MA; [b]University of California, Berkeley, CA

ABSTRACT

Two-stage randomization is a powerful design for estimating treatment effects in the presence of interference; that is, when one individual's treatment assignment affects another individual's outcomes. Our motivating example is a two-stage randomized trial evaluating an intervention to reduce student absenteeism in the School District of Philadelphia. In that experiment, households with multiple students were first assigned to treatment or control; then, in treated households, one student was randomly assigned to treatment. Using this example, we highlight key considerations for analyzing two-stage experiments in practice. Our first contribution is to address additional complexities that arise when household sizes vary; in this case, researchers must decide between assigning equal weight to households or equal weight to individuals. We propose unbiased estimators for a broad class of individual- and household-weighted estimands, with corresponding theoretical and estimated variances. Our second contribution is to connect two common approaches for analyzing two-stage designs: linear regression and randomization inference. We show that, with suitably chosen standard errors, these two approaches yield identical point and variance estimates, which is somewhat surprising given the complex randomization scheme. Finally, we explore options for incorporating covariates to improve precision. We confirm our analytic results via simulation studies and apply these methods to the attendance study, finding substantively meaningful spillover effects.

1. Introduction

A common assumption in causal inference is "no interference" between units: one individual's outcomes are unaffected by another individual's treatment assignment (Cox 1958; Rubin 1980). However, this assumption does not hold in settings ranging from infectious diseases to education to labor markets (for a recent review, see, Halloran and Hudgens 2016). In many of these cases, such interference is of direct substantive interest.

Two-stage randomization is a powerful design for estimating causal effects involving interference. In the setting we consider, first whole clusters (e.g., households, schools, or graph partitions) are assigned to treatment or control. Second, units within each treated cluster are randomly assigned to treatment or control, as if each treated cluster were a separate, individually-randomized experiment. This design allows researchers to assess spillover effects either by comparing untreated units in treated clusters with pure control units in control clusters or by comparing units across clusters with different proportions assigned to treatment (Hudgens and Halloran 2008).

Our motivating example is a large randomized evaluation of an intervention targeting student absenteeism among elementary and high school students in the School District of Philadelphia (Rogers and Feller 2018). In the original study, parents of at-risk students were randomly assigned to a direct mail intervention with tailored information about their students' attendance over the course of the year. In treated households with multiple eligible students, one student was selected at random to be the subject of the mailings, following a two-stage

randomization. Substantively, this is a rare opportunity to study intra-household dynamics around student behavior. Methodologically, this presents a rich test case for understanding how to analyze two-stage experiments in practice.

There has been substantial interest in two-stage randomization in recent years, with prominent examples in economics (Crépon et al. 2013), education (Somers et al. 2010), political science (Sinclair, McConnell, and Green 2012), and public health (Hudgens and Halloran 2008), as well as closely related variants in the context of large-scale social networks (Ugander et al. 2013). Such designs have become especially common in development economics (Angelucci and Di Maro 2016). There is also a small but growing methodological literature on analyzing two-stage experiments, including Hudgens and Halloran (2008), Liu and Hudgens (2014), and Rigdon and Hudgens (2015) in statistics; and Sinclair, McConnell, and Green (2012) and Baird et al. (2018) in the social sciences.

We build on this literature by addressing three practical issues that arise in analyzing the attendance study. First, school districts are typically interested in the intervention's impact on students rather than on households; that is, districts give equal weight to each individual rather than equal weight to each household. Similarly, public health researchers administering treatment to villages of different sizes might be interested in the impact on the overall population rather than on village-level averages, especially if the treatment is more effective in larger villages. With the exception of Sinclair, McConnell, and Green (2012), however, existing approaches focus either on equal

---

weights for households (e.g., Hudgens and Halloran 2008) or side-step the issue by assuming households are of equal size (e.g., Baird et al. 2018). We propose unbiased estimators for a broad class of individual- and household-weighted estimands, with corresponding theoretical and estimated variances. We also derive the bias of a simple difference in means for estimating individual-weighted estimands. Since researchers typically estimate these two estimands with different precision (see, e.g., Athey and Imbens 2017), we recommend that researchers report both in practice.

Second, we connect two common approaches for analyzing two-stage designs: linear regression, which is more common in the social sciences, and randomization inference, which is more common in epidemiology and public health. We show that, with suitably chosen standard errors, regression and randomization inference yield identical point and variance estimates. These results hold for a broad class of weighted estimands. We believe this equivalence will be important in practice, since the vast majority of applied papers in this area take a "regression first" approach to analysis that can obfuscate key inferential issues.

Finally, we explore options for incorporating covariates to improve precision, with a focus on post-stratification and model-assisted estimation. We then confirm our analytic results via simulation studies and apply these methods to the attendance study. Overall, we find strong evidence of a spillover effect that is (depending on the scale of the outcome) roughly 60–80% as large as the primary effect. This holds across different estimands as well as with and without covariate adjustment. Accounting for spillovers therefore dramatically improves the cost effectiveness of the intervention, from around $5 per additional day to around $3 per additional day.

This article proceeds as follows. Section 2 defines the two-stage randomization, sets up the notation, and discusses the relevant assumptions. Section 3 defines the estimands of interest both for constant and varying household sizes. Sections 4 and 5 deal with unbiased estimators and their variances. Section 6 demonstrates how we can use regression with appropriate standard errors to obtain the randomization-based estimators. Section 7 explores covariate adjustment. Section 8 reports the results of extensive simulation studies. Section 9 analyzes the student attendance experiment. Section 10 concludes and offers directions for future work. The supplementary materials contains additional technical material and all proofs.

1.1. Motivating Example: Student Absenteeism

More than 10% of public school students in the United States—over five million students—are chronically absent each year, defined as missing 18 or more days of the roughly 180-day school year (ED Office for Civil Rights 2016). Rogers and Feller (2018) recently conducted the first randomized evaluation of an intervention aimed at reducing student absenteeism for this population. This intervention delivered targeted information to parents of at-risk students in the School District of Philadelphia via five pieces of direct mail over the 2014–2015 School Year. The mailing clearly stated the student's number of absences that year ("Your student has been absent 16 days this school year"), included a simple bar chart showing the same information graphically, and gave additional text on the importance of attending school. Rogers and Feller (2018) found

Table 1. Number of households by size, and proportion of treated households for each size

	2	3	4–7
Total N	3169	557	150
Proportion assigned to treatment	0.66	0.65	0.65

that the treatment reduces chronic absenteeism by over 10% relative to control. The approach is extremely cost-effective, costing around $5 per additional day of student attendance—more than an order of magnitude more cost-effective than the current best-practice intervention.

A key practical challenge in implementing the original study was that the mailings were designed to provide information about a single student. Students were eligible to be the target of the intervention if they met certain prespecified criteria, including type of school, home language, and no perfect attendance in the previous year. In households with multiple eligible students, one student was randomly selected to be the focal student. (The study excludes other members of the household, such as noneligible siblings.) Rogers and Feller (2018) addressed possible spillover only briefly in the original study, largely because the focus was on the primary effect of the intervention and because households with multiple students were around 15 percent of all households in the sample.

In this article, we consider a subset of $N = 3876$ households with between $n_i = 2$ and $n_i = 7$ eligible students in each household and $n^+ = 8654$ total students. Table 1 shows the distribution of household size. The vast majority of these households (82%) have only two students; only one percent (35 households) have five or more students. Our goal is to estimate the primary and spillover effects on attendance for this finite sample. We are also interested in the extent to which these estimates differ depending on whether we give equal weight to each household or to each individual. The original study estimated these effects with a simple difference-in-means estimator, which could be biased in practice; see Section 4.2.

This experimental design presents a rare opportunity to assess intra-household spillovers. There is substantial evidence across fields that such intra-household spillovers are meaningful in magnitude. For example, several voter mobilization studies have found spillover effects that are between one-third and two-thirds as large as the primary effect (Nickerson 2008; Sinclair, McConnell, and Green 2012). We are interested in spillover in the attendance study for two key reasons. First, ignoring the spillover effect under-states the overall impact of the intervention. For example, an important metric is the cost of each additional student day; ignoring spillover artificially lowers the corresponding cost-effectiveness estimates. Second, the research team faced a practical question of whether to develop a separate intervention for households with multiple eligible students, which would be costly to implement and test. If the spillover effect is comparable in magnitude to the primary effect, such development is unnecessary. This is similar to decisions around interventions targeting infectious diseases (Hudgens and Halloran 2008) and in economics (Baird et al. 2018).

2. Setup and Assumptions

We now review the setup and assumptions for a two-stage experiment in the presence of interference. The discussion

closely follows Hudgens and Halloran (2008), modifying their terminology slightly to better fit our applied example and to recognize some small differences in emphasis in the social science literature. We first define potential outcomes and state the relevant assumptions, then follow with a description of two-stage randomized designs. We postpone the formal introduction of our estimands to Section 3. For additional reviews on causal inference under interference, see, among others, Sinclair, McConnell, and Green (2012), Bowers, Fredrickson, and Panagopoulos (2013), VanderWeele, Tchetgen Tchetgen, and Halloran (2014), Halloran and Hudgens (2016), Aronow and Samii (2017), and Athey and Imbens (2017).

2.1. Potential Outcomes and Relaxing SUTVA

We use the potential outcomes framework to describe the problem (Neyman 1923; Rubin 1974). Consider N households $i = 1, \ldots, N$ with n_i individuals in household i, and where $n^+ \equiv \sum n_i$ is the total number of individuals. To be consistent with the existing literature on two-stage experiments, we use the double-index notation, such that $\cdot_{ij}$ denotes the individual j in household i. For household i, let $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{in_i})$ denote the assignment vector for the n_i units in that household, where $Z_{ij} = 1$ if the jth individual in household i is assigned to treatment, and $Z_{ij} = 0$ otherwise. Similarly, define $\boldsymbol{Z}_{i,-j}$ as the sub-vector of $\boldsymbol{Z}_i$ that excludes the jth value. Finally, aggregate all household-level assignments via $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N\}$. Let Y_{ij}^{obs} denote the observed outcome for individual j in household i, which will be either binary or continuous in our motivating example. In general, let $\boldsymbol{Y}_i(\boldsymbol{Z}) = (\boldsymbol{Y}_{i1}(\boldsymbol{Z}), \ldots, \boldsymbol{Y}_{in_i}(\boldsymbol{Z}))$ be the vector of potential outcomes for household i, and $\boldsymbol{Y}(\boldsymbol{Z}) = \{\boldsymbol{Y}_1(\boldsymbol{Z}), \ldots, \boldsymbol{Y}_N(\boldsymbol{Z})\}$ be the list of potential outcome vectors for all households.

At this stage, practical inference is infeasible without imposing additional restrictions on the structure of potential outcomes. The standard stable unit treatment value assumption (SUTVA; Rubin 1980), which implies that there is no interference between units, is inappropriate in our context. We instead focus on putting structure on two types of interference: between-household and within-household interference.

