Optimal Design of Experiments in the Presence of Interference[*]

Sarah Baird[†], J. Aislinn Bohren[‡], Craig McIntosh[§], Berk Özler[¶]

November 2017

We formalize the optimal design of experiments when there is interference between units, i.e. an individual's outcome depends on the outcomes of others in her group. We focus on randomized saturation designs, two-stage experiments that first randomize treatment saturation of a group, then individual treatment assignment. We map the potential outcomes framework with partial interference to a regression model with clustered errors, calculate standard errors of randomized saturation designs, and derive analytical insights about the optimal design. We show that the power to detect average treatment effects declines precisely with the ability to identify novel treatment and spillover effects.

KEYWORDS: Experimental Design, Causal Inference

JEL: C93, O22, I25

[*] We are grateful for the useful comments received from Frank DiTraglia, Elizabeth Halloran, Cyrus Samii, Patrick Staples, and especially Peter Aronow. We also thank seminar participants at Caltech, Cowles Econometrics Workshop, Econometric Society Australiasian Meeting, Fred Hutch workshop, Harvard T.H. Chan School of Public Health, Monash, Namur, Paris School of Economics, Stanford, University of British Columbia, UC Berkeley, University of Colorado, University of Melbourne, the World Bank, Yale Political Science and Yale School of Public Health. We thank the Global Development Network, Bill & Melinda Gates Foundation, National Bureau of Economic Research Africa Project, World Bank's Research Support Budget, and several World Bank trust funds (Gender Action Plan, Knowledge for Change Program, and Spanish Impact Evaluation fund) for funding.

[†] George Washington University, sbaird@gwu.edu

[‡] University of Pennsylvania, abohren@sas.upenn.edu

[§] University of California, San Diego, ctmcintosh@ucsd.edu

[¶] World Bank, bozler@worldbank.org

1 Introduction

The possibility of interference in experiments, where the treatment status of an individual affects the outcomes of others, gives rise to a plethora of important questions. How does the benefit of treatment depend on the intensity of treatment within a population? What if a program benefits some by diverting these benefits from others? Does the study have an unpolluted counterfactual? Further, in the presence of interference, a full understanding of the policy environment requires a measure of spillover effects that are not captured, or are even a source of bias, in standard experimental designs. This is critical to determine the overall program impact.

Empirical researchers across multiple academic disciplines have become increasingly interested in bringing such spillover effects under the lens of experimental investigation. Over the past decade, a new wave of experimental studies relax the assumptions around interference between units. These studies have used a variety of methods, including using experimental variation across treatment groups, leaving some members of a group untreated, exploiting exogenous variation in within-network treatments, and intersecting an experiment with pre-existing networks.[1] Compared to experiments with no interference, these experiments often seek to measure a larger set of effects and involve more complex design choices. Therefore, researchers interested in using experiments to study interference face a novel set of design questions.

In this paper, we study experimental design in the presence of interference. We focus on settings with *partial interference*, in which individuals are split into mutually exclusive clusters, such as villages or schools, and interference occurs between individuals within a cluster but not across clusters. As established in Hudgens and Halloran (2008), a two-stage randomization procedure, in which first each cluster is randomly assigned a treatment saturation, and second, individuals within each cluster are randomly assigned to treatment according to the realized treatment saturation, can identify treatment and spillover effects when there is partial interference.[2,3] Hudgens and Halloran (2008); Liu and Hudgens (2014); Tchet-

[1] (i) Bobba and Gignoux (2016); Miguel and Kremer (2004); (ii) Barrera-Osorio, Bertrand, Linden and Perez-Calle (2011); Lalive and Cattaneo (2009); (iii) Babcock and Hartman (2010); Beaman (2012); Conley and Udry (2010); Duflo and Saez (2002); Munshi (2003); (iv) Banerjee, Chandrasekhar, Duflo and Jackson (2013); Chen, Humphries and Modi (2010); Macours and Vakis (2008); Oster and Thornton (2012).

[2] Many recent empirical papers use this randomization procedure, including Banerjee, Chattopadhyay, Duflo, Keniston and Singh (2012); Busso and Galiani (2014); Crepon, Duflo, Gurgand, Rathelot and Zamora (2013); Gine and Mansuri (forthcoming); Sinclair, McConnell and Green (2012).

[3] *Partial population* experiments (Moffitt 2001), in which clusters are assigned to treatment or control,

gen Tchetgen and VanderWeele (2010) define causal estimands, find unbiased estimators of these estimands, and characterize the distributions of these estimators in such designs.[4] But a key questions remains: how should a researcher designing such a *randomized saturation* (RS) experiment select the set of treatment saturations and the share of clusters to assign to each saturation? In this paper, we explore the trade-offs involved in these design choices, from the perspective of how they affect the standard errors of the estimators of different treatment and spillover effects.

Our first contribution is to provide a foundation for the regression models commonly used by economists to analyze RS experiments. We map a potential outcomes model with partial interference into a regression model with intra-cluster correlation, which provides a bridge between the causal inference literature and the methods used to analyze RS designs in practice. This mapping requires two restrictions on the population distribution of potential outcomes: (i) the population average potential outcome only depends on an individual's treatment status; and (ii) the share of treated individuals in the cluster, and the variance-covariance matrix of the population distribution of potential outcomes is block-diagonal.[5] Athey and Imbens (2017) perform a similar derivation for a model with uncorrelated observations and no interference. Our derivation is an extension of their approach that allows for intra-cluster correlation and partial interference.

We show that using this regression model to analyze data from an RS experiment identifies a set of novel estimands: not only can the researcher identify an unbiased estimate of the usual intention-to-treat effect, but she can also observe spillover effects on treated and untreated individuals, and understand how the intensity of treatment drives spillover effects on these groups. These are the infinite population analogues of the estimands that Hudgens and Halloran (2008) show can be consistently estimated in a finite population model. The estimate of the average effect on all individuals in treated clusters, which we refer to as the

and a subset of individuals in treatment clusters are offered treatment, also identify certain treatment and spillover effects when there is partial interference. But they provide no exogenous variation in treatment saturation to identify whether these effects vary with the intensity of treatment. Most extant partial population experiments feature cluster-level saturations that are either endogenous (Mexico's conditional cash transfer program, PROGRESA/Oportunidades (Alix-Garcia, McIntosh, Sims and Welch 2013; Angelucci and De Giorgi 2009; Bobonis and Finan 2009)) or fixed and typically set at 50% (Duflo and Saez 2003).

[4] Aronow and Samii (forthcoming) and Manski (2013) study identification and variance estimation under more general forms of interference.

[5] Hudgens and Halloran (2008) make the stronger assumption of stratified interference to estimate variances in a setting with partial interference. Graham, Imbens and Ridder (2010) relax this assumption with one of observational symmetry, i.e. exchangeability.

Total Causal Effect, provides the policy maker with a very simple tool to understand how the intensity of treatment will drive outcomes for a representative individual.

Next we illustrate the power trade-offs that exist in designing RS experiments, i.e. choosing the set of saturations and the share of clusters to assign to each saturation. We derive closed-form expressions for the standard errors (SEs) of the OLS estimates of various treatment and spillover estimands. Using these expressions, we derive properties of the optimal designs to measure different sets of estimands. The ability to identify novel estimands, such as slope effects, comes at a cost: decreased statistical power to measure intention-to-treat effects pooled across all saturations. In other words, the same variation in treatment saturation that permits measurement of how treatment and spillover effects vary with the intensity of treatment is detrimental to the power of the simple experimental comparison of treatment to pure control. By placing RS designs in the clustered error regression framework, we provide the closest possible analogue to the familiar power calculations in cluster randomized trials. This makes the design trade-offs present in RS experiments transparent. In related work, Hirano and Hahn (2010) study the power of a partial population experiment to analyze a linear-in-means model with no intra-cluster correlation.

We conclude with numerical simulations of hypothetical and published RS designs. First, we calculate the optimal designs for objective functions that include different sets of individual saturation, slope and pooled estimands. This demonstrates how the optimal design depends on the set of estimands that the researcher would like to identify and estimate precisely. Second, we calculate the standard errors for several RS designs used in published papers. This illustrates how design choices affect the standard errors of different estimators. To compute these numerical results, we use software that we developed to assist researchers in designing RS experiments. It is publicly available at http://pdel.ucsd.edu/solutions/index.html.

The remainder of the paper is structured as follows. Section 2 sets up the potential outcomes framework, formalizes an RS design and defines estimands related to spillovers. Section 3 connects the potential outcomes framework to a regression model with clustered errors, presents closed-form expressions for the standard errors and derives properties of the optimal RS design to measure different sets of estimands. Section 4 presents the numerical application. All proofs are in Appendix A.

2 Causal Inference with Partial Interference

2.1 Potential Outcomes

A researcher seeks to draw inference on the outcome distribution of an infinite population $\mathcal{I}$ under different treatment allocations. The population is partitioned into equal-sized, non-overlapping groups, or clusters, of size n.[6] Individual i in cluster c has response function $Y_{ic} : \{0,1\}^n \to \mathcal{Y}$ that maps each potential cluster treatment vector $\mathbf{t} = (t_1, ..., t_n) \in \{0,1\}^n$ into potential outcome $Y_{ic}(\mathbf{t}) \in \mathcal{Y}$, where $t \in \{0,1\}$ is a binary treatment status in which $t = 1$ corresponds to being offered treatment and $t = 0$ corresponds to not being offered treatment, and $\mathcal{Y} \subset \mathbb{R}$ is a set of potential outcomes. The response function is independent of the treatment vectors for all clusters $d \neq c$ – spillovers may flow within a cluster, but do not flow between clusters. Thus, we relax the stable unit treatment value assumption (SUTVA) within clusters, but maintain it across clusters. This set-up is referred to as *partial interference* (Sobel 2006).[7]

A random sample is drawn from this infinite population and randomly assigned treatment according to a prespecified experimental design. Our goal is to study the power of different experimental designs to detect treatment and spillover effects by comparing the standard errors of estimands across designs. In order to characterize these standard errors, we make two assumptions on the mean and the variance-covariance matrix of the population distribution of potential outcomes.

First, we assume that the *expected potential outcome* $E[Y_{ic}(\mathbf{t})]$ at potential treatment vector $\mathbf{t} \in \{0,1\}^n$, where the expectation is with respect to the population distribution of potential outcomes, only depends on individual treatment status t_i and the treatment saturation $p(\mathbf{t}) \equiv \frac{1}{n}\sum_{j=1}^{n} t_j$. In other words, it is independent of the identity of the other individuals who receive treatment.

Assumption 1. *There exists a* $\overline{Y} : \{0,1\} \times (0,1) \to \text{co}(\mathcal{Y})$*, where* $\text{co}(\mathcal{Y})$ *is the convex hull of the set of potential outcomes, such that for all treatment vectors* $\mathbf{t} \in \{0,1\}^n \setminus \{0^n, 1^n\}$ *with* $p(\mathbf{t}) = p$*, the expected potential outcome for an individual with treatment status* $t \in \{0,1\}$ *is* $\overline{Y}(t,p)$.

[6] We assume clusters are equal in size to simplify the analysis. In practice, datasets may have significant variation in the size of the cluster and the researcher may want to group clusters into different sized bins, i.e. rural and urban clusters.

[7] The assumption of no interference across groups is testable. For example, see Miguel and Kremer (2004).

To maintain consistent notation, define $\overline{Y}(0,0) \equiv E[Y_{ic}(0^n)]$ and $\overline{Y}(1,1) \equiv E[Y_{ic}(1^n)]$ as the expected potential outcomes when no individuals and all individuals within a cluster, respectively, are treated. Assumption 1 allows for a characterization of the standard errors of estimands without possessing information about the underlying network structure within a cluster.[8] Note that it does not preclude the *realized* potential outcomes from depending on the identity of the individuals who receive treatment. Therefore, it is weaker than the stratified interference assumption proposed by Hudgens and Halloran (2008), which assumes that the realized potential outcomes of an individual are independent of the identity of the other individuals assigned to treatment.

Second, we make an assumption about the variance-covariance matrix of the distribution of potential outcomes. Clustering of outcomes can be due to either (i) the extent to which outcomes are endogenously driven by the treatment of others in the same cluster, which is interference between units, or (ii) a statistical random effect in outcomes that is correlated between individuals – *correlated effects* (Manski 1993) – which does not stem from interference between units. To capture (ii), we allow potential outcomes to be correlated across individuals within the same cluster, while maintaining no correlation across clusters.

Assumption 2. *There exist $\sigma^2 > 0$ and $\tau^2 \geq 0$ such that for all $\mathbf{t}, \mathbf{t}' \in \{0,1\}^n$, the variance-covariance matrix for the distribution of potential outcomes satisfies:*

1. $\text{Var}(Y_{ic}(\mathbf{t})) = \sigma^2 + \tau^2$,
2. $\text{Cov}(Y_{ic}(\mathbf{t}), Y_{jc}(\mathbf{t})) = \tau^2$ *for* $i \neq j$,
3. $\text{Cov}(Y_{ic}(\mathbf{t}), Y_{jd}(\mathbf{t}')) = 0$ *for* $c \neq d$.

Assumption 2 imposes homoskedasticity across all potential outcomes for a given individual and across potential outcomes between two individuals in the same cluster. In other words, the variance and covariance of the distribution of potential outcomes do not depend on the treatment status of an individual or the treatment saturation of a cluster.[9] We will often use $\rho \equiv \tau^2/(\tau^2 + \sigma^2)$ to denote the intra-cluster correlation (ICC).

[8] In the absence of this assumption, a researcher would need to observe the complete network structure in each cluster, understand the heterogeneity in networks across clusters, and use a model of network-driven spillovers to simulate the variance in outcomes that could be generated by these networks. This is not an issue when there is no interference.

[9] The analysis can allow for heteroskedasticity. The standard errors are less tractable to characterize analytically, and hence, optimal design results are also less tractable. We view the homoskedastic case as a natural benchmark to establish how randomizing treatment saturation impacts power.

Assumption 2 allows us to connect the potential outcomes framework to a regression model with a block-diagonal error structure. Our goal is to provide a bridge between the theoretical literature and the use of field experiments in economics to measure spillover effects. To this end, it is natural to impose a variance structure on potential outcomes that maps to the regression model typically used for power calculations when there is no interference.[10] It enables a direct comparison of the power of RS designs to the power of the canonical individually-randomized (blocked) and cluster-randomized (clustered) designs, making explicit the impact that randomizing saturation has on power. A regression model with a block-diagonal structure is also the model underlying the use of OLS with clustered standard errors to analyze resulting data, the method commonly used for analysis.

