Optimal Design of Experiments in the Presence of Interference*

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

September 2016

Abstract

In recent years, empirical researchers have become increasingly interested in studying settings with interference between units, in which an individual's outcome depends on the behavior and outcomes of others in her group. This paper formalizes the optimal design and analysis of two-stage randomized controlled trials to measure causal estimands in the presence of interference. Our main contributions are to map the potential outcomes framework for causal inference to a regression model with clustered errors, calculate the power of different two-stage designs and derive analytical insights for the optimal design of such experiments. We show that the power to detect average treatment effects declines precisely with the ability to identify novel treatment and spillover effects. We provide software for optimal design.

KEYWORDS: Experimental Design, Networks, Cash Transfers
JEL: C93, O22, I25

*We are grateful for the useful comments received from Peter Aronow, Frank DiTraglia, Elizabeth Hollaran, Cyrus Samii, Patrick Staples as well as 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, and Yale. 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

In randomized experiments, the possibility of interference between individuals gives rise to a plethora of important questions. What if a program creates benefits to some only by diverting these benefits from others? How do treatment benefits depend on the intensity of treatment within a population? Does the study even have an unpolluted counterfactual? In many contexts, a full understanding of the policy environment requires us to measure spillover and threshold effects that are not captured by (or, worse, are sources of bias in) standard experimental designs, as understanding spillover effects is critical to arriving at a sensible view of overall program impacts.

For this reason, empirical researchers in multiple academic disciplines have become increasingly interested in bringing spillover effects under the lens of experimental investigation. A new wave of empirical work has emerged in the past decade that relaxes the assumptions around interference between units. This literature includes studies that uncover network effects using experimental variation across treatment groups,[1] leave some members of a group untreated,[2] exploit plausibly exogenous variation in within-network treatments,[3] or intersect an experiment with pre-existing networks.[4] Further progress has been made by exploiting *partial population* experiments (Robert A. Moffitt 2001), in which clusters are assigned to treatment or control, and a subset of individuals are offered treatment within clusters assigned to treatment. This design partially overcomes the challenge of allowing for interference, but provides no exogenous variation in treatment saturation to estimate the extent to which program effects are driven by the intensity of treatment in a cluster.[5]

The recent interest in interference between individuals has also spawned a rich econometrics literature. Peter M. Aronow and Cyrus Samii (2015) and Charles F. Manski (2013)

[1] Matteo Bobba and Jeremie Gignoux (2013); Edward Miguel and Michael Kremer (2004).

[2] Felipe Barrera-Osorio, Marianne Bertrand, Leigh Linden and Francisco Perez-Calle (2011); Rafael Lalive and M. A. Cattaneo (2009).

[3] Philip S. Babcock and John L. Hartman (2010); Lori A. Beaman (2012); Timothy G. Conley and Christopher R. Udry (2010); Esther Duflo and Emmanuel Saez (2002); Kaivan Munshi (2003).

[4] Abhijit Banerjee, Arun G. Chandrasekhar, Esther Duflo and Matthew O. Jackson (2013); Jiehua Chen, Macartan Humphries and Vijay Modi (2010); Karen Macours and Renos Vakis (2008); Emily Oster and Rebecca Thornton (2012).

[5] Most extant partial population experiments feature cluster-level saturations that are either endogenous (Oportunidades) or fixed ((Esther Duflo and Emmanuel Saez 2003), where they are typically set at 50%). PROGRESA/Oportunidades (Mexico) is perhaps the most-studied example of a partial population experiment. This program 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 (Jennifer Alix-Garcia, Craig McIntosh, Katharine R. E. Sims and Jarrod R. Welch 2013; Manuela Angelucci and Giacomo De Giorgi 2009; Gustavo J. Bobonis and Frederico Finan 2009). Other partial population experiments include Duflo and Saez (2003) and Peter Kuhn, Peter Kooreman, Adriaan Soetevent and Arie Kapteyn (2011).

consider the most general settings, in which there are "arbitrary forms of independence and treatment assignment dependencies. Eric J. Tchetgen Tchetgen and Tyler VanderWeele (2010) and Lan Liu and Michael G. Hudgens (2014) follow Michael Hudgens and Elizabeth Halloran (2008), both of which assume stratified interference to estimate causal effects under the assumption of structural symmetry. Bryan S Graham, Guido W Imbens and Geert Ridder (2010) relax this assumption with one of observational symmetry, aka exchangeability.

In this paper, we focus on settings with *partial interference*, in which individuals are split into mutually exclusive groups, such as villages or schools, and interference occurs between individuals within a group but not across groups. The most direct way of using experimental design to study spillover effects in this environment is to conduct a two-level randomization in which the share of individuals assigned to treatment within a cluster is directly randomized, after which individuals within each cluster are randomly assigned to treatment according to the realized cluster-level intensity in the first stage.[6] Following Hudgens and Halloran (2008), we provide a foundation for these *randomized saturation* (RS) designs using Rubin's potential outcomes framework and maintaining stratified interference. We then depart from the previous literature by assuming the variance-covariance matrix of the population distribution of potential outcomes is block-diagonal. This allows us to map the potential outcomes model to a regression model with clustered standard errors, which is the traditional method used to analyze RS experiments in economics. We derive closed-form expressions for the variance of various treatment and spillover estimands, which can be used for statistical power calculations and to derive the optimal design of experiments to measure spillover effects. In related work, Keisuke Hirano and Jinyong Hahn (2010) study the power of a partial population experiment to analyze a linear-in-means model. By studying the design of RS experiments within the environment that underlies the typical considerations of cluster-randomized controlled trials, one of our main contributions is to illustrate the power tradeoffs that exist in choosing the set of saturations and share of clusters to assign to each saturation, as well as providing a convenient tool to calculate statistical power.

The advantage of a two-level *randomized saturation* experiment is that it allows the researchers to identify a set of novel estimands: not only can the researcher consistently identify the usual intention-to-treat effect and the spillover effect on the non-treated, but she can also observe spillover effects on treated units, and understand how the intensity of treatment drives spillover effects for the treated and the untreated alike. The experimental estimate of the Total Causal Effect on all units within treatment clusters provides the poli-

[6] Abhijit Banerjee, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston and Nina Singh (2012); Matias Busso and Sebastian Galiani (2014); Bruno Crepon, Esther Duflo, Marc Gurgand, Roland Rathelot and Philippe Zamora (2013); Xavier Gine and Ghazala Mansuri (2012); Betsy Sinclair, Margaret McConnell and Donald P. Green (2012).

cymaker a very simple tool to understand how altering the intensity of implementation will drive outcomes for the representative individual.

However, the ability to identify novel estimands comes with a cost, namely decreased statistical power. We show that the same variation that permits measurement of saturation effects is directly detrimental to the power of the simple experimental comparison of treatment to pure control. By placing RS designs in the random effects environment, we provide the closest possible analog to the familiar power calculations in cluster randomized trials, thereby making the design tradeoffs present in RS experiments as transparent as possible. We illustrate these tradeoffs with an application of our results to different cluster-RCT designs and provide code to allow researchers to calculate the power of different potential RS designs.

The remainder of the paper is structured as follows. Section 2 sets up the potential outcomes framework, defines a randomized saturation 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 variance of the estimates and derives properties of the optimal randomized saturation design to detect different effects. Section 4 presents an application of optimal design results to actual field experiments. All proofs not contained in the body of the paper are in Appendix A. Appendix B presents additional possibilities that an RS design can identify.

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 treatments. Represent the target population by probability space $(\mathcal{I}, \Omega, \mathbb{P})$ with outcome set Ω and individuals in $\mathcal{I}$ partitioned into equal-sized, non-overlapping groups, or clusters, of size n.[7]

Individual i in cluster c has response function $Y_{ic}(\cdot) : \{0,1\}^n \to \mathcal{Y}$ that maps the potential cluster treatment vectors $\mathbf{t}_c = (t_{1c}, ..., t_{nc}) \in \{0,1\}^n$ into potential outcomes $Y_{ic}(\mathbf{t}_c) \in \mathcal{Y}$, where $t_{ic} \in \{0,1\}$ is a binary treatment status in which $t_{ic} = 1$ corresponds to being offered treatment and $t_{ic} = 0$ corresponds to not being offered treatment. Note that the response function is independent of t_{jd} for all $d \neq c$ and $j = 1, ..., n$; spillovers may flow within a cluster, but do not flow between clusters. Thus we relax the stable unit treatment value

[7]We assume clusters are equal in size to simplify the analysis. The results easily extend to unequally sized clusters. 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.

assumption (SUTVA) within clusters, but maintain it across clusters. This set-up is referred to as *partial interference* (Michael E. Sobel 2006).[8]

Our goal is to study the power of different experimental designs to detect treatment and spillover effects. Thus, we seek to characterize the variance of the estimands measuring these effects. This requires additional assumptions on the structure of interference. The first assumption restricts how the identity of individuals receiving treatment impacts the outcome of other individuals in the same cluster. We use the *stratified interference* assumption proposed by Hudgens and Halloran (2008), which assumes that the outcome of an individual is independent of the identity of the other individuals assigned to treatment. Let $\mathbf{t}_{-i,c}$ denote the cluster treatment vector $\mathbf{t}_c$ with the ith entry removed and let $\mathcal{P}(\mathbf{t}_{-i,c})$ denote the set of permutations of $\mathbf{t}_{-i,c}$.

Assumption 1 (Stratified Interference)**.** *For any treatment vector* $\mathbf{t}_{-i,c} \in \{0,1\}^{n-1}$ *and permutation* $\mathbf{t}' \in \mathcal{P}(\mathbf{t}_{-i,c})$, $Y_{ic}(t_{ic}, \mathbf{t}') = Y_{ic}(t_{ic}, \mathbf{t}_{-i,c})$ *for all* $t_{ic} \in \{0,1\}$, $i = 1, ..., n$ *and* $c = 1, ..., C$.

