Optimal Design of Experiments in the Presence of Interference*

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

July 2017

Abstract

This paper formalizes the optimal design of randomized controlled trials (RCTs) in the presence of interference between units, where an individual's outcome depends on the behavior and outcomes of others in her group. We focus on randomized saturation (RS) designs, which are two-stage RCTs that first randomize the treatment saturation of a group, then randomize individual treatment assignment. Our main contributions are to map the potential outcomes framework with partial interference to a regression model with clustered errors, calculate the statistical power of different RS designs, and derive analytical insights for how to optimally design an RS experiment. We show that the power to detect average treatment effects declines precisely with the ability to identify novel treatment and spillover estimands, such as how effects vary with the intensity of treatment. We provide software that assists researchers in designing RS experiments.

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 – settings in which 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 even have an unpolluted counterfactual? In the presence of interference, a full understanding of the policy environment requires a measure of spillover effects that are not captured by (or worse, are sources 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 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. Researchers have used a variety of methods, including (i) using experimental variation across treatment groups, (ii) leaving some members of a group untreated, (iii) exploiting exogenous variation in within-network treatments, and (iv) intersecting an experiment with pre-existing networks.[1]

The recent interest in interference between individuals has also spawned a rich econometrics literature. Aronow and Samii (forthcoming) and Manski (2013) consider the most general settings, in which there are arbitrary forms of independence and treatment assignment dependencies. In this paper, we study 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. *Partial population* experiments (Moffitt 2001), in which clusters are assigned to treatment or control, and a subset of individuals are offered treatment within treatment clusters, partially overcome the challenge of allowing for partial interference. But they provide no exogenous variation in treatment saturation to estimate the extent to which program effects are driven by the intensity of treatment.[2] To identify whether and how treatment and spillover effects vary with the in-

[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] Most extant partial population experiments feature cluster-level saturations that are either endogenous

tensity of treatment in a cluster, we use *randomized saturation* (RS) experiments, a two-stage randomization procedure in which first, the share of individuals assigned to treatment within a cluster is randomized and second, individuals within each cluster are randomly assigned to treatment according to the realized cluster-level saturation from the first stage.[3] Hudgens and Halloran (2008), Tchetgen Tchetgen and VanderWeele (2010) and Liu and Hudgens (2014) also study settings with partial interference using a two-stage design.

Our first contribution is to provide a foundation for the regression models commonly used to analyze RS designs by setting up a potential outcomes model with partial interference and mapping it into a regression model. We place two restrictions on the population distribution of potential outcomes – the population average potential outcome only depends on an individual's treatment status and the share of treated individuals in the cluster, and the variance-covariance matrix of the population distribution of potential outcomes is block-diagonal.[4] These assumptions allow us to map the potential outcomes model to a regression model with clustered standard errors, which provides a bridge between the causal inference literature and the methods used to analyze randomized saturation (RS) designs in practice. 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 RS designs identify a set of novel estimands: not only can the researcher consistently identify the usual intention-to-treat effect, but she can also observe spillover effects on treated and untreated units, and understand how the intensity of treatment drives spillover effects for the treated and the untreated alike. These are similar to the estimands that Hudgens and Halloran (2008) show can be consistently estimated in a finite population

(Mexico's conditional cash transfer program, PROGRESA/Oportunidades) or fixed (Duflo and Saez 2003) and typically set at 50%. PROGRESA is perhaps the most-studied example – it features a treatment decision at the cluster (village) level and an objective poverty eligibility threshold at the household level, so both eligible and ineligible individuals in treatment villages can be compared to their counterparts in the pure control group. PROGRESA has been used to examine spillover effects in several contexts (Alix-Garcia, McIntosh, Sims, and Welch 2013; Angelucci and De Giorgi 2009; Bobonis and Finan 2009).

[3]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).

[4]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.

model. The experimental estimate of the average effect on all individuals in treated clusters, which we refer to as the Total Causal Effect, provides the policy maker with a very simple tool to understand how altering the intensity of implementation will drive outcomes for a representative individual.

Our second main contribution is to illustrate the power tradeoffs that exist in designing RS experiments – that is, 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 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, namely, decreased statistical power to measure intention-to-treat effects pooled across all saturations. 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 framework, we provide the closest possible analog to the familiar power calculations in cluster randomized trials. This makes the design tradeoffs present in RS experiments as transparent as possible. 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 an application that uses numerical simulations to illustrate the theoretical tools we develop using hypothetical and published study designs. First, we explicitly define and estimate optimal designs for objective functions that include different individual saturation, slope and pooled estimands. We use the SE calculations to demonstrate the power trade-offs that arise based on which estimands the researcher would like to identify and estimate. We also calculate standard errors for randomized saturation designs in published papers and show how these designs affect the power trade-off between different estimands. These power calculations and numerical optimizations are conducted using software we developed specifically for designing RS experiments, which is available for researchers 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, defines a 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 an application illustrating the optimal design results. 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 a population under different treatment allocations. The target population $\mathcal{I}$ is partitioned into equal-sized, non-overlapping groups, or clusters, of size n.[5] 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 $\mathcal{Y}$ 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).[6]

A random sample is taken 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 the estimands in different designs. In order to characterize these standard errors, we make two assumptions on the distribution of potential outcomes. First, we assume that the *population average potential outcome* $E[Y_{ic}(\mathbf{t})]$ at potential treatment vector $\mathbf{t} \in \{0,1\}^n$ depends only on individual treatment status t_i and cluster treatment saturation $p(\mathbf{t}) \equiv \frac{1}{n}\sum_{i=1}^{n} t_i$, where the expectation is with respect to the population distribution of potential outcomes.

Assumption 1. *There exists a function* $\overline{Y} : \{0,1\}\times[0,1] \to \text{co}(\mathcal{Y})$*, where* $\text{co}(\mathcal{Y})$ *is the convex*

[5] 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 – for example, rural and urban clusters.

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

hull of the set of potential outcomes, such that the population average potential outcome for an individual with treatment status t *is* $E[Y_{ic}(\mathbf{t})] = \overline{Y}(t,p)$ *at all treatment vectors* $\mathbf{t} \in \{0,1\}^n$ *with treatment saturation* $p(\mathbf{t}) = p$.

Assumption 1 says that in expectation, the impact an individual receiving treatment has on the outcomes of others in the same cluster is independent of the treated individual's identity. This allows for a characterization of the variance of estimands without possessing information about the underlying network structure within a cluster.[7] Assumption 1 is weaker than the stratified interference assumption proposed by Hudgens and Halloran (2008), which assumes that the realized potential outcomes of an individual is independent of the identity of the other individuals assigned to treatment.

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 a type of 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. In order to also allow for (ii), we assume a variance-covariance structure for the distribution of potential outcomes that allows potential outcomes to be correlated across individuals within the same cluster. We assume there is no correlation across clusters.

Assumption 2. *Given* $\sigma^2 > 0$ *and* $\tau^2 \geq 0$, *the variance-covariance structure for the population distribution of potential outcomes is:*

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* $c \neq d$

for all $\mathbf{t}, \mathbf{t}' \in \{0,1\}^n$.

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,

[7] 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 within clusters, as each unit has only two potential outcomes.

the variance and covariance of the distribution of potential outcomes do not depend on the treatment status of an individual or the share treated in a cluster.[8]

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.[9] 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. We will often use $\rho \equiv \tau^2/(\tau^2 + \sigma^2)$ to denote the intra-cluster correlation (ICC).

2.2 A Randomized Saturation Design

Suppose a researcher draws a sample of C clusters of size n.[10] 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 $f : \Pi \to [0, 1]$, 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. Each 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

[8]The analysis could be extended to allow for heteroskedasticity. In this case, the standard errors would be less tractable to characterize analytically, and hence, optimal design results would also be less tractable. We view the homoskedastic case as a natural benchmark to establish general insights about how RS designs impact power.

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

[10]The RS design and studies 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.

treatment (i.e. $\sum_{i=1}^{n} T_{ic} = nP_c$ for all c). Let T_c denote the realized treatment vector. An RS design is completely characterized by the pair $\{\Pi, f\}$. The RS design nests several common experimental designs, including the clustered, blocked, and partial population designs.[11]

We refer to individuals assigned to treatment as *treated* individuals, individuals in clusters assigned saturation zero as *pure controls* and individuals who are not assigned to treatment but are in clusters with treated individuals as *within-cluster controls*. Let $S_{ic} = \mathbb{1}\{T_{ic} = 0, P_c > 0\}$ be the random variable that denotes whether individual ic is a within-cluster control and $C_{ic} = \mathbb{1}\{T_{ic} = 0, P_c = 0\}$ be the random variable that denotes whether individual ic is a pure control. An RS design has share of treated individuals $\mu \equiv \sum_{p\in\Pi} pf(p)$, share of within-cluster control individuals $\mu_S \equiv 1 - \mu - \psi$, and share of control individuals $\psi \equiv f(0)$. A RS design has a pure control if $\psi > 0$.

In order to identify treatment and spillover effects, we must place a restriction on the support of the RS design. We say a 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 blocked and clustered designs are trivial, and it is not possible to identify any spillover effects in these designs, while the partial population design is non-trivial and it is possible to identify spillover effects on the untreated.

An RS design introduces correlation between the treatment statuses of two individuals in the same cluster,

$$\text{Cor}(T_{ic}, T_{jc}) = \left(\frac{1}{n-1}\right)\left(\frac{n\eta^2}{\mu(1-\mu)} - 1\right),$$

where $\eta^2 \equiv \sum_{p\in\Pi} p^2 f(p) - \mu^2$ denotes the variance of the cluster-level treatment saturation. This variance in treatment saturation will play a key role in determining the power of an RS design when there is correlation between the potential outcomes of individuals in the same cluster, $\rho > 0$. At one extreme, a clustered design has perfect correlation between the treatment statuses of individuals in the same cluster, $r = 1$, while at the other extreme, a

[11] Fixing the share of treated individuals at 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 $f(P) = \mu/P$.

blocked design has slightly negative correlation, $r = -1/(n-1)$.[12] These two designs bracket the continuum of RS designs, so it is natural that RS designs have an intermediate level of correlation.