2.2 A Randomized Saturation Design

Suppose a researcher draws a sample of C clusters of size n. A *randomized saturation* (RS) design is a two-stage treatment assignment mechanism that specifies how to assign treatment to these $N \equiv nC$ individuals. The first stage randomizes the treatment saturation of each cluster. Let $\Pi \subset [0,1]$ be a finite set of treatment saturations. Each cluster c is randomly assigned a treatment saturation $P_c \in \Pi$ according to the distribution $f \in \Delta(\Pi)$, which specifies the share of clusters assigned to each saturation. The second stage randomizes the treatment status of each individual in the cluster, according to the realized saturation of the cluster. Individual i in cluster c is randomly assigned treatment $T_{ic} \in \{0,1\}$, where the realized cluster treatment saturation P_c specifies the share of individuals assigned to treatment (i.e. $\sum_{i=1}^{n} T_{ic} = nP_c$). Let T_c denote the realized treatment vector for cluster c. An RS design is completely characterized by the pair $\{\Pi, f\}$.[11]

We refer to individuals assigned to treatment as *treated* individuals, individuals in clusters assigned saturation zero as *pure controls*, and individuals in treated clusters who are not assigned to treatment as *within-cluster controls*. Let $S_{ic} = \mathbb{1}\{T_{ic} = 0, P_c > 0\}$ and $C_{ic} = \mathbb{1}\{T_{ic} = 0, P_c = 0\}$ denote whether individual ic is a within-cluster control or pure control, respectively. An RS design has share of treated individuals $\mu \equiv \sum_{p \in \Pi} pf(p)$, share of within-

[10] See Duflo, Glennerster and Kremer (2007) for power expressions when there is no interference.

[11] The framework discussed here use a simple, spatially defined definition of a cluster that is mutually exclusive and exhaustive. This is distinct from determining how to assign treatment in overlapping social networks (Aronow 2012), which requires a more complex sequential randomization routine (Toulis and Kao 2013). An additional benefit of an RS design is that it also creates exogenous variation in the saturation of any overlapping network in which two individuals in the same cluster have a higher probability of being linked than two individuals in different clusters.

cluster control individuals $\mu_S \equiv 1-\mu-\psi$, and share of control individuals $\psi \equiv f(0)$. There is a pure control if $\psi > 0$. We say an RS design is *non-trivial* if it has at least two saturations, at least one of which is strictly interior. Multiple saturations guarantee a comparison group to determine whether effects vary with treatment saturation, and an interior saturation guarantees the existence of within-cluster controls to identify spillovers on the untreated.

The RS design nests several common experimental designs, including the clustered, blocked, and partial population designs.[12] The blocked and clustered designs are trivial, and it is not possible to identify any spillover effects in these designs. The partial population design is non-trivial, and it is possible to identify whether there are spillover effects on the untreated.

Variation in the treatment saturation introduces correlation between the treatment statuses of individuals in the same cluster. Fixing μ and defining $\eta^2 \equiv \sum_{p\in\Pi} p^2 f(p)$, the variance in treatment saturation $\eta^2 - \mu^2$ and correlation in treatment status for an RS design is bracketed between that of a clustered design, which has the maximum possible variance in treatment saturation and perfect correlation between the treatment statuses of individuals in the same cluster, and that of a blocked design, which has no variance in treatment saturation and slightly negative correlation in treatment status (due to sampling without replacement). We will show that when there is also correlation between the potential outcomes of individuals in the same cluster, $\rho > 0$, the interaction of these two correlations play a key role in determining the power of an RS design.

Discussion. We implicitly assume that all individuals who are part of the spillover network within a cluster are included in the sample. If spillovers occur on individuals outside of the sampling frame, either because there is a 'gateway to treatment' within the cluster and not all eligible individuals are sampled, or because not all individuals in a cluster's spillover network are eligible for treatment, then it is necessary to distinguish between the *true* treatment saturation (the share of treated individuals in the cluster) and the *assigned* treatment saturation (the share of treated individuals out of the sampled individuals in the cluster).[13] If the sampling rate and share of the cluster eligible for treatment are constant

[12] Fixing μ, the clustered design corresponds to $\Pi = \{0, 1\}$ and $f(1) = \mu$, the blocked design corresponds to $\Pi = \{\mu\}$ and $f(\mu) = 1$, and the partial population design corresponds to $\Pi = \{0, P\}$ and $f(P) = \mu/P$.

[13] For example, Gine and Mansuri (forthcoming) sample every fourth household in a neighborhood, and randomly offer treatment to 80 percent of these households. This causes the true treatment saturation to be 20 percent rather than the assigned 80 percent. Other examples include: unemployed individuals on official unemployment registries form a small portion all unemployed individuals in an administrative

across clusters, the true saturation is proportional to the assigned saturation. If sampling rates are driven by cluster characteristics, or the share of the cluster that is eligible for treatment varies across clusters, then the true saturation is endogenous. In this case, the researcher can instrument for the true saturation with the assigned saturation. To streamline the analysis, we maintain the assumption that the assigned and true saturations coincide.

Our framework can be applied to settings with perfect compliance, or to identify intention to treat effects in settings with imperfect compliance. While non-compliance does not bias intention to treat estimands, it presents a second channel for interference – treatment and spillover effects may vary with treatment saturation due to compliance effects, in addition to the direct impact of an individual's treatment on others outcomes. Exploring extensions that allow compliance to depend on treatment saturation – thereby defining a response function that depends on whether individuals comply with assigned treatment – is an important avenue for future research.

2.3 Treatment and Spillover Estimands

Next we define a set of estimands for treatment and spillover effects. We focus on average effects across all individuals in the population. Recall that $\overline{Y}(t,p)$ is the *expected potential outcome* at individual treatment t and saturation p.

Individuals offered treatment will experience a direct treatment effect from the program, as well as a spillover effect from the treatment of other individuals in their cluster. Let $\underline{p} \equiv 1/n$ denote the treatment saturation corresponding to a cluster with a single treated individual. The *Treatment on the Uniquely Treated* (TUT) measures the intention to treat an individual, absent any spillover effects, $TUT \equiv \overline{Y}(1,\underline{p}) - \overline{Y}(0,0)$, and the *Spillover on the Treated* (ST) measures the spillover effect at saturation p on individuals offered treatment, $ST(p) \equiv \overline{Y}(1,p) - \overline{Y}(1,\underline{p})$. The familiar *Intention to Treat* (ITT) is the sum of these two effects, $ITT(p) = TUT + ST(p)$. Individuals not offered treatment experience only a spillover effect. The *Spillover on the Non-Treated* (SNT) is the analogue of the ST for individuals not offered treatment, $SNT(p) \equiv \overline{Y}(0,p) - \overline{Y}(0,0)$.[14] Given these definitions, there are *spillover*

region (Crepon et al. 2013); neighborhoods eligible for infrastructure investments comprise only 3 percent of all neighborhoods (McIntosh, Alegria, Ordonez and Zenteno 2013); and malaria prevention efforts target vulnerable individuals, who account for a small share of total cluster population (Killeen, Smith, Ferguson, Mshinda, Abdulla et al. 2007).

[14] If an RS design does not have a pure control, one could define analogous estimands relative to the lowest saturation in the design. For example, if clusters have a base saturation of share p_0 of individuals receiving treatment before an intervention, one could define estimands relative to p_0.

effects on the treated (non-treated) if there exists a p such that $ST(p) \neq 0$ ($SNT(p) \neq 0$).

We can also measure the slope of spillovers with respect to treatment saturation. The *Slope of Spillovers on the Treated* measures the rate of change in the spillover effect on treated individuals between saturations p and p', $DT(p, p') \equiv (ST(p') - ST(p))/(p' - p)$. If spillover effects are affine, then this is a measure of the slope; otherwise, it is a first order approximation. Let $DNT(p, p')$ denote the analogue for individuals not offered treatment.

In the presence of spillovers, the true effectiveness of a program is measured by the total effect of treatment on both treated and untreated individuals. The *Total Causal Effect* (TCE) measures this overall cluster-level effect on clusters treated at saturation p, compared to pure control clusters, $TCE(p) \equiv pITT(p) + (1-p)SNT(p)$. We say that treatment effects are *diversionary* at saturation p if the benefits to treated individuals are offset by negative externalities imposed on untreated individuals in the same cluster, $ITT(p) > 0$ and $TCE(p) < pITT(p)$. Diversionary treatment effects redistribute units of the outcome within a cluster to treated individuals, and the true effectiveness of the program is muted compared to the intention to treatment effect.[15] If the TCE is negative, the program causes an aggregate reduction in the average outcome, even though treatment effects may be positive. This highlights one reason why it is imperative to use the TCE, rather than the ITT, to inform policy in the presence of spillovers. The ITT may misrepresent the true effectiveness of the program.

We can also measure the direct impact of being assigned to treatment at a given saturation. The *Value of Treatment* (VT) measures the individual value of receiving treatment at saturation p, $VT(p) \equiv \overline{Y}(1, p) - \overline{Y}(0, p)$. If $VT(p)$ is decreasing in p, then the value of treatment is decreasing in the share of other individuals treated and spillover effects *substitute* for treatment, while if the VT is increasing in p, then the value of treatment is increasing in the share of other individuals treated and spillover effects *complement* treatment.

Hudgens and Halloran (2008) also study causal inference in the presence of partial interference, and define a similar set of estimands for a finite population. The estimands defined above are the infinite population analogues.[16]

[15] This does not say anything about the welfare implications of diversionary effects. To do so requires a welfare criterion specifying the social value of different distributions of the outcome within a cluster.

[16] The ST and SNT defined in our paper are the infinite population analogues of the indirect causal effects defined in their paper, the ITT is the analogue of their total causal effect, the TCE is the analogue of their overall causal effect and the VT is the analogue of their direct causal effect.

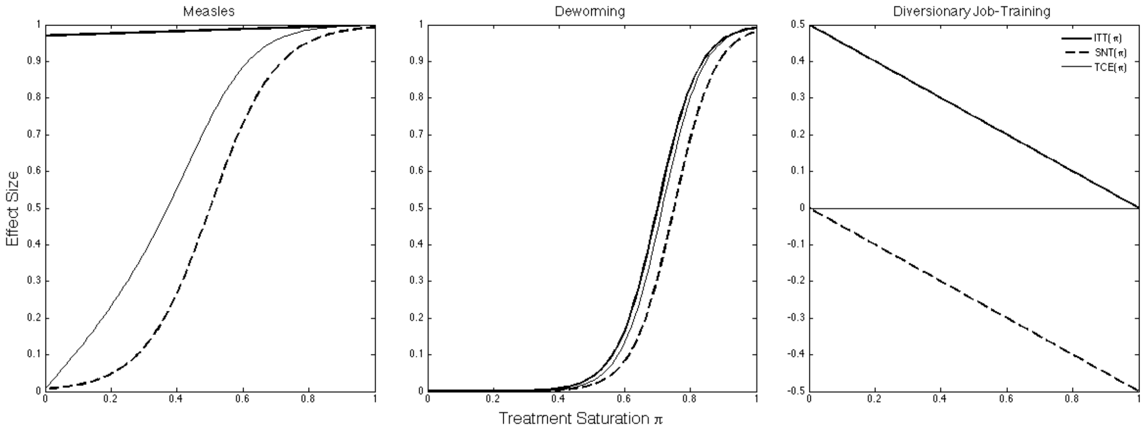

Figure 1. Examples of Spillovers

2.4 Examples of Spillovers

We illustrate the subtlety and importance of measuring spillover effects with three stylized examples: measles vaccinations, deworming interventions and job training programs. Consider an intervention that vaccinates a share p of individuals in a cluster. The TUT measures the efficacy of the vaccination in isolation. The vaccination almost fully protects vaccinated individuals, independent of the treatment saturation – the $ITT(p)$ is flat with respect to p, and spillovers on treated individuals, $ST(p)$, are small. However, the protection to the non-treated only becomes sizeable when the saturation is high enough to provide herd immunity – the $SNT(p)$ increases in p. Thus, the value of receiving the vaccination, $VT(p)$, is very large when vaccination rates are low, and approaches zero at high vaccination rates, since the unvaccinated are protected by herd immunity. Positive spillovers on unvaccinated individuals creates a free-rider problem that may diminish the salience of vaccinations in populations with very high overall treatment levels. This is illustrated in the left panel of Figure 1.

Deworming provides a more challenging case. Reinfection rates are proportional to the population prevalence of worm infections, which means that individuals who have received deworming treatment will quickly become reinfected in environments with high prevalence. The population saturation of deworming treatment drives long-term outcomes for both treated and non-treated individuals, and effective deworming requires near universal treatment. The poignant irony of such a program is that the $VT(p)$ is close to zero at all saturations even though deworming can be effective if applied universally. The key feature of this setting is the positive externality of treatment on both non-treated and other treated individuals. This is illustrated in the center panel of Figure 1.

Another example is a job training program in which the training has no effect on the overall supply of jobs – treatment simply diverts benefits from non-treated to treated individuals, but provides little net benefit (Crepon et al. 2013). Similar examples are tutoring programs for admissions to college or grant-writing workshops that improve specific proposals for a fixed funding pool. This type of diversionary treatment effect will have a $TCE(p)$ that is zero for all p, even though the $ITT(p)$ and the $VT(p)$ are strictly positive. In the face of diversionary effects, an RS design is imperative to identify the total policy effect, which is zero. A blocked design that uses within-cluster controls as counterfactuals will yield the mistaken conclusion that the overall impact of a program is positive. This is illustrated in the right panel of Figure 1.

3 Standard Errors and Optimal Design

This section maps the potential outcomes framework from Section 2.1 into a regression model that identifies the estimands defined in Section 2.3, derives analytical expressions for the standard errors of the OLS estimates, and characterizes properties of the optimal RS design for several sets of estimands. We begin with the individual saturation and slope estimands, and follow with complementary results for a model that estimates average effects across multiple saturations (pooled estimands). We conclude with an illustration of the power trade-off between measuring slope and pooled estimands.