Stratified interference allows for a characterization of the variance without possessing information about the underlying network structure within a cluster.[9] Tchetgen Tchetgen and VanderWeele (2010) also maintain stratified interference, and it is similar in spirit to the anonymous interactions assumption in Manski (2013). Given Assumption 1, we can simplify the response function to $Y_{ic}(\cdot) : \{0,1\} \times [0,1] \to \mathcal{Y}$, where the potential outcome $Y_{ic}(t_{ic}, p_c)$ depends on individual treatment status t_{ic} and cluster treatment saturation $p_c = \frac{1}{n}\sum_{i=1}^{n} t_{ic}$.

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, or (ii) a statistical random effect in outcomes that is correlated between individuals, or *correlated effects* (Charles Manski 1993). Our definition of the response function allows for (i), which is a type of interference between units. We allow for (ii) by assuming a variance-covariance structure for the distribution of potential outcomes that allows units within the same cluster to have correlated potential outcomes. Note (ii) does not stem from interference between units.

[8]The assumption of no interference across groups is testable. For example, in the cluster-RCT evaluating the Zomba Cash Transfer Program in Malawi (Sarah Baird, Craig McIntosh and Berk Özler 2011), the cluster unit is a census enumeration area (EA). Each EA contains an average of 250 households spanning several contiguous villages. EAs were selected as the clusters because they provide sampling frames with clearly delinieated official boundaries. Given the population and geographic size of an EA, it is plausible that SUTVA will hold between EAs. We tested this assumption and present the findings in Table A1.

[9]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.

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

1. $\text{Var}(Y_{ic}(t_1, p_1)) = \sigma^2 + \tau^2$,
2. $\text{Cov}(Y_{ic}(t_1, p_1), Y_{jc}(t_2, p_2)) = \tau^2$ *for* $i \neq j$,
3. $\text{Cov}(Y_{ic}(t_1, p_1), Y_{jd}(t_2, p_2)) = 0$ *for* $c \neq d$

for all $t_1, t_2 \in \{0, 1\}$ *and* $p_1, p_2 \in [0, 1]$.

This variance-covariance structure allows potential outcomes to be correlated across individuals within the same cluster, but assumes potential outcomes are uncorrelated across clusters. Let $\rho \equiv \tau^2/(\tau^2 + \sigma^2)$ denote the intra-cluster correlation (ICC). It imposes homoskedasticity across all potential outcomes for a given individual (i.e. all treatment statuses and saturations), and across potential outcomes between two individuals in the same cluster.

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 yields the regression model typically used for power calculations when there is no interference.[10] It enables a direct comparison of the power of RS designs to the power of the canonical individually-randomized (blocked) and cluster-randomized (clustered) designs, making explicit the impact that randomizing saturation has on power. A regression model with a block-diagonal structure is also the model underlying the use of OLS with clustered standard errors to analyze resulting data, the method commonly used for analysis.

2.2 A Randomized Saturation Design

Suppose a researcher draws a sample of C clusters of size n.[11] 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. Each cluster c is assigned a treatment saturation $P_c \in \Pi \subset [0, 1]$ according

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

[11] The randomized saturation design and studies discussed here use a simple, spatially defined definition of 'cluster' that is mutually exclusive and exhaustive. This is distinct from determining how to assign treatment in overlapping social networks (Peter Aronow 2012), which requires a more complex sequential randomization routine (Panos Toulis and Edward Kao 2013). An additional benefit of a randomized saturation 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.

to the distribution f, where P_c is a random variable with finite support Π. 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 assigned treatment $T_{ic} \in \{0, 1\}$, where the realized cluster treatment saturation specifies the probability of treatment, $P(T_{ic} = 1|P_c) = P_c$. 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.[12]

We refer to individuals assigned to treatment as *treated* individuals, individuals in clusters assigned saturation $P_c = 0$ 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 the following ex ante distribution over treatment outcomes:

Treated	μ	$\equiv$	$\sum_{p\in\Pi} pf(p)$
Pure Control	ψ	$\equiv$	$f(0)$
Within-cluster Control	μ_S	$\equiv$	$1 - \mu - \psi$

A randomized saturation design has a pure control if $\psi > 0$.

An RS design introduces correlation between the treatment statuses of two individuals in the same cluster, $r \equiv \text{Cor}(T_{ic}, T_{jc}) = \eta^2/(\mu(1-\mu))$, 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 blocked design has no correlation, $r = 0$. These two designs bracket the continuum of RS designs, so it is natural that RS designs have an intermediate level of correlation.

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.

Definition 1 (Non-Trivial Design)**.** *A randomized saturation design is* non-trivial *if the support of* Π *contains at least* 2 *saturations and* $\exists\ p \in \Pi$ *such that* $p \in (0, 1)$*.*

Multiple saturations guarantee a comparison group to determine whether effects vary with

[12]Fixing the probability of treatment 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 $f(P) = \mu/P$.

treatment saturation, and an interior saturation guarantees the existence of within-cluster controls to identify spillovers on the untreated. Note that the blocked and clustered designs are trivial, and it is not possible to identify spillover effects in these designs, while the partial population design is non-trivial and it is possible to identify spillover effects on the untreated.

Remark. We implicitly assume that the researcher samples all individuals who are part of the spillover network in every sampled cluster. If this is not the case and spillovers occur on individuals outside of the study sample, either because there is a 'gateway to treatment' and not all eligible individuals are sampled or because not all individuals in a cluster 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 remainder of the theoretical analysis, we assume that the assigned and true saturations coincide.

2.3 Treatment and Spillover Estimands

Next we define a set of estimands for treatment and spillover effects, both at specific saturations and pooled across multiple saturations. We focus on average effects across all individuals in the population. Define the *population average potential outcome* at individual treatment assignment $t \in \{0, 1\}$ and saturation $p \in [0, 1]$ as $\overline{Y}(t, p) = E[Y_{ic}(t, p)]$.

Individual Saturation Effects. 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*

[13] For example, Gine and Mansuri (2012) 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 (Craig McIntosh, Tito Alegria, Gerardo Ordonez and Rene Zenteno 2013); and malaria prevention efforts target vulnerable individuals, who account for a small share of total cluster population (GF Killeen, TA Smith, HM Ferguson, H Mshinda, S Abdulla et al. 2007).

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 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 inter-

[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.

ference, 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.

The relationship between the VT and the slopes of the ST and SNT captures the trade-off a policy maker faces in determining which saturation maximizes the TCE. Taking the derivative of TCE with respect to p and rearranging terms,

$$\frac{dTCE}{dp} = VT(p) + p\left(\frac{dST}{dp}\right) + (1-p)\left(\frac{dSNT}{dp}\right).$$

The total marginal impact of adding another individual to treatment at saturation p, $dTCE/dp$, is equal to the sum of the direct benefit of treatment to the individual, $VT(p)$, and the marginal spillover effect on others in the cluster from treating this additional individual. If treatment is directly beneficial and exhibits positive externalities, increasing the saturation will increase the TCE. Faced with negative externalities, increasing the saturation will increase the TCE only if the marginal individual benefit exceeds the marginal social cost.

Pooled Effects. Combining observations from clusters with different saturations yields pooled effects, which are weighted sums of the effects at each individual saturation. For RS design (Π, f), we define the pooled ITT as the difference between the expected outcome for individuals offered treatment and the expected outcome for pure control individuals, giving *equal* weight to each saturation in the support of Π,

$$\overline{ITT} \equiv \frac{1}{|\Pi|-1}\sum_{\Pi\setminus\{0\}} \overline{Y}(1,p) - \overline{Y}(1,\underline{p}) = \frac{1}{|\Pi|-1}\sum_{\Pi\setminus\{0\}} ITT(p),$$

where $|\Pi|-1$ is the number of non-zero saturations. The definitions for $\overline{ST}$, $\overline{SNT}$, $\overline{TCE}$ and $\overline{VT}$ are analogous, substituting 0 for $\underline{p}$. Defining the pooled ST and SNT to place equal weight on each saturation-specific effect ensures that they are comparable.[16] The pooled VT can be expressed as $\overline{VT} = \overline{ITT} - \overline{SNT}$. However, it is not possible to express the pooled

[16] Other weighting schemes yield pooled estimates that are not easily comparable. For example, if each saturation is weighted according to the share of individuals at that saturation, then saturation p receives weight $pf(p)$ in $\overline{ITT}$ and weight $(1-p)f(p)$ in $\overline{SNT}$. Thus, high saturations are given more weight for the pooled ITT, while low saturations are given more weight for the pooled SNT, and a comparison of the two measures does not have a natural interpretation.

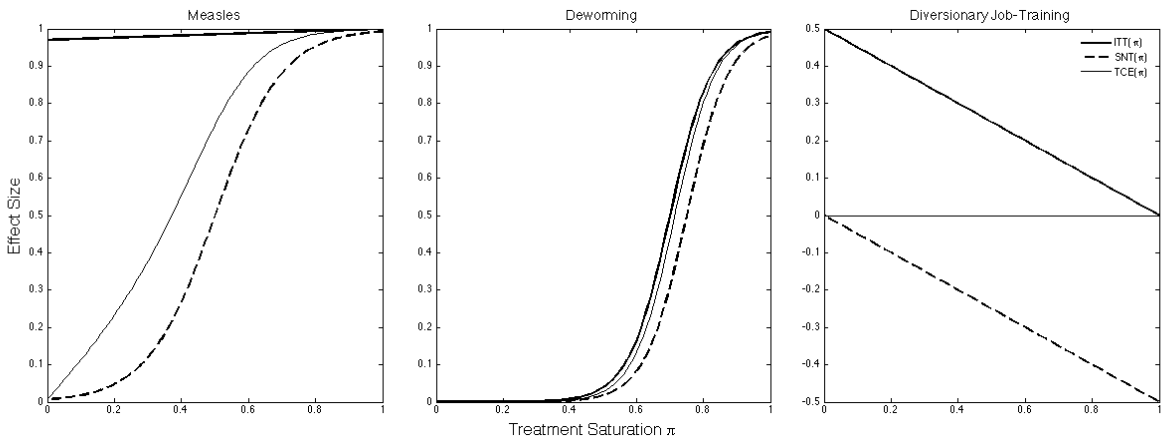

FIGURE 1. Examples

TCE as a function of $\overline{ITT}$ and $\overline{SNT}$, since

$$\overline{TCE} = \frac{1}{|\Pi| - 1} \sum_{\Pi \setminus \{0\}} TCE(p) = \frac{1}{|\Pi| - 1} \sum_{\Pi \setminus \{0\}} \left(pITT(p) + (1-p)SNT(p) \right).$$

The pooled measures can also be used to test for the presence of spillovers, externalities and diversionary effects. For example, a sufficient test for the presence of spillovers on treated individuals is $\overline{ST} \neq 0$.