Discussion. We implicitly assume that all individuals who are part of the spillover network in a cluster are included in the sample. If this is not the case and 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 sampled individuals in the cluster).[13] If the sampling rate and share of the cluster eligible for treatment are constant 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 assume 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 saturation due to compliance effects or the direct impact of an individual's treatment on others. Exploring extensions considering compliance as a function

[12] There are two sources of correlation in treatment status between individuals in the same cluster: (i) the correlation introduced by the variation in treatment saturation across clusters; (ii) when the share of treated individuals in a cluster is fixed, then when an individual is selected for treatment, this reduces the probability that another individual is selected for treatment i.e. conditional on realized treatment share P_c, $\text{Cor}(T_{ic}, T_{jc}|P_c) = -1/(n-1)$. In a blocked design, the negative correlation stems from the fact that (ii) is the only source of correlation. If instead each individual in a cluster was independently treated with probability P_c (sampling with replacement), then (ii) would not exist.

[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 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).

of the assigned treatment saturation – and 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 the *population average potential outcome* at individual treatment assignment $t \in \{0,1\}$ and saturation $p \in [0,1]$ is $\overline{Y}(t,p)$.

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 $\underline{p} \equiv 1/n$ corresponds 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, $SNT(p) \equiv \overline{Y}(0,p) - \overline{Y}(0,0)$. Given these definitions, there are *spillover 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 rate of change in spillovers. The *Slope of Spillovers on the Treated* measures the rate of change of the spillover effect on treated individuals between saturations p_j and p_k, $DT(p_j,p_k) \equiv (ST(p_k) - ST(p_j))/(p_k - p_j)$. If spillover effects are affine, then this is a measure of $dST(p)/dp$; otherwise, it is a first order approximation of the slope. The analogue slope effect for individuals not offered treatment is denoted $DNT(p_j,p_k)$.

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 value within a cluster to treated individuals, and the true effectiveness of the program is muted compared to the direct treatment effect captured in the ITT.[14] If the TCE is negative, the program causes an aggregate reduction in the average potential outcome, even though treatment effects may be positive. In the presence of spillovers, it is imperative to use the TCE, rather than the ITT, to inform policy, as 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 treatment is *complementary* with spillover effects.[15]

Hudgens and Halloran (2008) also study causal inference in the presence of partial interference, and define a set of estimands for a finite population. The ST and SNT defined above are the infinite population analogues of the indirect causal effects defined in their paper, the ITT is the analogue of the total causal effect, the TCE is the analogue of their overall causal effect and the VT is the analogue of their direct causal effect.

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 a cluster. The TUT measures the efficacy of the vaccination in isolation. The vaccination almost fully protects vaccinated individuals independent of the treatment saturation, which means 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, which means the $SNT(p)$ varies from zero to one. Thus, the value of receiving the

[14] Of course, to say anything about the welfare implications of diversionary effects requires a welfare criterion specifying the social value of different distributions of the outcome variable within a cluster.

[15] If a RS design does not include a pure control, one could define analogous estimands for the ITT, SNT, TCE and VT relative to the lowest saturation in the study. For example, if clusters have a base saturation of share p_0 individuals receiving a treatment before an intervention, a researcher could use estimands that are defined relative to p_0.

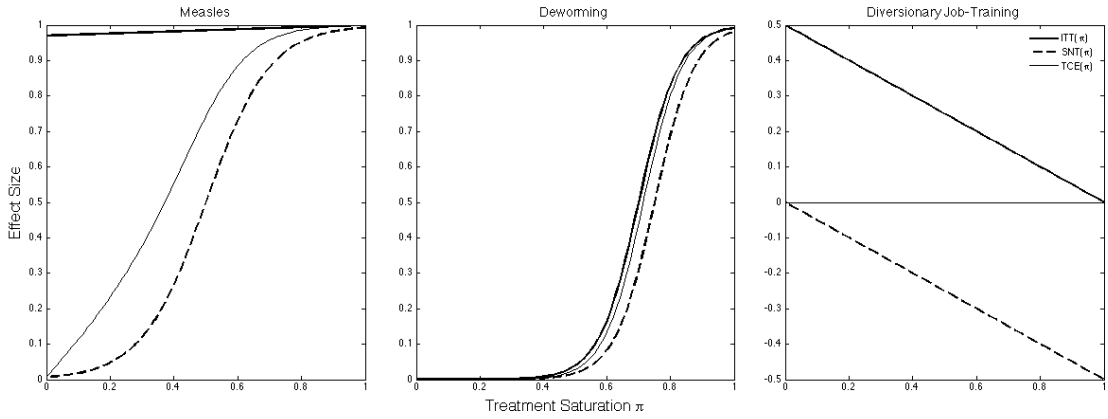

FIGURE 1. Examples

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 from treatment create a free-rider problem that may diminish the salience of vaccinations in populations that have 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 especially 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. Using within-cluster controls as counterfactuals will yield mistaken conclusions 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 developed in 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 to detect different sets of effects. We begin with the individual saturation and slope estimands, and follow with complementary results for a model that pools multiple saturations. The section concludes with an illustration of the power trade-off between measuring slope and pooled effects.

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

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

where $Y_{ic}^{obs} \equiv Y_{ic}(T_c)$ denotes the observed outcome for individual ic. To map the potential outcomes framework into this 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 $\beta_{2p} \equiv \overline{Y}(0,p) - \overline{Y}(0,0)$.[16] Define the error as

$$\begin{aligned} \varepsilon_{ic} \equiv \sum_{\mathbf{t} \in \{0,1\}^n \setminus 0^n} & \mathbb{1}_{T_c = \mathbf{t}} \left(T_{ic}\{Y_{ic}(\mathbf{t}) - \overline{Y}(1, p(\mathbf{t}))\} + S_{ic}\{Y_{ic}(\mathbf{t}) - \overline{Y}(0, p(\mathbf{t}))\} \right) \\ & + C_{ic}\{Y_{ic}(0^n) - \overline{Y}(0,0)\}, \end{aligned} \tag{2}$$

where $p(\mathbf{t})$ is the share of treated individuals in treatment vector $\mathbf{t}$. Athey and Imbens (2017) build a similar connection for a potential outcomes model with no interference and no intra-cluster correlation. The following lemma characterizes the distribution of the error

[16] Note that we are not assuming a constant treatment effect when estimating (1) – $\hat{\boldsymbol{\beta}}$ is the *average* treatment or spillover effect.

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* $E[\varepsilon_{ic}\varepsilon_{jd}|T_c] = 0$ *for* $c \neq d$*.*

Given Lemma 1, the OLS estimate of (1) will yield an unbiased estimate of $\boldsymbol{\beta}$. For any RS design with an interior saturation and a pure control, this estimate identifies $\hat{ITT}(p) = \hat{\beta}_{1p}$, $\hat{SNT}(p) = \hat{\beta}_{2p}$, $\hat{TCE}(p) = p\hat{\beta}_{1p} + (1-p)\hat{\beta}_{2p}$ and $\hat{VT}(p) = \hat{\beta}_{1p} - \hat{\beta}_{2p}$ for each $p \in \Pi \setminus \{0\}$. Hudgens and Halloran (2008) present similar estimators for finite population estimands and show these estimators are unbiased (Theorems 1 and 2).[17] Tests for the presence of treatment and spillover effects at saturation p are $\hat{\beta}_{1p} \neq 0$ and $\hat{\beta}_{2p} \neq 0$. A one-tailed test of the sign of $\hat{\beta}_{2p}$ determines whether treatment creates a negative or positive externality on untreated individuals, $\hat{\beta}_{1p} \neq \hat{\beta}_{2p}$ determines whether the value to treatment is non-zero and $\{\hat{\beta}_{1p} \geq 0, \hat{\beta}_{2p} \leq 0\}$ tests for diversionary effects at saturation p.

We can also use (1) to estimate the slope effects. Given saturations p_j and p_k, the slope effect on individuals offered treatment is $\delta_{jk}^T \equiv \left(\beta_{1p_k} - \beta_{1p_j}\right) / \left(p_k - p_j\right)$, with an analogous expression for the slope effect on within-cluster controls, δ_{jk}^S. A pure control is not required – any RS design with two interior saturations identifies the slope effect for both treatment and within-cluster control individuals. To estimate the slope effect in a design with no pure control, replace the control group with the within-cluster 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.[18]

Standard Errors. Our first result characterizes the standard errors (SEs) for the estimates of the individual saturation effects and slope effects from (1). This illustrates how the standard errors depend on the RS design.[19]

[17] In Hudgens and Halloran (2008), the sample is equal to the population and uncertainty stems from unobserved potential outcomes. Our model is defined for an infinite population, and uncertainty stems from both unobserved potential outcomes and sampling uncertainty. Minor technical modifications to their proofs establish the analogous results in our setting.

[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] 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).

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

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

for each $p \in \Pi$*. For any RS design* (Π, f) *with* $\kappa \geq 2$ *interior saturations, the SE for the slope effect on treated individuals between saturations* $p_j > 0$ *and* $p_k > 0$ *is*

$$\text{SE}_{DT}(p_j, p_k) = \frac{1}{p_k - p_j} \sqrt{\frac{\tau^2 + \sigma^2}{nC} * \left(n\rho \left(\frac{1}{f(p_j)} + \frac{1}{f(p_k)} \right) + (1-\rho) \left(\frac{1}{p_j f\left(p_j\right)} + \frac{1}{p_k f\left(p_k\right)} \right) \right)}$$

Substituting $(1-p)f(p)$ *for* $pf(p)$ *yields analogous expressions for untreated individuals, denoted* $\text{SE}_{SNT}(p)$ *(for* $p \in (0,1)$*) and* $\text{SE}_{DNT}(p_j, p_k)$*.*

Theorem 1 illustrates the relationship between the correlation structure of outcomes and the precision of estimates in an RS design. At one extreme, if there is no correlation ($\rho = 0$), the variation in $\hat{ITT}(p)$ depends on the share of *treated* individuals at saturation p, $pf(p)$, and the share of control individuals, ψ. There is no correlation between potential outcomes within a cluster, so observing $Y_{ic}(1,p)$ for treated individual i provides no information about the potential outcome $Y_{jc}(1,p)$ for untreated individual j and the share of within-cluster control individuals at saturation p is irrelevant for SE_{ITT}. At the other extreme, if there is perfect correlation ($\rho = 1$), the variation in $\hat{ITT}(p)$ depends on the *total* share of individuals at saturation p, $f(p)$, and the share of control individuals. Within a cluster, there is perfect correlation between individuals' potential outcomes and observing $Y_{ic}(1,p)$ for treated individual i provides perfect information about the potential outcome $Y_{jc}(1,p)$ for untreated individual j. At intermediate levels of correlation, SE_{ITT} depends on a weighted average of the share of treated individuals and the total share of individuals at saturation p.