2.1.1. Between-Household Interference

First, we assume that there is no between-household interference, which Sobel (2006) referred to as *partial interference*.

Assumption 2.1 (No Interference Across Households). Interference occurs only within a household. That is, $\boldsymbol{Y}_i(\boldsymbol{Z}) = \boldsymbol{Y}_i(\boldsymbol{Z}_i)$.

This is effectively a "between household SUTVA" assumption and greatly reduces the complexity of the problem. In the context of the attendance study, this assumption states that students in different households do not affect each others' attendance. This assumption is violated if, for instance, friends skip school together. Nonetheless, we view spillovers within households as far more important than spillovers between households and thus consider Assumption 2.1 to be a useful approximation.

2.1.2. Within-Household Interference

Even with the partial interference assumption, practical inference remains challenging. The key complication is that, without additional restrictions, the potential outcomes depend on the identity of the treated individual. To see this, consider household i with three students in which only one student is assigned to treatment. Under partial interference, the oldest student, $j = 1$, has three potential outcomes, $Y_{i1}(1, 0, 0)$, $Y_{i1}(0, 1, 0)$, and $Y_{i1}(0, 0, 1)$, which correspond to assigning the oldest, middle, and youngest student to treatment, respectively, as well as $Y_{i1}(0, 0, 0)$ if none receive treatment. Thus, the oldest student actually has *two* different spillover effects, $Y_{i1}(0, 1, 0) - Y_{i1}(0, 0, 0)$ and $Y_{i1}(0, 0, 1) - Y_{i1}(0, 0, 0)$, depending on which other student in the household receives the treatment.

As Hudgens and Halloran (2008) argued, this makes inference difficult, especially with respect to variance calculations (see also Tchetgen Tchetgen and VanderWeele 2012). Instead, Hudgens and Halloran (2008) proposed the *stratified interference* assumption, which states that the precise identity of the treated individual in the treated cluster does not matter for untreated individuals in the same cluster (see also Manski 2013).

Assumption 2.2 (Stratified Interference).

$$Y_{ij}(\boldsymbol{Z}_{i,-j}, Z_{ij} = 0) = Y_{ij}(\boldsymbol{Z}'_{i,-j}, Z_{ij} = 0)$$

$$\forall \boldsymbol{Z}_{i,-j}, \boldsymbol{Z}'_{i,-j} \quad \text{s.t.} \quad \sum_j^{n_i} Z_{ij} = \sum_j^{n_i} Z'_{ij} = 1. \quad (1)$$

Heuristically, this imposes additional structure on the problem by assuming that potential outcomes are only a function of the number (or, depending on context, proportion) of individuals assigned to treatment within each household.

2.2. Assignment Mechanism and Observed Outcomes

Two-stage randomization is a special case of a multi-stage, nested randomization (see, e.g., Sobel 2006) that is used to assign treatment to units in a nested structure (throughout, we will refer to individuals nested within household). Figure 1 highlights the sequential nature of the two-stage design we consider. Specifically, we consider designs in which each stage follows complete randomization; that is, randomizations in which a fixed number of units are assigned to treatment at each stage (Imbens and Rubin 2015). See also Liu and Hudgens (2014), who refer to this as a permutation randomization, and Tchetgen Tchetgen and VanderWeele (2012), who contrast completely randomized and Bernoulli designs in the second stage.

Formally, let $\boldsymbol{H} = (H_1, \ldots, H_N)$ be the vector of treatment assignments at the household level, such that $H_i = 1$ if household i is assigned to treatment and $H_i = 0$ otherwise. For the first stage of randomization, we assume that a fixed integer of households $N_1 \in \{1, \ldots, N - 1\}$ are assigned to treatment, with $P(\boldsymbol{H}) = 1/\binom{N}{N_1}$ for all $\boldsymbol{H}$ such that $\sum_i^N H_i = N_1$ and $P(\boldsymbol{H}) = 0$ otherwise. Analogously define $N_0 = N - N_1$ as the number of households assigned to control. For the second stage, individuals are assigned to treatment conditional on the realized value of the first stage of randomization. For individuals in households
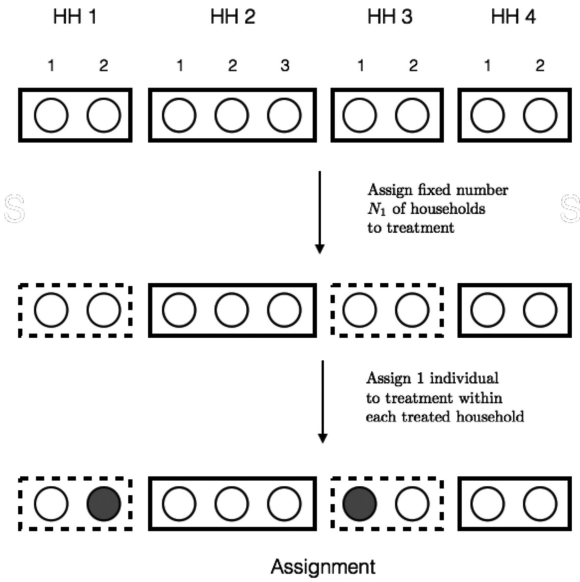
Figure 1. Schematic of the two-stage randomized design we consider. Treated households are represented by dashed bold rectangles. Treated individuals are filled circles.

with $H_i = 1$, we assume that exactly one individual is assigned to treatment, with $P(\mathbf{Z}_i \mid H_i = 1) = \frac{1}{n_i}$ if $\sum_j^{n_i} Z_{ij} = 1$ and $P(\mathbf{Z}_i \mid H_i = 1) = 0$ otherwise. For individuals in households with $H_i = 0$, $Z_{ij} = 0$ for all j. Thus, all individuals assigned to treatment are in households assigned to treatment.

Given the assumptions in Section 2.1 and the type of assignments allowed by the two-stage randomization mechanism we consider, the potential outcomes for individual j in household i simplify to $Y_{ij}(\mathbf{Z}) = Y_{ij}(H_i = h, Z_{ij} = z)$. There are three possible combinations: $Y_{ij}(1, 1)$, $Y_{ij}(1, 0)$, and $Y_{ij}(0, 0)$. We regard these potential outcomes as fixed and define the observed outcome as a deterministic function of the treatment assignment and potential outcomes

$$Y_{ij}^{\text{obs}} = H_i Z_{ij} Y_{ij}(1,1) + H_i(1 - Z_{ij})Y_{ij}(1,0) + (1 - H_i)Y_{ij}(0,0),$$

where the randomness is entirely due to $\mathbf{H}$ and $\mathbf{Z}$. That is, unless otherwise stated, all expectations and variances are with respect to the randomization distribution; inference is fully justified by the randomization itself (Fisher 1935). Finally, we introduce the sets $\mathcal{T}_{hz} = \{(i, j) : H_i = h \text{ and } Z_{ij} = z\}$ to denote the set of households and individuals who are assigned to $H_i = h$ and $Z_{ij} = z$.

3. Estimands

We next discuss estimands of interest, closely following Hudgens and Halloran (2008). We start with the setting in which all households or clusters are of equal size and then turn to households of varying size.

3.1. Constant Household Size

3.1.1. Primary and Spillover Effects
We first assume that all households are of the same size. That is, $n_i = n$ for all households $i = 1, \ldots, N$. Next, we define average potential outcomes at the household level,

$\overline{Y}_i(h, z) = \frac{1}{n} \sum_j^n Y_{ij}(h, z)$, which average across all individuals $j = 1, \ldots, n$ within each household. Since all households are the same size, we can analogously define the average potential outcomes for the sample, $\overline{Y}(h, z) = \frac{1}{N} \sum_i^N \overline{Y}_i(h, z)$, which is the average household-level potential outcome across all households. Estimands are contrasts between these sample average potential outcomes. Unless otherwise stated, all estimands we consider here are finite sample estimands; that is, they are defined for the units in our sample.

Definition 3.1 (Estimands with equal-sized households). Define the average *primary effect* as follows

$$\tau^P = \frac{1}{Nn} \sum_i^N \sum_j^n (Y_{ij}(1,1) - Y_{ij}(0,0))$$

$$= \frac{1}{N} \sum_i^N (\overline{Y}_i(1,1) - \overline{Y}_i(0,0)) = \overline{Y}(1,1) - \overline{Y}(0,0), \quad (2)$$

and the average *spillover effect* as

$$\tau^S = \frac{1}{Nn} \sum_i^N \sum_j^n (Y_{ij}(1,0) - Y_{ij}(0,0))$$

$$= \frac{1}{N} \sum_i^N (\overline{Y}_i(1,0) - \overline{Y}_i(0,0)) = \overline{Y}(1,0) - \overline{Y}(0,0). \quad (3)$$

These estimands have various names in the literature. We take the terminology *primary effect* from Toulis and Kao (2013), but Baird et al. (2018) used the term *treatment on the uniquely treated* for an analogous quantity. Hudgens and Halloran (2008) call these estimands the *total* and *indirect* effects, respectively.

3.1.2. Other Estimands
Hudgens and Halloran (2008) proposed two additional estimands, the *direct* and *overall* effects. The direct effect is essentially the impact of the second stage, within-household randomization, which is $\tau^D = \overline{Y}(1,1) - \overline{Y}(1,0)$ with equal-sized households. Following the effect decomposition in Hudgens and Halloran (2008), the direct effect can be defined as the difference between the primary and spillover effects. Since the spillover effect is a more natural estimand in our setting, we do not discuss the direct effect further.

The overall effect is the impact of household-level random assignment. From a policy perspective, this quantity is of obvious interest to decision makers, providing a single number for the intervention's impact. From a statistical perspective, however, the overall effect has the somewhat awkward feature that it is only defined with respect to a given assignment mechanism

$$\tau^O = \frac{1}{Nn} \sum_i^N \sum_j^n (E[Y_{ij}(1, Z_{ij})] - Y_{ij}(0,0)), \quad (4)$$

where $E[Y_{ij}(1, Z_{ij})]$ is the expected potential outcome for individual j in household i

$$E[Y_{ij}(1, Z_{ij})] = P(Z_{ij} = 1 \mid H_i = 1)Y_{ij}(1,1)$$
$$+ P(Z_{ij} = 0 \mid H_i = 1)Y_{ij}(1,0).$$

Similar to VanderWeele and Tchetgen Tchetgen (2011), we can therefore rewrite the overall effect as a weighted average of the primary and spillover effects

$$\tau^O = P(Z_{ij} = 1 \mid H_i = 1)\tau^P + P(Z_{ij} = 0 \mid H_i = 1)\tau^S,$$

with weights equal to the second stage treatment probability. As a result, we focus on the primary and spillover effects throughout the main text and defer corresponding results for the overall effect to the supplementary materials.