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. First consider an intervention that vaccinates a share p of a cluster. The vaccination is almost fully protective of the vaccinated individual independent of the treatment saturation, which means the $ITT(p)$ is flat with respect to p. However, the protection to the non-treated only becomes sizeable as the saturation becomes high enough to provide herd immunity, and the $SNT(p)$ varies from zero to one. Thus, the $VT(p)$ 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 treatment saturation 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 near 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, a non-trivial 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 Minimum Detectable Effects and Optimal Design

This section maps the potential outcomes framework developed in Section 2.1 into a regression model to estimate the estimands defined in Section 2.3 and derives properties of the optimal RS design to detect different sets of effects. We begin with the individual saturation and slope estimands and then derive complementary results for the pooled estimands. The section concludes with an illustration of the power trade-off between measuring slope and pooled effects.

3.1 Individual Saturation and Slope Effects

Suppose that a researcher would like to draw inference about how treatment and spillover effects vary with treatment saturation.

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 $Y_{ic}^{obs} \equiv Y_{ic}(T_{ic}, P_c)$ 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 and treatment status. Recall $\overline{Y}(t,p) = E[Y_{ic}(t,p)]$ is the population average potential outcome at treatment t and saturation p. Let $\beta_0 \equiv \overline{Y}(0,0)$, $\beta_{1p} \equiv \overline{Y}(1,p) - \overline{Y}(0,0)$ and $\beta_{2p} \equiv \overline{Y}(0,p) - \overline{Y}(0,0)$. Define the error as

$$\begin{aligned} \varepsilon_{ic} \equiv \quad & \sum_{p \in \Pi} T_{ic} \mathbb{1}_{P_c=p} \{Y_{ic}(1,p) - \overline{Y}(1,p)\} + \sum_{p \in \Pi} S_{ic} \mathbb{1}_{P_c=p} \{Y_{ic}(0,p) - \overline{Y}(0,p)\} \\ & + (1 - T_{ic} - S_{ic}) \{Y_{ic}(0,0) - \overline{Y}(0,0)\}. \end{aligned} \tag{2}$$

S. Athey and G. Imbens (2016) 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 in terms of the distribution of potential outcomes.

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

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

Given Lemma 1, the OLS estimate of (1) will yield an unbiased estimate of $\boldsymbol{\beta}$. For any non-trivial RS design with 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.

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

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

$$\delta^T_{jk} \equiv \left(\beta_{1p_k} - \beta_{1p_j}\right) / \left(p_k - p_j\right),$$

with an analogous expression for the slope effect on within-cluster controls, δ^S_{jk}. 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]

Minimum Detectable Effects. The minimum detectable effect (MDE) is the smallest value of an estimand that it is possible to distinguish from zero (Howard S. Bloom 1995). Suppose that the true value of an estimand is β. Given statistical significance level α, the null hypothesis that $\beta = 0$ is rejected with probability γ (the power) for values of β that exceed:

$$\text{MDE} = (t_{1-\gamma} + t_\alpha) * \text{SE}(\hat{\beta}). \tag{3}$$

Therefore, given an RS design, in order to determine the MDEs of a set of estimands, we need to characterize the variance of the estimates of these estimands.

Our first result characterizes the MDEs for the individual saturation effects estimated in (1), using the block diagonal variance-covariance matrix derived in Lemma 1.[19]

Theorem 1 (Individual Saturation MDE)**.** *Assume Assumptions 1 and 2 and let* (Π, f) *be a randomized saturation design with a pure control. For each* $p \in \Pi$*, the MDE of* $ITT(p)$ *for statistical significance level* α *and power* γ *is:*

$$\text{MDE}^T(p) = (t_{1-\gamma} + t_\alpha)\sqrt{\frac{\tau^2 + \sigma^2}{nC} * \left\{(n-1)\,\rho\left(\frac{1}{f(p)} + \frac{1}{\psi}\right) + \left(\frac{1}{pf(p)} + \frac{1}{\psi}\right)\right\}}.$$

Substituting $(1-p)\,f\,(p)$ *for* $pf(p)$ *yields an analogous expression for the MDE of* $SNT(p)$*, denoted* $\text{MDE}^S(p)$*.*

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

[19] In general, the OLS estimator is inefficient when errors are correlated. Therefore, power calculations based on the variance of the OLS estimate will be conservative if GLS or another more efficient estimator is used to analyze the resulting data.

This expression illustrates the relationship between the clustered error structure and the power of the RS design. The first term in the brackets captures the variation in $\hat{ITT}(p)$ due to correlated variation within a cluster and the second term captures the variation in $\hat{ITT}(p)$ due to individual variation. Consider the two extreme cases of the error distribution, no correlation ($\tau^2 = 0$) and perfect correlation ($\sigma^2 = 0$). If $\tau^2 = 0$, the $\text{MDE}^T(p)$ depends on the share of treatment individuals at saturation p, $pf(p)$ and the share of control individuals, ψ. There is no correlation between individuals' potential outcomes within a cluster, so observing the potential outcome $Y_{ic}(1,p)$ for treated individual i provides no information about the unobserved potential outcome $Y_{jc}(1,p)$ for within-cluster control individual j. At the other extreme, if $\sigma^2 = 0$, the $\text{MDE}^T(p)$ depends on the *total* share of individuals at saturation p, $f(p)$, and the share of control individuals. There is perfect correlation between individuals' potential outcomes within a cluster, so observing the potential outcome $Y_{ic}(1,p)$ for treated individual i provides perfect information about the unobserved potential outcome $Y_{jc}(1,p)$ for within-cluster control individual j. Intermediate levels of correlation depend on a weighted average of the share of treated individuals total share of individuals at saturation p and the share of control individuals.[20]

Optimal Design: Individual Saturation Effects. The design choice for measuring individual saturation effects involves choosing the optimal share of clusters to allocate to each saturation bin, given a set of saturations Π and an objective function that minimizes a weighted sum of the MDEs at each saturation. If the researcher places equal weight on each effect, she chooses f to solve

$$\min_{f} \sum_{p \in \Pi} \left(\text{MDE}^T(p) + \text{MDE}^S(p) \right). \tag{4}$$

In general, a researcher will want to allocate more clusters to more extreme saturations, as these saturations have more uneven shares of treatment and within-cluster control individuals within a cluster. As the intra-cluster correlation increases, this asymmetry is muted since within-cluster control individuals provide information about treated individuals, and vice versa, and the uneven share has a smaller impact on power. For a given set of saturations Π, intra-cluster correlation ρ and cluster size n, it is straightforward to numerically optimize the share of clusters to assign to each saturation.

[20] Using this expression to inform experimental design requires estimates of τ^2 and σ^2. One could use existing observational data or conduct a small pilot experiment (Jinyong Hahn, Keisuke Hirano and Dean Karlan 2011). Note that the common cluster component of the error in observational data will include both endogenous effects and correlated effects, which would lead to an overestimation of τ^2 in the presence of spillover effects. A first-stage pilot experiment would not suffer from the same problem by separating these two effects.

The optimal size of the control group ψ^* lies in a relatively narrow range between approximately $1/(2|\Pi|-1)$ and $1/|\Pi|$. The total share of individuals allocated to the control is always larger than the share of individuals allocated to treatment or within-cluster control at any saturation, since the marginal impact of adding another individual to the control reduces all terms in (4). As ρ increases, the optimal control size increases since within-cluster control and treated individuals in the same cluster have more correlated outcomes, and the optimal f allocates fewer clusters to each positive treatment saturation.

Corollary 1 (Optimal Control Size)**.** *Let* Π *be the support of an RS design with at least one interior saturation and a pure control. Then the share of control clusters* ψ^* *that minimizes* (4) *is between* $1/(2|\Pi|-1)$ *and* $1/|\Pi|$ *and is increasing in* ρ*.*

Minimum Detectable Slope Effects. Similar to the MDE, the *Minimum Detectable Slope Effect* (MDSE) is the smallest slope between saturations p_j and p_k that it is possible to distinguish from zero with a given power γ,

$$\text{MDSE} = (t_{1-\gamma} + t_\alpha) * \text{SE}(\hat{\delta}_{jk}). \tag{5}$$

The following theorem characterizes the MDSEs for the slope effects estimated in (1).

Theorem 2 (MDSE)**.** *Assume Assumptions 1 and 2 and let* (Π, f) *be a randomized saturation design with* $\kappa \geq 2$ *interior saturations. The MDSE between saturations* p_j *and* p_k *for statistical significance level* α *and power* γ *is:*

$$\text{MDSE}^T(p_j, p_k) = \frac{(t_{1-\gamma} + t_\alpha)}{p_k - p_j}\sqrt{\frac{\tau^2 + \sigma^2}{nC} * \left((n-1)\,\rho\left(\frac{1}{f(p_j)} + \frac{1}{f(p_k)}\right) + \left(\frac{1}{p_j f\,(p_j)} + \frac{1}{p_k f\,(p_k)}\right)\right)}$$

Substituting $(1-p)\,f\,(p)$ *for* $pf\,(p)$ *yields an analogous expression for the MDSE of* $SNT(p)$*, denoted* $\text{MDSE}^S(p_j, p_k)$*.*

As the distance between two saturations increases, the $1/(p_k - p_j)$ term 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 MDSE between p_j and p_k, while the latter effect dominates at saturations close to zero or one, and spreading the saturations apart increases the MDSE between p_j and p_k. The degree to which the latter effect dominates depends on the correlation between outcomes. 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 MDSE between p_j and p_k and a more

equal distribution leads to a smaller MDSE. When ρ is small, the share of treatment and within-cluster control individuals assigned to each saturation, $p_j f(p_j)$ and $p_k f(p_k)$, are more important.