3.1 Individual Saturation and Slope Effects

A Regression Framework. A regression model to estimate treatment and spillover effects at each saturation in the support of an RS design (Π, f) is

$$Y_{ic}^{obs} = \beta_0 + \sum_{p\in\Pi\backslash\{0\}} \beta_{1p} T_{ic} * \mathbb{1}\{P_c = p\} + \sum_{p\in\Pi\backslash\{0\}} \beta_{2p} S_{ic} * \mathbb{1}\{P_c = p\} + \varepsilon_{ic}, \tag{1}$$

where $Y_{ic}^{obs} \equiv Y_{ic}(T_c)$ denotes the observed outcome for individual ic and ε_{ic} is an unobserved error. To map the potential outcomes framework into this regression model, we define the regression coefficients and error in terms of potential outcomes, population average potential outcomes and realized treatment status. Let $\beta_0 \equiv \overline{Y}(0,0)$, $\beta_{1p} \equiv \overline{Y}(1,p) - \overline{Y}(0,0)$ and

$\beta_{2p} \equiv \overline{Y}(0,p) - \overline{Y}(0,0)$.[17] Define the residual as

$$\varepsilon_{ic} \equiv \sum_{\mathbf{t}\in\{0,1\}^n} \mathbb{1}_{T_c=\mathbf{t}} \left(Y_{ic}(\mathbf{t}) - \overline{Y}(t_i, p(\mathbf{t}))\right), \tag{2}$$

where $p(\mathbf{t})$ is the share of treated individuals in treatment vector $\mathbf{t} = (t_1, ..., t_n)$. The following lemma characterizes the distribution of the error in terms of the distribution of potential outcomes.

Lemma 1. *Assume Assumptions 1 and 2. Then the error defined in (2) is strictly exogenous, $E[\varepsilon_{ic}|T_c] = 0$, and has a block-diagonal variance-covariance matrix with $E[\varepsilon_{ic}^2|T_c] = \sigma^2 + \tau^2$, $E[\varepsilon_{ic}\varepsilon_{jc}|T_c] = \tau^2$ for $i \neq j$ and $E[\varepsilon_{ic}\varepsilon_{jd}|T_c, T_d] = 0$ for $c \neq d$.*

Athey and Imbens (2017) derive a similar result for a potential outcomes model with no interference and no intra-cluster correlation.

Given Lemma 1, the OLS estimate of (1) yields an unbiased estimate of $\boldsymbol{\beta}$. For any RS design with an interior saturation and a pure control, this estimate identifies $I\hat{T}T(p) = \hat{\beta}_{1p}$, $S\hat{N}T(p) = \hat{\beta}_{2p}$, $T\hat{C}E(p) = p\hat{\beta}_{1p} + (1-p)\hat{\beta}_{2p}$ and $\hat{V}T(p) = \hat{\beta}_{1p} - \hat{\beta}_{2p}$ for each $p \in \Pi \setminus \{0\}$. Tests for the presence of treatment and spillover effects at saturation p are $\hat{\beta}_{1p} \neq 0$ and $\hat{\beta}_{2p} \neq 0$, $\hat{\beta}_{1p} \neq \hat{\beta}_{2p}$ tests whether the value to treatment is non-zero, a one-tailed test of the sign of $\hat{\beta}_{2p}$ determines whether treatment creates a negative or positive externality on untreated individuals, and $\{\hat{\beta}_{1p} \geq 0, \hat{\beta}_{2p} \leq 0\}$ tests for diversionary effects.[18] Hudgens and Halloran (2008) present similar estimators for finite population estimands and show that these estimators are unbiased.[19]

The OLS estimate of (1) also yields unbiased estimates of the slope estimands, $D\hat{T}(p,p') = (\hat{\beta}_{1p'} - \hat{\beta}_{1p})/(p'-p)$ for each $(p,p') \in \Pi \setminus \{0\}$, with an analogous expression for $D\hat{N}T(p,p')$. A pure control is not required to estimate the slope estimands – any RS design with two interior saturations identifies the slope effect for both treatment and within-cluster control individuals. If a design has no pure control, replace the control group with the within-cluster

[17] Note that we are not assuming a constant treatment effect in (1); $\boldsymbol{\beta}$ is the *average* effect.

[18] This model also allows for tests on the shape of the $ITT(p)$ and $SNT(p)$. For example, three interior saturations allows one to test for concavity or convexity.

[19] In Hudgens and Halloran (2008), the sample is equal to the population, and uncertainty stems from the unobserved potential outcomes for each individual. Our model has an infinite population, and uncertainty stems from both the unobserved potential outcomes for each individual and sampling uncertainty from observing a subset of the population. Minor technical modifications to their proofs establish the analogous unbiased results in our setting.

controls in the lowest saturation in the RS design, and redefine the coefficients in (1) to be relative to the population mean of untreated individuals at the lowest saturation.

Standard Errors. Our first result characterizes the standard errors (SEs) for the OLS estimator of the individual saturation and slope estimands from (1).[20]

Theorem 1. *Assume Assumptions 1 and 2. For any RS design* (Π, f) *with a pure control, the SE of the treatment effect at saturation* $p > 0$ *is*

$$\mathrm{SE}_{ITT}(p) = \sqrt{\frac{\tau^2+\sigma^2}{nC} * \left(n\rho\left(\frac{1}{f(p)} + \frac{1}{\psi}\right) + (1-\rho)\left(\frac{1}{pf(p)} + \frac{1}{\psi}\right)\right)}$$

for each $p \in \Pi$. *For any RS design* (Π, f) *with at least two interior saturations, the SE for the slope effect on treated individuals between saturations* $p > 0$ *and* $p' > p$ *is*

$$\mathrm{SE}_{DT}(p, p') = \frac{1}{p'-p}\sqrt{\frac{\tau^2+\sigma^2}{nC} * \left(n\rho\left(\frac{1}{f(p)} + \frac{1}{f(p')}\right) + (1-\rho)\left(\frac{1}{pf(p)} + \frac{1}{p'f(p')}\right)\right)}$$

Substituting $1-p$ *and* $1-p'$ *for* p *and* $p' < 1$, *respectively, yields analogous expressions for untreated individuals, denoted* $\mathrm{SE}_{SNT}(p)$ *and* $\mathrm{SE}_{DNT}(p, p')$.[21]

Theorem 1 illustrates how the precision of the OLS estimates depends on the the RS design and the correlation structure of outcomes. At one extreme, if there is no correlation ($\rho = 0$), the variation in $I\hat{T}T(p)$ is inversely proportional to the number of *treated individuals* at saturation p and the number of control individuals. There is no correlation between potential outcomes within a cluster, so two observations from the same cluster provide the same amount of information about $ITT(p)$ as two observations from different clusters. At the other extreme, if there is perfect correlation between potential outcomes within a cluster ($\rho = 1$), the variation in $I\hat{T}T(p)$ is inversely proportional to the number of *clusters* at saturation p and the number of control clusters. Observing $Y_{ic}(1, p)$ provides perfect information about $Y_{jc}(1, p)$, so a second observation from the same cluster provides no additional information about $ITT(p)$. At intermediate levels of correlation, SE_{ITT} depends on a weighted average of the number of treated individuals and the number of clusters at saturation p.

[20] In general, the OLS estimator is inefficient when errors are correlated. The standard errors characterized in Theorem 1 will be conservative if GLS or another more efficient estimator is used to analyze the resulting data. However, the OLS standard errors are useful for studying ex-ante design questions, due to their tractable analytical characterization.

[21] Using these expressions to inform experimental design requires estimates of τ^2 and σ^2. One could use existing observational data or conduct a small pilot experiment (Hahn, Hirano and Karlan 2011).

Next consider the standard error of the slope effect for treated individuals. As the distance between two saturations increases, $1/(p'-p)$ decreases, making it possible to detect smaller slope effects. At the same time, decreasing p decreases the number of treatment individuals at this saturation, which increases the SE. The former effect dominates when the saturations are close together, and spreading the saturations apart decreases the SE, while the latter effect dominates when p is close to zero, and further decreasing p will increase the SE. When ρ is large, the number of clusters assigned to each saturation play a larger role in determining the SE; a more equal distribution leads to a smaller SE. When ρ is small, the number of treated individuals assigned to each saturation is more important than the number of clusters; equalizing the number of treated individuals at each saturation reduces the SE.

Theorem 1 can be used to characterize the power of an RS design. The minimum detectable effect (MDE) is the smallest value of an estimand that it is possible to distinguish from zero (Bloom 1995). Given statistical significance level α, the null hypothesis of no treatment effect at saturation p is rejected with probability γ (the power) for values of $ITT(p)$ that exceed MDE $= (t_{1-\gamma} + t_\alpha)\,\mathrm{SE}_{ITT}(p)$. The expressions for the MDEs of the spillover effect on untreated individuals and the slope effects are analogous.

Optimal Design: Individual Saturation Effects. Given a set of saturations Π, the design choice involves choosing the share of clusters to allocate to each saturation. If the researcher places equal weight on estimating the treatment and spillover effect at each saturation in $\Pi \setminus \{0\}$, she chooses f to minimize the sum of standard errors,

$$\min_{f\in\Delta(\Pi)} \sum_{p\in\Pi\setminus\{0\}} \left(\mathrm{SE}_{ITT}(p) + \mathrm{SE}_{SNT}(p)\right).^{22,23} \tag{3}$$

First consider the choice of how many clusters to allocate to each positive saturation. By design, clusters assigned to extreme saturations have a more unequal number of treatment and within-cluster control individuals, relative to saturations closer to 0.5.[24] A researcher

[22] For ease of exposition, throughout the optimal design sections, we maintain that any saturation $p \in [0,1]$ and distribution $f \in \Delta(\Pi)$ are feasible. In other words, we ignore the indivisibility of individuals or clusters. The properties of the optimal design extend in a straightforward way to the case where the set of feasible saturations and distributions are discrete.

[23] This objective is equivalent to maximizing the probability of rejection for a test of the null of no effect i.e. minimizing the minimum detectable effect.

[24] For example, if $n = 20$, clusters treated at saturation 0.25 have 5 treated individuals and 15 within-cluster controls, whereas clusters treated at saturation 0.5 have 10 of each.

who places equal weight on measuring effects at each positive saturation will want to allocate a larger share of clusters to these more extreme saturations. This stems directly from the concavity of the SE with respect to the number of treated or within-cluster control individuals at that saturation. As ρ increases, the share of clusters at a given saturation has a larger impact on the SE, relative to the share of treated or within-cluster control individuals at that saturation. Therefore, the asymmetry of the optimal f decreases with ρ.

Next, consider the optimal control group size. The marginal impact of adding another cluster to the control reduces all SEs in (3), while the marginal impact of adding another cluster to an interior saturation only reduces the SEs at that saturation. Therefore, when outcomes within a cluster are perfectly correlated, the optimal design allocates more clusters to the control group than to each positive treatment saturation. When $\rho < 1$, the number of treated and within-cluster control individuals are also important, and more extreme saturations have more unequal numbers of treated and untreated individuals. In the optimal design, the number of treated individuals at saturations close to one may be larger than the number of pure control individuals, since a larger share of clusters are allocated to these high saturations to guarantee enough within-cluster controls. Similar intuition holds for saturations close to zero. However, since an additional control individual reduces all of the SEs in (3), the minimum of the number of treated and within-cluster control individuals at each saturation is smaller than the number of control individuals. Proposition 1 formalizes these insights.

Proposition 1 (Optimal Shares)**.** *Assume Assumptions 1 and 2 and fix a set of saturations* Π*. Let* f^* *minimize* (3)*, with* $\psi^* \equiv f^*(0)$*.*

1. *When* $\rho < 1$*, a larger share of clusters are allocated to more extreme saturations, and the minimum of the share of treated and within-cluster control individuals at any positive treatment saturation is less than the share of pure control individuals: for any* $p, p' \in \Pi \setminus \{0\}$ *with* $|0.5 - p| > |0.5 - p'|$*,* $f^*(p) > f^*(p')$*, and* $\psi^* > \min\{p, 1-p\} f^*(p)$ *for all* $p \in \Pi \setminus \{0\}$*.*
2. *When* $\rho = 1$*, an equal share of clusters are allocated to each treatment saturation, and a larger share of clusters are allocated to the control group:* $\psi^* > f^*(p) = f^*(p')$ *for all* $p, p' \in \Pi \setminus \{0\}$*.*

For a given intra-cluster correlation ρ and cluster size n, it is straightforward to numerically solve for the optimal share of clusters to assign to each saturation.

Optimal Design: Slope Effects. There are two steps to the design choice to measure slope effects: choosing the set of saturations and choosing the share of clusters to allocate to each saturation. Suppose a researcher places equal weight on estimating the slope effect for treated and untreated individuals, and believes that both slope effects are monotonic. Then she chooses an RS design with two saturations to solve

$$\min_{p_1,p_2,f(p_1)\in(0,1)^3} \mathrm{SE}_{DT}(p_1,p_2)+\mathrm{SE}_{DNT}(p_1,p_2). \tag{4}$$

The optimal saturations are symmetric about 0.5 in order to equalize the share of treated individuals in the smaller saturation and the share of untreated individuals in the larger saturation. The optimal distance between saturations is increasing in ρ, as the unequal share of treated and untreated individuals at extreme saturations has a smaller impact on the SEs when outcomes are more correlated. An equal share of clusters are allocated to each saturation, irrespective of ρ, since both saturations are equally extreme (i.e. the same distance from 0.5). This design equalizes the standard errors, $\mathrm{SE}_{DT}(p_1^*,p_2^*)=\mathrm{SE}_{DNT}(p_1^*,p_2^*)$.

Proposition 2 (Optimal Saturations)**.** *Assume Assumptions 1 and 2. The RS design that minimizes (4) equally divides clusters between two saturations that are symmetric about 0.5, $p_1^*=(1-\Delta)/2$ and $p_2^*=(1+\Delta)/2$, where the optimal distance between saturations $\Delta\in[\sqrt{2}/2,1)$ is increasing in ρ and n and satisfies*

$$\frac{n\rho}{8(1-\rho)}=\frac{2\Delta^2-1}{(1-\Delta^2)^2}.$$

If $\rho=0$, then $\Delta=\sqrt{2}/2$ for all n, and if $\rho=1$, then $\Delta\approx 1$.[25]

Figure 2 plots the optimal treatment saturations as a function of ρ.