3.2. Varying Household Size

We now generalize these results to allow for varying household size. Broadly, there are now two types of estimands, *household-weighted* estimands (HW) that assign equal weight to households, regardless of the number of individuals in each household; and *individual-weighted* estimands (IW) that assign equal weight to individuals, regardless of the distribution across households. A substantial literature on cluster-randomized trials addresses related questions; see, among others, Donner and Klar (2000), Imai, King, and Nall (2009), Schochet (2013), and Middleton and Aronow (2015). Specifically, we generalize the equal-sized household estimands to allow for *two-stage weights*.

Definition 3.2 (Two-stage weighted estimands). Define the average *primary effect* as follows

$$\tau_W^P = \sum_{i=1}^N w_i^* \sum_{j=1}^{n_i} (Y_{ij}(1,1) - Y_{ij}(0,0)), \qquad (5)$$

and the average *spillover effect* as

$$\tau_W^S = \sum_{i=1}^N w_i^* \sum_{j=1}^{n_i} (Y_{ij}(1,0) - Y_{ij}(0,0)), \qquad (6)$$

where $w_i^* = \frac{1}{N n_i}$ corresponds to household-weighted estimands and $w_i^* = \frac{1}{n^+}$ corresponds to individual-weighted estimands.

When $n_i = n$ for all i, both HW and IW estimands are identical to the equal-sized household estimands. When n_i is not constant, the resulting household weighted estimands are

$$\tau_{\text{HW}}^P = \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij}(1,1) - Y_{ij}(0,0)) \quad \text{and}$$

$$\tau_{\text{HW}}^S = \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij}(1,0) - Y_{ij}(0,0)).$$

These are the estimands in Hudgens and Halloran (2008). The corresponding individual weighted estimands are

$$\tau_{\text{IW}}^P = \frac{1}{n^+} \sum_i^N \sum_j^{n_i} (Y_{ij}(1,1) - Y_{ij}(0,0)) \quad \text{and}$$

$$\tau_{\text{IW}}^S = \frac{1}{n^+} \sum_i^N \sum_j^{n_i} (Y_{ij}(1,0) - Y_{ij}(0,0)),$$

where n^+ is the total number of individuals.

4. Estimation

Next, we turn to estimating these quantities of interest. First, we generalize the results of Hudgens and Halloran (2008) to allow for unbiased estimation of any two-stage weighted estimand. We then discuss additional complications that arise when estimating individual-weighted effects.

4.1. Unbiased Estimation

We now define two-stage weights for estimation, of which household and individual weights are special cases.

Proposition 4.1 (Unbiasedness of two-stage weighted estimators). Define two-stage inverse probability weights $w_i^{(00)}$, $w_i^{(10)}$, and $w_i^{(11)}$ as follows

$$w_i^{(11)} = \frac{1}{P(H_i = 1)} \frac{1}{P(Z_{ij} = 1 \mid H_i = 1)},$$

$$w_i^{(10)} = \frac{1}{P(H_i = 1)} \frac{1}{P(Z_{ij} = 0 \mid H_i = 1)},$$

$$w_i^{(00)} = \frac{1}{P(H_i = 0)}.$$

Consider two-stage estimand weights, w_i^*, as in Definition 3.2. The weighted primary and spillover effect estimators $\hat{\tau}_W^P$ and $\hat{\tau}_W^S$,

$$\hat{\tau}_W^P = \sum_{(i,j) \in \mathcal{T}_{11}} w_i^{(11)} w_i^* Y_{ij}^{\text{obs}}(1,1) - \sum_{(i,j) \in \mathcal{T}_{00}} w_i^{(00)} w_i^* Y_{ij}^{\text{obs}}(0,0),$$

$$\hat{\tau}_W^S = \sum_{(i,j) \in \mathcal{T}_{10}} w_i^{(10)} w_i^* Y_{ij}^{\text{obs}}(1,0) - \sum_{(i,j) \in \mathcal{T}_{00}} w_i^{(00)} w_i^* Y_{ij}^{\text{obs}}(0,0),$$

are unbiased for their corresponding estimands with respect to the randomization distribution. That is $\mathbb{E}[\hat{\tau}_W^P] = \tau_W^P$ and $\mathbb{E}[\hat{\tau}_W^S] = \tau_W^S$.

The proof is given in the supplementary materials and follows closely from Hudgens and Halloran (2008). These estimators have a simple difference-in-means form either when household size is constant or with equal weight on households, as in Hudgens and Halloran (2008),

$$\hat{\tau}_{\text{HW}}^P = \frac{1}{N_1} \sum_{i \in \mathcal{T}_{11}} \overline{Y}_i^{\text{obs}}(1,1) - \frac{1}{N_0} \sum_{i \in \mathcal{T}_{00}} \overline{Y}_i^{\text{obs}}(0,0) \quad \text{and}$$

$$\hat{\tau}_{\text{HW}}^S = \frac{1}{N_1} \sum_{i \in \mathcal{T}_{10}} \overline{Y}_i^{\text{obs}}(1,0) - \frac{1}{N_0} \sum_{i \in \mathcal{T}_{00}} \overline{Y}_i^{\text{obs}}(0,0). \qquad (7)$$

The unbiased, individual weighted estimators have the form of Horvitz–Thompson estimators modified for our two-stage randomization. Thus the unbiased estimators for τ_{IW}^P and τ_{IW}^S are

$$\hat{\tau}_{\text{IW}}^P = \frac{1}{n^+} \frac{N}{N_1} \sum_{(i,j) \in \mathcal{T}_{11}} \frac{n_i}{1} Y_{ij}^{\text{obs}}(1,1) \qquad - \frac{1}{n^+} \frac{N}{N_0} \sum_{(i,j) \in \mathcal{T}_{00}} Y_{ij}^{\text{obs}}(0,0),$$

$$\hat{\tau}_{\text{IW}}^S = \frac{1}{n^+} \frac{N}{N_1} \sum_{(i,j) \in \mathcal{T}_{10}} \frac{n_i}{n_i - 1} Y_{ij}^{\text{obs}}(1,0) \qquad - \frac{1}{n^+} \frac{N}{N_0} \sum_{(i,j) \in \mathcal{T}_{00}} Y_{ij}^{\text{obs}}(0,0),$$

with household-level assignment probabilities N_1/N and N_0/N and (conditional) individual-level probabilities $1/n_i$ and $(n_i - 1)/n_i$. Since these are inverse probability weight estimators and since the probability of treatment assignment is a function of household size, the primary effect estimator up-weights larger households, with weights proportional to $n_i = \{1/n_i\}^{-1}$. Similarly, the spillover effect estimator down-weights larger households, with weights proportional to $n_i/(n_i - 1) = \{(n_i - 1)/n_i\}^{-1}$.

As is common with Horvitz–Thompson estimators, unbiased estimation typically comes at the price of additional variance. In practice, researchers can often reduce this variance by first normalizing the weights (i.e., Hájek weights), which introduces some small bias. See the supplementary materials for more details on the Hájek estimator in this context.

4.2. Bias of the Simple Difference Estimator

Contrast the unbiased estimator with a simple difference estimator, that is, the difference-in-means across individuals ignoring households

$$\widehat{\tau}_{sd}^{P} = \frac{1}{n_{11}^{+}} \sum_{(i,j) \in \mathcal{T}_{11}} Y_{ij}^{obs}(1,1) - \frac{1}{n_{00}^{+}} \sum_{(i,j) \in \mathcal{T}_{00}} Y_{ij}^{obs}(0,0) \quad (8)$$

$$\widehat{\tau}_{sd}^{S} = \frac{1}{n_{10}^{+}} \sum_{(i,j) \in \mathcal{T}_{10}} Y_{ij}^{obs}(1,0) - \frac{1}{n_{00}^{+}} \sum_{(i,j) \in \mathcal{T}_{00}} Y_{ij}^{obs}(0,0), \quad (9)$$

where $n_{11}^{+} = \sum_i \mathbb{1}(H_i = 1) \sum_j \mathbb{1}(Z_{ij} = 1)$, $n_{10}^{+} = \sum_i \mathbb{1}(H_i = 1) \sum_j \mathbb{1}(Z_{ij} = 0)$, and $n_{00}^{+} = \sum_i \mathbb{1}(H_i = 0)n_i$. These are essentially the estimators used in Rogers and Feller (2018).

Despite its intuitive appeal, this estimator can be biased in practice. There are two main sources of bias. First, echoing results from Middleton and Aronow (2015), when household sizes vary, the quantities n_{11}^{+}, n_{10}^{+}, and n_{00}^{+} are themselves random variables. Thus, both the numerator and denominator of each group average are random; and the mean of a ratio is not, in general, equal to the ratio of means. Second, individual-level treatment probabilities vary by household size; in the design we consider here, the probability of treatment assignment conditional on being in a treated household is $\mathbb{P}\{Z_{ij} = 1 \mid H_i = 1\} = 1/n_i$. Thus, ignoring n_i—and, by extension, the varying treatment probability—can lead to biased estimates. We derive the exact form of the bias in the following proposition.

Proposition 4.2. The simple difference estimators, $\widehat{\tau}_{sd}^{P}$ and $\widehat{\tau}_{sd}^{S}$, defined in Equations (8) and (9), have the following bias for their respective estimands.

$$\text{bias}\left(\widehat{\tau}_{sd}^{P}\right) = \frac{1}{N\bar{n}} \sum_i \left(\frac{\bar{n}}{n_i} - 1\right) \sum_j Y_{ij}(1,1)$$

$$+ \frac{1}{N_0\bar{n}} \text{cov}\left(\frac{\sum_{\mathcal{T}_{00}} Y_{ij}(0,0)}{n_{00}^{+}}, n_{00}^{+}\right) \quad (10)$$

and

$$\text{bias}\left(\widehat{\tau}_{sd}^{S}\right) = \frac{1}{N\bar{n}} \sum_i \left(\frac{\frac{\bar{n}}{\bar{n}-1}}{\frac{n_i}{n_i-1}} - 1\right) \sum_j Y_{ij}(1,0)$$

$$+ \left(\frac{1}{N_0\bar{n}} \text{cov}\left(\frac{\sum_{\mathcal{T}_{00}} Y_{ij}(0,0)}{n_{00}^{+}}, n_{00}^{+}\right)\right.$$

$$\left. - \frac{1}{N_1(\bar{n}-1)} \text{cov}\left(\frac{\sum_{\mathcal{T}_{10}} Y_{ij}(1,0)}{n_{10}^{+}}, n_{10}^{+}\right)\right). \quad (11)$$

If household size is constant, all of these terms are zero. If the covariance between household size and potential outcomes is zero, only the first term of each equation remains. In simulations in Section 8, we show that the overall bias can be large if household sizes vary and treatment effects also vary by household size.