Next consider the standard error of the slope effect. As the distance between two saturations increases, $1/(p_k - p_j)$ decreases, making it possible to detect smaller slope effects. At the same time, increasing the spread of saturations makes the number of treatment (within-cluster control) individuals very small at low (high) saturations. The former effect dominates at saturations close to 1/2, and spreading the saturations apart decreases the SE, while the latter effect dominates at saturations close to zero or one, and spreading the saturations apart increases the SE. When ρ is large, the share of clusters assigned to each saturation,

$f(p_j)$ and $f(p_k)$, play a larger role in determining the SE; a more equal distribution leads to a smaller SE. When ρ is small, the share of treatment (within-cluster control) individuals assigned to each saturation, $p_j f(p_j)$ and $p_k f(p_k)$ ($(1-p_j)f(p_j)$ and $(1-p_k)f(p_k)$), are more important; a more equal share leads to a smaller 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 β_{1p} that exceed MDE $= (t_{1-\gamma} + t_\alpha) \times \text{SE}(\hat{\beta}_{1p})$. The expressions for the MDEs of the spillover effect on untreated individuals and the slope effects are analogous.

In general, the OLS estimator is inefficient when errors are correlated. The variance of the OLS estimate characterized in Theorem 1 will be conservative if GLS or another more efficient estimator is used to analyze the resulting data.

Optimal Design: Individual Saturation Effects. Given a set of saturations Π, the design choice for measuring individual saturation effects involves choosing the share of clusters to allocate to each saturation. If the researcher places equal weight on the treatment and spillover effect at each saturation, she chooses f to minimize the sum of standard errors,[20]

$$\min_f \sum_{p \in \Pi \setminus \{0\}} \left(\text{SE}_{ITT}(p) + \text{SE}_{SNT}(p) \right). \tag{3}$$

First consider the choice of how many clusters to allocate to each positive saturation. By design, extreme saturations have more uneven shares of treatment and within-cluster control individuals, relative to saturations closer to 1/2. A researcher who places equal weight on measuring effects at each saturation in Π will want to allocate a larger share of clusters to these more extreme saturations. This stems directly from the concavity of the SE. As ρ increases, this asymmetry in the optimal f is muted since within-cluster control individuals provide information about treated individuals, and vice versa, and the uneven shares of treatment and within-cluster control individuals has a smaller impact on the precision of estimates

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

Next, consider the optimal control group size. The marginal impact of adding another cluster to the control reduces all terms in (3), while the marginal impact of adding another cluster to an interior saturation only reduces the SEs at that saturation. Therefore, the optimal share of individuals allocated to the control group is always larger than the smallest share of individuals at any treatment saturation. The optimal control size increases with ρ – when the outcomes of treated and within-cluster control individuals are more correlated, the optimal f allocates a smaller share of clusters to each positive treatment saturation. Corollary 1 formalizes these insights.

Corollary 1. *Fix a set of saturations* Π*. Let* f^* *minimize* (3)*, with* $\psi^* \equiv f^*(0)$*.*

1. *A larger share of clusters are allocated to more extreme saturations: given* $p_1, p_2 \in \Pi \setminus \{0\}$ *such that* $|0.5 - p_1| > |0.5 - p_2|$*,* $f^*(p_1) > f^*(p_2)$*. The difference in the share of clusters at each saturation,* $f^*(p_1) - f^*(p_2)$*, is decreasing in* ρ*.*

2. *The share of individuals assigned to pure control is larger than the share of treated or within-cluster control individuals at any positive treatment saturation,* $\psi^* > \min\{p, 1-p\}f^*(p)$ *for all interior saturations* $p \in \Pi$*.*

3. *For each interior saturation* $p \in \Pi$*,* $f^*(p)$ *is decreasing in* ρ*, and* ψ^* *is increasing in* ρ*.*

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: selecting the set of saturations Π and choosing the share of clusters to allocate to each saturation. Suppose a researcher is interested in measuring the slope effect, she places equal weight on estimating the slope effect for treated and untreated individuals, and she believes that the slope effects are monotonic. Then she chooses an RS design to solve

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

The optimal saturations are symmetric about one half, an equal share of clusters are allocated to each saturation, and the optimal distance between saturations is increasing in ρ.

Corollary 2 (Optimal Saturations)**.** *The RS design that minimizes* (4) *equally divides clusters between two saturations symmetric about 1/2,* $p_1^* = (1-\Delta)/2$ *and* $p_2^* = (1+\Delta)/2$*, where*

the optimal distance between saturations $\Delta \in \left(\sqrt{2}/2, 1\right)$ *satisfies*

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

If $\rho = 0$*, then* $\Delta = \sqrt{2}/2$ *for all* n*,* Δ *is increasing in* ρ *and* n*, and if* $\rho = 1$*, then* $\Delta = 1 - 2/n$*.*

Note that this optimal design equalizes the standard errors, $\text{SE}_{DT}(p_j^*, p_k^*) = \text{SE}_{DNT}(p_j^*, p_k^*)$.

More generally, if a researcher is interested in identifying individual saturation or slope effects at or between more than two saturations, Theorem 1 can be used to answer questions like what is the optimal spacing of saturations and what share of clusters should be assigned to each saturation. For example, suppose a researcher would like to test for linearity by including three saturations. To minimize the analogue of (4), one saturation should be 1/2 and the other two should be spaced symmetrically about one half. A larger share of clusters should be allocated to the extreme saturations relative to saturation 1/2.

3.2 Pooled Effects

Suppose the researcher would like to combine observations from clusters with different saturations to measure an average of different estimands across all saturations in the RS design. What we refer to as 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 ITT 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 and treatment status. Let $\overline{Y}(1) \equiv \frac{1}{\mu} \sum_{p \in \Pi \setminus \{0\}} p f(p) \overline{Y}(1, p)$ and $\overline{Y}(0) \equiv \frac{1}{\mu_S} \sum_{p \in \Pi \setminus \{0\}} (1-p) f(p) \overline{Y}(0, p)$ be the population average potential outcome averaged across all non-zero saturations in the RS design, when $t = 1$ 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 error as

$$\varepsilon_{ic} \equiv T_{ic}\{Y_{ic}(1, T_{-i,c}) - \overline{Y}(1)\} + S_{ic}\{Y_{ic}(0, T_{-i,c}) - \overline{Y}(0)\} + C_{ic}\{Y_{ic}(0^n) - \overline{Y}(0,0)\}, \quad (6)$$

where $T_{-i,c}$ is the treatment vector for individuals $j \neq i$ in cluster c. The following Lemma characterizes the distribution of the error in terms of the distribution of potential outcomes and the structure of the RS design.

Lemma 2. *Under Assumptions 1 and 2, the error defined in (6) is strictly exogenous, $E[\varepsilon_{ic}|T_c] = 0$, uncorrelated across clusters, and has within-cluster variance-covariance matrix specified as follows:*

1. *Treated clusters: the variance for treated individuals is $\text{Var}(\varepsilon_{ic}^2) = \sigma^2 + \tau^2 + \phi_T$ and for untreated individuals is $\text{Var}(\varepsilon_{ic}^2) = \sigma^2 + \tau^2 + \phi_S$. The covariance is $\text{Cov}(\varepsilon_{ic}, \varepsilon_{jc}) = \tau^2 + \phi_{TT}$ between treated individuals, $\text{Cov}(\varepsilon_{ic}, \varepsilon_{jc}) = \tau^2 + \phi_{SS}$ between untreated individuals, and $\text{Cov}(\varepsilon_{ic}, \varepsilon_{jc}) = \tau^2 + \phi_{TS}$ between a treated and untreated individual.*

2. *Control clusters: the variance is $\text{Var}(\varepsilon_{ic}) = \sigma^2$ and the covariance is $\text{Cov}(\varepsilon_{ic}, \varepsilon_{jc}) = \tau^2$.*

where $\phi_T \equiv \frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}} pf(p)\overline{Y}(1,p)^2 - \overline{Y}(1)^2$ captures the variation in $\overline{Y}(1,\cdot)$ across saturations in the RS design, with analogous definitions for ϕ_S, ϕ_{TT}, ϕ_{TS} and ϕ_{SS}.

By Lemma 2, the OLS estimate of (5) will yield an unbiased estimate of $\boldsymbol{\beta}$ for any RS design with an interior saturation and a pure control.

The interpretation of $\boldsymbol{\beta}$ is somewhat subtle and depends on the RS design. 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 is an estimate of 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 is an estimate of the pooled SNT with weight $w(p) = (1-p)f(p)$. Due to these different weights, a 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.[21]

[21] 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

Pooling observations across multiple saturations introduces the possibility of heteroskedasticity. Lemma 2 characterizes the precise form of this heteroskedasticity, which depends on the expected potential outcome at each saturation in the RS design. When $ITT(p)$ and $SNT(p)$ are relatively flat, the heteroskedasticity will be small, whereas when these estimands vary with the intensity of treatment, the heteroskedasticity will be more significant. The error is homoskedastic precisely when the expected potential outcomes do not vary with the treatment saturations in the RS design.

Definition 1. *Treatment and spillover effects are* constant *on* Π *if for all* $p_j, p_k \in \Pi$, $\overline{Y}(1, p_j) = \overline{Y}(1, p_k)$ *and* $\overline{Y}(0, p_j) = \overline{Y}(0, p_k)$.

Corollary 3. *Given saturations* Π*, the error has a block-diagonal variance-covariance matrix if and only if treatment and spillover effects are constant on* Π.

Generally, cluster robust standard errors should be used in two-level experiments due to the design effect. This corollary provides an additional argument for doing so when estimating (5) due to the variation in treatment and spillover effects at different saturations.

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 are slope 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 that reduces the precision of estimates. 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.

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 assigned to treatment 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 $p = 2/3$ clusters.