More generally, if a researcher is interested in identifying individual saturation or slope effects at more than two saturations, Theorem 1 can be used to derive the optimal spacing of saturations and the optimal share of clusters to assign to each saturation. For example, a design to test for linearity has three saturations. The optimal design to test for linearity would ensure that the three saturations are sufficiently far apart, and the two extreme saturations are not too large or small.

[25] When $\rho=1$ and n is finite, Δ is the largest distance such that there is at least one treatment and one within-cluster control individual at each saturation, i.e. $\Delta=1-2/n$.

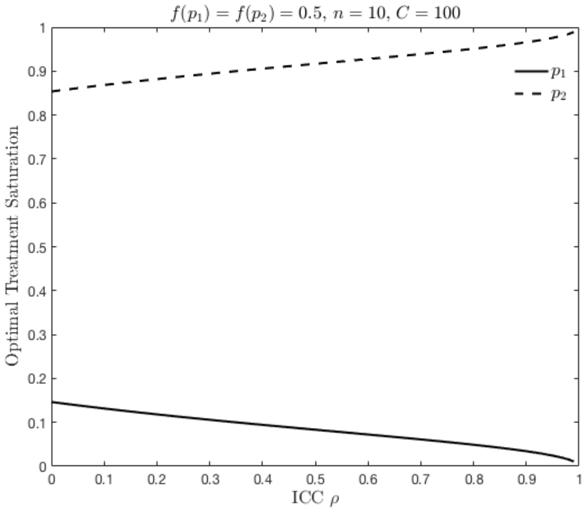

Figure 2. Optimal Design: Slope Effects

3.2 Pooled Effects

Suppose a researcher would like to combine observations from treated clusters to measure an average effect across all saturations in the RS design. A *pooled* estimand is a weighted sum of the estimand at each individual saturation. Given design (Π, f) and vector of weights $w : \Pi \to [0,1]$, a pooled treatment effect that assigns weight $w(p)$ to $ITT(p)$ is $\overline{ITT} \equiv \sum_{\Pi\setminus\{0\}} w(p)ITT(p)$. The definitions for $\overline{ST}$, $\overline{SNT}$, $\overline{TCE}$ and $\overline{VT}$ are analogous.

A Regression Framework. A regression model to estimate pooled effects is

$$Y_{ic}^{obs} = \beta_0 + \beta_1 T_{ic} + \beta_2 S_{ic} + \varepsilon_{ic}. \tag{5}$$

As in Section 3.1, we map the potential outcomes framework into this model by defining the regression coefficients and error in terms of potential outcomes, population average potential outcomes and realized treatment status. In the pooled case, the definition of the regression coefficients also depends on the RS design. Given an RS design (Π, f), let $\overline{Y}(1) \equiv \sum_{p\in\Pi\setminus\{0\}} pf(p)\overline{Y}(1,p)/\mu$ and $\overline{Y}(0) \equiv \sum_{p\in\Pi\setminus\{0\}}(1-p)f(p)\overline{Y}(0,p)/\mu_S$ be the population average potential outcome, averaged across all non-zero saturations in the RS design, for $t = 1$ and $t = 0$, respectively. Let $\beta_0 \equiv \overline{Y}(0,0)$, $\beta_1 \equiv \overline{Y}(1) - \overline{Y}(0,0)$ and $\beta_2 \equiv \overline{Y}(0) - \overline{Y}(0,0)$. Define the residual as

$$\varepsilon_{ic} \equiv C_{ic}(Y_{ic}(\mathbf{0}) - \overline{Y}(0,0)) + \sum_{\mathbf{t}\in\{0,1\}^n\setminus\{\mathbf{0}\}} \mathbb{1}_{T_c=\mathbf{t}} \left(Y_{ic}(\mathbf{t}) - \overline{Y}(t_i)\right). \tag{6}$$

The following lemma establishes that the OLS estimate of (5) is unbiased.

Lemma 2. *Assume Assumption 1. For any RS design with an interior saturation and a pure control, the OLS estimate $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$.*

The interpretation of $\hat{\boldsymbol{\beta}}$ is somewhat subtle. When observations are pooled across saturations, $\hat{\beta}_1$ places a disproportionate weight on treated individuals in high saturation clusters, relative to low saturation clusters – it identifies the pooled ITT with weight $w(p) = pf(p)$. Similarly, $\hat{\beta}_2$ places a disproportionate weight on untreated individuals in low saturation clusters, relative to high saturation clusters – it identifies the pooled SNT with weight $w(p) = (1-p)f(p)$. Due to these different weights, the comparison of the two pooled measures does not have a natural interpretation. Additionally, one must be careful when combining these estimates to identify other effects. For example, $\hat{\beta}_1 + \hat{\beta}_2$ is a pooled measure of the TCE with weight $w(p) = f(p)$, but $\hat{\beta}_1 - \hat{\beta}_2$ is not a pooled measure of the VT.[26]

Pooling observations across multiple saturations introduces the possibility of heteroskedasticity. The form of this heteroskedasticity depends on the RS design and the population average potential outcome at each positive saturation. When $ITT(p)$ and $SNT(p)$ are relatively flat with respect to $p \in \Pi \setminus \{0\}$, the heteroskedasticity will be small, whereas when these estimands significantly vary with the intensity of treatment, the heteroskedasticity will be large. The error is homoskedastic precisely when the $ITT(p)$ and $SNT(p)$ are constant with respect to the positive treatment saturations in the RS design.

Definition 1. *Treatment and spillover effects are* constant *on a set of saturations* Π *if for all* $p, p' \in \Pi \setminus \{0\}$, $\overline{Y}(1,p) = \overline{Y}(1,p')$ *and for all* $p, p' \in \Pi \setminus \{1\}$, $\overline{Y}(0,p) = \overline{Y}(0,p')$.

Lemma 3. *Assume Assumptions 1 and 2. Given an RS design* (Π, f)*, the error defined in* (6) *has homoskedastic variance and within-cluster covariance if and only if treatment and spillover effects are constant on the set of positive saturations* $\Pi \setminus \{0\}$.

[26] What we call *saturation weights*, which have a similar interpretation to sampling weights, can be used to adjust for the different probability of being assigned to treatment at each saturation. To estimate a pooled ITT and SNT that places equal weight $w(p) = 1/|\Pi|$ on the treatment or spillover estimand at each saturation, estimate (5) with weights $s_{ic} = 1/P_c f(P_c)$ for treated individuals and weight $s_{ic} = 1/(1-P_c)f(P_c)$ for within-cluster controls. Using these weights, $\hat{\beta}_1 - \hat{\beta}_2$ is now a pooled measure of the VT that places equal weight on each saturation, but $\hat{\beta}_1 + \hat{\beta}_2$ is no longer a pooled measure of the TCE. For example, consider a design with three saturations, $\Pi = \{0, 1/3, 2/3\}$ and an equal share of clusters assigned to each saturation, $f(p) = 1/3$ for each $p \in \Pi$. An individual in a cluster assigned $p = 2/3$ is twice as likely to be treated as a cluster assigned $p = 1/3$. Weighting the treated individuals in clusters assigned $p = 1/3$ and $p = 2/3$ by $s_{ic} = 3$ and $s_{ic} = 3/2$, respectively, allows one to calculate the pooled estimate that places equal weight on both clusters, rather than twice as much weight on the clusters treated at saturation $2/3$.

Generally, cluster robust standard errors should be used in two-level experiments due to the correlated outcomes within clusters. This proposition provides an additional argument for doing so when estimating (5), due to the variation in treatment and spillover effects at different saturations. Lemma 4 in Appendix A characterizes the precise form of this heteroskedasticity.

Standard Errors. Since an RS design opens the door to a novel set of questions about how treatment and spillover effects vary with intensity of treatment, and still identifies pooled treatment and spillover effects, it may be tempting to conclude that there is no reason *not* to run an RS design. If there is variation in treatment and spillover effects, then the heteroskedastic errors in the pooled regression are not an important issue, as the researcher is more interested in the individual saturation model (1), while if no slope effects emerge, then the pooled model is homoskedastic and there is no need to worry about multiple treatment saturations introducing heteroskedasticity. However, this line of reasoning misses a crucial piece of the story. Next, we show that including multiple treatment saturations increases the standard errors of pooled estimates, *even* when the treatment and spillover effects are constant, so that the error in (5) is homoskedastic.

In order to isolate the impact that multiple positive treatment saturations have on the SEs for the pooled estimands, we focus on the case where treatment and spillover effects are constant across all positive saturations in the RS design. Let η_T^2 be the variance in treatment saturation across treated clusters,

$$\eta_T^2 \equiv \sum_{p \in \Pi \setminus \{0\}} \frac{p^2 f(p)}{1-\psi} - \left(\frac{\mu}{1-\psi}\right)^2 = \frac{\eta^2}{1-\psi} - \frac{\mu^2}{(1-\psi)^2}, \tag{7}$$

where $f(p)/(1-\psi)$ is the share of treated clusters assigned to saturation $p > 0$ (recall $\eta^2 - \mu^2$ is the total variance in treatment saturation). Trivially, $\eta_T^2 = 0$ when there is a single positive saturation. Theorem 2 characterizes the SEs for the OLS estimator of the pooled estimands in (5).

Theorem 2. *Assume Assumptions 1 and 2. Let* (Π, f) *be an RS design with at least one interior saturation and a pure control, and suppose that treatment and spillover effects are* *constant on $\Pi \setminus \{0\}$. The SE of the pooled treatment effect is*

$$\overline{\mathrm{SE}}_{ITT} = \sqrt{\frac{\tau^2+\sigma^2}{nC} * \left(n\rho \left(\frac{1}{(1-\psi)\,\psi} + \left(\frac{1-\psi}{\mu^2} \right) \eta_T^2 \right) + (1-\rho) \left(\frac{1}{\mu} + \frac{1}{\psi} \right) \right)}.$$

Substituting μ_S for μ yields an analogous expression for the SE of the pooled spillover effect on the untreated, denoted $\overline{\mathrm{SE}}_{SNT}$.

The SE for the pooled treatment effect depends on the number of treated and control individuals and the variance in treatment saturation across treated clusters, η_T^2. Crucially, when outcomes within a cluster are correlated, the SE is strictly increasing in η_T^2, and introducing multiple treatment saturations reduces precision. The SE is minimized in a partial population design, in which there is a single positive saturation and a pure control. This design has no variation in treatment saturation across treated clusters, $\eta_T^2 = 0$.

Corollary 1 (Optimality of Partial Population Design)**.** *Assume Assumptions 1 and 2. For any $(\psi, \mu) \in (0,1) \times (0, 1-\psi)$, the partial population design with treatment saturation $p = \mu/(1-\psi)$ simultaneously minimizes $\overline{\mathrm{SE}}_{ITT}$ and $\overline{\mathrm{SE}}_{SNT}$.*[27,28]

If a researcher a priori believes that slope effects are small and intra-cluster correlation is high, she is best off selecting a partial population design. Moving away from the partial population design to a design with multiple treatment saturations, the variance of the pooled treatment effect increases linearly with respect to η_T^2. The rate at which this variance increases is proportional to ρ. Therefore, the power loss is more severe for settings with higher intra-cluster correlation.

Optimal Partial Population Design. Next, we characterize the optimal treatment saturation and control size for a partial population design. In a partial population design with saturation p, the pooled effects are equivalent to the individual effects at p. The SE of the ITT decreases with p, while the SE of the SNT increases with p, as illustrated in the left panel of Figure 3. The relative importance a researcher places on estimating these two effects will determine the optimal choice of p. If a researcher places equal weight on each effect,

$$\min_{(p,\psi)\in(0,1)^2} \mathrm{SE}_{ITT}(p) + \mathrm{SE}_{SNT}(p), \tag{8}$$

[27] Corollary 1 holds even when treatment and spillover effects are not constant on positive treatment saturations. In this case, reducing the variation in treatment saturation leads to more precise standard errors through two channels: that discussed above, as well as the resulting reduction in heteroskedasticity.

[28] The feasible range of μ is $(0, 1-\psi)$ because there does not exist an RS design with $\mu > 1 - \psi$.

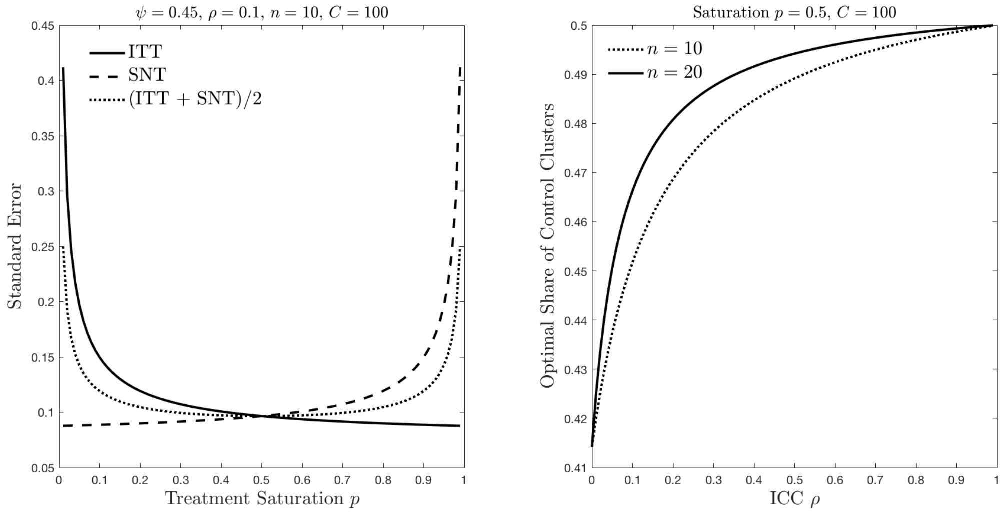

Figure 3. Optimal Partial Population Design

then the optimal saturation $p^* = 0.5$ creates equally sized treatment and within-cluster control groups, and equalizes the SEs, $\text{SE}_{ITT}(0.5) = \text{SE}_{SNT}(0.5)$. The optimal share of control clusters depends on ρ. As ρ increases, the number of *clusters* at each saturation becomes more important than the number of *individuals* in each treatment group. Similar to Proposition 1, the optimal share of control clusters increases in ρ. It is always optimal to allocate more than a third of clusters to the pure control, since a control individual serves as a counterfactual for both treated individuals and within-cluster controls. When $\rho = 0$, designating about 41% of clusters as pure controls is optimal, while when $\rho = 1$, 50% is optimal. The right panel of Figure 3 illustrates the optimal share of control clusters for a partial population design with $p^* = 0.5$. Proposition 3 summarizes these results.