Optimal Design: Slope Effects. There are two steps to the design choice to measure slope effects: selecting the set of saturations Π and deciding to share of clusters to allocate to each saturation bin, f. If the researcher places equal weight on the MDSE for treated and untreated individuals, she chooses an RS design with two saturations to solve

$$\min_{p_j, p_k} \text{MDSE}^T(p_j, p_k) + \text{MDSE}^S(p_j, p_k). \tag{6}$$

The optimal saturations are symmetric about one half and the optimal distance between saturations is increasing in ρ.

Corollary 2 (Optimal Saturations)**.** *The saturations that minimize* (6) *are* $p_j^* = (1-\Delta^*)/2$ *and* $p_k^* = (1+\Delta^*)/2$*, where* $\Delta^*(\rho, n) \in \left(\sqrt{2}/2, 1\right)$ *is the optimal distance between saturations. If* $\rho = 0$*, then* $\Delta^*(0, n) = \sqrt{2}/2$ *for all* n *and if* $\rho > 0$*, then* $\lim_{n\to\infty} \Delta^*(1, n) = 1$*. The optimal distance* $\Delta^*(\rho, n)$ *is increasing in* ρ *and* n*.*

Given more than two saturations, a researcher could use Theorem 2 to answer questions like what is the optimal spacing of saturations or what share of clusters should be assigned to each bin?

3.2 Pooled Effects

Suppose the researcher would like to pool all treated individuals and all within-cluster controls to measure the pooled ITT and SNT. These pooled estimates provide the most powerful tests for the presence of treatment and spillover effects, as the minimum detectable pooled effect is always smaller than the minimum detectable effect at any individual saturation.

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

$$Y_{ic}^{obs} = \beta_0 + \beta_1 T_{ic} + \beta_2 S_{ic} + \varepsilon_{ic}. \tag{7}$$

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 at treatment t, averaged across all non-zero saturations in the RS

design. 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

$$\varepsilon_{ic} \equiv T_{ic}\{Y_{ic}(1,P_c) - \overline{Y}(1)\} + S_{ic}\{Y_{ic}(0,P_c) - \overline{Y}(0)\} + C_{ic}\{Y_{ic}(0,0) - \overline{Y}(0,0)\}. \tag{8}$$

The following lemma characterizes the distribution of the error in terms of the distribution of potential outcomes.

Lemma 2. *Assume Assumptions 1 and 2. Then the error defined in (8) is strictly exogenous, $E[\varepsilon_{ic}|T_{ic}, S_{ic}] = 0$, and has a block-diagonal variance-covariance matrix with no correlation across clusters, heteroskedastic individual variance $E[\varepsilon_{ic}^2|T_{ic} = 1] = \sigma^2 + \tau^2 + \phi_T$, $E[\varepsilon_{ic}^2|S_{ic} = 1] = \sigma^2 + \tau^2 + \phi_S$ and $E[\varepsilon_{ic}^2|C_{ic} = 1] = \sigma^2 + \tau^2$, and heteroskedastic within-cluster covariance $E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic} = T_{jc} = 1] = \tau^2 + \phi_{TT}$, $E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic} = S_{jc} = 1] = \tau^2 + \phi_{TS}$, $E[\varepsilon_{ic}\varepsilon_{jc}|S_{ic} = S_{jc} = 1] = \tau^2 + \phi_{SS}$ and $E[\varepsilon_{ic}\varepsilon_{jc}|C_{ic} = C_{jc} = 1] = \tau^2$ for $i \neq j$, where $\phi_T = \frac{1}{\mu}\sum_{p\in\Pi\setminus\{0\}} pf(p)\overline{Y}(1,p)^2 - \overline{Y}(1)^2$ 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}.*

Proof. The expected value of the error for treated individuals is $E[\varepsilon_{ic}|T_{ic} = 1] = E[Y_{ic}(1,P_c) - \overline{Y}(1)|T_{ic} = 1] = \sum_{p\in\Pi\setminus\{0\}} Pr(P_c = p|T_{ic} = 1)E[Y_{ic}(1,p)] - \overline{Y}(1) = \sum_{p\in\Pi\setminus\{0\}} g_1(p)\overline{Y}(1,p) - \overline{Y}(1) = 0$ where $g_1(p) \equiv pf(p)/\mu$ 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 $E[\varepsilon_{ic}^2|T_{ic} = 1] = E[(Y_{ic}(1,P_c) - \overline{Y}(1))^2|T_{ic} = 1] = \sum_{p\in\Pi\setminus\{0\}} g_1(p)(\tau^2 + \sigma^2 + \overline{Y}(1,p)^2) - 2\sum_{p\in\Pi\setminus\{0\}} g_1(p)\overline{Y}(1,p)\overline{Y}(1) + \overline{Y}(1)^2 = \tau^2 + \sigma^2 + \sum_{p\in\Pi\setminus\{0\}} g_1(p)\overline{Y}(1,p)^2 - \overline{Y}(1)^2$. Similarly, the variance of the error for within-cluster controls is $E[\varepsilon_{ic}^2|S_{ic} = 1] = \tau^2 + \sigma^2 + \sum_{p\in\Pi\setminus\{0\}} g_0(p)\overline{Y}(0,p)^2 - \overline{Y}(0)^2$ where $g_0(p) \equiv (1-p)f(p)/\mu_S$, 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 $E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic} = T_{jc} = 1] = \sum_{p\in\Pi\setminus\{0\}} g_{11}(p)E[(Y_{ic}(1,p) - \overline{Y}(1))(Y_{jc}(1,p) - \overline{Y}(1))] = \tau^2 + \sum_{p\in\Pi\setminus\{0\}} g_{11}(p)(\overline{Y}(1,p) - \overline{Y}(1))^2$, where $g_{11}(p) \equiv p^2 f(p)/(\eta^2 + \mu^2)$. Similarly, $E[\varepsilon_{ic}\varepsilon_{jc}|T_{ic} = S_{jc} = 1] = \tau^2 + \sum_{p\in\Pi\setminus\{0\}} g_{10}(p)(\overline{Y}(1,p) - \overline{Y}(1))(\overline{Y}(0,p) - \overline{Y}(0))$, where $g_{10}(p) \equiv p(1-p)f(p)/\sum_{p\in\Pi\setminus\{0\}} p(1-p)f(p)$, $E[\varepsilon_{ic}\varepsilon_{jc}|S_{ic} = S_{jc} = 1] = \tau^2 + \sum_{p\in\Pi\setminus\{0\}} g_{00}(p)(\overline{Y}(0,p) - \overline{Y}(0))^2$, where $g_{00}(p) \equiv (1-p)^2 f(p)/\sum_{p\in\Pi\setminus\{0\}}(1-p)^2 f(p)$, 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. □

Therefore, the OLS estimate of (7) will yield an unbiased estimate of $\boldsymbol{\beta}$.

Saturation Weights. When individuals are pooled across saturations, this unintentionally places a disproportionate weight on treated individuals in high saturation clusters and non-treated individuals in low saturation clusters. Similar to sampling weights, what we call

saturation weights correct for the different probability of being assigned to treatment or within-cluster-control in each saturation bin, and are necessary to estimate the pooled ITT and SNT we define in Section 2.3 [21]

Definition 2 (Saturation Weights)**.** *Saturation weights apply weight* $s_{ic} = 1/(P_c f(P_c))$ *to treated individuals* $(T_{ic} = 1)$ *and weight* $s_{ic} = 1/(1 - P_c)f(P_c)$ *to non-treated individuals in treated clusters* $(S_{ic} = 1)$.

Estimating Equation (7) with these saturation weights will yield OLS coefficients that place equal weight on the treatment or spillover effect at each saturation.

For any non-trivial RS design with a pure control, the OLS estimate of $\boldsymbol{\beta}$ in (7) with saturation weights identifies consistent estimates of the pooled effects defined in Section 2.3, including $\widehat{\overline{ITT}} = \hat{\beta}_1$, $\widehat{\overline{SNT}} = \hat{\beta}_2$ and $\widehat{\overline{VT}} = \hat{\beta}_1 - \hat{\beta}_2$. Sufficient tests for the presence of spillover effects are $\hat{\beta}_1 \neq 0$ and $\hat{\beta}_2 \neq 0$, while a one-tailed test of the sign of $\hat{\beta}_2$ determines whether treatment creates a negative or positive externality on non-treated individuals and $\hat{\beta}_1 \neq \hat{\beta}_2$ determines whether there is direct value to treatment.

Different saturation weights are necessary to estimate the pooled TCE because, by definition, this effect weights treated individuals proportional to p and non-treated individuals proportional to $1 - p$ at each saturation p. The OLS estimate of $\boldsymbol{\beta}$ in (7) with saturation weights $s'_c = 1/f(P_c)$ yields a consistent estimate of the pooled TCE,

$$\widehat{\overline{TCE}} = \left(\frac{1}{|\Pi|}\right)\left(\hat{\beta}_1 \sum_{\Pi\setminus\{0\}} p + \hat{\beta}_2 \sum_{\Pi\setminus\{0\}} (1-p)\right).$$

Note (7) doesn't identify the $\overline{ST}$ or TUT.