Let η_T^2 be the component of the variance in treatment saturation across clusters that arises from multiple non-zero saturations,

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

where $f(p)/(1-\psi)$ is the distribution of treatment saturation conditional on $p > 0$, with support $\Pi \setminus \{0\}$, and η^2 is the total variance in treatment saturation. Trivially, $\eta_T^2 = 0$ when there is a single non-zero saturation.

In order to isolate the impact that variance in treatment saturation has on the SE for the pooled ITT and SNT, we focus on the case where spillover and treatment effects are constant across all saturations in the RS design.

Theorem 2 (Pooled SE)**.** *Let* (Π, f) *be an RS design with an interior saturation and a pure control. Assume Assumptions 1, 2 and treatment and spillover effects are constant on* Π*. The SE of* $\overline{ITT}$ *is:*

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

Substituting μ_S *for* μ *yields an analogous expression for the SE of* $\overline{SNT}$*, denoted* $\overline{\text{SE}}_{SNT}$*.*

The SE for the pooled estimands depends on the size of the treatment and control groups and the within-cluster variation in treatment status. Crucially, when outcomes within clusters are correlated ($\rho > 0$), the SE is strictly increasing in the variation in treatment saturation η_T^2 and introducing multiple treatment saturations reduces precision. Standard errors are minimized in a partial population design in which there is a single treatment saturation and a pure control. This design has no variation in treatment saturation, $\eta_T^2 = 0$.

Corollary 4 (Optimality of Partial Population Design)**.** *Suppose* $\rho > 0$*. For any* (μ, ψ)*, the partial population design with treatment saturation* $p = \mu/(1-\psi)$ *and a pure control simultaneously minimizes* $\overline{\text{SE}}_{ITT}$ *and* $\overline{\text{SE}}_{SNT}$*.*

Moving away from the partial population design to a design with variation in the treatment saturation, the power loss is more severe for settings with higher intra-cluster correlation. The variance of $\hat{\boldsymbol{\beta}}$ increases linearly with respect to η_T^2 and the rate at which this variance increases is proportional to ρ. Therefore, if the researcher a priori believes that slope effects

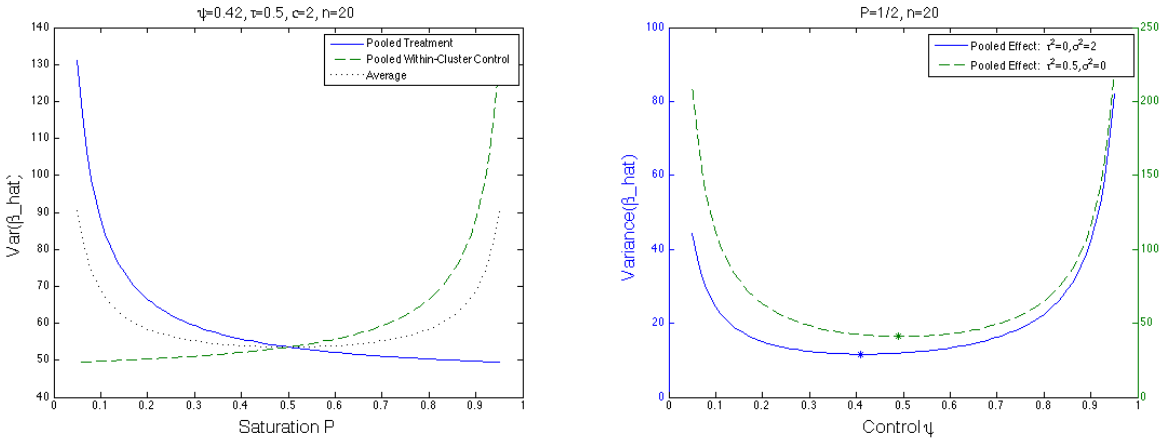

FIGURE 2. Partial Population Design

are small and intra-cluster correlation is high, she is best off selecting a partial population design. The next subsection explores which partial population design to choose.

Optimal Partial Population Design. Consider the optimal treatment saturation p and control size ψ for a partial population design. The SE of the ITT decreases with p, while the SE of the SNT increases with p. The relative importance of detecting these two effects, as well as their expected magnitudes, will determine the optimal choice of p. If a researcher places equal weight on each effect,

$$\min_{(p,\psi)} \overline{\text{SE}}_{ITT}(p,\psi) + \overline{\text{SE}}_{SNT}(p,\psi), \tag{8}$$

then the optimal saturation creates equally sized treatment and within-cluster control groups by choosing $p^* = 0.5$. The left panel of Figure 2 illustrates the SEs in a partial population design, as a function of p. Note $\overline{\text{SE}}_{ITT}(0.5,\psi) = \overline{\text{SE}}_{SNT}(0.5,\psi)$.

The optimal share of control clusters depends on ρ and n. As ρ increases, the optimal share of control clusters also increases – within-cluster controls and treated individuals provide more information about each other, and the total number of *clusters* in each treatment group becomes more important for statistical power than the total number of *individuals* in each treatment group. In a partial population design with saturation 0.5, it is always optimal to allocate more than a third of clusters to the pure control as the control serves as the counterfactual for both treatment and spillover groups; designating about 41% of clusters as pure controls yields the smallest SE when $\rho = 0$, while designating 50% is preferable when

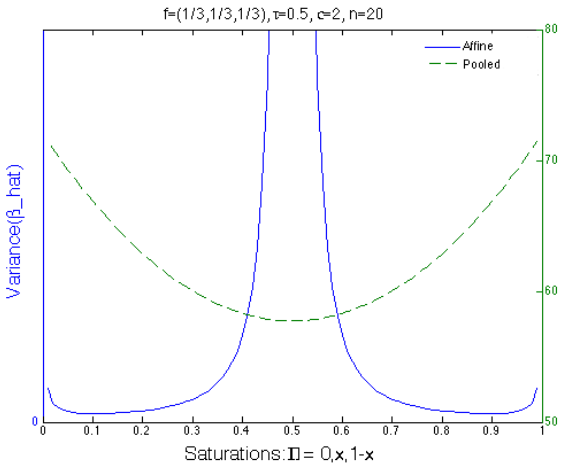

FIGURE 3. Trade-off between SEs of Pooled and Slope Estimands

$\rho = 1$. Corollary 5 summarizes these results.

Corollary 5. *The partial population design that minimizes* (8) *has saturation* $p^* = 0.5$ *and allocates share*

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

of clusters to pure control for $\rho \in [0,1)$ *and* $\psi^* = 1/2$ *for* $\rho = 1$. *The optimal share of control clusters* ψ^* *is increasing in* ρ *and* n.

The right panel of Figure 2 illustrates how the SEs in a partial population design with saturation 0.5 (note they are equal) depend on the control group size when $\rho = 0$ and $\rho = 1$. The minimum in each case is marked with an asterisk. Corollary 5 is similar in spirit to Hirano and Hahn (2010). They show that a partial population design identifies the ITT and SNT in a linear-in-means model when $\rho = 0$, and establish that the optimal treatment saturation is $p^* = 0.5$ and the optimal control group size is $\psi^* = \sqrt{2} - 1$.

3.3 The Design Trade-off.

Taken together, the results in Sections 3 provide important insights on experimental design. Clustering of outcomes can be due to either correlated effects or interference between units. Theorems 1 and 2 show that the source of clustering plays an important role in determining the power of different designs. The optimal design depends crucially on the degree of intra-cluster correlation and the degree to which individual effects vary with intensity of treatment – precisely the two underlying factors that drive clustering of outcomes.

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 treatment saturations will reduce statistical power without yielding novel insights and the researcher is better off running a partial population design. However, partial population designs have the drawback that they only measure effects at a single saturation. When a researcher seeks to identify or rule out slope effects, she will need to introduce variation in the treatment saturation. A graphical representation of the tradeoff in precision between measuring pooled and slope effects is presented in Figure 3.

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. Thus, the optimal RS design for a slope analysis stands in sharp contrast to that for an analysis at individual saturations or a pooled analysis. If the researcher seeks to identify both slope and individual effects, the optimal design will depend on the relative importance that the researcher places on each effect.

4 Application

This section illustrates our results by characterizing the optimal design for different hypothetical objective functions and calculating 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 as a tool for researchers.[22]

First, suppose a researcher uses a clustered design to identify the average treatment effect. She selects $C = 100$ clusters, each of which contain $n = 10$ individuals, and is interested in the precision of her estimate of the ITT. She implements the optimal clustered design, which assigns 50% of the clusters to the control group and 50% to the treatment group and identifies $ITT(1)$. The SE of $ITT(1)$ depends on the intra-cluster correlation, ρ, and is measured in standard deviations. 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.

[22]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 to Detect Pooled Effects

	BENCHMARK: CLUSTERED DESIGN			PARTIAL POPULATION DESIGN				
Objective Function	**min SE_ITT**			**min SE_ITT + SE_SNT**			**min SE_ITT + 2*SE_SNT**	**min SE_SNT s.t. SE_ITT≤0.09**
ICC: ρ	0	0.1	1	0	0.1	1	0.1	0.1
Optimal saturation 1: pure control	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Optimal saturation 2: p2	1.00	1.00	1.00	0.50	0.50	0.50	0.41	0.78
Optimal share in pure control: ψ	0.50	0.50	0.50	0.41	0.45	0.50	0.45	0.47
Optimal share in p2	0.50	0.50	0.50	0.59	0.55	0.50	0.55	0.53
SE of pooled ITT	0.063	0.087	0.200	0.076	0.097	0.200	0.100	0.090
SE of pooled SNT	.	.	.	0.076	0.097	0.200	0.094	0.117
Other parameters: C=100, n=10								

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 4, the optimal design is a partial population experiment (PPE), $\Pi = \{0, p\}$ and $f = \{\psi, 1-\psi\}$. This design identifies $ITT(p)$ and $SNT(p)$. From Corollary 5, we know that when the researcher places equal weight on minimizing the SE_{ITT} and SE_{SNT}, the optimal treatment saturation is $p^* = 0.5$, meaning that half of the individuals in each treatment cluster are assigned to treatment, and the optimal share of control clusters ranges from $\psi^* = 41\%$ to 50% as ρ increases from 0 to 1 (Table 1, Columns 4-6). The SE for the ITT and SNT are equal, $\text{SE}_{ITT}(0.5) = \text{SE}_{SNT}(0.5)$, and range from 0.076 to 0.200 as ρ increases from 0 to 1. Hence, when the researcher wants to detect spillovers on untreated individuals, the SE for the ITT rises. The source of this 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 within-cluster control individuals provide more information about treated individuals for high ρ.