Proposition 3. *Assume Assumptions 1 and 2. The partial population design that minimizes (8) has saturation $p^* = 0.5$ and allocates share of clusters*

$$\psi^* = \frac{-\kappa + \sqrt{\kappa^2 + (1-\rho)\kappa}}{1-\rho} \in [\sqrt{2} - 1, 0.5)$$

to pure control for $\rho \in [0, 1)$, where $\kappa \equiv 1 + (n-1)\rho$, and $\psi^* = 0.5$ for $\rho = 1$. The optimal control size ψ^* is increasing in ρ and n.

This result is similar in spirit to Hirano and Hahn (2010). They show that a partial population design identifies the VT and SNT in a linear-in-means model, and characterize the standard errors for the case of no intra-cluster correlation. When $\rho = 0$, the optimal design in Proposition 3 is equivalent to the optimal design in their sequential optimization case.

3.3 The Design Trade-off.

Taken together, the results in Section 3 illustrate a novel design trade-off: introducing variation in the intensity of treatment identifies spillover estimands, but reduces the precision of estimates of pooled effects, particularly when intra-cluster correlation is high. If the researcher has a strong prior belief that spillover effects are relatively flat with respect to treatment intensity, but ρ is high, then choosing an RS design with multiple positive treatment saturations will reduce precision without yielding novel insights, and the researcher is better off running a partial population design. However, partial population designs have the drawback that they cannot identify or rule out spillover effects – to do so, the researcher needs sufficient variation in the intensity of treatment.

Moreover, if the researcher is primarily interested in identifying slope effects, a design with no pure control is optimal. But such a design cannot identify treatment and spillover effects at any individual saturation or pooled across saturations. Thus, the optimal RS design for a slope analysis stands in sharp contrast to that for an individual saturation or pooled analysis. If the researcher seeks to identify both slope and individual or pooled effects, the optimal design will depend on the relative importance that the researcher places on each estimand, as well as the level of intra-cluster correlation.

Figure 4 depicts the trade-off between measuring pooled and slope effects for an RS design with a pure control and two interior saturations that are symmetric about 0.5. The precision of the pooled estimate is increasing as the two interior saturations approach 0.5, which corresponds to a partial population design, while the precision of the slope estimate first decreases and then increases as the interior saturations approach 0.5, capturing the non-monotonic effect of the distance between saturations on the precision of the slope estimand.

4 Application

To illustrate our results, we numerically characterize the optimal design for several objective functions and calculate the power of RS designs from published studies in economics and political science. These examples quantify the power trade-offs that arise between measuring individual, slope and pooled effects. The calculations are conducted using code we developed

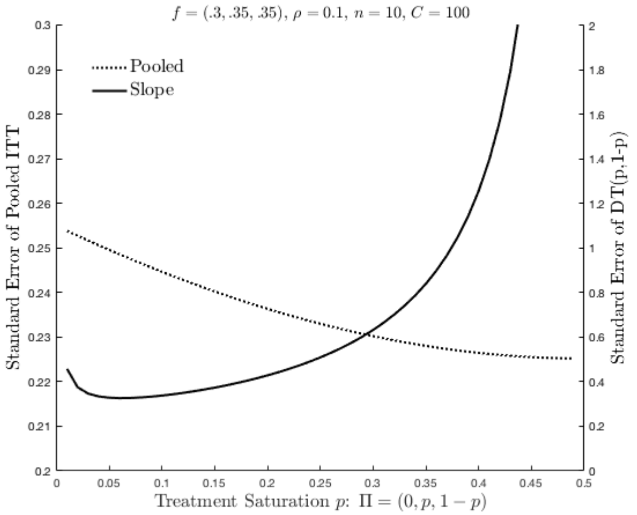

Figure 4. Trade-off between SEs of Pooled and Slope Estimands

as a tool for researchers.[29]

Suppose a researcher selects a sample of $C = 100$ clusters, each of which contain $n = 10$ individuals. As a benchmark, suppose the researcher uses a clustered design to identify the average treatment effect, $ITT(1)$. She implements the optimal clustered design, which assigns 50% of the clusters to the control group and 50% to the treatment group. The standard error of her estimate will depend on the intra-cluster correlation, ρ. We measure this standard error in terms of standard deviations of the distribution of potential outcomes (or equivalently, assume total variance is $\sigma^2 + \tau^2 = 1$). When $\rho = 0$, $\text{SE}_{ITT}(1) = 0.063$. It increases with ρ, rising to 0.087 when $\rho = 0.1$ and 0.200 when $\rho = 1$ (Table 1, Columns 1-3). The researcher cannot identify any spillover effects on treated or untreated individuals.

Next, suppose that the researcher also would like to measure spillover effects on untreated individuals and cares equally about the precision of the estimates of the pooled ITT and pooled SNT. Applying Corollary 1, the optimal design is a partial population experiment. From Proposition 3, we know that the optimal treatment saturation assigns 50% of the individuals in each treatment cluster to treatment, and the optimal share of control clusters ranges from 41% to 50% as ρ increases from 0 to 1 (Table 1, Columns 4-6). The optimal design equalizes the SEs for the ITT and SNT, which range from 0.076 to 0.200 as ρ increases from 0 to 1. These SEs are larger than the SEs for the ITT in the clustered design. The

[29] We created a Graphical User Interface (GUI) to answer many optimal design questions and calculate power for a given RS design. Code in R and Python is also available to conduct numerical optimization for more complex design questions. All code is available at http://pdel.ucsd.edu/solutions/index.html.

Table 1. Optimal Design for Pooled Estimands

	Clustered Design			Partial Population Design				
Objective Function	$\min_{p,\psi} \mathrm{SE}_{ITT}(p)$			$\min_{p,\psi}$ $\mathrm{SE}_{ITT}(p)$ $+$ $\mathrm{SE}_{SNT}(p)$			$\min_{p,\psi}$ $\mathrm{SE}_{ITT}(p)$ $+$ $2\,\mathrm{SE}_{SNT}(p)$	$\min_{p,\psi}$ $\mathrm{SE}_{SNT}(p)$; $\mathrm{SE}_{ITT}(p)$ $\leq .09$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ICC ρ	0.0	0.1	1.0	0.0	0.1	1.0	0.1	0.1
Pure control	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Optimal p	1.00	1.00	1.00	0.50	0.50	0.50	0.41	0.78
Optimal ψ	0.50	0.50	0.50	0.41	0.45	0.50	0.45	0.47
Optimal $f(p)$	0.50	0.50	0.50	0.59	0.55	0.50	0.55	0.53
$\mathrm{SE}_{ITT}(p)$	0.063	0.087	0.200	0.076	0.097	0.200	0.100	0.090
$\mathrm{SE}_{SNT}(p)$	.	.	.	0.076	0.097	0.200	0.094	0.117

Sample Size: $C = 100, n = 10$

source of the power loss is obvious: it stems from reassigning some treatment and control individuals to serve as within-cluster controls. The power loss is decreasing in ρ, as the share of clusters at each saturation becomes more important for precision, and this share approaches that of the clustered design.

Now suppose that the researcher cares more about estimating the pooled SNT, relative to the pooled ITT. A partial population experiment remains optimal, but the optimal treatment saturation decreases. If she places twice as much weight on the $\mathrm{SE}_{SNT}(p)$ in her objective function, relative to the $\mathrm{SE}_{ITT}(p)$, then for $\rho = 0.1$, the optimal design assigns 41% of the individuals in each treatment cluster to treatment and 45% of clusters to the control group (Table 1, Column 7). This produces SEs of 0.100 and 0.094 for the ITT and SNT, respectively.[30] Alternatively, the design that minimizes the SE of the SNT while maintaining a SE of 0.09 for the ITT (approximately the SE in the clustered design) assigns 78% of the individuals in each treatment cluster to treatment and 47% of clusters to the control group (Table 1, Column 8). This yields a SE of 0.117 for the SNT.

If a researcher wishes to estimate the slope effect for treated and untreated individuals, and does not care about identifying individual or pooled effects, then from Proposition 2, the optimal design will equally divide clusters between two interior saturations that are

[30] Moving to a more extreme objective that places nine times as much weight on the SNT does not substantially alter the share of clusters allocated to pure control (47%), but does significantly reduce the optimal treatment saturation (23%).

Table 2. Optimal Design for Slope Estimands and Precision in Existing Studies

	Optimal RS Designs				Existing Studies			
Objective Function	$\min_{p_1,p_2,f}$ $\text{SE}_{DT}(p_1,p_2)+$ $\text{SE}_{DNT}(p_1,p_2)$			$\min_{p_2,p_3,f}$ $\overline{\text{SE}}_{ITT}+$ $\overline{\text{SE}}_{SNT}+$ $\text{SE}_{DT}(p_2,p_3)+$ $\text{SE}_{DNT}(0,p_3)$	Banerjee et al.; Crepon et al.	Sinclair et al.	Baird et al.	Baird et al. II; $\min_f \overline{\text{SE}}_{SNT}$ s.t. $\overline{\text{SE}}_{ITT}$ $\leq .095$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ICC ρ	0.0	0.1	1.0	0.1	0.1	0.1	0.1	0.1
p_1	0.15	0.13	0.10	0.00	0.00	0.00	0.00	0.00
p_2	0.85	0.87	0.90	0.21	0.25	0.10	0.33	0.33
p_3	.	.	.	0.88	0.50	0.50	0.67	0.67
p_4	.	.	.	.	0.75	1.00	1.00	1.00
p_5	.	.	.	.	1.00	.	.	.
$f(p1)$	0.50	0.50	0.50	0.28	0.20	0.25	0.55	0.45
$f(p2)$	0.50	0.50	0.50	0.33	0.20	0.25	0.15	0.21
$f(p3)$	.	.	.	0.39	0.20	0.25	0.15	0.21
$f(p4)$	.	.	.	.	0.20	0.25	0.15	0.13
$f(p5)$	.	.	.	.	0.20	.	.	.
$\overline{\text{SE}}_{ITT}$	.	.	.	0.104	0.113	0.109	0.095	0.095
$\overline{\text{SE}}_{SNT}$	.	.	.	0.109	0.120	0.111	0.115	0.106
SE_{DT}	0.179	0.191	0.250	0.217	0.240	0.242	0.289	0.269
SE_{DNT}	0.179	0.191	0.250	0.192	0.240	0.274	0.251	0.221

Sample Size: $C = 100, n = 10$

symmetric about one half, and will not have a pure control. When $\rho = 0$, the optimal design assigns either 15% and 85% of individuals in each cluster to treatment, and the SE of the slope effect is 0.179 for both treated and non-treated individuals (Table 2, Column 1). Increasing ρ moves the optimal saturations further apart and increases the SEs for the slope effects (Table 2, Columns 2 - 3). When outcomes within a cluster are perfectly correlated, the optimal saturations are as far apart as possible while still maintaining at least one treated and one non-treated individual in each treatment cluster. This corresponds to saturations $1/n$ and $(n-1)/n$.

However, few researchers will likely be interested in designing an experiment to maximize the precision of slope estimates, at the expense of sacrificing the ability to identify standard estimands, such as the ITT. Suppose a researcher places equal weight on the precision of the pooled and slope estimates, and chooses a design with a control group $p_1 = 0$ and two positive saturations $p_3 > p_2 > 0$ to minimize

$$\min_{p_2, p_3, f} \overline{\text{SE}}_{ITT} + \overline{\text{SE}}_{SNT} + \text{SE}_{DT}(p_2, p_3) + \text{SE}_{DNT}(0, p_3).^{31} \tag{9}$$

When $\rho = 0.1$, the optimal design assigns either 21% and 88% of individuals in treated clusters to treatment, and allocates 33% of clusters to the low treatment saturation, 39% to the high treatment saturation, and the remaining 28% to the control group (Table 2, Column 4).[32] The SEs of 0.104 and 0.109 for the pooled ITT and SNT, respectively, are 7-13% larger than the SEs in the optimal partial population design (Table 1, Column 5).[33] The precision loss in moving from a partial population design to an RS design arises from the variation in treatment saturation. But this variation is precisely what enables the identification of slope effects. This illustrates the design trade-off discussed in Section 3.3.

Finally, we calculate the standard errors for RS designs used by three published studies. To facilitate comparability with the optimal designs discussed above, we use the same number of clusters ($C = 100$), individuals per cluster ($n = 10$) and intra-cluster correlation ($\rho = 0.1$) as in our examples, rather than the actual values from each study.[34]

We begin with the RS design used in Banerjee et al. (2012) and Crepon et al. (2013). Clusters were assigned to a pure control group and four equally spaced treatment saturations in equal shares, $\Pi = \{0, 0.25, 0.50, 0.75, 1\}$ and $f = \{0.2, 0.2, 0.2, 0.2, 0.2\}$. By virtue of having a pure control group and more than two interior saturations, this study design can identify the ITT and SNT (pooled and saturation-specific) effects and slope effects. Our power cal-

[31] With a control group, the saturations used to measure the slope for non-treated individuals (0 and p_3) are further apart than the saturations used to measure the slope for treated individuals (p_2 and p_3). This is because there are no treated individuals in the control group, so $\text{SE}_{DT}(0, p_3)$ is undefined.

[32] The optimal interior saturations are not symmetric around one half. Since the distance between saturations for the slope effect on non-treated individuals is greater than the distance between saturations for the slope effect on treated individuals, the small share of non-treated individuals at p_3 has less of an effect on the SE_{DNT}, relative to the effect of the small share of treated individuals at p_2 on the SE_{DT}.

[33] The SE_{DT} is larger than in the optimal slope design in Column 2 because the distance between saturations for treated individuals is smaller in this design, and fewer clusters are allocated to the saturations that identify the DT. The SE_{DNT} is approximately the same as in the optimal slope design in Column 2 – the distance between the saturations for untreated individuals is larger, but fewer clusters are allocated to the highest saturation.