Minimum Detectable Pooled Effects. The pooled MDE depends on the variance in treatment saturation across clusters, η^2. It is useful to separate the component of η^2 that arises from multiple non-zero saturations,

$$\eta_T^2 \equiv \sum_{p\in\Pi\setminus\{0\}} \frac{p^2 f(p)}{1-\psi} - \left(\frac{\mu}{1-\psi}\right)^2 = \left(\frac{1}{1-\psi}\right)\eta^2 - \left(\frac{\psi}{(1-\psi)^2}\right)\mu^2 \tag{9}$$

[21] For example, suppose an RS design assigns clusters to three saturations, $\Pi = \{0, 1/3, 2/3\}$ with equal probability, $f(p) = 1/3$ for each $p \in \Pi$. In a cluster assigned $p = 2/3$, an individual is twice as likely to be assigned to treatment as a cluster with $p = 1/3$. Weighting the treated individuals by $s^T_{2/3} = 3/2$ and $s^T_{1/3} = 3$ allows one to calculate the pooled estimate we define in Section 2.3, which places equal weight on both clusters, rather than twice as much weight on the $p = 2/3$ clusters.

where $f(p)/(1-\psi)$ is the distribution of treatment saturation, conditional on $p > 0$, with support $\Pi\backslash\{0\}$. Note $\eta_T^2 = 0$ for the blocked, clustered and partial population designs.

The next result characterizes the MDE for the pooled ITT and SNT when the spillover effects are constant across all saturations in the RS design.

Assumption 3. *For all $p_j, p_k \in \Pi$, $\overline{Y}(1, p_j) = \overline{Y}(1, p_k)$ and $\overline{Y}(0, p_j) = \overline{Y}(0, p_k)$.*

Given Assumption 3, there is no heteroskedasticity in the error.

Theorem 3 (Pooled MDE)**.** *Assume Assumptions 1, 2 and 3 and let (Π, f) be a non-trivial randomized saturation design with a pure control. The MDE of $\overline{ITT}$ for statistical significance level α and power γ is:*

$$\overline{\text{MDE}}^T = (t_{1-\gamma} + t_\alpha)\sqrt{\frac{\tau^2 + \sigma^2}{nC}\left((n-1)\,\rho\left(\frac{1}{(1-\psi)\,\psi} + \left(\frac{1-\psi}{\mu^2}\right)\eta_T^2\right) + \left(\frac{1}{\mu} + \frac{1}{\psi}\right)\right)}.$$

Substituting μ_S for μ yields an analogous expression for the MDE of $\overline{SNT}$, denoted $\overline{\text{MDE}}^S$.

The MDE for the pooled effects depends on the size of the treatment and control groups and the within-cluster variation in treatment status, η_T^2. The first term in the brackets captures the variation in $\hat{\beta}$ due to the common cluster component of the error term, and the second term captures the variation in $\hat{\beta}$ due to individual variation. Introducing randomization into the treatment saturation of clusters results in a power loss when there is a common cluster component to the error. Otherwise, if $\tau^2 = 0$, the standard error only depends on the size of the treatment and control groups, but is independent of how treatment is distributed across clusters.

Optimal Design: Pooled Effects. Next, we derive the optimal RS design to test for the presence of treatment effects and spillover effects on the untreated. When outcomes within clusters are correlated, variance in treatment saturations across clusters leads to a power loss (this follows directly from Theorem 3). Consider the partial population design in which all treated clusters all have the same treatment saturation p. This design minimizes the variation in treatment saturation. Therefore, for any (μ, ψ), the partial population design with $p = \mu/(1-\psi)$ minimizes the MDE for pooled treatment and spillover effects.

Corollary 3 (Optimality of Partial Population Design)**.** *Suppose $\tau^2 > 0$. Then, for any (μ, ψ), the partial population design $\{\{0, \mu/(1-\psi)\}, \{\psi, 1-\psi\}\}$ simultaneously minimizes $\overline{\text{MDE}}^T$ and $\overline{\text{MDE}}^S$.*

Note that the definition of $\overline{ITT}$ and $\overline{SNT}$ depend on the saturations in the support of the RS design, and therefore, different supports Π yield different pooled estimands (unless the $ITT(p)$ and $SNT(p)$ are constant with respect to p). Thus, if the researcher wants to fix the definition of a set of pooled estimands and find the optimal RS design to detect these estimands, she will select a set of saturations Π and chooses the distribution of saturations to minimize

$$\min_{f} \overline{\text{MDE}}^{T} + \overline{\text{MDE}}^{S}.$$

For any Π, it is straightforward to numerically solve for the optimal f.

Consider the optimal treatment saturation p and control size ψ for a partial population experiment. With a single interior saturation, the pooled MDEs are equivalent to the individual MDEs at saturation p. Choosing the optimal p involves a trade-off. The power of the ITT increases with p, while the power of the SNT decreases with p. The relative importance of detecting these two effects, as well as their expected magnitudes, will determine the optimal p. Suppose a researcher places equal weight on each MDE,

$$\min_{(p,\psi)} \text{MDE}^{T}(p) + \text{MDE}^{S}(p). \tag{10}$$

Then the optimal saturation is $p^* = 0.5$, which creates equally sized treatment and within-cluster control groups. The left panel of Figure 2 illustrates the variance of $ITT(p)$ and $SNT(p)$ in a partial population design, as a function of p, and shows that (10) is minimized at $p = 0.5$. The optimal size of the control group depends on the relative magnitude of cluster and individual variation in outcomes, and the size of the cluster, and lies in a relatively narrow range between approximately 0.41 and 0.5.[22] Corollary 4 summarizes these results.

Corollary 4. *Suppose $\tau^2 > 0$. Then a partial population experiment with $p^* = 0.5$ and $\psi^* \in (\sqrt{2} - 1, \sqrt{n(1+n)} - n)$ minimizes (10). If $\rho = 0$, then $\psi^* = \sqrt{2} - 1$ for all n, while if $\rho = 1$, then $\psi^* = \sqrt{n(1+n)} - n$. In this design, $\text{MDE}^{T}(p) = \text{MDE}^{S}(p)$.*

Designating about 41% of individuals as pure controls yields the smallest sum of standard errors when there is no common cluster component to the error, while designating close to 50% is preferable when there is no individual component to error. It is always optimal to have the control be larger than a third of clusters because it serves as the counterfactual for both treatment and spillover groups. As τ increases, the optimal number of control clusters increases. This comparative static arises because the variance in $\hat{\beta}$ due to individual error is proportional to the total number of *individuals* in each treatment group, while the variance in $\hat{\beta}$ due to correlated error is proportional to the total number of *clusters* in each treatment

[22] It is straightforward to numerically compute the optimal control size for a specific (τ, σ, n).

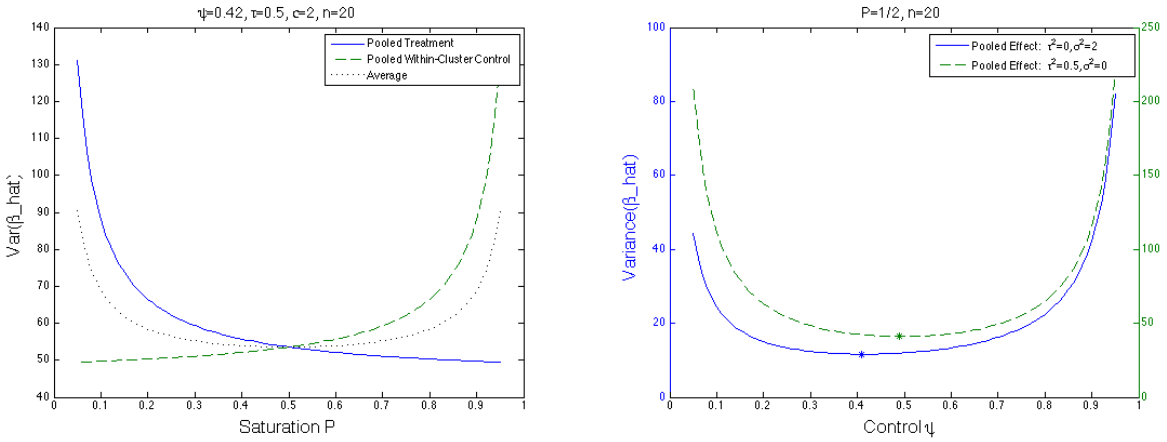

FIGURE 2. Partial Population Design

group. The right panel of Figure 2 illustrates the standard error of $I\hat{T}T(0.5)$ and $S\hat{N}T(0.5)$ in a partial population design (note they are equal), as a function of the control group size ψ, for (i) no intra-cluster correlation ($\tau = 0$), and (ii) perfect intra-cluster correlation ($\sigma = 0$). The minimum in each case is marked with an asterisk.

Corollary 4 is similar in spirit to Hirano and Hahn (2010). They show how a partial population design can identify spillover effects in a linear-in-means model of with no ICC ($\tau^2 = 0$). Their spillover effect of interest is the average spillover on all individuals in a treatment cluster (using our definitions, $pST(p) + (1-p)SNT(p)$) and their treatment effect of interest is the average marginal impact of treatment, conditional on being in a treatment cluster ($ITT(p) - pST(p) - (1-p)SNT(p)$). In this framework, they also show that the optimal partial population saturation for a fixed size control group is $p^* = 1/2$ and that the optimal control group size is $\psi^* = \sqrt{2} - 1$.

Moving away from the partial population design to a design with variation in the treatment saturation, $\eta_T > 0$, leads to a power loss in the ability to detect pooled effects. This power loss increases more rapidly with respect to η_T^2 for settings with higher ICC, and designs with smaller control groups and treatment groups. Corollary 5 characterizes the rate at which this power loss occurs.