Now suppose the researcher wants to detect a pooled SNT that is smaller or larger than the pooled ITT. A partial population experiment remains optimal, but now the optimal treatment saturation and control group size minimizes

$$\min_{p,\psi} \theta \text{SE}_{ITT}(p) + (1-\theta)\text{SE}_{SNT}(p),$$

where $\theta \in [0, 1]$ is the relative weight that the researcher places on detecting treatment versus spillover effects. When $\rho = 0.1$ and $\theta = 1/3$, the optimal treatment saturation is $p^* = 0.41$,

TABLE 2. Optimal Design to Detect Slope Effects and SE in Existing Studies

	OPTIMAL RS DESIGNS				EXISTING STUDIES			
Objective Function	**min SE_DT + SE_DNT**			**min SE_ITT + SE_SNT + SE_DT + SE_DNT**	**Banerjee et al. & Crepon et al.**	**Sinclair et al.**	**Baird et al.**	**Baird et al. sat; shares solve min SE_SNT s.t. SE_ITT≤.095**
ICC: ρ	0	0.1	1	0.1	0.1	0.1	0.1	0.1
Saturation 1: p1	0.15	0.13	1/n	0.00	0.00	0.00	0.00	0.00
Saturation 2: p2	0.85	0.87	(n-1)/n	0.21	0.25	0.10	0.33	0.33
Saturation 2: p3	.	.	.	0.88	0.50	0.50	0.67	0.67
Saturation 2: p4	.	.	.	.	0.75	1.00	1.00	1.00
Saturation 2: p5	.	.	.	.	1.00	.	.	.
Share in p1	0.50	0.50	0.50	0.28	0.20	0.25	0.55	0.45
Share in p2	0.50	0.50	0.50	0.33	0.20	0.25	0.15	0.21
Share in p3	.	.	.	0.39	0.20	0.25	0.15	0.21
Share in p4	.	.	.	.	0.20	0.25	0.15	0.13
Share in p5	.	.	.	.	0.20	.	.	.
SE of pooled ITT	.	.	.	0.104	0.113	0.109	0.095	0.095
SE of pooled SNT	.	.	.	0.109	0.120	0.111	0.115	0.106
SE of Treated slope	0.179	0.191	0.204	0.217	0.240	0.242	0.289	0.269
SE of Untreated slope	0.179	0.191	0.204	0.192	0.240	0.274	0.251	0.221
Other parameters: C=100, n=10								

meaning 41% of individuals in a treatment cluster are assigned to treatment (Table 1, Column 7). The optimal share of clusters allocated to the control group is $\psi^* = .45$, as was the case for $\theta = 1/2$ and $\rho = 0.1$. This produces $\text{SE}_{ITT}(.41) = 0.100$ and $\text{SE}_{SNT}(.41) = 0.094$.[23] Alternatively, if the researcher wants to minimize the SE of the SNT, while maintaining a SE for the pooled ITT of 0.09 or lower (approximately the SE in the clustered design), then she would use treatment saturation $p^* = 0.78$ and share of control clusters $\psi^* = .47$. This would yield a higher standard error for the spillover effect on the non-treated, 0.117 (Table 1, Column 8).

Next, suppose the researcher wishes to estimate the slope effect for treated and untreated individuals, and does not care about identifying the individual or pooled ITT and SNT. Then the optimal design will have two interior saturations and no pure control group. She chooses a design to solve

$$\min_{p_1,p_2,f} \text{SE}_{DT}(p_1, p_2, f) + \text{SE}_{DNT}(p_1, p_2, f).$$

By Corollary 2, the optimal saturations are symmetric about 0.5 and clusters equally divided between these two saturations. When $\rho = 0$, the optimal saturations are $p_1^* = 0.15$ and $p_2^* = 0.85$. This produces a SE of 0.179 for both the treated and non-treated individuals (Table 2, Column 1). Increasing ρ moves the optimal saturations further apart and increases

[23] Moving to a more extreme $\theta = 1/10$ does not alter the share of clusters allocated to pure control substantively ($\psi^* = 47\%$), but significantly reduces the optimal treatment saturation ($p^* = 0.23$).

the SEs for the slope effects (Table 2, Columns 2 - 3). At the extreme, when outcomes within a cluster are perfectly correlated ($\rho = 1$), the optimal saturations are as far apart as possible while still maintaining at least one treated and one within-cluster control individual in each cluster. This corresponds to saturations $p_1^* = 1/n$ and $p_2^* = (n-1)/n$.

However, not many researchers are interested in designing an experiment to maximize the precision of slope estimands, at the expense of not being able to identify standard estimands, such as the ITT. To give a sense of the optimal design when the researcher would like to have a pure control group along with two interior saturations, we consider an objective function that puts equal weights on both the pooled and slope effects,

$$\min_{p_1,p_2,f} \overline{\text{SE}}_{ITT} + \overline{\text{SE}}_{SNT} + \text{SE}_{DT}(p_1, p_2, f) + \text{SE}_{DNT}(0, p_2, f).$$

When $\rho = 0.1$, it is optimal to allocate 33% of the clusters to saturation $p_1^* = 0.21$, 39% to saturation $p_2^* = 0.88$, and the remaining $\psi^* = 28\%$ to the pure control group, i.e. $\Pi^* = \{0, 0.21, 0.88\}$ and $f^* = \{0.28, 0.33, 0.39\}$ (Table 2, Column 4).[24] The calculated SEs of 0.104 and 0.109 for the pooled ITT and SNT, respectively, indicate an 7-13% increase in the SEs compared to the optimal partial population design using the same parameters (Table 1, Column 5).[25] Unlike the precision loss that arises when moving from a clustered to a partial population design, the precision loss in moving from a partial population design to a RS design with two interior saturations arises from the increased variance of treatment saturations, rather than a reduction in sample size. It is precisely this variance in treatment saturation that enables identification of slope effects.

Finally, we calculate the standard errors in 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 ($\rho = 0.1$)

[24]The careful reader might note that the optimal interior saturations are not symmetric about 0.5, as would be the case if we were solely interested in minimizing detectable slope effects. In this example, a pure control group is included to identify the ITT and SNT. Furthermore, at 27%, the size of the optimal control group is smaller than the control group size that minimizes the sum of the individual SEs for the ITT and SNT (Corollary 1).

[25]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* i.e. 0.21 and 0.88. 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, 0.88.

as in our examples, rather than the actual values from each study.[26]

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, slope effects and test for the shapes of $ITT(p)$ and $SNT(p)$. The cell with 100% treatment saturation allows for examination of general equilibrium effects when everyone in the target population is treated, compared with the partial equilibrium effects in lower saturation cells. Our power calculations for this design yield $\overline{\text{SE}}_{ITT} = 0.113$, $\overline{\text{SE}}_{SNT} = 0.120$, and $\text{SE}_{DT} = \text{SE}_{DNT} = 0.240$ (Table 2, Column 5). All of these figures are higher than their counterparts under the optimal design for minimizing the sum of these four variables (Table 2, Column 4), demonstrating the power loss that arises from having a richer, more granular design that can, for example, test for concavity of $ITT(p)$ and $SNT(p)$.

Our next example is the design used by Sinclair et al. (2012). They randomized nine-digit zip codes in a congressional district in Illinois 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.[27] 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 $ST(p)$ for $p = 0.5$ and $p = 1$. Our power calculations for this design yield $\overline{\text{SE}}_{ITT} = .109$, $\overline{\text{SE}}_{SNT} = 0.111$, $\text{SE}_{DT} = 0.242$ and $\text{SE}_{DNT} = 0.274$, respectively (Table 2, Column 6). The pooled SEs are quite similar to their counterparts under the optimal design for minimizing the sum of these four quantities (Table 2, Column 4), but the slope effect SEs are substantially higher, particularly for the non-treated (0.274 vs. 0.192), because the largest saturation

[26]The pooled SEs in columns 4-8 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 treatment slope effects larger than 0.62. Suppose the true slope is 0.5. It will not be possible to reject the null hypothesis that the slope is zero, but there will still be heteroskedasticity and the pooled SE for treated individuals will be strictly larger than 0.104 which is the pooled SE for a ITT effect of zero. To account for this, researchers should build some sample size cushion into their designs.

[27]The saturation of 0.5 is approximate, as one core household plus half of the remaining households were randomly assigned to treatment in clusters assigned to that saturation.

containing within-cluster controls is 0.5.

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 also 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 SEs for the pooled ITT that are smaller than those in 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 SEs for the pooled ITT and SNT, we can ask whether there is a way to allocate clusters to this set of saturations that leads to both lower pooled and slope SEs. Consider the objective function that minimizes the SE of the SNT, subject to the constraint that the SE of the ITT remains below its value in the original study design, $\min_f \overline{\text{SE}}_{SNT}$ subject to $\overline{\text{SE}}_{ITT} \leq 0.095$. The optimal distribution of treatment saturations for this objective allocates a lower share of clusters to the pure control group and to saturation $p = 1$, 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 also decreases the SEs of the slope effects. As we kept Π fixed, 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, which treatment allocation will maximize the total benefit? Small policy trials conducted on a subset of the population can miss important scale or congestion effects that will accompany the full-scale implementa-

tion of a program. 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. Varying the cluster-level saturation can lead to differential impacts on prices, norms, and congestion effects. The RS design presented here provides an experimental framework to inform these policy questions and bolster both external and internal validity.

In this paper, we attempt to formalize the optimal design and analysis of RS designs. Building on the previous multidisciplinary literature, we map the potential outcomes framework to a clustered error regression model, which allows us to gain analytical insights for the optimal design of such experiments and derive ex-ante power calculations. The benefit of randomizing treatment saturations is the ability to generate direct experimental evidence on the nature of spillover and threshold effects both for treated and non-treated individuals. The cost of doing so is the precision of these estimates. Having laid out the assumptions necessary to estimate both the mean and variance of spillover effects, we derive analytical closed-form expressions for the standard errors. The SEs 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 in that randomizing saturations allows the researcher to identify novel estimands but comes at the cost of the precision of the estimates of more basic estimands. This is an inherent feature of RS designs.