[34] The pooled SEs are calculated for a model with constant treatment and spillover effects, which implies homoskedastic errors. These are lower bounds for the pooled SEs when treatment and spillover effects are not constant, and therefore, errors are heteroskedastic. Even if it is not possible to reject the null hypothesis of a zero slope effect, there may still be a small slope effect that creates heteroskedasticity. For example, in column 4, the design is powered to detect slope effects on treated individuals that are larger than 0.62. Suppose the true slope is 0.5. Then the design is not powered to detect an effect this small, but there will still be heteroskedasticity, and the pooled SE for treated individuals will be strictly larger than 0.104. To account for this, researchers should build some sample size cushion into their designs.

culations yield $\overline{\text{SE}}_{ITT} = 0.113$, $\overline{\text{SE}}_{SNT} = 0.120$ and $\text{SE}_{DT}(0.25, 1) = \text{SE}_{DNT}(0, .75) = 0.240$ (Table 2, Column 5). All of these SEs are higher than their counterparts in the design that minimizes the sum of these four SEs (Table 2, Column 4). This illustrates the power loss that arises from having a richer design that can, for example, test for the concavity of $ITT(p)$ and $SNT(p)$.

Our next example is the design used by Sinclair et al. (2012). They randomized clusters into a pure control and three different saturations, $\Pi = \{0, 1/n, 0.50, 1\}$ and $f = \{0.25, 0.25, 0.25, 0.25\}$, where $1/n$ is the saturation in which only one household is treated.[35] In addition to the estimands that can be identified in Banerjee et al. (2012) and Crepon et al. (2013), this design can also identify the TUT and the $ST(p)$. Our power calculations yield $\overline{\text{SE}}_{ITT} = .109$, $\overline{\text{SE}}_{SNT} = 0.111$, $\text{SE}_{DT}(1/n, 1) = 0.242$ and $\text{SE}_{DNT}(0, 0.5) = 0.274$ (Table 2, Column 6). The pooled SEs are quite similar to their counterparts in the design that minimizes the sum of these four SEs (Table 2, Column 4). However, the slope effect SEs are substantially higher, particularly for the non-treated (0.274 vs. 0.192). This is because the largest saturation containing within-cluster controls is 0.50, so the saturations used to identify the slope effect on non-treated individuals are too close together.

Our final example is Baird, McIntosh and Özler (2011), which has a pure control and three positive saturations, $\Pi = \{0, 0.33, 0.67, 1\}$ and $f = \{0.55, 0.15, 0.15, 0.15\}$. While the saturations in this design are equally spaced, they are not equally sized: the pure control group, at 55% of clusters, is much larger than the share assigned to any treatment saturation. The combination of having a larger control group and smaller variation in treatment saturations produces a smaller SE for the pooled ITT, relative to Banerjee et al. (2012) and Crepon et al. (2013), but higher SEs for the slope effects, particularly for treated individuals (Table 2, Column 7). The SE for the pooled SNT is 2 percentage points (or 21%) higher than that for the ITT, indicating that the pooled spillover effects on the untreated are underpowered, relative to the pooled treatment effects.

Given this large difference between the SEs for the pooled ITT and SNT, we explore whether it is possible to allocate clusters to this set of saturations in a way that reduces the SE of the pooled SNT, while maintaining the SE of the pooled ITT at 0.095. The optimal distribution of treatment saturations allocates a lower share of clusters to the pure

[35] The saturation of 0.5 is approximate, as one core household plus half of the remaining households were assigned to treatment in these clusters. For the purposes of our calculations, we use 0.5.

control group and saturation 1, and a higher share to the two interior saturations, 1/3 and 2/3 (Table 2, Column 8). Such a design dominates the original study design, as it not only lowers the SE for the pooled SNT, but it also decreases the SEs of both slope effects. The improved precision comes from redistributing clusters more efficiently between different treatment saturations, particularly by reallocating clusters from the pure control to interior saturations.

5 Conclusion

In recent years, empirical researchers have become increasingly interested in studying interference between subjects. Experiments designed to rigorously estimate spillovers open up a fascinating set of research questions and provide policy-relevant information about program design. For example, if a vaccination or a bed net distribution program with fixed resources can either treat 50% of all villages or 100% of half of them, measuring spillover effects will determine which treatment allocation maximizes the total benefit. Variation in the intensity of treatment can determine whether there are important scale or congestion effects that lead to differential impacts on prices, norms or behavior. Further, RCTs that fail to account for spillovers can produce biased estimates of intention-to-treat effects, while finding meaningful treatment effects but failing to observe deleterious spillovers can lead to misconstrued policy conclusions. The RS design presented here provides an experimental framework that can inform these policy questions and bolster both internal and external validity.

We formalize the design and analysis of such RS designs. We show that varying the treatment saturation across clusters generates direct experimental evidence on the nature of spillover effects for both treated and non-treated individuals. Having laid out the assumptions necessary to identify average treatment and spillover effects, we derive analytical closed-form expressions for the standard errors. This allows us to gain analytical insights into the optimal design of such experiments and derive ex-ante power calculations. The standard errors for the pooled intention-to-treat effect and spillover effect on the non-treated are directly related to the variation in treatment saturation. A design trade-off emerges: varying the treatment saturation allows the researcher to identify novel estimands, but this variation comes with a cost – it reduces the precision of the estimates of more basic estimands. This is an inherent feature of RS designs.

References

Alix-Garcia, Jennifer, Craig McIntosh, Katharine R. E. Sims, and Jarrod R. Welch, "The Ecological Footprint of Poverty Alleviation: Evidence from Mexico's Oportunidades Program," *The Review of Economics and Statistics*, May 2013, *95* (2), 417–435.

Angelucci, Manuela and Giacomo De Giorgi, "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?," *American Economic Review*, March 2009, *99* (1), 486–508.

Aronow, Peter, "A General Method for Detecting Interference in Randomized Experiments," *Sociological Methods Research*, 2012, *41* (1), 3–16.

Aronow, Peter M. and Cyrus Samii, "Estimating Average Causal Effects Under General Interference," *Annals of Applied Statistics*, forthcoming.

Athey, S. and G.W. Imbens, "The Econometrics of Randomized Experiments," *Handbook of Economic Field Experiments*, 2017, *1*, 73 – 140.

Babcock, Philip S. and John L. Hartman, "Networks and Workouts: Treatment Size and Status Specific Peer Effects in a Randomized Field Experiment," Working Paper 16581, National Bureau of Economic Research December 2010.

Baird, Sarah, Craig McIntosh, and Berk Özler, "Cash or Condition? Evidence from a Cash Transfer Experiment," *The Quarterly Journal of Economics*, 2011, *126* (4), 1709–1753.

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson, "The Diffusion of Microfinance," *Science*, July 2013, *341* (6144).

_ , Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh, "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training," Working Paper 17912, NBER March 2012.

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

Beaman, Lori A., "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S.," *The Review of Economic Studies*, 2012, *79* (1), 128–161.

Bloom, Howard S., "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs," *Evaluation Review*, October 1995, *19* (5), 547–556.

Bobba, Matteo and Jeremie Gignoux, "Neighborhood Effects in Integrated Social Policies," *World Bank Economic Review*, 2016.

Bobonis, Gustavo J. and Frederico Finan, "Neighborhood Peer Effects in Secondary School Enrollment Decisions," *The Review of Economics and Statistics*, November 2009, *91* (4), 695–716.

Busso, Matias and Sebastian Galiani, "The Causal Effect of Competition on Prices and Quality: Evidence from a Field Experiment," NBER Working Papers 20054, National Bureau of Economic Research, Inc April 2014.

Chen, Jiehua, Macartan Humphries, and Vijay Modi, "Technology Diffusion and Social Networks: Evidence from a Field Experiment in Uganda," 2010. Working Paper.

Conley, Timothy G. and Christopher R. Udry, "Learning about a New Technology: Pineapple in Ghana," *American Economic Review*, March 2010, *100* (1), 35–69.

Crepon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment," *The Quarterly Journal of Economics*, 2013, *128* (2), 531–580.

Duflo, Esther and Emmanuel Saez, "Participation and investment decisions in a retirement plan: the influence of colleagues' choices," *Journal of Public Economics*, July 2002, *85* (1), 121–148.

_ and _ , "The Role Of Information And Social Interactions In Retirement Plan Decisions: Evidence From A Randomized Experiment," *The Quarterly Journal of Economics*, August 2003, *118* (3), 815–842.

_ , Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit," Technical Report, C.E.P.R. Discussion Papers Jan 2007. CEPR Discussion Papers.

Gine, Xavier and Ghazala Mansuri, "Together We Will : Experimental Evidence on Female Voting Behavior in Pakistan," *AEJ: Microeconomics*, forthcoming.

Graham, Bryan S, Guido W Imbens, and Geert Ridder, "Measuring the effects of segregation in the presence of social spillovers: a nonparametric approach," 2010.

Hahn, Jinyong, Keisuke Hirano, and Dean Karlan, "Adaptive Experimental Design Using the Propensity Score," *Journal of Business & Economic Statistics*, January 2011, *29* (1), 96–108.

Hirano, Keisuke and Jinyong Hahn, "Design of randomized experiments to measure social interaction effects," *Economics Letters*, January 2010, *106* (1), 51–53.

Hudgens, Michael and Elizabeth Halloran, "Towards Causal Inference with Interference," *Journal of the American Statistical Association*, 2008, *103* (482), 832–842.

Killeen, GF, TA Smith, HM Ferguson, H Mshinda, S Abdulla et al., "Preventing childhood malaria in Africa by protecting adults from mosquitoes with insecticide-treated nets," *PLoS Med*, 2007, *4* (7), e229.

Lalive, Rafael and M. A. Cattaneo, "Social Interactions and Schooling Decisions.," *The Review of Economics and Statistics*, 2009, *91* (3), 457–477.

Liu, Lan and Michael G. Hudgens, "Large sample randomization inference of causal effects in the presence of interference," *Journal of the American Statistical Association*, 2014, *109* (505), 288–301.

Macours, Karen and Renos Vakis, "Changing Households' Investments and Aspirations through Social Interactions: Evidence from a Randomized Transfer Program in a Low-Income Country," Technical Report 2008. World Bank Working Paper 5137.

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

Manski, Charles F., "Identification of treatment response with social interactions," *The Econometrics Journal*, February 2013, *16* (1), S1–S23.

McIntosh, Craig, Tito Alegria, Gerardo Ordonez, and Rene Zenteno, "Infrastructure Impacts and Budgeting Spillovers: The Case of Mexico's Habitat Program," 2013. Working Paper.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 01 2004, *72* (1), 159–217.

Moffitt, Robert A., "Policy Interventions, Low-Level Equilibria And Social Interactions," in Steven Durlauf and Peyton Young, eds., *Social Dynamics*, MIT Press 2001, pp. 45–82.

Munshi, Kaivan, "Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market," *Quarterly Journal of Economics*, 2003, *118* (2), 549–599.

Oster, Emily and Rebecca Thornton, "Determinants of Technology Adoption: Peer Effects in Menstrual Cup Take-Up," *Journal of the European Economic Association*, 2012, *10* (6), 1263–1293.

Sinclair, Betsy, Margaret McConnell, and Donald P. Green, "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments," *American Journal of Political Science*, 2012, *56* (4), 1055–1069.

Sobel, Michael E., "What Do Randomized Studies of Housing Mobility Demonstrate?: Causal Inference in the Face of Interference," 2006.

Tchetgen, Eric J. Tchetgen and Tyler VanderWeele, "On Causal Inference in the Presence of Interference," *Statistical Methods in Medical Research*, 2010, *21* (1), 55–75.

Toulis, Panos and Edward Kao, "Estimation of Causal Peer Influence Effects," *Journal of Machine Learning Research*, 2013, *28.* Proceedings of the 30th International Conference on Machine Learning Research.

A Appendix: Proofs from Section 3

A.1 Preliminary Calculations

This section provides background material used in the proofs of Lemma 2 and Theorems 1 and 2. Consider the OLS estimate of

$$Y_{ic} = X'_{ic}\boldsymbol{\beta} + \varepsilon_{ic}, \tag{10}$$

where X_{ic} is a vector of treatment status covariates and ε_{ic} is an error term with a block-diagonal variance-covariance matrix. Given $X'_c = [X_{1c} \text{ ... } X_{nc}]$ and $\varepsilon'_c = [\varepsilon_{1c} \text{ ... } \varepsilon_{nc}]$, let $E[\varepsilon_c \varepsilon'_c | X_c] = \sigma^2 \mathbf{I}_n + \tau^2 \mathbf{1}_n$ denote the within-cluster variance-covariance matrix, where $\mathbf{1}_n$ is the $n \times n$ matrix of ones. Between clusters, $E[\varepsilon_{ic}\varepsilon_{jd}|\mathbf{X}] = 0$ for all $c \neq d$, where $\mathbf{X}' = [X'_1 \text{ ... } X'_C]$. Let $\mathbf{Y}$ denote the vector of observed potential outcomes. The estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = A^{-1}\mathbf{X}'\mathbf{Y} \tag{11}$$

and the exact finite sample variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= A^{-1}\left(\sum_{c=1}^{C} X'_c E[\varepsilon_c \varepsilon'_c | X_c] X_c\right) A^{-1} \\
&= A^{-1}\left(\sum_{c=1}^{C} X'_c (\sigma^2 \mathbf{I}_n + \tau^2 \mathbf{1}_n) X_c\right) A^{-1} \\
&= \sigma^2 A^{-1} + \tau^2 A^{-1} B A^{-1}, \qquad (12)
\end{aligned}$$

where

$$\begin{aligned}
A &\equiv \sum_{c=1}^{C} X'_c X_c = \sum_{c=1}^{C}\sum_{i=1}^{n} X_{ic}X'_{ic} \qquad (13) \\
B &\equiv \left(\sum_{c=1}^{C} X'_c \mathbf{1}_n X_c\right). \qquad (14)
\end{aligned}$$

A.2 Proofs of Lemmas 1, 2 and 3

Proof of Lemma 1. Suppose the realized treatment vector for cluster c is $\mathbf{t} = (t_1, ..., t_n)$. Then $E[\varepsilon_{ic}|T_c = \mathbf{t}] = E[Y_{ic}(\mathbf{t}) - \overline{Y}(t_i, p(\mathbf{t}))] = 0$. The variance of the error is $E[\varepsilon_{ic}^2|T_c = \mathbf{t}] = E[(Y_{ic}(\mathbf{t}) - \overline{Y}(t_i, p(\mathbf{t})))^2] = \tau^2 + \sigma^2$. The covariance of the error between individuals in the same cluster is $E[\varepsilon_{ic}\varepsilon_{jc}|T_c = \mathbf{t}] = E[(Y_{ic}(\mathbf{t}) - \overline{Y}(t_i, p(\mathbf{t})))(Y_{jc}(\mathbf{t}) - \overline{Y}(t_j, p(\mathbf{t})))] = \tau^2$. Errors across clusters are not correlated, since outcomes across clusters are not correlated.