4.3. Stratification and Post-Stratification by Household Size

Finally, we consider stratification and post-stratification by household size. If household-level randomization is stratified by household size, inference for the individual-weighted estimand is immediate. In particular, let τ_k^P and τ_k^S be the stratum-specific estimands for the stratum with household size $n_i = k$,

$$\tau_k^P = \frac{1}{N^{(k)}} \sum_i^N \mathbb{1}(n_i = k) \frac{1}{k} \sum_j^k (Y_{ij}(1,1) - Y_{ij}(0,0)) \quad \text{and}$$

$$\tau_k^S = \frac{1}{N^{(k)}} \sum_i^N \mathbb{1}(n_i = k) \frac{1}{k} \sum_j^k (Y_{ij}(1,0) - Y_{ij}(0,0)),$$

where $N^{(k)}$ is the number of households of size k. Since household size is constant within each stratum, the corresponding household- and individual-weighted estimands are equivalent. We can therefore rewrite the overall individual-weighted estimands as weighted averages of the stratum-specific effects,

$$\tau_{IW}^P = \sum_{k=2}^K \frac{n^{(k)+}}{n^+} \tau_k^P \quad \text{and} \quad \tau_{IW}^S = \sum_{k=2}^K \frac{n^{(k)+}}{n^+} \tau_k^S,$$

where $n^{(k)+} = \sum_{i:n_i=k} n_i$ and where we assume (without essential loss of generality) that household sizes range from $k = 2, \ldots, K$. Plugging in $\widehat{\tau}_k^P$ and $\widehat{\tau}_k^S$ gives the corresponding unbiased estimate.

To modify the above results for household-weighted estimates, simply replace the weight $n^{(k)+}/n^+$ with $N^{(k)}/N$. In other words, weight each stratum by the number of households in that stratum, rather than the number of individuals. While stratification is not necessary to obtain unbiased estimates of household-weighted estimands, stratification will generally improve precision so long as household size is predictive of the outcome.

In practice, it is not always possible or feasible to stratify randomization by household size. Fortunately, researchers can often post-stratify by household size; that is, the researcher can analyze the experiment as if randomization had been stratified by size. In the supplementary materials, we extend the theoretical guarantees from Miratrix, Sekhon, and Yu (2013) to

two-stage randomization, and include additional discussion of the technical details.

Finally, if there are relatively small samples or the distribution of household sizes varies widely, researchers might want to "mix and match" among possible strategies. In the attendance study, there are 3169 households of size $n_i = 2$ but only two households of size $n_i = 7$. Thus, it is unreasonable to post-stratify precisely on household size. Instead, we post-stratify by dividing household size into $n_i \in \{2, 3, 4-7\}$, using the unbiased IW estimator for households of size four to seven. This is inherently a bias-variance tradeoff and will depend on the particular context. If desired, we could also adjust for n_i via regression, as discussed in Section 7 (see also Middleton and Aronow 2015). Of course, researchers should pre-specify such procedures whenever possible.

5. Variance

We next provide the theoretical variance of the unbiased, weighted estimators of Section 4.1 as well as a conservative estimator of that variance. We conclude with a brief discussion of inference given a point estimate and its estimated variance.

5.1. Theoretical and Estimated Variance

We give general results for the variance of the unbiased, weighted estimator, of which the household and individual weights are special cases. Given estimand weights w_i^*, define the transformed potential outcomes as $Y_{ij}^w(h, z) \equiv N n_i w_i^* Y_{ij}(h, z)$. For the household weights, the transformed and original potential outcomes are identical, $Y_{ij}^w(h, z) = Y_{ij}(h, z)$. For the individual weights, the transformed potential outcomes are rescaled by the relative household size, $Y_{ij}^w(h, z) = (n_i/\bar{n}) \cdot Y_{ij}(h, z)$, where $\bar{n}$ is the average household size.

We now define several useful terms, effectively decomposing the overall variance of the transformed potential outcomes into a within- and between-household variance. Let $\sigma_{i,hz}^{2,w} = 1/n_i \sum_j (Y_{ij}^w(h, z) - \bar{Y}_i^w(h, z))^2$ be the within-household potential outcome variances for $Y_{ij}^w(h, z)$, and let $\Sigma_{11}^w = \frac{1}{N} \sum_i \sigma_{i,11}^{2,w}$ and $\Sigma_{10}^w = \frac{1}{N} \sum_i \frac{1}{(n_i-1)^2} \sigma_{i,10}^{2,w}$ be the (rescaled) average within-cluster variances for $Y_{ij}^w(1, 1)$ and $Y_{ij}^w(1, 0)$ respectively. Finally, define the between-cluster variance of cluster-level averages

$$V_{hz}^w = \frac{1}{N-1} \sum_i^N (\bar{Y}_i^w(h, z) - \bar{Y}^w(h, z))^2,$$

$$V_P^w = \frac{1}{N-1} \sum_i^N ([\bar{Y}_i^w(1, 1) - \bar{Y}_i^w(0, 0)] \\ - [\bar{Y}^w(1, 1) - \bar{Y}^w(0, 0)])^2,$$

$$V_S^w = \frac{1}{N-1} \sum_i^N ([\bar{Y}_i^w(1, 0) - \bar{Y}_i^w(0, 0)] \\ - [\bar{Y}^w(1, 0) - \bar{Y}^w(0, 0)])^2,$$

where V_{hz}^w is the between-cluster variance of the average cluster-level potential outcome, $\bar{Y}_i^w(h, z)$. V_P^w and V_S^w are the (unidentifiable) cluster-level treatment variances of the primary and spillover effects, respectively.

Proposition 5.1 (Theoretical variance of the two-stage weighted estimators). The two-stage weighted estimators have the following variances under the randomization distribution

$$\mathrm{var}(\hat{\tau}_W^P) = \frac{\Sigma_{11}^w + V_{11}^w}{N_1} + \frac{V_{00}^w}{N_0} - \frac{V_P^w}{N}$$

and

$$\mathrm{var}(\hat{\tau}_W^S) = \frac{\Sigma_{10}^w + V_{10}^w}{N_1} + \frac{V_{00}^w}{N_0} - \frac{V_S^w}{N}.$$

This variance has the same form as the standard Neymanian variance. However, the increased variance due to the two-level randomization is reflected in the first numerator, which has two terms instead of one. Intuitively, this is a decomposition of the marginal variance of potential outcomes into Σ_{hz}^w, the average of the within-household variances, and V_{hz}^w, the variance of the household-level average potential outcomes.

We can obtain an estimated variance that is a "conservative" estimate for the true variance (in the sense of being too wide in expectation) with respect to the randomization distribution. Let $s_{hz}^{2,w}$ be the cluster-level sample variance for the cluster-level average transformed potential outcomes, $\bar{Y}_i^{\mathrm{obs}, w}(h, z)$. That is,

$$s_{hz}^{2,w} = \frac{1}{N_h - 1} \sum_i^N \mathbb{1}(H_i = h) \left(\bar{Y}_i^{\mathrm{obs}, w}(h, z) - \bar{Y}^{\mathrm{obs}, w}(h, z) \right)^2,$$

where $\bar{Y}^{\mathrm{obs}, w}(h, z)$ is the average transformed observed outcome for the set $\mathcal{T}_{hz}$, and where $\bar{Y}_i^{w, \mathrm{obs}}(h, z)$ is the average observed outcome for the set $\mathcal{T}_{hz}$ in household i.

Theorem 5.1 (Estimated variance of the two-stage weighted estimators). Consider the variance estimators $\widehat{\mathrm{var}}(\hat{\tau}_W^P)$ and $\widehat{\mathrm{var}}(\hat{\tau}_W^S)$

$$\widehat{\mathrm{var}}(\hat{\tau}_W^P) = \frac{s_{11}^{2,w}}{N_1} + \frac{s_{00}^{2,w}}{N_0}, \tag{12}$$

$$\widehat{\mathrm{var}}(\hat{\tau}_W^S) = \frac{s_{10}^{2,w}}{N_1} + \frac{s_{00}^{2,w}}{N_0}. \tag{13}$$

The proposed estimators are conservative estimates of their respective estimands. That is, $\mathbb{E}(\widehat{\mathrm{var}}(\hat{\tau}_W^P)) \geq \mathrm{var}(\hat{\tau}_W^P)$ and $\mathbb{E}(\widehat{\mathrm{var}}(\hat{\tau}_W^S)) \geq \mathrm{var}(\hat{\tau}_W^S)$. $\widehat{\mathrm{var}}(\hat{\tau}_W^P)$ and $\widehat{\mathrm{var}}(\hat{\tau}_W^S)$ are unbiased if $V_P^w = 0$ and $V_S^w = 0$, respectively.

The results in Proposition 5.1 and Theorem 5.1 can be applied to HW and IW estimators by simply plugging the appropriated weights defined in Proposition 4.1. In particular, plugging in the HW weights recovers the results of Hudgens and Halloran (2008). Consistent with their results, the estimated variance for the weighted estimator is unbiased if the treatment effects are constant.

5.2. Inference

We briefly discuss inference for these quantities given an estimator and its estimated variance, following the setup in Liu and Hudgens (2014). For additional discussion, see Tchetgen Tchetgen and VanderWeele (2012) and Rigdon and Hudgens (2015).

For general outcomes, Liu and Hudgens (2014) derived both Chebyshev and Wald confidence intervals (CIs) for household weighted estimands under two asymptotic regimes. In the first regime, the number of households (i.e., N) remains fixed, while the size of each household grows large, that is, $\min(n_1, \ldots, n_N) \to \infty$. Inference in this regime either relies on Chebyshev CIs, which are typically too wide to be practically useful, or Wald CIs, which require several additional conditions that are unlikely to hold in our example, including homogeneity of impacts across households. In the second regime, the size of each hold (i.e., n) remains fixed while the number of households grows large, that is, $N \to \infty$. At a high level, valid Wald CIs in this regime require a standard Lindeberg condition on the estimators and some restrictions on the within- and between-household variances, as well as requiring that N_0/N_1 remains relatively constant as $N \to \infty$. We rely on this approach here, as it seems like a reasonable asymptotic approximation in our setting. Thus, an asymptotic $1 - \gamma$ CI for τ_{HW}^P is

$$\widehat{\tau}_{\mathrm{HW}}^P \pm z_{1-\gamma/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\tau}_{\mathrm{HW}}^P)},$$

where $z_{1-\gamma/2}$ is the $1 - \gamma/2$ quantile of the standard Normal distribution. We caution that this asymptotic approximation might have poor performance with a small number of households (see, e.g., Pustejovsky and Tipton 2016). In the supplementary materials, we discuss the corresponding assumptions for individual weighted estimands. We argue that we can obtain valid Wald CIs via the $N \to \infty$ asymptotic regime of Liu and Hudgens (2014) separately for each household size stratum.

6. Regression-Based Estimation

We now connect these randomization-based results with more familiar regression-based methods. Our key result is that, with the appropriate standard errors, conventional linear regression estimates are equivalent to the randomization-based estimates. This approach regards regression as a convenient tool (sometimes known as a *derived linear model*; see Hinkelmann and Kempthorne 2012), and does not equate the regression with a specific generative model. In other words, this approach does not impose a model for the potential outcomes. See Baird et al. (2018) for additional discussion of regression for two-stage randomized designs.