Corollary 5 (Linearity of Power Loss)**.** *Fix μ and ψ. Then* $\text{Var}(\hat{\beta}_1)$ *and* $\text{Var}(\hat{\beta}_2)$ *increase linearly with respect to η_T^2, with slope proportional to $\tau^2(1-\psi)/\mu^2$ and $\tau^2(1-\psi)/\mu_S^2$, respectively.*

Taken together, these corollaries provide important insights on experimental design. If the researcher is only interested in detecting treatment effects and spillover effects on the non-treated, then a partial population experiment has the smallest MDE, and Corollary 4 specifies the optimal control group size. However, partial population designs have the

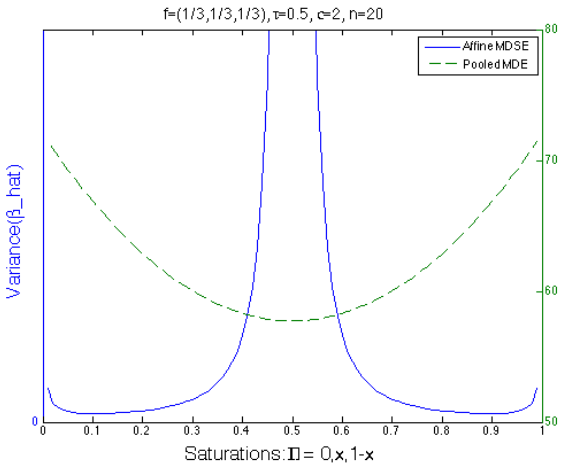

FIGURE 3. Trade-off between Pooled MDE and MDSE

drawback that they only measure effects at a single saturation. When researchers care about the effects at multiple saturations, they will need to introduce variation in the treatment saturation. Corollary 5 establishes the rate at which the power of the pooled effects declines from this increase in treatment saturation variance.

3.3 The Trade-off Between Estimating Pooled and Slope Effects

The optimal RS design for a slope analysis stands in sharp contrast to that for a pooled analysis, most obviously in the size of the pure control and the extent of variation in treatment saturation. Although a partial population design with a saturation of $p = 1/2$ is optimal for detecting pooled effects, this design does not identify slope effects. Similarly, a design with no pure control is optimal for detecting slope effects, but does not identify pooled effects. A graphical representation of the tradeoff between detecting pooled and slope effects is presented in Figure 3. The optimal RS design to identify both slope and pooled effects will depend on the relative importance that the researcher places on each effect, as well as the expected size of each effect.

4 Application

We now provide examples that illustrate our results on experimental design and quantify the power tradeoffs that arise between measuring pooled, individual and slope effects. We characterize the optimal design for different hypothetical objective functions and calculate the power of existing RS designs from published studies in economics and political science.

First, suppose a researcher uses a clustered design to identify the average treatment effect

TABLE 1. Optimal Design to Detect Pooled Effects

	CLUSTERED DESIGN			PARTIAL POPULATION DESIGN				
Objective Function	**min_(Π,f) MDE_T**			**min_(Π,f) MDE_T + MDE_S**			**min_(Π,f) MDE_T + 2MDE_S**	**min_(Π,f) MDE_S s.t. MDE_T <= 0.25**
ρ	0	0.1	1	0	0.1	1	0.1	0.1
Optimal saturation 1: pure control	0	0	0	0	0	0	0	0
Optimal saturation 2: π2	1	1	1	0.50	0.50	0.50	0.41	0.85
Optimal share in pure control: ψ	0.50	0.50	0.50	0.41	0.45	0.49	0.45	0.48
Optimal share in π2	0.50	0.50	0.50	0.59	0.55	0.51	0.55	0.52
MDE_T	0.18	0.24	0.56	0.21	0.27	0.57	0.29	0.25
MDE_S	.	.	.	0.21	0.27	0.57	0.26	0.38
Other parameters: C=100, n=10								

of an intervention for a continuous outcome measure. She conducts her study in $C = 100$ clusters, each of which contains $n = 10$ individuals, and is interested in the MDE of the ITT at significance level $\alpha = 0.05$ and power $\kappa = 0.80$. 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 MDE of $ITT(1)$ depends on the intra-cluster correlation, ρ, and is measured in standard deviations (SD). With no intra-cluster correlation, $\rho = 0$, $\text{MDE}^T(1) = 0.18$. It increases with ρ, rising to 0.24 when $\rho = 0.1$ and 0.56 when $\rho = 1$ (Table 1, Columns 1-3). In this design, the researcher cannot identify any spillover effects on treated or untreated individuals.

Next, suppose that the researcher is also interested in detecting spillover effects on untreated individuals and cares equally about the MDE for the pooled ITT and pooled SNT. Applying Corollary 3, the optimal design is a partial population experiment (PPE) $(\Pi, f) = (\{0, p\}, \{\psi, 1-\psi\})$. This design identifies $ITT(p)$ and $SNT(p)$. From Corollary 4, we know that when the researcher places equal weight on minimizing the MDE^T and MDE^S, 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 clusters in the control group ranges from $\psi^* = 41\%$ to 49% as ρ increases from 0 to 1 (Table 1, Columns 4-6). The MDE for the ITT and SNT are equal, $\text{MDE}^T(0.5) = \text{MDE}^S(0.5)$, and range from 0.21 to 0.57 as ρ increases from 0 to 1. Hence, when the researcher wants to detect spillovers on non-treated individuals, the MDE 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. Th 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

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

	OPTIMAL RS DESIGNS				EXISTING STUDIES			
Objective Function	min_(Π,f) MDSE_T + MDSE_S			min_(Π,f) MDSE_T + MDSE_S +MDE_T + MDE_S	Banerjee et al. & Crepon et al.	Sinclair et al.	Baird et al.	Baird et al.: min_f MDE_S s.t. MDE_T<=.27
ρ	0	0.1	1	0.1	0.1	0.1	0.1	0.1
Saturation 1: π1	0.15	0.13	0.08	0	0	0	0	0
Saturation 2: π2	0.85	0.87	0.92	0.21	0.25	0.10	0.33	0.33
Saturation 3: π3				0.88	0.50	0.50	0.67	0.67
Saturation 4: π4					0.75	1	1	1
Saturation 5: π5					1	.	.	.
Share in π1	0.50	0.50	0.50	0.27	0.20	0.25	0.55	0.46
Share in π2	0.50	0.50	0.50	0.34	0.20	0.25	0.15	0.21
Share in π3				0.39	0.20	0.25	0.15	0.21
Share in π4					0.20	0.25	0.15	0.12
Share in π5					0.20	.	.	.
MDE_T	.	.	.	0.30	0.32	0.31	0.27	0.27
MDE_S	.	.	.	0.31	0.34	0.31	0.33	0.30
MDSE_T	0.50	0.55	0.84	0.62	0.69	0.71	0.83	0.78
MDSE_S	0.50	0.55	0.84	0.56	0.69	0.78	0.72	0.63
Other parameters: C=100, n=10								

treatment saturation and control group size minimizes

$$\min_{p,\psi} \theta \text{MDE}^T(p) + (1-\theta)\text{MDE}^S(p)$$

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 $p^* = 0.41$, 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 $\text{MDE}^T(.41) = 0.29$ and $\text{MDE}^S(.41) = 0.26$.[23] Alternatively, if the researcher wants to minimize the MDE of the SNT, while maintaining a MDE for the pooled ITT of 0.25 or lower (approximately the MDE in the clustered design), then she would use a treatment saturation of $p^* = 0.85$, assigning 85% of the individuals in treated clusters to treatment, and would only be able to detect a spillover effect on the non-treated that is greater than 0.38 (Table 1, Column 8).

Suppose the researcher wishes to estimate a slope effect, and does not care about identifying the individual and pooled ITT and SNT. Then the optimal design will have two interior saturations and no pure control group, and by Corollary 2, the optimal spacing of the two interior saturations is symmetric about 0.5. Solving

$$\min_{p_1,p_2,f} \text{MDSE}^T(p_1,p_2) + \text{MDSE}^S(p_1,p_2),$$

yields optimal saturations $p_1^* = 0.15$ and $p_2^* = 0.85$ when $\rho = 0$ and clusters equally divided

[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 ($p^* = 0.23$).

between these two saturations, $f^*(0.15) = f^*(0.85)$. This produces a MDSE of 0.50 for both the treated and non-treated individuals (Table 2, Column 1). Increasing ρ moves the optimal saturations further apart (Table 2, Columns 2 - 3) and increases the MDSEs, but it remains optimal to equally divide clusters between the two saturations.

However, not many researchers are interested in designing an experiment to minimize the MDSE 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 alter the objective function to put equal weights on both the MDEs and the MDSEs,

$$\min_{p_1,p_2,f} \overline{\text{MDE}}^T + \overline{\text{MDE}}^S + \text{MDSE}^T(p_1, p_2) + \text{MDSE}^S(0, p_2).$$

When $\rho = 0.1$, it is optimal to allocate 34% of the clusters to saturation $p_1^* = 0.21$, 39% to saturation $p_2^* = 0.88$, and the remaining $\psi^* = 27\%$ to the pure control group, i.e. $\Pi^* = \{0, 0.21, 0.88\}$ and $f^* = \{0.27, 0.34, 0.39\}$ (Table 2, Column 4).[24] The calculated MDEs of 0.30 and 0.31 for the pooled ITT and SNT, respectively, indicate an 8-13% (2-4 pp) increase in the MDEs compared to the optimal PPE using the same parameters (Table 1, Column 5).[25] Unlike the power loss that arises when moving from a clustered to a PPE design, the power loss in moving from a PPE design to a design with two interior saturations arises due to 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.

We conclude this section by calculating the power of RS designs used in three published studies, which allow us to further demonstrate the power tradeoff between various objective functions. To facilitate comparability with the optimal designs discussed above, we use the same number of clusters ($C = 100$), individuals per cluster ($n = 10$) and the intra-cluster correlation ($\rho = 0.1$) as in our examples, rather than the actual numbers from the study. We present the implied MDEs and MDSEs from these existing studies and compare them with those from our example in column 4, which uses an objective function that puts equal weights on both the MDEs and the MDSEs.