Proof of Lemma 2. The error defined in (6) is not strictly exogenous, so we need to establish directly that $\hat{\boldsymbol{\beta}}$ is unbiased for (10) when $X'_{ic} = [\ 1 \quad T_{ic} \quad S_{ic}\ ]$. From $\hat{\boldsymbol{\beta}} = A^{-1}\mathbf{X}'\mathbf{Y}$, with

$$A = \sum_{c=1}^{C}\sum_{i=1}^{n} \begin{bmatrix} 1 & T_{ic} & S_{ic} \\ T_{ic} & T_{ic}^2 & T_{ic}S_{ic} \\ S_{ic} & T_{ic}S_{ic} & S_{ic}^2 \end{bmatrix} = nC \begin{bmatrix} 1 & \mu & \mu_S \\ \mu & \mu & 0 \\ \mu_S & 0 & \mu_S \end{bmatrix} \tag{15}$$

and

$$\mathbf{X}'\mathbf{Y} = \sum_{c=1}^{C}\sum_{i=1}^{n} \begin{bmatrix} Y_{ic} & T_{ic}Y_{ic} & S_{ic}Y_{ic} \end{bmatrix}',$$

$$\hat{\boldsymbol{\beta}} = \frac{1}{nC}\sum_{c=1}^{C}\sum_{i=1}^{n} \begin{bmatrix} \frac{1}{\psi}C_{ic}Y_{ic} & \frac{1}{\mu}T_{ic}Y_{ic} - \frac{1}{\psi}C_{ic}Y_{ic} & \frac{1}{\mu_S}S_{ic}Y_{ic} - \frac{1}{\psi}C_{ic}Y_{ic} \end{bmatrix}'.$$

Therefore, $E[\hat{\beta}_0|\mathbf{X}] = \overline{Y}(0,0)$,

$$E[\hat{\beta}_1|\mathbf{X}] = \frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}} pf(p)\overline{Y}(1,p) - \overline{Y}(0,0) = \overline{Y}(1) - \overline{Y}(0,0)$$

and

$$E[\hat{\beta}_2|\mathbf{X}] = \frac{1}{\mu_S}\sum_{p\in\Pi\setminus\{0\}} (1-p)f(p)\overline{Y}(0,p) - \overline{Y}(0,0) = \overline{Y}(0) - \overline{Y}(0,0),$$

which establishes that $\hat{\boldsymbol{\beta}}$ is unbiased.

Lemma 4. *Assume Assumptions 1 and 2. Given RS design* (Π, f)*, the error defined in* (6) *has individual variances*

$$\begin{aligned} E[\varepsilon_{ic}^2|T_{ic} = 1, T_c = \mathbf{t}] &= \sigma^2 + \tau^2 + (\overline{Y}(1, p(\mathbf{t})) - \overline{Y}(1))^2 \\ E[\varepsilon_{ic}^2|S_{ic} = 1, T_c = \mathbf{t}] &= \sigma^2 + \tau^2 + (\overline{Y}(0, p(\mathbf{t})) - \overline{Y}(0))^2 \\ E[\varepsilon_{ic}^2|C_{ic} = 1, T_c = \mathbf{t}] &= \sigma^2 + \tau^2 \end{aligned}$$

and within-cluster covariances

$$\begin{aligned}
E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic} = T_{jc} = 1, T_c] &= \tau^2 + (\overline{Y}(1, p(\mathbf{t})) - \overline{Y}(1))^2 \\
E[\varepsilon_{ic}\varepsilon_{jc}|S_{ic} = S_{jc} = 1, T_c] &= \tau^2 + (\overline{Y}(0, p(\mathbf{t})) - \overline{Y}(0))^2 \\
E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic} = S_{jc} = 1, T_c] &= \tau^2 + (\overline{Y}(0, p(\mathbf{t})) - \overline{Y}(0))(\overline{Y}(1, p(\mathbf{t})) - \overline{Y}(1)) \\
E[\varepsilon_{ic}\varepsilon_{jc}|C_{ic} = C_{jc} = 1, T_c] &= \tau^2.
\end{aligned}$$

The errors are uncorrelated across clusters, $E[\varepsilon_{ic}\varepsilon_{jd}|T_c, T_d] = 0$ *for* $c \neq d$.

Proof. The variance of the error for treated individuals is

$$\begin{aligned}
E[\varepsilon_{ic}^2|T_{ic} = 1, T_c = \mathbf{t}] &= E[(Y_{ic}(1, \mathbf{t}_{-i}) - \overline{Y}(1))^2] \\
&= E[Y_{ic}(1, \mathbf{t}_{-i})^2 - 2\overline{Y}(1)Y_{ic}(1, \mathbf{t}_{-i}) + \overline{Y}(1)^2] \\
&= \tau^2 + \sigma^2 + \overline{Y}(1, p(\mathbf{t}))^2 - 2\overline{Y}(1)\overline{Y}(1, p(\mathbf{t})) + \overline{Y}(1)^2 \\
&= \tau^2 + \sigma^2 + (\overline{Y}(1, p(\mathbf{t})) - \overline{Y}(1))^2.
\end{aligned}$$

The covariance of the error between treated individuals in the same cluster is

$$\begin{aligned}
E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic} = T_{jc} = 1, T_c = \mathbf{t}] &= E[(Y_{ic}(1, \mathbf{t}_{-i}) - \overline{Y}(1))(Y_{jc}(1, \mathbf{t}_{-j}) - \overline{Y}(1))] \\
&= \tau^2 + \overline{Y}(1, p(\mathbf{t}))^2 - 2\overline{Y}(1)\overline{Y}(1, p(\mathbf{t})) + \overline{Y}(1)^2 \\
&= \tau^2 + (\overline{Y}(1, p(\mathbf{t})) - \overline{Y}(1))^2.
\end{aligned}$$

The other variances and covariances are analogous. Errors across clusters are not correlated since outcomes across clusters are not correlated. □

Proof of Lemma 3. Suppose treatment effects are constant on $\Pi \setminus \{0\}$. Then $\overline{Y}(1, p) = \overline{Y}(1)$ and $\overline{Y}(0, p) = \overline{Y}(0)$ for all $p \in \Pi \setminus \{0\}$. From Lemma 4, $E[\varepsilon_{ic}^2|T_{ic} = 1, T_c = \mathbf{t}] = \sigma^2 + \tau^2 + (\overline{Y}(1, p(\mathbf{t})) - \overline{Y}(1))^2 = \sigma^2 + \tau^2$, with similar calculations for the other variances and covariances. Therefore, the variance-covariance matrix reduces to a block-diagonal structure with variance $\sigma^2 + \tau^2$ and covariance τ^2.

For the other direction, suppose the variance-covariance matrix is block-diagonal with variance $\sigma^2 + \tau^2$ and covariance τ^2. Then $\overline{Y}(1, p(\mathbf{t})) - \overline{Y}(1) = 0$ for all $\mathbf{t} \neq \mathbf{0}$ that arise in (Π, f). But then there must be no variation in the population average potential outcome across positive saturations for treated individuals. Similarly, there must be no variation in the population average potential outcome across positive saturations for within-cluster controls. Therefore, treatment effects are constant on $\Pi \setminus \{0\}$.

A.3 Proof of Theorems 1 and 2

Proof of Theorem 1. Assume Assumptions 1 and 2. Consider an RS design with two interior saturations, p_1 and p_2, and a pure control. Let $T_{kic} \equiv T_{ic} * \mathbb{1}\{P_c = p_k\}$ and $S_{1ic} \equiv S_{ic} * \mathbb{1}\{P_c = p_k\}$ for $k = 1, 2$. We want to compute $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ for (10) when

$$X'_{ic} = [\begin{array}{ccccc} 1 & T_{1ic} & S_{1ic} & T_{2ic} & S_{2ic} \end{array}].$$

By Lemma 1, the error distribution is block-diagonal. Let $\mu_k \equiv p_k f(p_k)$, $s_k \equiv (1-p_k)f(p_k)$, $\eta_k \equiv p_k^2 f(p_k)$ and $q_k \equiv (1-p_k)^2 f(p_k) = s_k - \mu_k + \eta_k$. From (12), $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2 A^{-1} + \tau^2 A^{-1} B A^{-1}$, with

$$A = \sum_{c=1}^{C}\sum_{i=1}^{n} \begin{bmatrix} 1 & T_{1ic} & S_{1ic} & T_{2ic} & S_{2ic} \\ T_{1ic} & T_{1ic}^2 & S_{1ic}T_{1ic} & T_{2ic}T_{1ic} & S_{2ic}T_{1ic} \\ S_{1ic} & T_{1ic}S_{1ic} & S_{1ic}^2 & T_{2ic}S_{1ic} & S_{2ic}S_{1ic} \\ T_{2ic} & T_{1ic}T_{2ic} & S_{1ic}T_{2ic} & T_{2ic}^2 & S_{2ic}T_{2ic} \\ S_{2ic} & T_{1ic}S_{2ic} & S_{1ic}S_{2ic} & T_{2ic}S_{2ic} & S_{2ic}^2 \end{bmatrix} = nC \begin{bmatrix} 1 & \mu_1 & s_1 & \mu_2 & s_2 \\ \mu_1 & \mu_1 & 0 & 0 & 0 \\ s_1 & 0 & s_1 & 0 & 0 \\ \mu_2 & 0 & 0 & \mu_2 & 0 \\ s_2 & 0 & 0 & 0 & s_2 \end{bmatrix}$$

and

$$B = \sum_{c=1}^{C} \left(\begin{bmatrix} n \\ \sum_{i=1}^n T_{1ic} \\ \sum_{i=1}^n S_{1ic} \\ \sum_{i=1}^n T_{2ic} \\ \sum_{i=1}^n S_{2ic} \end{bmatrix} * \begin{bmatrix} n \\ \sum_{i=1}^n T_{1ic} \\ \sum_{i=1}^n S_{1ic} \\ \sum_{i=1}^n T_{2ic} \\ \sum_{i=1}^n S_{2ic} \end{bmatrix}' \right) = n^2 C \begin{bmatrix} 1 & \mu_1 & s_1 & \mu_2 & s_2 \\ \mu_1 & \eta_1 & \mu_1 - \eta_1 & 0 & 0 \\ s_1 & \mu_1 - \eta_1 & q_1 & 0 & 0 \\ \mu_2 & 0 & 0 & \eta_2 & \mu_2 - \eta_2 \\ s_2 & 0 & 0 & \mu_2 - \eta_2 & q_2 \end{bmatrix},$$

where the second equalities follow from $\sum_{i=1}^n T_{kic} = np_k$, $\sum_{i=1}^n S_{kic} = n(1-p_k)$, $\sum_{c=1}^C \sum_{i=1}^n T_{kic} = np_k \times Cf(p_k) = nC\mu_k$, $\sum_{c=1}^C \sum_{i=1}^n S_{kic} = n(1-p_k) \times Cf(p_k) = nCs_k$, $T_{kic}^2 = T_{kic}$, $\sum_{c=1}^C (\sum_{i=1}^n T_{kic})^2 = n^2 p_k^2 \times Cf(p_k) = n^2 C\eta_k$, $\sum_{c=1}^C (\sum_{i=1}^n T_{kic} \times \sum_{i=1}^n S_{kic}) = n^2 p_k(1-p_k) \times Cf(p_k) = n^2 C(\mu_k - \eta_k)$, $(\sum_{i=1}^n T_{1ic})(\sum_{i=1}^n S_{2ic}) = 0$, and other analogous calculations. Taking the diagonal entries of

$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2 A^{-1} + \tau^2 A^{-1} B A^{-1}$ yields

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_{1p_j}) &= \frac{1}{nC} * \left\{ n\tau^2 \left(\frac{1}{f(p_j)} + \frac{1}{\psi} \right) + \sigma^2 \left(\frac{1}{\mu_j} + \frac{1}{\psi} \right) \right\} && (16) \\ \mathrm{Var}(\hat{\beta}_{2p_j}) &= \frac{1}{nC} * \left\{ n\tau^2 \left(\frac{1}{f(p_j)} + \frac{1}{\psi} \right) + \sigma^2 \left(\frac{1}{s_j} + \frac{1}{\psi} \right) \right\} && (17) \end{aligned}$$

for each $p_j \in \Pi \setminus \{0\}$. Taking the square root yields the standard errors.

To compute the SE_{DT}, note $\mathrm{Var}(\hat{DT}(p_1, p_2)) = \mathrm{Var}(\hat{\beta}_{1p_2} - \hat{\beta}_{1p_1}) / (p_2 - p_1)^2$ and

$$\mathrm{Cov}(\hat{\beta}_{1p_1}, \hat{\beta}_{1p_2}) = \frac{n\tau^2 + \sigma^2}{\psi n C},$$

where the expression for the covariance comes from the matrix $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$. Therefore,

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_{1p_2} - \hat{\beta}_{1p_1}) &= \mathrm{Var}(\hat{\beta}_{1p_1}) + \mathrm{Var}(\hat{\beta}_{1p_2}) - 2\,\mathrm{Cov}(\hat{\beta}_{1p_1}, \hat{\beta}_{1p_2}) \\ &= \frac{1}{nC} * \left\{ n\tau^2 \left(\frac{1}{f(p_1)} + \frac{1}{f(p_2)} \right) + \sigma^2 \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} \right) \right\}, && (18) \end{aligned}$$

where $\mathrm{Var}(\hat{\beta}_{1p_k})$ follows from (22). Similarly, $\mathrm{Var}(D\hat{N}T(p_1, p_2)) = \mathrm{Var}(\hat{\beta}_{2p_2} - \hat{\beta}_{2p_1}) / (p_2 - p_1)^2$, where

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_{2p_2} - \hat{\beta}_{2p_1}) &= \mathrm{Var}(\hat{\beta}_{2p_2}) + \mathrm{Var}(\hat{\beta}_{2p_1}) - 2\,\mathrm{Cov}(\hat{\beta}_{2p_1}, \hat{\beta}_{2p_2}) \\ &= \frac{1}{nC} * \left\{ n\tau^2 \left(\frac{1}{f(p_1)} + \frac{1}{f(p_2)} \right) + \sigma^2 \left(\frac{1}{s_1} + \frac{1}{s_2} \right) \right\}. && (19) \end{aligned}$$

Dividing (18) and (19) by $(p_2 - p_1)^2$ and taking the square root yields the $SE_{DT}(p_1, p_2)$ and $SE_{DNT}(p_1, p_2)$. It is straightforward to extend these expressions to more than two interior saturations.