We consider two basic regression approaches, an individual-level regression and a household-level regression. For simplicity, we start with the equal-sized household case and then show that these results generalize to any two-stage weights. We then demonstrate the dangers of using standard errors that ignore the two-stage structure. This section builds on existing results for robust and cluster-robust standard errors, especially Bell and McCaffrey (2002). See also Cameron and Miller (2015), Imbens and Kolesar (2016), and Pustejovsky and Tipton (2016).

6.1. Individual-Level Regression

First, we construct the individual-level linear model,

$$Y_{ij}^{\mathrm{obs}} = \alpha + \beta^P H_i Z_{ij} + \beta^S H_i (1 - Z_{ij}) + \varepsilon_{ij}, \qquad (14)$$

where the uncertainty in ε_{ij} is entirely due to randomization. It is straightforward to show that standard OLS estimates for β^P and β^S are identical to the randomization-based estimators in Section 4.1. The theorem below states that the randomization-based standard errors are equivalent to a particular cluster-robust generalization of heteroscedasticity-consistent standard errors, known as *HC2* (MacKinnon and White 1985).

Theorem 6.1. For equal-sized households, let Y be the vector containing all the observed outcomes Y_{ij}^{obs}; let X denote the appropriate design matrix (formally defined in the supplementary materials) with columns corresponding to the intercept, HZ, and $H(1 - Z)$; and let $\beta = (\alpha, \beta^P, \beta^S)$. The linear model in Equation (14) can therefore be rewritten as $Y = X\beta + \varepsilon$, with corresponding least squares estimate, $\widehat{\beta}^{\mathrm{ols}}$. These estimates are unbiased for their corresponding estimands.

Further, define the cluster-robust generalization of HC2 standard errors as

$$\widehat{\mathrm{var}}_{hc2}^{\mathrm{clust}}(\widehat{\beta}^{\mathrm{ols}})$$

$$= (X^t X)^{-1} \sum_{s=1}^{S} X_s^t (I_{N_s} - P_{ss})^{-1/2} \widehat{\varepsilon}_s \widehat{\varepsilon}_s (I_{N_s} - P_{ss})^{-1/2} X_s (X^t X)^{-1}$$

where X_s and $\widehat{\varepsilon}_s$ are the subsets of X and $\widehat{\varepsilon}$ corresponding to household s, and P_{ss} is defined as $P_{ss} = X_s (X^t X)^{-1} X_s^t$.

Then,

$$\widehat{\mathrm{var}}_{hc2}^{\mathrm{clust}}(\widehat{\beta}^{P,\mathrm{ols}}) = \widehat{\mathrm{var}}(\widehat{\tau}^P) \quad \text{and} \quad \widehat{\mathrm{var}}_{hc2}^{\mathrm{clust}}(\widehat{\beta}^{S,\mathrm{ols}}) = \widehat{\mathrm{var}}(\widehat{\tau}^S)$$

where $\widehat{\mathrm{var}}_{hc2}^{\mathrm{clust}}(\widehat{\beta}^{P,\mathrm{ols}}) = (\widehat{\mathrm{var}}_{hc2}^{\mathrm{clust}}(\widehat{\beta}^{\mathrm{ols}}))_{22}$ and $\widehat{\mathrm{var}}_{hc2}^{\mathrm{clust}}(\widehat{\beta}^{S,\mathrm{ols}}) = \widehat{\mathrm{var}}_{hc2}^{\mathrm{clust}}(\widehat{\beta}^{ols}))_{33}$.

In short, Theorem 6.1 confirms that we can obtain the same randomization-based point- and variance-estimators via the individual-level linear model in Equation 14 with HC2 cluster-robust standard errors. This is similar to results obtained with heteroscedastic-robust standard errors in simpler designs (e.g., Samii and Aronow 2012; Imbens and Rubin 2015). In Section 6.4, we demonstrate the effect of failing to account for clustering on standard errors. Researchers can estimate these standard errors directly in R via, for example, the clubSandwich package. See Pustejovsky and Tipton (2016) for additional discussion on the performance of clustered standard errors with a relatively small number of clusters.

6.2. Household-Level Regression

We now consider regression at the household level. This is a common strategy in cluster-randomized trials and yields identical inference to individual-level regression with clustered standard errors (see, e.g., Cameron and Miller 2015; Athey and Imbens 2017).

We separately aggregate treated and control units within each treated household, thus considering three types of household-level aggregates for household i. Each treated household has two household-level averages, $\overline{Y}_i^{\text{obs}}(1,1)$ and $\overline{Y}_i^{\text{obs}}(1,0)$; each control household has one household-level average, $\overline{Y}_i^{\text{obs}}(0,0)$. We can therefore assemble a vector of household-average outcomes, $\overline{Y}_k^{\text{obs}}$ of length $2N_1 + N_0$. We introduce the indicators $H_k^{(11)}$ and $H_k^{(10)}$; $H_k^{(11)} = 1$ if the aggregate is over the treated units in treated households and $H_k^{(11)} = 0$ otherwise; $H_k^{(10)} = 1$ if the aggregate is over the control units in treated households and $H_k^{(10)} = 0$ otherwise. We then consider the following linear model

$$\overline{Y}_k^{\text{obs}} = \alpha + \beta^P H_k^{(11)} + \beta^S H_k^{(10)} + \varepsilon_k'. \tag{15}$$

We now show that we can obtain the randomization-based point and variance estimates via the linear model estimates with standard (i.e., noncluster) heteroscedastic-robust standard errors.

Theorem 6.2. For equal-sized households, the OLS estimates for β^P and β^S in Equation 15 are unbiased estimators of the corresponding estimands. Define the heteroscedastic-robust estimator of the variances

$$\widehat{\text{var}}_{hc2}(\hat{\tau}^P) \equiv \widehat{\text{var}}_{hc2}(\widehat{\beta}^P) = \frac{\sum_{k:H_k^{(10)}=0} \widehat{\varepsilon}_k^2 (H_k^{(11)} - \overline{H})^2}{(\sum_{k:H_k^{(10)}=0} (H_k^{(11)} - \overline{H}^{(11)})^2)^2}$$

and

$$\widehat{\text{var}}_{hc2}(\hat{\tau}^S) \equiv \widehat{\text{var}}_{hc2}(\widehat{\beta}^S) = \frac{\sum_{k:H_k^{(11)}=0} \widehat{\varepsilon}_k^2 (H_k^{(10)} - \overline{H})^2}{(\sum_{k:H_k^{(11)}=0} (H_k^{(10)} - \overline{H}^{(10)})^2)^2}.$$

We have,

$$\widehat{\text{var}}_{hc2}(\hat{\tau}^P) = \widehat{\text{var}}(\hat{\tau}^P) \quad \text{and} \quad \widehat{\text{var}}_{hc2}(\hat{\tau}^S) = \widehat{\text{var}}(\hat{\tau}^S),$$

where the $\widehat{\varepsilon}_k$'s are the *HC2* residuals (see supplementary materials for exact definition).

In short, Theorem 6.2 states that we can aggregate to the household level and proceed as if this were a standard completely randomized trial (Imbens and Rubin 2015). Intuitively, the aggregation at the household level is another way of accounting for the household structure in the two-stage randomization scheme. Since the definition for the heteroscedastic-robust estimator of the variance is standard, it is straightforward to fit Equation 15 and the corresponding variance in R via, for example, the `vcovHC` function with the `HC2` option in the `sandwich` package (Zeileis 2004).

6.3. Results for Weighted Estimands

Finally, we can modify the household-level regression to estimate any two-stage weighted estimand. Thus, we can use this approach to recover the results in Sections 4 and 5. The key idea is to run an unweighted regression on transformed outcomes.

Theorem 6.3. Let w_i^* be two-stage estimand weights, with transformed potential outcomes, $Y_{ij}^{\text{obs},w}(h,z) = Nn_i w_i^* Y_{ij}^{\text{obs}}(h,z)$. Then, the results of Theorem 6.2 hold when applied to the

transformed potential outcomes. That is: $\widehat{\beta}^P = \hat{\tau}_W^P$, $\widehat{\beta}^S = \hat{\tau}_W^S$, $\widehat{\text{var}}_{hc2}(\widehat{\beta}^P) = \widehat{\text{var}}(\hat{\tau}_W^P)$, and $\widehat{\text{var}}_{hc2}(\widehat{\beta}^S) = \widehat{\text{var}}(\hat{\tau}_W^S)$.

This approach is subtly different from using weighted least squares, which would also reweight the design matrix.

6.4. Failing to Account for Clustering

It is instructive to consider the consequences of "naively" analyzing a two-stage experiment as if it were a completely randomized experiment, ignoring the household structure. With equal-sized households, the point estimates will be the same as for the appropriate analysis, but the standard errors will differ. In particular, let $\widehat{\text{var}}_{hc2}^{\text{het}}(\hat{\tau}^P)$ be the (noncluster) HC2 robust standard error for the primary effect; that is, these are the variances from Equation (15) for the household aggregates incorrectly applied to the individual level. We show in the supplementary materials that,

$$\mathbb{E}\left[\widehat{\text{var}}_{hc2}^{\text{het}}(\hat{\tau}^P)\right] - \text{var}(\hat{\tau}^P) = (\Sigma_{00} + V_{00})$$
$$\times \left\{ \frac{1}{nN_0 - 1}\left(1 - n\rho_{00}\right) - \frac{1}{N}\frac{V_P}{\Sigma_{00} + V_{00}} \right\},$$

where V_{00} and Σ_{00} are the between- and within-household variances, respectively, of the control potential outcomes, $Y_{ij}(0,0)$, and $\rho_{00} \equiv V_{00}/(\Sigma_{00} + V_{00})$ is the intraclass correlation (ICC). This quantity is negative—that is, the variance is anti-conservative in expectation—if

$$\rho_{00} > \frac{1}{n} - \left(\frac{nN_0 - 1}{nN}\right)\left(\frac{V_P}{\Sigma_{00} + V_{00}}\right).$$

Since the last term is nonnegative, we can build intuition by setting that term to zero. For example, consider the special case of $V_P = 0$, which would occur if there is a constant additive effect. Under these conditions, the estimated variance is anti-conservative if $\rho_{00} > 1/n$. In the social sciences, typical values of ICC range from 0.1 to 0.3 (e.g., Gelman and Hill 2006). Thus, even with households of size 4 or 5, the estimated variance could be anti-conservative. We see this behavior in the simulations in Section 8.

7. Covariate Adjustment

Finally, we explore how to incorporate individual- and cluster-level covariates in a two-stage experiment. There is an extensive literature on the use of covariates in randomized trials (see, for example, Imbens and Rubin 2015). In this section, we briefly address stratification, post-stratification, and model-assisted estimation.