We begin with the RS design used in Banerjee et al. (2012) and Crepon et al. (2013), in which clusters were assigned to a pure control group and four equally spaced treatment

[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 MDEs for the ITT and SNT (Corollary 1).

[25] The MDSE^T is larger than in the optimal slope design in Column 2 because the distance between saturations for treated individuals is smaller in this design.

saturations in equal numbers: $\Pi = \{0, 0.25, 0.50, 0.75, 1\}$ 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 $\text{MDE}^T = 0.32$, $\text{MDE}^S = 0.34$, and $\text{MDSE}^T = \text{MDSE}^S = 0.69$ (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 comes from Sinclair, McConnell and Green (2012), which randomized nine-digit zip codes in a congressional district in Illinois into a pure control and three different saturations: $\Pi = \{0, \underline{p}, 0.50, 1\}$ and $f = \{0.25, 0.25, 0.25, 0.25\}$, where $\underline{p}$ is the saturation in which only one household is treated.[26] 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 quantify $ST(p)$ for $p = 0.5$ and $p = 1$. Our power calculations for this design yield $\text{MDE}^T = \text{MDE}^S = 0.31$, $\text{MDSE}^T = 0.71$ and $= \text{MDSE}^S = 0.78$, respectively (Table 2, Column 6). The pooled MDEs are quite similar to their counterparts under the optimal design for minimizing the sum of these four variables (Table 2, Column 4), but the MDSEs are substantially higher, particularly for the non-treated (0.78 vs. 0.56) because the largest saturation containing within-cluster controls is 0.5.

Our final example is Baird, McIntosh and Özler (2011), in which three different saturations of schoolgirls were offered cash transfers, along with a pure control group: $\Pi = \{0, 0.33, 0.67, 1\}$ and $f = \{0.55, 0.15, 0.15, 0.15\}$. The main difference in this design is that while the saturations are equally spaced like the other studies discussed above, they are not equally sized: the pure control group, at 55% of the 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 MDEs for the pooled ITT and SNT that are smaller than those in Banerjee et al. (2012) and Crepon et al. (2013), but higher MDSEs, particularly for treated individuals (Table 2, Column 7).

In Baird, McIntosh and Özler (2011), the MDE for the pooled SNT is 6 percentage points (or 20%) higher than that for the ITT, indicating that the pooled spillover effects on the non-treated are underpowered relative to the direct treatment effects. Given this large

[26] 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.

difference between MDEs for the pooled ITT and SNT, we can ask whether there is a way to allocate clusters to the same set of saturations that leads to lower MDEs and MDSEs. We consider the objective $\min_f \overline{\text{MDE}}^S$ subject to $\overline{\text{MDE}}^T \leq 0.27$, which minimizes the MDE of the SNT, subject to the constraint that the MDE of the ITT remains below its value in the original study design. Our calculations show that the researchers should have allocated a lower share of the clusters to the pure control group and to saturation 1, and a higher share to the two interior saturations (Table 2, Column 8). Such a design would dominate the original study design as it would not only substantially lower the MDE for the pooled SNT, but also considerably decrease the MDSE for the treated and non-treated. As we kept the set of saturations fixed in this alternate design, the improved statistical power comes simply 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. Research designs and 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. In this paper, we attempt to formalize the optimal design and analysis of two-stage cluster-randomized controlled trials, which we term *randomized saturation* 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.

More specifically, 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 statistical power. Having laid out the assumptions necessary to estimate both the mean and variance of spillover effects, we derive analytical closed-form expressions for minimum detectable effects (MDE). The MDEs 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 power to detect more basic estimands. This is an inherent feature of RS designs. The previous section provided applications of our main results to illustrate design choices

and power implications for studies with different researcher objectives. The optimization for each of these applications is conducted using code we developed, which is available for researchers at http://

The framework presented here serves as an important guide to inform policy questions. 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 implementation of a program. To the extent that varying the cluster-level saturation leads to differential impacts on prices, norms, and congestion effects, the RS design provides an experimental framework that can bolster both external and internal validity. Furthermore, RS experiments inform future studies: determining whether or not observed baseline intra-cluster correlation is due to interference provides guidance to other researchers in how to design follow-up experiments.

References

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

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

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

Aronow, Peter M., and Cyrus Samii. 2015. "Estimating Average Causal Effects Under Interference Between Units."

Athey, S., and G. Imbens. 2016. "The Econometrics of Randomized Experiments."

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

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

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

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2012. "Can Institutions be Reformed from Within? Evidence from a Randomized Experiment with the Rajasthan Police." National Bureau of Economic Research, Inc NBER Working Papers 17912.

Barrera-Osorio, Felipe, Marianne Bertrand, Leigh Linden, and Francisco 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(2): 167–195.

Beaman, Lori 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(1): 128–161.

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

Bobba, Matteo, and Jeremie Gignoux. 2013. "Policy Evaluation in the Presence of Spatial Externalities: Reassesing the Progresa Program." Working Paper.

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

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

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

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

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

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

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

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

Gine, Xavier, and Ghazala Mansuri. 2012. "Together we will : experimental evidence on female voting behavior in Pakistan." Working Paper.

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

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

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

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

Killeen, GF, TA Smith, HM 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(7): e229.

Kuhn, Peter, Peter Kooreman, Adriaan Soetevent, and Arie Kapteyn. 2011. "The Effects of Lottery Prizes on Winners and Their Neighbors: Evidence from the Dutch Postcode Lottery." *American Economic Review*, 101(5): 2226–2247.

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

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

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

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

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

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

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

Moffitt, Robert A. 2001. "Policy Interventions, Low-Level Equilibria And Social Interactions." 45–82. MIT Press.

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

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

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

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

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

Toulis, Panos, and Edward 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 Mathematical Appendix

A.1 Preliminary Calculations

This section provides background material used in Theorems 3, 2, 5 and 4.

Variance of OLS. Consider the OLS estimate of

$$y_{ic} = x_{ic}'\boldsymbol{\beta} + \varepsilon_{ic}, \tag{11}$$

where x_{ic} is a vector of treatment status covariates and ε_{ic} is an error with block-diagonal structure such that $E[\varepsilon_{ic}^2|X] = \tau^2 + \sigma^2$, $E[\varepsilon_{ic}\varepsilon_{jc}|X] = \tau^2$ if $i \neq j$ and $E[\varepsilon_{ic}\varepsilon_{jd}|X] = 0$ if $c \neq d$. Let $X_c'X_c = \sum_{i=1}^{n} x_{ic}x_{ic}'$ and $\varepsilon_c' = [\ \varepsilon_{1c} \quad ... \quad \varepsilon_{nc}\ ]$. Then

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \frac{1}{nC}A^{-1}BA^{-1}$$

where

$$A \equiv \text{plim}\,\frac{1}{nC}\sum_{c=1}^{C} X_c'X_c \qquad \text{and} \qquad B \equiv \text{plim}\,\frac{1}{nC}\sum_{c=1}^{C} X_c'\varepsilon_c\varepsilon_c'X_c.$$

Given that all clusters are identical ex-ante, $\frac{1}{nC}\sum_{c=1}^{C} E[X_c'X_c] = E[x_{ic}x_{ic}']$ and $\frac{1}{nC}\sum_{c=1}^{C} E[X_c'\varepsilon_c\varepsilon_c'X_c] = \frac{1}{n}E[X_c'\varepsilon_c\varepsilon_c'X_c]$. Therefore, $A = E[x_{ic}x_{ic}']$ and $B = E[X_c'\varepsilon_c\varepsilon_c'X_c]/n$. We will utilize this expression to calculate $\text{Var}(\hat{\boldsymbol{\beta}})$ for different RS designs and vectors of treatment covariates.

A.2 Proof of Theorems

Proof of Theorem 1. Consider an RS design with two interior saturations and a pure control. We want to compute $\text{Var}(\hat{\boldsymbol{\beta}})$ for (11) when $x_{ic}' = [\ 1 \quad T_{1ic} \quad S_{1ic} \quad T_{2ic} \quad S_{2ic}\ ]$, where $T_{1ic} \equiv T_{ic} * \mathbb{1}\{P_c = p_1\}$, $S_{1ic} \equiv S_{ic} * \mathbb{1}\{P_c = p_1\}$, and so forth. By Lemma 1, the error distribution is block-diagonal. Therefore, from Section A.1, $\text{Var}(\hat{\boldsymbol{\beta}}) = A^{-1}BA^{-1}/nC$ where $A = E[x_{ic}x_{ic}']$ and $B = E[X_c'\varepsilon_c\varepsilon_c'X_c]/n$. 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 $q_k \equiv (1-p_k)^2 f(p_k) = s_k - \mu_k + \eta_k$. Then

$$A = E\begin{bmatrix} 1 & T_{1ic} & S_{1ic} & T_{2ic} & S_{2ic} \\ T_{1ic} & T_{1ic}^2 & S_{1ic}T_{1ic} & T_{2ic}T_{1ic} & S_{2ic}T_{1ic} \\ S_{1ic} & T_{1ic}S_{1ic} & S_{1ic}^2 & T_{2ic}S_{1ic} & S_{2ic}S_{1ic} \\ T_{2ic} & T_{1ic}T_{2ic} & S_{1ic}T_{2ic} & T_{2ic}^2 & S_{2ic}T_{2ic} \\ S_{2ic} & T_{1ic}S_{2ic} & S_{1ic}S_{2ic} & T_{2ic}S_{2ic} & S_{2ic}^2 \end{bmatrix} = \begin{bmatrix} 1 & \mu_1 & s_1 & \mu_2 & s_2 \\ \mu_1 & \mu_1 & 0 & 0 & 0 \\ s_1 & 0 & s_1 & 0 & 0 \\ \mu_2 & 0 & 0 & \mu_2 & 0 \\ s_2 & 0 & 0 & 0 & s_2 \end{bmatrix}$$