References

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

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

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

Aronow, P. M. and C. Samii (forthcoming): "Estimating Average Causal Effects Under General Interference," *Annals of Applied Statistics*.

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

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

Baird, S., C. McIntosh, and B. Özler (2011): "Cash or Condition? Evidence from a Cash Transfer Experiment," *The Quarterly Journal of Economics*, 126, 1709–1753.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The Diffusion of Microfinance," *Science*, 341.

Banerjee, A., R. Chattopadhyay, E. Duflo, D. Keniston, and N. Singh (2012): "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training," Working Paper 17912, NBER.

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

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

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

Bobba, M. and J. Gignoux (2016): "Neighborhood Effects in Integrated Social Policies," *World Bank Economic Review*.

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

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

Chen, J., M. Humphries, and V. Modi (2010): "Technology Diffusion and Social Networks: Evidence from a Field Experiment in Uganda," Working Paper.

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

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

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using Randomization in Development Economics Research: A Toolkit," Tech. rep., C.E.P.R. Discussion Papers, cEPR Discussion Papers.

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

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

Gine, X. and G. Mansuri (forthcoming): "Together We Will : Experimental Evidence on Female Voting Behavior in Pakistan," *AEJ: Microeconomics*.

Graham, B. S., G. W. Imbens, and G. Ridder (2010): "Measuring the effects of segregation in the presence of social spillovers: a nonparametric approach," .

Hahn, J., K. Hirano, and D. Karlan (2011): "Adaptive Experimental Design Using the Propensity Score," *Journal of Business & Economic Statistics*, 29, 96–108.

Hirano, K. and J. Hahn (2010): "Design of randomized experiments to measure social interaction effects," *Economics Letters*, 106, 51–53.

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

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

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

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

Macours, K. and R. Vakis (2008): "Changing Households' Investments and Aspirations through Social Interactions: Evidence from a Randomized Transfer Program in a Low-Income Country," Tech. rep., world Bank Working Paper 5137.

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

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

McIntosh, C., T. Alegria, G. Ordonez, and R. Zenteno (2013): "Infrastructure Impacts and Budgeting Spillovers: The Case of Mexico's Habitat Program," Working Paper.

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

Moffitt, R. A. (2001): "Policy Interventions, Low-Level Equilibria And Social Interactions," in *Social Dynamics*, ed. by S. Durlauf and P. Young, MIT Press, 45–82.

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

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

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

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

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

Toulis, P. and E. Kao (2013): "Estimation of Causal Peer Influence Effects," *Journal of Machine Learning Research*, 28, proceedings of the 30th International Conference on Machine Learning Research.

A Proofs from Section 3

A.1 Proof of Theorems 1 and 2

Preliminary Calculations This section provides background material used in the proofs of Theorems 1 - 4. Consider the OLS estimate of

$$Y_{ic} = X_{ic}'\boldsymbol{\beta} + \varepsilon_{ic}, \tag{9}$$

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 $X_c' = [X_{1c} \ ... \ X_{nc}]$ and $\varepsilon_c' = [\varepsilon_{1c} \ ... \ \varepsilon_{nc}]$, 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, $E[\varepsilon_{ic}\varepsilon_{jd}|X] = 0$ for all $c \neq d$. Then the exact finite sample variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= A^{-1}\left(\sum_{c=1}^{C} X_c' E[\varepsilon_c\varepsilon_c'|X_c]X_c\right)A^{-1} \\
&= A^{-1}\left(\sum_{c=1}^{C} X_c'(\sigma^2\mathbf{I}_n + \tau^2\mathbf{1}_n)X_c\right)A^{-1} \\
&= \sigma^2 A^{-1} + \tau^2 A^{-1}BA^{-1},
\end{aligned} \tag{10}$$

where

$$A \equiv \sum_{c=1}^{C}\sum_{i=1}^{n} X_{ic}X_{ic}' \tag{11}$$

$$B \equiv \left(\sum_{c=1}^{C} X_c'\mathbf{1}X_c\right). \tag{12}$$

Proof of Theorem 1. 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 (9) when

$$X_{ic}' = [\ 1 \quad T_{1ic} \quad S_{1ic} \quad T_{2ic} \quad S_{2ic}\ ].$$

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 Section A.1, $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) =$

$\sigma^2 A^{-1} + \tau^2 A^{-1} B A^{-1}$, 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 $\sum_{i=1}^n T_{kic} = np_k$, $\sum_{i=1}^n S_{kic} = n(1-p_k)$, $\sum_{c=1}^C \sum_{i=1}^n T_{kic} = np_k \times Cf(p_k) = nC\mu_k$, $\sum_{c=1}^C \sum_{i=1}^n S_{kic} = n(1-p_k) \times Cf(p_k) = nCs_k$, $\sum_{c=1}^C (\sum_{i=1}^n T_{kic})^2 = n^2 p_k^2 \times Cf(p_k) = n^2 C\eta_k$, $\sum_{c=1}^C (\sum_{i=1}^n T_{kic} \times \sum_{i=1}^n S_{kic}) = n^2 p_k(1-p_k) \times Cf(p_k) = n^2 C(\mu_k - \eta_k)$, $(\sum_{i=1}^n T_{1ic})(\sum_{i=1}^n S_{2ic}) = 0$, and other analogous calculations. Taking the diagonal entries of $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ yields

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

and

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

for each $p_j \in \Pi$. Taking the square root yields the results for Theorem 1. It is straightforward to extend the result to more than two interior saturations.

To compute the SE_{DT}, note $\mathrm{Var}(\delta_{jk}^T) = \mathrm{Var}(\beta_{1p_k} - \beta_{1p_j}) / (p_k - p_j)^2$, $\mathrm{Cov}(\beta_{1p_k}, \beta_{1p_j}) =$

$(n\tau^2+\sigma^2)/\psi nC$, and

$$\begin{aligned}\mathrm{Var}(\beta_{1p_k}-\beta_{1p_j}) &= \mathrm{Var}(\beta_{1p_j})+\mathrm{Var}(\beta_{1p_k})-2\,\mathrm{Cov}(\beta_{1p_k},\beta_{1p_j})\\ &= \frac{1}{nC}*\left\{n\tau^2\left(\frac{1}{f(p_j)}+\frac{1}{f(p_k)}\right)+\sigma^2\left(\frac{1}{\mu_j}+\frac{1}{\mu_k}\right)\right\},\end{aligned}$$

where $\mathrm{Var}(\beta_{1p_k})$ follows from Theorem 1. Similarly, $\mathrm{Var}(\delta^S_{jk})=\mathrm{Var}(\beta_{2p_k}-\beta_{2p_j})/\left(p_k-p_j\right)^2$, where

$$\begin{aligned}\mathrm{Var}(\beta_{2p_k}-\beta_{2p_j}) &= \mathrm{Var}(\beta_{2p_j})+\mathrm{Var}(\beta_{2p_k})-2\,\mathrm{Cov}(\beta_{2p_k},\beta_{2p_j})\\ &= \frac{1}{nC}*\left\{n\tau^2\left(\frac{1}{f(p_j)}+\frac{1}{f(p_k)}\right)+\sigma^2\left(\frac{1}{s_j}+\frac{1}{s_k}\right)\right\}.\end{aligned}$$

Taking the square root yields the result for the slope effect.

Proof of Theorem 2. Consider an RS design with at least one interior saturation and a pure control. We want to compute $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ for (9), when

$$X'_{ic}=[\ 1\quad T_{ic}\quad S_{ic}\ ].$$

By Lemma 2, the error distribution is block-diagonal. From Section A.1, $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2A^{-1}+\tau^2A^{-1}BA^{-1}$, with

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

and

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

where the second equalities follow from $\sum_{i=1}^n T_{ic}=nP_c$, $\sum_{i=1}^n S_{ic}=n(1-P_c)$ if $P_c>0$, $\sum_{c=1}^C nP_c = n\sum_{p\in\Pi}pCf(p)=nC\mu$, $\sum_{c=1}^C(\sum_{i=1}^n T_{ic})^2=\sum_{c=1}^C n^2P_c^2=n^2C(\eta^2+\mu^2)$,

$\sum_{c=1}^{C}(\sum_{i=1}^{n} T_{ic} \times \sum_{i=1}^{n} S_{ic}) = \sum_{c=1}^{C} n^2 P_c(1-P_c) = n^2 C(\mu - \eta^2 - \mu^2)$, and other analogous calculations. Taking the diagonal entries of $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$, and plugging in (7) to relate η^2 and η_T^2 yields the result for Theorem 2.

A.2 Proofs of Lemmas 1 and 2

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

Proof of Lemma 2. The expected value of the error for treated individuals is

$$\begin{aligned} E[\varepsilon_{ic}|T_{ic}=1] &= E[Y_{ic}(1,T_{-i,c}) - \overline{Y}(1)|T_{ic}=1] \\ &= \sum_{p\in\Pi\setminus\{0\}} Pr(P_c = p|T_{ic}=1)\overline{Y}(1,p) - \overline{Y}(1) \\ &= \frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}} pf(p)\overline{Y}(1,p) - \overline{Y}(1) \\ &= 0, \end{aligned}$$

since from the perspective of a treated individual, $Pr(P_c = p|T_{ic}=1) = pf(p)/\mu$. Similarly, $E[\varepsilon_{ic}|S_{ic}=1] = 0$ and $E[\varepsilon_{ic}|T_{ic} = S_{ic} = 0] = 0$. The variance of the error for treated individuals is

$$\begin{aligned} E[\varepsilon_{ic}^2|T_{ic}=1] &= E[(Y_{ic}(1,T_{-i,c}) - \overline{Y}(1))^2|T_{ic}=1] \\ &= \frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}} pf(p)(\tau^2+\sigma^2+\overline{Y}(1,p)^2) - \frac{2}{\mu}\sum_{p\in\Pi\setminus\{0\}} pf(p)\overline{Y}(1,p)\overline{Y}(1) + \overline{Y}(1)^2 \\ &= \tau^2+\sigma^2 + \frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}} pf(p)\overline{Y}(1,p)^2 - \overline{Y}(1)^2. \end{aligned}$$