Stratification and post-stratification. As with household size in Section 4.3, the simplest way to account for covariates is to incorporate them into the randomization by stratifying on them. In general, the researcher can partition households into discrete strata, regard each stratum as a separate "mini experiment," and estimate a pooled effect by averaging across strata. This will improve the precision of the treatment effect estimate so long as the stratifying covariate is predictive of the outcome. Researchers cannot, however, stratify by a covariate that

varies within household, as this could destroy the nested structure in the data by assigning different individuals in the same household to different "household-level" treatments. For example, we cannot stratify the two-stage randomization by gender, as some houses will have both boys and girls; instead, we could stratify by whether all students in the household are boys, which is an aggregate version of the individual-level covariate. Finally, as with household size, researchers can also post-stratify on household-level covariates. Of course, it is possible to combine stratification and post-stratification; for example, first stratify by household size and then, for each household size, post-stratify by whether the household speaks English as the primary language.

Model-assisted estimation. We also consider model-assisted estimation to incorporate covariates (Cochran 1977; Rosenbaum 2002; Hansen and Bowers 2009; Aronow and Middleton 2013). Following the setup in Hansen and Bowers (2009), consider K covariates $x^{(1)}, \ldots, x^{(K)}$ (which typically include a constant) with corresponding coefficient vector, $\gamma = (\gamma_1, \ldots, \gamma_K)$, such that $r(\gamma) = \{r_{ij}(\{x^{(k)}\}_k, \gamma)\}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ is a function mapping covariates to predictions. To simplify notation, we will let $x = \{x^{(k)}\}_k$. In practice, the coefficients in γ are typically coefficients from a linear regression. In this case, let

$$r_{ij}(x, \gamma) = \sum_{k}^{K} x_{ij}^{(k)} \gamma_k,$$

where x_{ij} is a vector of covariates associated with unit j in cluster i. We regard $r_{ij}(x, \gamma)$ as fixed and known for all units, rather than estimated from the data. We then define the (residualized) potential outcome as

$$e_{ij}^{\gamma}(h, z) = Y_{ij}(h, z) - r_{ij}(x, \gamma). \qquad (16)$$

As with the corresponding potential outcomes, $Y_{ij}(h, z)$, the residualized potential outcomes, $e_{ij}^{\gamma}(h, z)$, are assumed to be fixed and are only observed if $H_i = h$ and $Z_{ij} = z$. We can then substitute $e_{ij}^{\gamma}(h, z)$ for $Y_{ij}(h, z)$ in defining the primary effect

$$\begin{aligned} \tau^P &= \overline{Y}(1, 1) - \overline{Y}(0, 0) \\ &= [\bar{e}^{\gamma}(1, 1) + \bar{r}(x, \gamma)] - [\bar{e}^{\gamma}(0, 0) + \bar{r}(x, \gamma)] \\ &= \bar{e}^{\gamma}(1, 1) - \bar{e}^{\gamma}(0, 0), \end{aligned}$$

rewriting τ^S in an analogous way.

Given $r_{ij}(\gamma)$, model-assisted estimation of τ^P is immediate via substituting the observed values of the residualized outcomes, $e_{ij}^{\gamma, \text{obs}}$, in place of the unadjusted outcomes, Y_{ij}^{obs}. The resulting difference-in-means estimator is unbiased regardless of the exact values of $r_{ij}(\gamma)$; that is, there is no need to appeal to a "correctly specified" linear model to obtain the particular coefficient vector, γ. So long as the covariates are predictive of the outcome, the variance of $e_{ij}^{\gamma, \text{obs}}$ will generally be smaller than the variance of Y_{ij}^{obs}, and the resulting model-assisted estimator will also have smaller estimated variance (so long as γ does not include extreme values).

Finally, the above derivations assume that γ is fixed and known; in practice, we must find some way to determine γ. The most straightforward approach is to generate a random hold-out sample, and estimate γ via a regression of Y on X for this group. While not always possible, the attendance study effectively has a hold-out sample that we use for this purpose; see Section 9. See Hansen and Bowers (2009) and Aronow and Middleton (2013) for additional discussion.

8. Simulations

8.1. Failing to Account for the Cluster Structure

We now turn to simulations assessing the importance of accounting for the cluster structure. Reflecting the nested structure in the actual experiment, we generate potential outcomes in two stages. First, we simulate household-level average potential outcomes via

$$\begin{aligned} \overline{Y}_i(0, 0) &\sim N\left(\mu_{00}, \sigma_{\mu}^2\right) \\ \tau_i^P &\sim N\left(\bar{\tau}^P, \sigma_{\tau^P}^2\right) \\ \tau_i^S &\sim N\left(\bar{\tau}^S, \sigma_{\tau^S}^2\right), \end{aligned}$$

where $\overline{Y}_i(1, 1) = \overline{Y}_i(0, 0) + \tau_i^P$ and $\overline{Y}_i(1, 0) = \overline{Y}_i(0, 0) + \tau_i^S$. Then, conditional on these values, we generate individual-level potential outcomes via $Y_{ij}(h, z) \sim N(\overline{Y}_i(h, z), \sigma_y^2)$, for each h and z. Across all simulations, we fix the mean potential outcomes, with $\mu_{00} = 2$, $\bar{\tau}^S = 0.7$, and $\bar{\tau}^P = 1.5$, and fix household size at $n_i = 4$ for all households. For convenience, we also restrict the household-level variance terms to be equal to each other, such that $\sigma_{\mu} = \sigma_{\tau^P} = \sigma_{\tau^S} = \sigma_c$, where σ_c is the common standard deviation. Thus, we vary three main parameters, $N \in \{50, 100, 500, 1000\}$ (we always set $N_1 = N/2$) and $\sigma_c, \sigma_y \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For each combination of parameters, we consider three methods: cluster-robust standard errors; non-cluster, robust standard errors; and nominal standard errors. For each, we compute the coverage of the associated 95% confidence interval, averaging over 2000 random draws of the assignment vector.

Table 2 shows the overall coverage for 95% confidence intervals, averaged across all values of the simulation parameters. As expected, coverage with the cluster-robust standard errors is slightly larger than 95% coverage. By contrast, the noncluster and nominal standard errors have below 95% coverage. First, this coverage pattern is quite stable across different sample sizes. At the same time, coverage strongly depends on the within- and between-household variances. Figure 2 shows the relationship between coverage and the intraclass correlation among control potential outcomes, defined as $\sigma_c^2/(\sigma_c^2 + \sigma_y^2)$ in this simulation. Consistent with the results in Section 6.4, the coverage for non-clustered standard errors grows increasingly poor as the ICC increases.

Table 2. Average coverage of 95% confidence intervals over 2000 replications

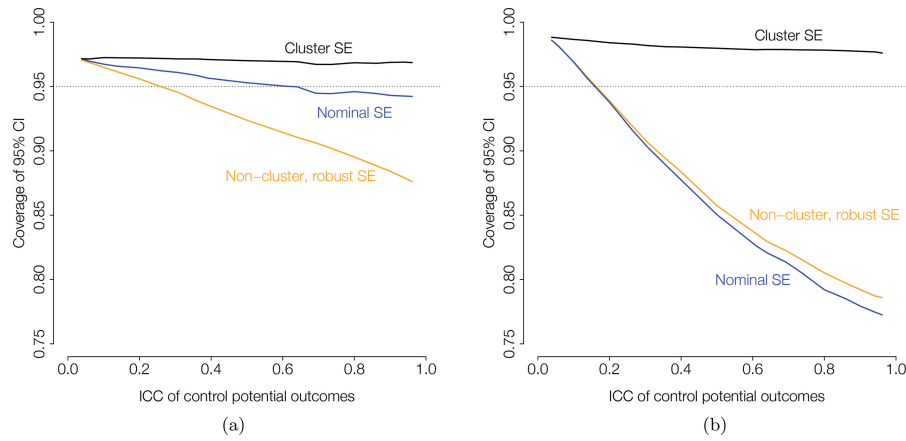|	τ^P	$\hat{\tau}^S$
Cluster robust SEs	0.97	0.98
Non-cluster robust SEs	0.93	0.87
Nominal SEs	0.95	0.86

Figure 2. Coverage for 95% confidence intervals for clustered SEs, noncluster robust SEs, and nominal SEs, with respect to the intraclass correlation of control potential outcomes, $\sigma_c^2/(\sigma_c^2 + \sigma_y^2)$. (a) Primary effect. (b) Spillover effect.

8.2. Comparing the Three Estimators for the IW Estimand

We now focus on the IW estimand and compare the unbiased estimator, the difference-in-means estimator, and the post-stratified estimator. We consider two scenarios: (a) when treatment effects are uncorrelated with household size; and (b) when treatment effects are correlated with household size. The data-generating process is the same as above, with a balanced household-level randomization with $N = 200$, $N_1 = 100$, and fixed $\sigma_c = \sigma_y = 0.3$. We generate households of size 2, 3, and 4 with equal probability, and introduce the parameters $\overline{\tau}_k^P, \overline{\tau}_k^S, \mu_{00}^{(k)}$ for $k = 2, 3, 4$. For scenario (a), we set $\overline{\tau}_k^P = 1.5, \overline{\tau}_k^S = 0.7, \mu_{00}^{(k)} = 2$ for all $k = 2, 3, 4$. For scenario (b), we allow the effects to vary by household size, as follows: $\overline{\tau}_2^P = 1.5, \overline{\tau}_3^P = 0.75, \overline{\tau}_4^P = 0.37,\quad \overline{\tau}_2^S = 0.7, \overline{\tau}_3^S = 0.35, \overline{\tau}_4^S = 0.17$, and $\mu_{00}^{(2)} = 2, \mu_{00}^{(3)} = 1, \mu_{00}^{(4)} = 0.5$.

The results are presented in Table 3. We see that when the treatment effect is uncorrelated with household size, the bias of all three estimators is negligible, but the Monte Carlo standard error of the unbiased estimator is an order of magnitude larger than that of the other two estimators. When treatment effect is correlated with household size, the biases of the unbiased and the post-stratified estimators are still very small, but the bias of the simple difference estimator is substantial—roughly the same size as the standard error. Again, the standard errors are smallest for the post-stratified estimator; overall, the post-stratified estimator clearly dominates in terms of RMSE.

Table 3. Bias and SE for different estimators for the IW estimand over 2,000 replications. '—' denotes $\leq 10^{-3}$

Estimator	hh size uncorrelated with effect		hh size correlated with effect	
	Avg. |bias|	Monte Carlo SE	Avg. |bias|	Monte Carlo SE
Primary effect				
Unbiased	—	0.11	—	0.07
Simple difference	—	0.04	0.18	0.12
Post-stratified	—	0.04	—	0.04
Spillover effect				
Unbiased	—	0.12	—	0.07
Simple difference	—	0.04	0.12	0.13
Post-stratified	—	0.04	—	0.04

9. Student Absenteeism in the School District of Philadelphia

9.1. Setup and Covariate Balance

We now apply these methods to our motivating example, the Rogers and Feller (2018) attendance intervention in the School District of Philadelphia. The original study included three active treatment arms and one control arm, with $N = 28,080$ total households. The first treatment arm merely reminded parents of the importance of attendance and did not provide any student-specific information. The second and third arms provided parents with different types of student-specific information. Rogers and Feller (2018) found weak impacts of assignment to the first arm and large impacts for the second and third arms (all relative to control). They also found minimal differences between the second and third arms.