Proof of Theorem 2. Assume Assumptions 1 and 2. Consider an RS design (Π, f) with at least one interior saturation and a pure control, and suppose that treatment and spillover effects are constant on $\Pi \setminus \{0\}$. We want to compute $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ for (10) when $X'_{ic} = [\ 1 \quad T_{ic} \quad S_{ic}\ ]$. By Lemma 2, the error distribution is block-diagonal. From (12), $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2 A^{-1} + \tau^2 A^{-1} B A^{-1}$, with

$$A = \sum_{c=1}^{C} \sum_{i=1}^{n} \begin{bmatrix} 1 & T_{ic} & S_{ic} \\ T_{ic} & T_{ic}^2 & T_{ic} S_{ic} \\ S_{ic} & T_{ic} S_{ic} & S_{ic}^2 \end{bmatrix} = nC \begin{bmatrix} 1 & \mu & \mu_S \\ \mu & \mu & 0 \\ \mu_S & 0 & \mu_S \end{bmatrix} \qquad (20)$$

and

$$\begin{aligned} B &= \sum_{c=1}^{C} \begin{bmatrix} n^2 & n\sum_{i=1}^n T_{ic} & n\sum_{i=1}^n S_{ic} \\ n\sum_{i=1}^n T_{ic} & \left(\sum_{i=1}^n T_{ic}\right)^2 & \left(\sum_{i=1}^n T_{ic}\right)\left(\sum_{i=1}^n S_{ic}\right) \\ n\left(\sum_{i=1}^n S_{ic}\right) & \left(\sum_{i=1}^n T_{ic}\right)\left(\sum_{i=1}^n S_{ic}\right) & \left(\sum_{i=1}^n S_{ic}\right)^2 \end{bmatrix} \\ &= n^2 C \begin{bmatrix} 1 & \mu & \mu_S \\ \mu & \eta^2 & \mu - \eta^2 \\ \mu_S & \mu - \eta^2 & \mu_S - \mu + \eta^2 \end{bmatrix}, \end{aligned} \tag{21}$$

where the second equalities follow from $\sum_{i=1}^n T_{ic} = nP_c$, $\sum_{i=1}^n S_{ic} = n(1-P_c)\mathbb{1}_{P_c>0}$, $\sum_{c=1}^C nP_c = nC\sum_{p\in\Pi} pf(p) = nC\mu$, $\sum_{c=1}^C(\sum_{i=1}^n T_{ic})^2 = n^2\sum_{c=1}^C P_c^2 = n^2C\eta^2$, $\sum_{c=1}^C(\sum_{i=1}^n T_{ic}\sum_{i=1}^n S_{ic}) = n^2\sum_{c=1}^C P_c(1-P_c) = n^2C(\mu-\eta^2)$ and $\sum_{c=1}^C(\sum_{i=1}^n S_{ic})^2 = n^2\sum_{c=1}^C(1-P_c)^2\mathbb{1}_{P_c>0} = n^2C(1-\psi-2\mu+\eta^2) = n^2C(\mu_S-\mu+\eta^2)$. Taking the diagonal entries of $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ yields

$$\text{Var}(\hat{\beta}_1) = \frac{1}{nC} * \left\{ n\tau^2\left(\frac{\eta^2}{\mu^2} + \frac{1}{\psi}\right) + \sigma^2\left(\frac{1}{\mu} + \frac{1}{\psi}\right)\right\} \tag{22}$$

$$\text{Var}(\hat{\beta}_2) = \frac{1}{nC} * \left\{ n\tau^2\left(\frac{\mu_S - \mu + \eta^2}{\mu_S^2} + \frac{1}{\psi}\right) + \sigma^2\left(\frac{1}{\mu_S} + \frac{1}{\psi}\right)\right\}. \tag{23}$$

Substituting $\rho = \tau^2/(\tau^2+\sigma^2)$ and the expression relating η^2 and η_T^2 defined in (7), then taking the square root yields the standard errors.

A.4 Proofs of Optimal Design Results

Proof of Proposition 1. Suppose that there are two interior saturations, p_1 and p_2. Without loss of generality, let $p_1 > p_2 \geq 0.5$. Let f denote $f(p_1)$. Then $f(p_2) = 1-\psi-f$. First fix ψ and consider the optimal $f \in [0, 1-\psi]$ to minimize (3). If $\rho = 1$, then $\text{SE}_{ITT}(p) = \text{SE}_{SNT}(p)$, and the first order condition is

$$\sqrt{\frac{1}{f}+\frac{1}{\psi}}\left(\sqrt{\frac{1}{1-\psi-f}+\frac{1}{\psi}}\right)^{-1} = f^2(1-\psi-f)^{-2}.$$

Therefore, $f^* = 1 - f^* - \psi$ and an equal share of clusters are allocated to each treatment saturation, $f(p_1) = f(p_2) = (1-\psi)/2$. This allocation equalizes all four SEs. If $\rho = 0$, then for $p > 0.5$, $\text{SE}_{ITT}(p) < \text{SE}_{SNT}(p)$, so it will not be possible to equalize the SEs. The first order condition is

$$\frac{1}{p_1 f^2\sqrt{\frac{1}{p_1 f}+\frac{1}{\psi}}}+\frac{1}{(1-p_1)f^2\sqrt{\frac{1}{(1-p_1)f}+\frac{1}{\psi}}}-\frac{1}{p_2(1-\psi-f)^2\sqrt{\frac{1}{p_2(1-\psi-f)}+\frac{1}{\psi}}}$$
$$-\frac{1}{(1-p_2)(1-\psi-f)^2\sqrt{\frac{1}{(1-p_2)(1-\psi-f)}+\frac{1}{\psi}}}=0. \tag{24}$$

For any $f\in(0,1)$, the function

$$\left(x\sqrt{\frac{1}{xf}+\frac{1}{\psi}}\right)^{-1}$$

is decreasing and strictly convex in x for all $x\in(0,1)$. By assumption, $p_1>p_2>1-p_2>1-p_1$. Therefore, if $f=1-\psi-f$, (24) is positive. The marginal value, measured in terms of the marginal reduction in the SE, from an additional individual in a given group is concave. When $f=1-\psi-f$, it is highest for within-cluster controls at saturation p_1, followed by within-cluster controls at p_2, then treated individuals at p_2, and finally, treated individuals at saturation p_1. Given this concavity, when $f=1-\psi-f$, the marginal value of an additional *cluster* at saturation p_1 is higher than the marginal value of an additional cluster at saturation p_2, and it must be that $f^*>1-f^*-\psi$. Similarly, for $\rho\in(0,1)$, it must be that $f^*>1-f^*-\psi$.

Next we consider the optimal share of control clusters. An additional control cluster reduces the SEs for every term in (3), whereas an additional cluster at saturation p reduces the SEs for only that saturation. If $\rho=0$, then non-control clusters are divided evenly between each treatment saturation, $f=(1-\psi)/2$, and the SEs are equalized. Therefore, the objective simplifies to

$$\min_{\psi\in[0,1]}\sqrt{\frac{2}{1-\psi}+\frac{1}{\psi}},$$

which has solution $\psi^*>f^*$. If $\rho>0$, then the optimal control size depends on the saturations. Suppose p_1 is close to one. Then when $\psi=f$, the marginal value of an additional cluster at saturation p_1 can be higher than the marginal value of an additional control cluster, due to the low number of within-cluster controls in each p_1 cluster. Therefore, it is possible that $\psi^*<f^*$. There are a large number of treated individuals at extreme p_1, so it may also be the case that $\psi^*<(1-p_1)f^*$. However, it can never be that $\psi^*<(1-p_1)f^*$, as it would be optimal to reallocate a cluster from p_1 to the control.

Proof of Proposition 2. Consider an RS design with two interior saturations and let $p_2 > p_1$. Denote the size of saturation p_1 by $f(p_1) = f$, and $f(p_2) = 1 - f$. Let $\Delta \equiv p_2 - p_1$ denote the distance between the two saturations. The SE_{DT} and SE_{DNT} are concave in p_1, p_2 and $1-p_1, 1-p_2$, respectively, and symmetric about one half. Therefore, the optimal design will equalize the SEs and minimizing (4) is equivalent to solving:

$$\min_{f,\Delta,p_1\in[0,1]^3} \frac{1}{\Delta^2}\left(n\rho\left(\frac{1}{f}+\frac{1}{1-f}\right)\right)+$$
$$\frac{1}{\Delta^2}\left((1-\rho)\left(\frac{1}{fp_1}+\frac{1}{(1-f)(p_1+\Delta)}+\frac{1}{f(1-p_1)}+\frac{1}{(1-f)(1-\Delta-p_1)}\right)\right).$$

Fixing Δ, the FOC wrt p_1 and f are

$$\frac{f}{1-f} = \frac{p_1^2(1-p_1)^2(2(p_1+\Delta)-1)}{(p_1+\Delta)^2(1-p_1-\Delta)^2(1-2p_1)}$$
$$\left(\frac{1}{(1-f)^2}\right)\left(\frac{1-\rho}{(p_1+\Delta)(1-\Delta-p)}+n\rho\right) = \left(\frac{1}{f^2}\right)\left(\frac{1-\rho}{p_1(1-p_1)}+n\rho\right)$$

The solution to this FOC is $p_1 = (1-\Delta)/2$ and $f = 0.5$, which implies $p_2 = p_1 + \Delta = (1+\Delta)/2$. Therefore, the optimal size of each saturation bin is equal and the optimal saturations are symmetric about 0.5. The Δ that minimizes (4) is equivalent to solving:

$$\min_{\Delta} \frac{1}{\Delta^2}\left(n\rho + 8(1-\rho)\left(\frac{1}{1-\Delta^2}\right)\right).$$

The optimal Δ^* solves:

$$\frac{n\rho}{8(1-\rho)} = \frac{2\Delta^2-1}{(1-\Delta^2)^2}.$$

If $\rho = 0$, then $2\Delta^2 - 1 = 0$, yielding $\Delta^* = \sqrt{2}/2$. Note that $(2\Delta^2-1)/(1-\Delta^2)^2$ is monotonically increasing for $\Delta \in [0,1)$, and strictly positive for $\Delta > \sqrt{2}/2$. The left hand side is increasing in ρ and n, and strictly positive when $\rho > 0$. Therefore, $\Delta^* > \sqrt{2}/2$ for $\rho > 0$, and Δ^* is increasing in ρ and n. If $\rho > 0$, then the left hand side converges to ∞ as $n \to \infty$, which requires $\Delta^* \to 1$. At the extreme, when $\rho = 1$, the optimal saturations are the furthest apart saturations that maintain one treated individual and one with-cluster control individual in each saturation, $p_1^* = 1/n$ and $p_2^* = (n-1)/n$.

Proof of Corollary 1. Fixing $(\mu, \psi) \in (0,1)$, $\overline{\mathrm{SE}}_{ITT}$ and $\overline{\mathrm{SE}}_{SNT}$ are both minimized at $\eta_T^2 = 0$. This corresponds to a partial population design with a control group of size ψ and a treatment saturation of $p = \mu/(1-\psi)$.

Proof of Proposition 3. Consider a partial population design with share of control clusters ψ and share of treated individuals μ. Then

$$\mathrm{SE}(\hat{\beta}_1) \;=\; \sqrt{\frac{\tau^2+\sigma^2}{nC} * \left\{ n\rho \left(\frac{1}{(1-\psi)\,\psi} \right) + (1-\rho) \left(\frac{1}{\mu} + \frac{1}{\psi} \right) \right\}}. \tag{25}$$

$\mathrm{SE}(\hat{\beta}_2)$ is analogous, replacing μ with $1-\mu-\psi$. Fixing ψ, the optimal treatment share solves $\min_\mu \mathrm{SE}(\hat{\beta}_1) + \mathrm{SE}(\hat{\beta}_2)$, which has solution $\mu = (1-\psi)/2$. This implies $\mu_S = \mu$, which corresponds to a partial population experiment with treatment saturation $p^* = 0.5$. Plugging $\mu = (1-\psi)/2$ into (25) yields $\mathrm{SE}(\hat{\beta}_1) = \mathrm{SE}(\hat{\beta}_2)$. Thus, the optimal share of control clusters solves

$$\min_{\psi} n\rho \left(\frac{1}{\psi(1-\psi)} \right) + (1-\rho) \left(\frac{1+\psi}{\psi(1-\psi)} \right). \tag{26}$$

When $\rho = 0$, (26) is minimized at $\psi^* = \sqrt{2}-1$. When $\rho = 1$, (26) is minimized at $\psi^* = 0.5$. When $\rho \in (0,1)$, the general FOC for (26) is

$$(1-\rho)(\psi^2 + 2\psi - 1) + n\rho(2\psi - 1) = 0. \tag{27}$$

Using the quadratic formula with $a = 1-\rho$, $b = 2(1-\rho+n\rho)$ and $c = -(1-\rho+n\rho)$ to solve for ψ yields the optimal control group size. Given that $(1+\psi)/\psi(1-\psi)$ and $1/\psi(1-\psi)$ are both convex and have unique minimums, any weighted sum of these functions is minimized at a value ψ^* that lies between the minimum of each function. Therefore, when $\rho \in (0,1)$, $\psi^* \in (\sqrt{2}-1, 1/2)$. Taking the derivative of (27) with respect to n yields

$$\frac{\partial \psi^*}{\partial n} = \frac{\rho(1-2\psi^*)}{2n\rho + 2(1-\rho)(1+\psi^*)} \geq 0$$

A similar calcuation establishes that $\frac{\partial \psi^*}{\partial \rho} > 0$.