Similarly, the variance of the error for within-cluster controls is

$$E[\varepsilon_{ic}^2|S_{ic}=1] = \tau^2+\sigma^2 + \frac{1}{\mu_S}\sum_{p\in\Pi\setminus\{0\}} (1-p)f(p)\overline{Y}(0,p)^2 - \overline{Y}(0)^2,$$

and the variance of the error for pure controls is $E[\varepsilon_{ic}^2|C_{ic}=1]=\tau^2+\sigma^2$. The covariance of the error between treated individuals in the same cluster is

$$E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic}=T_{jc}=1] \;=\; \tau^2+\frac{1}{(\eta^2+\mu^2)}\sum_{p\in\Pi\setminus\{0\}}p^2f(p)(\overline{Y}(1,p)-\overline{Y}(1))^2.$$

Similarly,

$$\begin{aligned}E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic}=S_{jc}=1] &= \tau^2+\frac{\sum_{p\in\Pi\setminus\{0\}}p(1-p)f(p)(\overline{Y}(1,p)-\overline{Y}(1))(\overline{Y}(0,p)-\overline{Y}(0))}{\sum_{p\in\Pi\setminus\{0\}}p(1-p)f(p)},\\ E[\varepsilon_{ic}\varepsilon_{jc}|S_{ic}=S_{jc}=1] &= \tau^2+\frac{\sum_{p\in\Pi\setminus\{0\}}(1-p)^2f(p)(\overline{Y}(0,p)-\overline{Y}(0))^2}{\sum_{p\in\Pi\setminus\{0\}}(1-p)^2f(p)},\end{aligned}$$

and $E[\varepsilon_{ic}\varepsilon_{jc}|C_{ic}=C_{jc}=1]=\tau^2$. Errors across clusters are not correlated since outcomes across clusters are not correlated.

Proof of Corollary 3. Recall $\overline{Y}(1)\equiv\frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}}pf(p)\overline{Y}(1,p)$. Suppose treatment effects are constant on Π. Then $\overline{Y}(1,p)=\overline{Y}(1)$ for all p. Therefore, $\phi_T\equiv\frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}}pf(p)\overline{Y}(1,p)^2-\overline{Y}(1)^2=\frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}}pf(p)\overline{Y}(1)^2-\overline{Y}(1)^2=0$. Similarly, $\phi_S=0$, $\phi_{TT}=0$, $\phi_{SS}=0$ and $\phi_{TS}=0$. 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 $\phi_T=0$. Therefore, $\frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}}pf(p)\overline{Y}(1,p)^2=\overline{Y}(1)^2$. But then there must be no variation in the population average potential outcome across saturations for treated individuals. Similarly, $\phi_S=0$ and there must be no variation in the population average potential outcome across saturations for within-cluster controls. Therefore, treatment effects are constant on Π.

A.3 Proofs of Optimal Design Results

The proof of Corollary 1 follows directly from Theorem 1.

Proof of Corollary 2. Consider an RS design with at least two interior saturations. Without loss of generality assume $p_k>p_j$ and fix the share of clusters allocated to these saturations at $f(p_k)+f(p_j)=F$. Denote the size of saturation bin j by $f(p_j)=f$ and bin k by $f(p_k)=F-f$. Let $p_j=p$ and denote the distance between the two saturations by

$\Delta \equiv p_k - p_j$. Then minimizing (4) is equivalent to solving:

$$\min_f \min_\Delta \min_p \frac{1}{\Delta^2}\left(n\rho\left(\frac{1}{f}+\frac{1}{F-f}\right)+(1-\rho)\left(\frac{1}{fp}+\frac{1}{(F-f)(p+\Delta)}+\frac{1}{f(1-p)}+\frac{1}{(F-f)(1-\Delta-p)}\right)\right)$$

For each Δ, the minimum occurs at the p that solves

$$p(1-p)f = (p+\Delta)(1-\Delta-p)(F-f). \tag{13}$$

For any (p, Δ, f) that satisfy (13) such that $f \neq F/2$, there exists a $(p', \Delta', F/2)$ that also satisfy (13) and strictly lower the objective function, since it lowers the term $\frac{1}{f}+\frac{1}{F-f}$ without affecting the term inside $(1-\rho)$. Therefore, the optimal size of each saturation bin is equal, $f = F/2$. Given Δ and $f = F/2$, the optimal saturations are $p_j = (1-\Delta)/2$ and $p_k = p + \Delta = (1+\Delta)/2$, which are symmetric about $1/2$. The Δ that minimizes (4) is equivalent to solving:

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

The optimal Δ^* solves:

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

If $\rho = 0$, then $2\Delta^2 - 1 = 0$, yielding $\Delta^* = \sqrt{2}/2$. Note that $(2\Delta^2-1)/(\Delta(1-\Delta^2)^2)$ 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 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, $p_j^* = 1/n$ and $p_k^* = (n-1)/n$.

Proof of Corollary 4. Fixing μ and ψ, $\text{Var}(\hat{\beta}_1)$ and $\text{Var}(\hat{\beta}_2)$ are both minimized at $\eta_T^2 = 0$. This corresponds to a partial population experiment with a control group of size ψ and a treatment saturation of $p = \mu/(1-\psi)$.

Proof of Corollary 5. A partial population design with share of control clusters ψ and share of treated individuals μ has SEs

$$\begin{aligned}
\text{SE}(\hat{\beta}_1) &= \sqrt{\frac{\tau^2+\sigma^2}{nC} * \left\{ n\rho \left(\frac{1}{(1-\psi)\,\psi} \right) + (1-\rho) \left(\frac{1}{\mu} + \frac{1}{\psi} \right) \right\}} \\
\text{SE}(\hat{\beta}_2) &= \sqrt{\frac{\tau^2+\sigma^2}{nC} * \left\{ n\rho \left(\frac{1}{(1-\psi)\,\psi} \right) + (1-\rho) \left(\frac{1}{1-\mu-\psi} + \frac{1}{\psi} \right) \right\}}.
\end{aligned}$$

Fixing ψ, the optimal treatment share solves

$$\min_{\mu} \text{SE}(\hat{\beta}_1) + \text{SE}(\hat{\beta}_2),$$

which has solution $\mu = (1-\psi)/2$. This implies $\mu_S = \mu$, which corresponds to a partial population experiment with treatment saturation $p = 1/2$. Plugging in $\mu = (1-\psi)/2$ yields

$$\text{SE}(\hat{\beta}_1) = \text{SE}(\hat{\beta}_2) = \sqrt{\frac{\tau^2+\sigma^2}{nC} * \left\{ n\rho \left(\frac{1}{\psi(1-\psi)} \right) + (1-\rho) \left(\frac{1+\psi}{\psi(1-\psi)} \right) \right\}}$$

Thus, the optimal share of control clusters solves

$$\min_{\psi} n\rho \left(\frac{1}{\psi(1-\psi)} \right) + (1-\rho) \left(\frac{1+\psi}{\psi(1-\psi)} \right). \tag{14}$$

When $\rho = 0$, (14) is minimized at $\psi^* = \sqrt{2}-1$. When $\rho = 1$, (14) is minimized at $\psi^* = 1/2$. When $\rho \in (0,1)$, the general FOC for (14) is

$$(1-\rho)(\psi^2 + 2\psi - 1) + n\rho(2\psi - 1) = 0.$$

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 $\rho \in (0,1)$, $\psi^* \in (\sqrt{2}-1, 1/2)$.

B Additional Analysis

This section presents additional uses of an RS design. First, we compute the power of an RS design to detect treatment effects when it is determined ex post that there are no spillover effects. We show that the SE of an RS design is nested between the SE of a blocked design

and the SE of clustered design. Second, we present a parametric linear model of spillovers and illustrate how an RS design can consistently estimate the pure control outcome. This is a useful result for situations in which institutional constraints prohibit including a pure control group.

B.1 Using Within-cluster Controls as Counterfactuals

Suppose there is no evidence of spillovers on untreated individuals – the estimate of $SNT(p)$ is a precise zero for all p. Then the within-cluster controls are not subject to interference from the treatment and they can be used as counterfactuals to increase the power of the treatment effect estimates.

Assumption 3. $\overline{Y}(0,p) = \overline{Y}(0,0)$ *for all* $p \in \Pi$.[28]

Given Assumption 3, the researcher can pool within-cluster and pure controls, and estimate a simpler model to measure treatment effects,

$$Y_{ic} \quad = \quad \beta_0 + \beta_1 T_{ic} + \varepsilon_{ic}. \tag{15}$$

This regression returns $\hat{\overline{ITT}} = \hat{\beta}_1$.[29] Power is significantly improved by the larger counterfactual, particularly when τ is high. Theorem 3 characterizes the pooled SE when the within-cluster controls are included in the counterfactual.

Theorem 3 (SE with Within-Cluster Controls)**.** *Let* (Π, f) *be a RS design. Assume Assumptions 1, 2 and 3. The SE of* $\overline{ITT}$ *is:*

$$\overline{\text{SE}}_{ITT} = \sqrt{\frac{\tau^2 + \sigma^2}{nC}\left(n\rho\left(\frac{\eta^2}{\mu^2(1-\mu)^2}\right) + (1-\rho)\left(\frac{1}{\mu(1-\mu)}\right)\right)}.$$

Theorem 3 nests the SE of this model between the more familiar expressions for the SE of the blocked and clustered designs. An immediate corollary is that the power of the pooled treatment effect in any RS design lies between the power of the treatment effect in the blocked and clustered designs.

[28] This assumption is testable using any RS design that yields a consistent estimate of $\hat{SNT}(p)$.

[29] Saturation weights are necessary if there are spillover effects on treated individuals, $ST(p) \neq 0$ for some $p \in \Pi$.

Corollary 6. *Let* $\overline{\text{SE}}^{RS}_{ITT}$ *be the standard error for an RS design with share of treated individuals* μ*. Then*

$$\text{SE}^{B}_{ITT} < \text{SE}^{RS}_{ITT} < \text{SE}^{C}_{ITT},$$

where $\text{SE}^{B}_{ITT} = \sqrt{\frac{1}{nC} * \frac{\sigma^2}{\mu(1-\mu)}}$ *is the SE in a blocked design with saturation* μ *and* $\text{SE}^{C}_{ITT} = \sqrt{\frac{1}{nC} * \frac{\sigma^2+n\tau^2}{\mu(1-\mu)}}$ *is the SE in a clustered design with share of treatment clusters* μ*.*

This follows directly from Theorem 3, noting that the blocked design corresponds to $\eta^2 = 0$ and the clustered design corresponds to $\eta^2 = \mu(1-\mu)$.