We now make three modifications that preserve the substantive questions from the original study but allow us to focus on the two-stage randomized design. First, we exclude households from the weak first treatment arm, instead using this group as the holdout sample for estimating the covariate adjustment model (see Section 7). Second, we combine the second and third treatment arms, since these are substantively very similar and have virtually identical impacts. Together, these modifications essentially create a two-arm trial: control versus combined second and third arms from the original study. Finally, we restrict the universe to the subset of households with two or more eligible students, which yields $N = 3,876$ total households, of which $N_1 = 2,568$ (66%) were assigned to treatment and $N_0 = 1,308$ (34%) were assigned to control, and $n^+ = 8,654$ total students.

We next assess covariate balance for this sample. While household-level randomization was not stratified by household size (see Section 4.3), the balance by household size is excellent, as shown in Table 1. Table 4 shows the covariate balance for each stage of randomization. The left bank shows balance for the first stage, household-level randomization, with covariates aggregated to the household level. The right bank shows balance for the second stage randomization, with individual-level covariates among households assigned to treatment ($H_i = 1$). Statistically, Table 4 shows that covariate balance is excellent for both stages of randomization, with all normalized

Table 4. Covariate means and normalized differences (Δ) by stage of randomization. For the first stage, balance is assessed for covariates aggregated to the household level. For the second stage, balance is assessed for individual-level covariates, restricted to households assigned to treatment ($H_i = 1$)

|	*First Stage* (all households) Household-level means			*Second Stage* ($H_i = 1$ only) Individual-level means		
	For $H_i = 1$	For $H_i = 0$	Δ	For $Z_{ij} = 1$	For $Z_{ij} = 0$	Δ
Female	0.53	0.54	−0.03	0.53	0.53	−0.01
Black/African-American	0.51	0.50	0.03	0.52	0.52	−0.01
English spoken at home	0.84	0.83	0.04	0.84	0.83	0.02
Limited English Proficiency	0.07	0.07	−0.01	0.07	0.08	−0.04
Free or Reduced Price Lunch	0.78	0.79	−0.03	0.78	0.79	−0.02
Prior year absences (days)	16.70	16.55	0.02	16.65	16.93	−0.03
Start-of-year absences (days)	1.20	1.14	0.05	1.22	1.21	0.01
Students per household (n_i)	2.2	2.2	−0.01	—		

differences (Imbens and Rubin 2015) below 0.05 in absolute value. Substantively, Table 4 emphasizes that the students come from largely disadvantaged households. Over three-quarters of these students qualify for Free or Reduced Price Lunch, which is only available to families at or near the Federal Poverty Line. Over 15% of households speak a language other than English at home, with 7% of students designated as Limited English Proficiency (LEP). Moreover, this is a very high-absence group, with an average of around 13 days absent in the previous school year (out of roughly 180 possible days); we also include number of absences prior to randomization in early October. We observe the grade for each student, which we treat as a discrete covariate and which ranges from first grade to high school senior. While we do not show balance by grade to conserve space, there is excellent balance across this covariate as well.

Finally, we are broadly interested in days absent as the outcome of interest. However, the distribution of absences has a long right tail; for example, several students in the sample are absent for over half the school year. As this greatly increases the variance, we consider two transformed outcomes of interest. First, we consider an indicator for whether a student is chronically absent, defined as missing 18 or more days during the school year, that is, $\mathbb{1}(\text{days} \geq 18)$; among students in the control group, 36% are chronically absent. Second, we consider log-absences, defined as $\log(\text{days} + 1)$, to allow for a continuous outcome without the very heavy right tail; baseline absences

among students in the control group are around 13 days or $\log(13 + 1) \approx 2.6$. For interpretability, we also report key point estimates in terms of raw days.

9.2. Results

Figure 3 shows the estimated impacts and corresponding 95% confidence intervals for the primary and spillover effects for both household- and individual-weighted estimands. In terms of chronic absenteeism (i.e., the binary outcome), the estimates for the HW and IW estimands are nearly identical: the unbiased estimates for the primary effects are around −4 percentage points (SE of 1.5 percentage points) for both τ_{HW}^P and τ_{IW}^P; the unbiased estimates for the spillover effects are around −3 percentage points (SE of 1.5 percentage points) for both τ_{HW}^S and τ_{IW}^S. These results are virtually unchanged when post-stratifying by household size, using the (conditionally) unbiased estimator within each post-stratification cell, defined by $n_i = 2$, $n_i = 3$, and $n_i \in \{4, \ldots, 7\}$. As discussed in Section 3.1.2 and in the supplementary materials, the overall effect is essentially a weighted average of the primary and spillover effects. Unsurprisingly, the unadjusted estimate of the overall effect is around 3.5 percentage points (SE of 1.4 percentage points) for both the HW and IW weighted estimands, with nearly identical results for the post-stratified estimate.
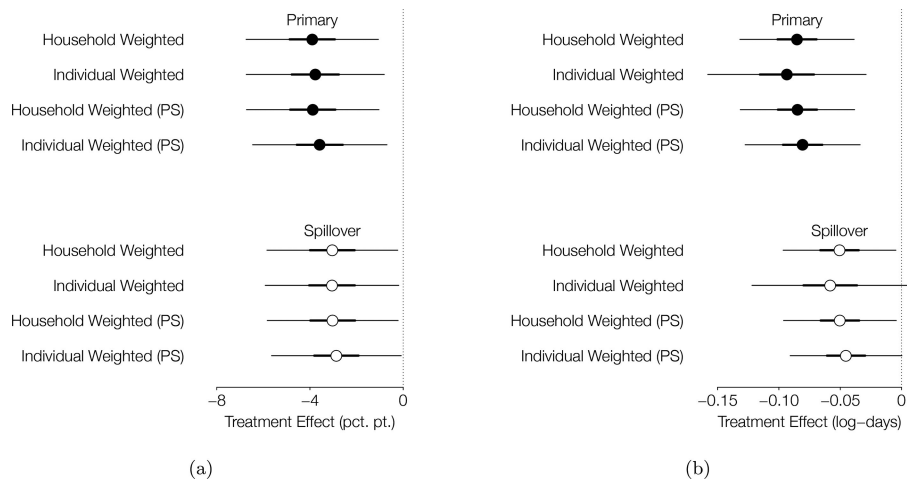
Figure 3. Treatment effect estimates and 95% confidence intervals for primary (filled-in circles) and spillover (open circles) effects, for household- and individual-weighted estimands with and without post-stratification (PS) by household size. (a) Binary outcome: chronically absent. (b) Continuous outcome: log(Days + 1).
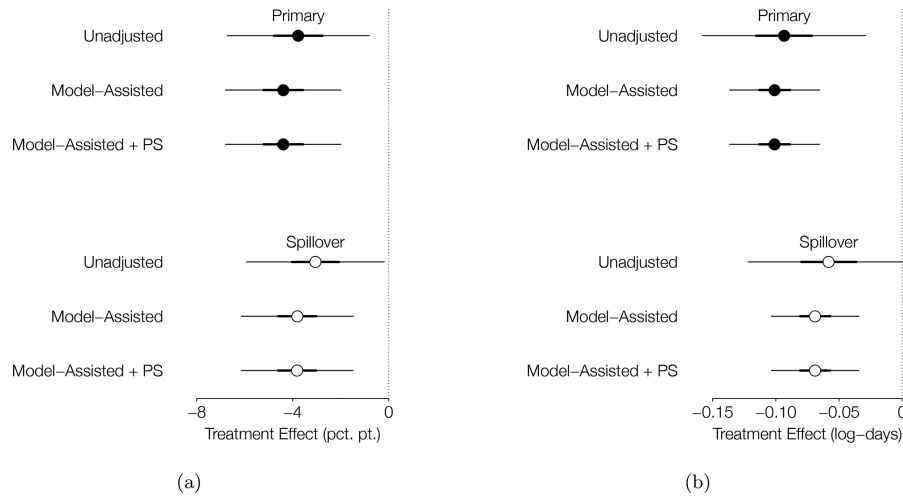
Figure 4. Treatment effect estimates and 95% confidence intervals for primary (filled-in circles) and spillover (open circles) effects for individual-weighted estimands, unadjusted, with model-assisted estimation, and with post-stratification (PS) by household size. (a) Binary outcome: chronically absent. (b) Continuous outcome: log(Days + 1).

The results are somewhat more variable for the impact on log-absences (i.e., the continuous outcome). The point estimates are quite close for the household-weighted and individual-weighted estimands: $\widehat{\tau}_{HW}^{P} = -0.085$ log-days and $\widehat{\tau}_{IW}^{P} = -0.093$ log-days for the primary effect, and $\widehat{\tau}_{HW}^{S} = -0.051$ log-days and $\widehat{\tau}_{IW}^{S} = -0.058$ log-days for the spillover effect. In terms of raw days, these are roughly -1.2 days for the primary effect and -0.7 days for the spillover effect. The point estimates are similarly close for the post-stratified estimator. The standard errors, however, are considerably larger for the unadjusted IW estimates: roughly 0.033 log-days for the IW estimands compared to 0.023 log-days for the HW estimands. Thus, the corresponding confidence intervals are roughly 50% larger for the IW estimands than for the HW esitmands. Post-stratification greatly reduces the standard errors for the IW estimand: for both IW and HW estimates, the standard errors are roughly 0.023 log-days, comparable to the standard errors for the HW estimate without post-stratification. As with the impact on chronic absenteeism, the estimates for the overall effect fall between the primary and spillover effect estimates, with estimates around -0.07 log-days (SE of 0.023 log-days) or around 1 day.

Despite minor differences between the sets of estimates, the pattern of effects is fairly clear: in general, we find that the spillover effect is between 60 and 80% as large as the primary effect, depending on the outcome. We also find few differences between the HW and IW estimates.

Next, Figure 4 shows covariate-adjusted estimates for individual-weighted estimands. First, we take advantage of the fact that there is a natural holdout sample in the experiment as analyzed; see Section 9. To obtain $\widehat{r}_{ij}(\gamma)$, we regress the outcome on covariates listed in Table 4 as well as student grade (categorical). Results do not appear sensitive to the particular choice of model. The resulting point estimates in Figure 4 are largely unchanged, if slightly larger in magnitude than the unadjusted estimates. The standard errors, however, are meaningfully smaller, especially for the continuous outcome: 0.018 log-days for the model-assisted estimator versus 0.033 log-days for the unadjusted estimator. Next, we can combine

model-assisted estimation with post-stratification, though the results are essentially identical. Finally, while do not have theoretical guarantees for covariate adjustment in a two-stage experiment, classical regression adjustment is nearly identical to the model-assisted estimation with the holdout sample.

In the end, we find strong evidence of intra-household spillover for the attendance intervention. This pattern holds with and without covariate adjustment, though the covariate-adjusted estimates are more precise. This underscores that merely focusing on the primary effect significantly underestimates the impact and cost effectiveness of the intervention. Rogers and Feller (2018) reported costs of around $6.60 per household. In our sample, the primary effect is around 1.2 days and the overall effect (i.e., the weighted average of the primary and spillover effects) is around 1 day. Thus, if we only consider the primary effect, the cost for each additional student day is $6.60 / 1.2 days $\approx$ $5.50 / day. By contrast, if we also consider spillovers, the cost for each additional student day is $6.60 / (2.2 \cdot 1 \text{ day}) \approx$ $3 / day, where $\bar{n} = n^{+}/N \approx 2.2$ is the average number of students per household in our sample.

10. Discussion

Two-stage randomizations are increasingly common designs in settings with interactions between units. This article addresses several practical issues in analyzing such designs. First, we address issues that arise when household sizes vary. Second, we demonstrate that regression can yield identical point- and variance-estimates to those derive from fully randomization-based methods. Methodologically, we believe that this is a useful addition to the literatures on both causal inference with interference and randomization-based inference. Substantively, we find convincing evidence of spillover effects of a large-scale attendance intervention.

There are several directions for future work. First, we are actively exploring covariate adjustment in this and other settings with more complex randomization schemes. The model-assisted approach is one such option, but many are possible (Lin 2013; Aronow and Middleton 2013). In particular,

extending the asymptotic results of Liu and Hudgens (2014) to incorporate covariates would be fruitful. Second, there is an open question of how to separately test the null hypotheses for no primary and no spillover effects in this type of design. Recent work from Athey, Eckles, and Imbens (2017) offers one promising direction. Third, it will be useful to extend these results to other, related designs. For example, Weiss, Lockwood, and McCaffrey (2016) discussed an interesting setting in which random assignment occurs at the individual level but individuals are then administered treatment in groups (such as in group therapy). Kang and Imbens (2016) proposed a "peer encouragement" design, which extends the two-stage randomization considered here to consider noncompliance. Fourth, we anticipate additional connections with nonrandomized studies that mimic a two-stage randomized design, such as Hong and Raudenbush (2006) and Perez-Heydrich et al. (2014).

Finally, we believe one promising direction for future work is to relax the stratified interference assumption while retaining partial interference. For example, Paluck, Shepherd, and Aronow (2016) collected detailed social network data for students within schools; the authors assume no interference between schools but leverage the within-school network structure for inference. Arpino and Mattei (2016) offered another possibility, using covariates to construct interference patterns between units. In the end, we hope that the results we give here will lead to increased use of two-stage randomized designs in practice.

Acknowledgments

References

Angelucci, M., and Di Maro, V. (2016), "Programme Evaluation and Spillover Effects," *Journal of Development Effectiveness*, 8, 22–43. [41]

Aronow, P. M., and Middleton, J. A. (2013), "A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments," *Journal of Causal Inference*, 1, 135–154. [50,53]

Aronow, P. M., and Samii, C. (2017), "Estimating Average Causal Effects Under General Interference, with an Application to a Social Network Experiment," *Annals of Applied Statistics*, 11, 1912–1947. [43]

Arpino, B., and Mattei, A. (2016), "Assessing the Causal Effects of Financial Aids to Firms in Tuscany Allowing for Interference," *The Annals of Applied Statistics*, 10, 1170–1194. [54]

Athey, S., Eckles, D., and Imbens, G. W. (2017), "Exact *P*-values for Network Interference," *Journal of the American Statistical Association*, Forthcoming. [54]

Athey, S., and Imbens, G. W. (2017), "The Econometrics of Randomized Experiments," in *Handbook of Field Experiments*, eds. A. Banerjee, and E. Duflo, Amsterdam: Elsevier. [42,43,48]

Baird, S., Bohren, J. A., McIntosh, C., and Özler, B. (2018), "Optimal Design of Experiments in the Presence of Interference," *The Review of Economics and Statistics*, forthcoming. [41,42,44,48]

Bell, R. M., and McCaffrey, D. F. (2002), "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, 28, 169–181. [48]

Bowers, J., Fredrickson, M. M., and Panagopoulos, C. (2013), "Reasoning About Interference Between Units: A General Framework," *Political Analysis*, 21, 97–124. [43]

Cameron, A. C., and Miller, D. L. (2015), "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50, 317–372. [48]

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley. [50]

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley. [41]

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., and Zamora, P. (2013), "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment," *The Quarterly Journal of Economics*, 128, 531–580. [41]

Donner, A., and Klar, N. (2000), *Design and Analysis of Cluster Randomization Trials in Health Research*, New York: Oxford University Press. [45]

ED Office for Civil Rights (2016), "Civil Rights Data Collection," available at *http://www2.ed.gov/about/offices/list/ocr/data.html*. [42]

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd. [44]

Gelman, A., and Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge, UK: Cambridge University Press. [49]

Halloran, M. E., and Hudgens, M. G. (2016), "Dependent Happenings: A Recent Methodological Review," *Current Epidemiology Reports*, 3, 1–9. [41,43]

Hansen, B. B., and Bowers, J. (2009), "Attributing Effects to a Cluster-Randomized Get-Out-The-Vote Campaign," *Journal of the American Statistical Association*, 104, 873–885. [50]

Hinkelmann, K., and Kempthorne, O. (2012), *Design and Analysis of Experiments, Special Designs and Applications* (Vol. 3), New York: Wiley. [48]

Hong, G., and Raudenbush, S. W. (2006), "Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data," *Journal of the American Statistical Association*, 101, 901–910. [54]

Hudgens, M. G., and Halloran, M. E. (2008), "Toward Causal Inference with Interference," *Journal of the American Statistical Association*, 103, 832–842. [41,42,43,44,45,47]

Imai, K., King, G., and Nall, C. (2009), "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation," *Statistical Science*, 24, 29–53. [45]

Imbens, G. W., and Kolesar, M. (2016), "Robust Standard Errors in Small Samples: Some Practical Advice," *Review of Economics and Statistics*, 98, 701–712. [48]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge, UK: Cambridge University Press. [43,48,49,52]

Kang, H., and Imbens, G. (2016), "Peer Encouragement Designs in Causal Inference with Interference," arXiv 1609.04464. [54]

Lin, W. (2013), "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *The Annals of Applied Statistics*, 7, 295–318. [53]

Liu, L., and Hudgens, M. G. (2014), "Large Sample Randomization Inference of Causal Effects in the Presence of Interference," *Journal of the American Statistical Association*, 109, 288–301. [41,43,48,54]

MacKinnon, J. G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325. [48]

Manski, C. F. (2013), *Public Policy in an Uncertain World: Analysis and Decisions*, Cambridge, MA: Harvard University Press. [43]

Middleton, J., and Aronow, P. M. (2015), "Unbiased Estimation of the Average Treatment Effect in Cluster Randomized Experiments," *Statistics, Politics and Policy*, 6, 39–75. [45,46,47]

Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013), "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments," *Journal of the Royal Statistical Society*, Series B, 75, 369–396. [46]

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments," *Statistical Science*, 5, 465–472. [43]

Nickerson, D. W. (2008), "Is Voting Contagious? Evidence from Two Field Experiments," *American Political Science Review*, 102, 49–57. [42]

Paluck, E. L., Shepherd, H., and Aronow, P. M. (2016), "Changing Climates of Conflict: A Social Network Experiment in 56 Schools," *Proceedings of the National Academy of Sciences*, 113, 566–571. [54]

Perez-Heydrich, C., Hudgens, M. G., Halloran, M. E., Clemens, J. D., Ali, M., and Emch, M. E. (2014), "Assessing Effects of Cholera Vaccination in the Presence of Interference," *Biometrics*, 70, 731–741. [54]

Pustejovsky, J. E., and Tipton, E. (2016), "Small Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models," arXiv 1601.01981. [48]

Rigdon, J., and Hudgens, M. G. (2015), "Exact Confidence Intervals in the Presence of Interference," *Statistics and Probability Letters*, 105, 130–135. [41,48]

Rogers, T., and Feller, A. (2018), "Reducing Student Absences at Scale by Targeting Parents. Misbeliefs," *Nature Human Behavior*, forthcoming. [41,42,46,51,53]

Rosenbaum, P. R. (2002), "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science*, 17, 286–327. [50]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688. [43]

——— (1980), "Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test"," *Journal of the American Statistical Association*, 75, 591–593. [41,43]

Samii, C., and Aronow, P. M. (2012), "On Equivalencies Between Design-Based and Regression-Based Variance Estimators for Randomized Experiments," *Statistics & Probability Letters*, 82, 365–370. [48]

Schochet, P. Z. (2013), "Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference," *Journal of Educational and Behavioral Statistics*, 38, 219–238. [45]

Sinclair, B., McConnell, M., and Green, D. P. (2012), "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments," *American Journal of Political Science*, 56, 1055–1069. [41,42]

Sobel, M. E. (2006), "What Do Randomized Studies of Housing Mobility Demonstrate? Causal Inference in the Face of Interference," *Journal of the American Statistical Association*, 101, 1398–1407. [43]

Somers, M.-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., and Zmach, C. (2010), "The Enhanced Reading Opportunities Study Final Report: The Impact of Supplemental Literacy Courses for Struggling Ninth-Grade Readers. NCEE 2010-4021," *National Center for Education Evaluation and Regional Assistance*. [41]

Tchetgen Tchetgen, E., and VanderWeele, T. (2012), "Estimation of Causal Effects in the Presence of Interference," *Statistical Methods for Medical Research*, 21, 55–75. [43,48]

Toulis, P., and Kao, E. (2013), "Estimation of Causal Peer Influence Effects," in *Proceedings of The 30th International Conference on Machine Learning*, pp. 1489–1497. [44]

Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013), "Graph Cluster Randomization: Network Exposure to Multiple Universes," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 329–337. [41]

VanderWeele, T. J., and Tchetgen Tchetgen, E. J. (2011), "Effect Partitioning under Interference in Two-Stage Randomized Vaccine Trials," *Statistics and Probability Letters*, 81, 861–869. [45]

VanderWeele, T. J., Tchetgen Tchetgen, E. J., and Halloran, M. E. (2014), "Interference and Sensitivity Analysis," *Statistical Science*, 29, 687–706. [43]

Weiss, M. J., Lockwood, J. R., and McCaffrey, D. F. (2016), "Estimating the Standard Error of the Impact Estimator in Individually Randomized Trials With Clustering," *Journal of Research on Educational Effectiveness*, 9, 421–444. [54]

Zeileis, A. (2004), "Econometric Computing with HC and HAC Covariance Matrix Estimators," *Journal of Statistical Software*, 11, 1–17. [49]