Corollary 6 provides context for a well-known result. Fixing the treatment share μ, the SE is decreasing in the variance of the treatment saturation η^2, and minimized when this variation is zero, which corresponds to the blocked design. Second, fixing η^2, the SE is minimized when $\mu(1-\mu)$ is maximized, which occurs at $\mu = 1/2$. As is well known, in the absence of spillovers, the optimal design is a blocked study with equal size treatment and control groups.

B.2 Inference in a Linear Model

It is also possible to measure slope effects by imposing a functional form on the shape of the spillover effects. For example, we could use an affine model to estimate the first order slope effect.

Assumption 4 (Linearity)**.** $\overline{Y}(t,p)$ *is affine in* p *for* $t \in \{0,1\}$*.*

Given Assumption 4, it is natural to estimate:

$$Y_{ic} \;=\; \alpha_0 + \alpha_1 T_{ic} + \delta_1 P_c + \delta_2 T_{ic} P_c + \varepsilon_{ic} \tag{16}$$

This regression identifies the TUT as the intercept of the treatment effect, $\hat{TUT} = \hat{\alpha}_1$. The coefficients δ_1 and δ_2 are slope terms estimating how spillover effects change with the saturation, $\hat{dST}/dp = \hat{\delta}_1 + \hat{\delta}_2$ and $\hat{dSNT}/dp = \hat{\delta}_1$. A test for $dST/dp = dSNT/dp$ is given by the hypothesis test $\delta_2 = 0$.[30]

[30] In order to test the linearity assumption, one could estimate

$$Y_{ic} = \alpha_0 + \alpha_1 T_{ic} + \alpha_2 S_{ic} + \delta_1 P_c + \delta_2 T_{ic} P_c + \varepsilon_{ic}. \tag{17}$$

The intercept δ_2 estimates the spillover effect on untreated individuals at saturation zero. This should be

Theorem 4 characterizes the standard errors of the slope estimands in an affine model, which is proportional to $\text{SE}(\hat{\delta_1} + \hat{\delta_2})$ for treated individuals and $\text{SE}(\hat{\delta_1})$ for untreated individuals.

Theorem 4 (Affine Slope SE)**.** *Assume Assumptions 1 and 2 and let* (Π, f) *be a randomized saturation design with* $\kappa \geq 2$ *interior saturations. The SE for the slope effect of treated individuals is:*

$$\text{SE}_{DT} = \sqrt{\frac{1}{nC} * \{n\tau^2 h_1 + \sigma^2 h_2\}}$$

where $m_x \equiv \frac{1}{C}\sum_{c=1}^{C} P_c^x = \sum_{p\in\Pi} p^x f(p)$ *and*

$$h_1 \equiv \left(\frac{(\eta^2+\mu^2)^2 - 2\mu(\eta^2+\mu^2)m_3 + \mu^2 m_4}{((\eta^2+\mu^2)^2 - \mu m_3)^2}\right)$$

and

$$h_2 \equiv \left(\frac{\eta^2+\mu^2}{(\eta^2+\mu^2)^2 - \mu m_3}\right).$$

An analogous expression characterizes the slope effect of untreated individuals, denoted SE_{DNT}.

B.3 Inference Without a Pure Control

The RS design opens up unique empirical possibilities in studies where there is no pure control group. This is particularly important for settings in which a pure control is not feasible due to regulatory requirements or other exogenous restrictions.[31] Without a pure control group, a study's counterfactual is subject to within-cluster spillovers. An RS design has the distinct advantage of allowing a researcher to test for the presence of spillover effects and estimate the unperturbed counterfactual. If the spillover effect is continuous at zero, the researcher can use the variation in treatment saturation to project what would happen to untreated individuals as the saturation approaches zero.[32] With this unperturbed counterfactual in hand, it is possible to correctly estimate the $\overline{\hat{ITT}}$.

zero, as $SNT(0) = 0$ by definition, so $\alpha_2 = 0$ serves as a hypothesis test for the linearity of the spillover relationship.

[31] For example, in McIntosh et al. (2013), a Mexican government rule required that each participating cluster (municipality) be guaranteed at least one treated sub-unit (neighborhood).

[32] Although continuity is a reasonable assumption, it is not universally applicable. Consider signalling in a ground-hog colony. Individuals are 'treated' by being alerted to the presence of a nearby predator, and the possible individual-level outcomes are 'aware' and 'not aware'. The animal immediately signals danger to the rest of the colony, and control outcomes will be universally 'aware' for any positive treatment saturation, but 'unaware' when the saturation is exactly zero.

Assumption 4 provides a simple way to estimate the pure control by assuming that the outcome variable is linear with respect to treatment saturation. Note that Theorem 4 requires at least two interior saturations, but does not require a pure control group.

Theorem 5 (Consistency with No Control)**.** *Assume 1, 2 and 4, and let* (Π, f) *be a randomized saturation design with* $\kappa \geq 2$ *interior saturations. Then the OLS estimates from* (16) *are consistent estimates of* $I\hat{T}T(p) = \hat{\alpha}_1 + (\hat{\delta}_1 + \hat{\delta}_2)p$ *and* $S\hat{N}T(p) = \hat{\delta}_1 p$*.*

Proof. Given Assumption 4, we can identify the slope of the ITT and SNT. The rest of the proof follows easily from the Law of Large Numbers. □

The hypothesis test $\delta_1 = 0$ determines whether there is a spillover effect on untreated individuals. If spillovers are present, then the counterfactual needs to be corrected. The coefficient $\hat{\alpha}_0$ is an estimate of the desired 'pure' control outcome, $\overline{Y}(0,0)$.

B.4 Proofs for Theorems in Appendix B

Proof of Theorem 3. We compute $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ for (9) when

$$x_{ic}' = \begin{bmatrix} 1 & T_{ic} \end{bmatrix}.$$

Recall from Section A.1 that $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2 A^{-1} + \tau^2 A^{-1} B A^{-1}$. Therefore,

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

and

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

where the second equalities follow from $\sum_{i=1}^{n} T_{ic} = nP_c$, $\sum_{c=1}^{C} nP_c = nC\mu$, and $\sum_{c=1}^{C} (\sum_{i=1}^{n} T_{ic})^2 = \sum_{c=1}^{C} n^2 P_c^2 = n^2 C(\eta^2 + \mu^2)$. This can be used to compute

$$\text{Var}(\hat{\beta}_1) = \frac{1}{nC} * \left[\left(\frac{\eta^2}{\mu^2(1-\mu)^2} \right) n\tau^2 + \left(\frac{1}{\mu(1-\mu)} \right) \sigma^2 \right].$$

Fixing μ, this expression is minimized at $\eta^2 = 0$.

Proof of Theorem 4. We want to compute $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ for (9) when

$$x_{ic}' = [\ 1 \quad T_{ic} \quad T_{ic}P_c \quad S_{ic} \quad S_{ic}P_c\ ].$$

Recall from Section A.1 that $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2 A^{-1} + \tau^2 A^{-1}BA^{-1}$. Therefore

$$\begin{aligned} A &= \sum_{c=1}^{C}\sum_{i=1}^{n} \begin{bmatrix} 1 & T_{ic} & T_{ic}P_c & S_{ic} & S_{ic}P_c \\ T_{ic} & T_{ic}^2 & T_{ic}^2P_c & T_{ic}S_{ic} & T_{ic}S_{ic}P_c \\ T_{ic}P_c & T_{ic}^2P_c & T_{ic}^2P_c^2 & T_{ic}S_{ic}P_c & T_{ic}S_{ic}P_c^2 \\ S_{ic} & T_{ic}S_{ic} & T_{ic}S_{ic}P_c & S_{ic}^2 & S_{ic}^2P_c \\ S_{ic}P_c & T_{ic}S_{ic}P_c & T_{ic}S_{ic}P_c^2 & S_{ic}^2P_c & S_{ic}^2P_c^2 \end{bmatrix} \\ &= nC \begin{bmatrix} 1 & \mu & \eta^2+\mu^2 & 1-\mu-\psi & \mu-\eta^2+\mu^2 \\ \mu & \mu & \eta^2+\mu^2 & 0 & 0 \\ \eta^2+\mu^2 & \eta^2+\mu^2 & m_3 & 0 & 0 \\ 1-\mu-\psi & 0 & 0 & 1-\mu-\psi & \mu-\eta^2+\mu^2 \\ \mu-\eta^2+\mu^2 & 0 & 0 & \mu-\eta^2+\mu^2 & \eta^2+\mu^2-m_3 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} B &= \sum_{c=1}^{C} \left(\begin{bmatrix} n \\ \sum_{i=1}^n T_{ic} \\ \sum_{i=1}^n T_{ic}P_c \\ \sum_{i=1}^n S_{ic} \\ \sum_{i=1}^n S_{ic}P_c \end{bmatrix} * \begin{bmatrix} n \\ \sum_{i=1}^n T_{ic} \\ \sum_{i=1}^n T_{ic}P_c \\ \sum_{i=1}^n S_{ic} \\ \sum_{i=1}^n S_{ic}P_c \end{bmatrix}' \right) \\ &= n^2C \begin{bmatrix} 1 & \mu & m_2 & 1-\mu-\psi & \mu-m_2 \\ \mu & m_2 & m_3 & \mu-m_2 & m_2-m_3 \\ m_2 & m_3 & m_4 & m_2-m_3 & m_3-m_4 \\ 1-\mu-\psi & \mu-m_2 & m_2-m_3 & 1-2\mu+m_2-\psi & \mu-2m_2+m_3 \\ \mu-m_2 & m_2-m_3 & m_3-m_4 & \mu-2m_2+m_3 & m_2-2m_3+m_4 \end{bmatrix} \end{aligned}$$

where $m_x \equiv \frac{1}{C}\sum_{c=1}^{C} P_c^x = \sum_{p\in\Pi} p^x f(p)$. Taking the diagonal entries of $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ yields the result, $\text{SE}_{DT} = \text{SE}(\hat{\delta}_3)$ and $\text{SE}_{DNT} = \text{SE}(\hat{\delta}_4)$.