$$\begin{aligned} B &= \frac{1}{n}E\left(\begin{bmatrix} \sum_{i=1}^n \varepsilon_{ic} \\ \sum_{i=1}^n T_{1ic}\varepsilon_{ic} \\ \sum_{i=1}^n S_{1ic}\varepsilon_{ic} \\ \sum_{i=1}^n T_{2ic}\varepsilon_{ic} \\ \sum_{i=1}^n S_{2ic}\varepsilon_{ic} \end{bmatrix} * \begin{bmatrix} (\sum_{i=1}^n \varepsilon_{ic}) \\ (\sum_{i=1}^n T_{1ic}\varepsilon_{ic}) \\ (\sum_{i=1}^n S_{1ic}\varepsilon_{ic}) \\ (\sum_{i=1}^n T_{2ic}\varepsilon_{ic}) \\ (\sum_{i=1}^n S_{2ic}\varepsilon_{ic}) \end{bmatrix}'\right) \\ &= (n-1)\tau^2 \begin{bmatrix} 1 & \mu_1 & s_1 & \mu_2 & s_2 \\ \mu_1 & \eta_1 & \mu_1 - \eta_1 & 0 & 0 \\ s_1 & \mu_1 - \eta_1 & q_1 & 0 & 0 \\ \mu_2 & 0 & 0 & \eta_2 & \mu_2 - \eta_2 \\ s_2 & 0 & 0 & \mu_2 - \eta_2 & q_2 \end{bmatrix} + \left(\tau^2 + \sigma^2\right) A \end{aligned}$$

Using mathematica to compute $\text{Var}(\hat{\boldsymbol{\beta}}) = \frac{1}{nC} * A^{-1}BA^{-1}$ and taking the diagonal entries yields

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

for each $p_j \in \Pi$. Similarly,

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

for $\text{Var}(\hat{\beta}_{2p_j})$. Plugging these variances into the expressions for the MDE, (3), yields the results for Theorem 1.

Proof of Theorem 2. To compute the MDSE, note $\text{Var}(\delta^T_{jk}) = \text{Var}(\beta_{1p_k} - \beta_{1p_j})/\left(p_k - p_j\right)^2$ and

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

since $\text{Cov}(\beta_{1p_k}, \beta_{1p_j}) = (n\tau^2 + \sigma^2)/\psi nC$ and $\text{Var}(\beta_{1p_j})$ and $\text{Var}(\beta_{1p_k})$ follow from Theorem 1. Similarly, $\text{Var}(\delta^S_{jk}) = \text{Var}(\beta_{2p_k} - \beta_{2p_j})/\left(p_k - p_j\right)^2$ and

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

Plugging these variances into the expressions for the MDSE, (5), yields the results for Theorem 2. Extending the result to more than two interior saturations is analogous.

Proof of Theorem 3. Consider an RS design with at least one interior saturation and a pure control. We want to compute $\text{Var}(\hat{\boldsymbol{\beta}})$ for (11) when $x'_{ic} = \begin{bmatrix} 1 & T_{ic} & S_{ic} \end{bmatrix}$. By Lemma 2, the error distribution is block-diagonal. Therefore, from Section A.1, $\text{Var}(\hat{\boldsymbol{\beta}}) = A^{-1}BA^{-1}/nC$ where $A = E[x_{ic}x'_{ic}]$ and $B = E[X'_c\varepsilon_c\varepsilon'_cX_c]/n$. Then

$$A = E\begin{bmatrix} 1 & T_{ic} & S_{ic} \\ T_{ic} & T^2_{ic} & T_{ic}S_{ic} \\ S_{ic} & T_{ic}S_{ic} & S^2_{ic} \end{bmatrix} = \begin{bmatrix} 1 & \mu & \mu_S \\ \mu & \mu & 0 \\ \mu_S & 0 & \mu_S \end{bmatrix}$$

$$\begin{aligned} B &= \frac{1}{n}E\begin{bmatrix} \left(\sum_{i=1}^n \varepsilon_{ic}\right)^2 & \left(\sum_{i=1}^n \varepsilon_{ic}\right)\left(\sum_{i=1}^n T_{ic}\varepsilon_{ic}\right) & \left(\sum_{i=1}^n \varepsilon_{ic}\right)\left(\sum_{i=1}^n S_{ic}\varepsilon_{ic}\right) \\ \left(\sum_{i=1}^n \varepsilon_{ic}\right)\left(\sum_{i=1}^n T_{ic}\varepsilon_{ic}\right) & \left(\sum_{i=1}^n T_{ic}\varepsilon_{ic}\right)^2 & \left(\sum_{i=1}^n T_{ic}\varepsilon_{ic}\right)\left(\sum_{i=1}^n S_{ic}\varepsilon_{ic}\right) \\ \left(\sum_{i=1}^n \varepsilon_{ic}\right)\left(\sum_{i=1}^n S_{ic}\varepsilon_{ic}\right) & \left(\sum_{i=1}^n T_{ic}\varepsilon_{ic}\right)\left(\sum_{i=1}^n S_{ic}\varepsilon_{ic}\right) & \left(\sum_{i=1}^n S_{ic}\varepsilon_{ic}\right)^2 \end{bmatrix} \\ &= (n-1)\tau^2\begin{bmatrix} 1 & \mu & \mu_S \\ \mu & \eta^2 + \mu^2 & \mu - \mu^2 - \eta^2 \\ \mu_S & \mu - \mu^2 - \eta^2 & \mu_S - \mu + \eta^2 + \mu^2 \end{bmatrix} + \left(\tau^2 + \sigma^2\right)A \end{aligned}$$

Using mathematica to compute $\text{Var}(\hat{\boldsymbol{\beta}}) = \frac{1}{nC} * A^{-1}BA^{-1}$, taking the diagonal entries and plugging in (9) to relate η^2 and η^2_T yields the result for Theorem 3.

A.3 Proof of Corollaries

Proof of Corollary 2. Fixing the size of each saturation bin $f(p_j) = f_j$ and $f(p_k) = f_k$ and the distance between two saturations $\Delta \equiv p_k - p_j$, minimizing $\text{MDSE}^T(p_j, p_k)$ $+\text{MDSE}^S(p_j, p_k)$ is equivalent to solving:

$$\min_{p_j} \left(\frac{1}{f_j p_j} + \frac{1}{f_k(p_j + \Delta)} + \frac{1}{f_j(1 - p_j)} + \frac{1}{f_k(1 - \Delta - p_j)} \right)$$

The minimum occurs at the p_j^* that solves $p_j^*(1 - p_j^*)f_j = (p_j^* + \Delta)(1 - \Delta - p_j^*)f_k$. When $f_j = f_k$, $p_j^* = (1 - \Delta)/2$ and $p_k^* = p_j^* + \Delta = (1 + \Delta)/2$, which is symmetric about 1/2.

Fixing $f_j = f_k$, the Δ that minimizes $\text{MDSE}^T(p_j, p_k) + \text{MDSE}^S(p_j, p_k)$ is equivalent to solving:

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

The optimal Δ^* solves:

$$\frac{(n-1)\tau^2}{2(\tau^2 + \sigma^2)} = \frac{2(\Delta^*)^2 - 1}{(1 - (\Delta^*)^2)^2}$$

If $\tau^2 = 0$, then $2(\Delta^*)^2 - 1 = 0$, yielding $\Delta^* = \sqrt{2}/2$. Note that $(2\Delta^2 - 1)/((1 - \Delta^2)^2)$ is monotonically increasing for $\Delta \in [0, 1)$, and strictly positive for $\Delta > \sqrt{2}/2$. When $\tau > 0$, $((n-1)\tau^2)(2(\tau^2 + \sigma^2))$ is also strictly positive, increasing in τ^2 and decreasing in σ^2. Therefore, $\Delta^* \in \left(\sqrt{2}/2, 1\right)$ for $\tau^2 > 0$ and finite n, Δ^* is increasing in τ^2 and n, and decreasing in σ^2. If $\tau^2 > 0$, then the left hand side converges to ∞ as $n \to \infty$, which requires $\Delta^* \to 1$.

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 4. Fixing ψ, a partial population design has the smallest variance, for any treatment size μ. Therefore, we can restrict attention to the set of partial population designs, and the expression for the MDEs simplify to

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

The objective is to minimize

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

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

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

Thus, it is sufficient to minimize

$$\min_{\psi} \mathrm{SE}(\hat{\beta}_1)$$

to find the optimal control group size. When $\tau^2 = 0$, $\mathrm{SE}(\hat{\beta}_1)$ simplifies to

$$\sqrt{\frac{\sigma^2}{nC} * \left(\frac{\psi+1}{(1-\psi)\psi} \right)}$$

which is minimized at $\psi^* = \sqrt{2} - 1$. When $\sigma^2 = 0$, $\mathrm{SE}(\hat{\beta}_1)$ simplifies to

$$\sqrt{\frac{\tau^2}{nC} * \left(\frac{n+\psi}{(1-\psi)\,\psi} \right)}$$

which is minimized at $\psi^* = \sqrt{n(1+n)} - n$. Note $\lim_{n\to\infty} \sqrt{n(1+n)} - n = 1/2$. Given that $(\psi+1)/((1-\psi)\psi)$ and $(\psi+n)/((1-\psi)\psi)$ are both convex with unique minimums, any weighted sum of these functions is minimized at a value ψ^* that lies between the minimum of each function. Therefore, when $\tau^2 > 0$ and $\sigma^2 > 0$, $\psi^* \in (\sqrt{2}-1, \sqrt{n(1+n)} - n)$.

Proof of Corollary 5. Follows directly from Theorem 3.

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 MDE of an RS design is nested between the MDE of a blocked and 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 4. *$Y(0,p) = Y(0,0)$ for all $p \in \Pi$.*[27]

Given Assumption 4, the researcher can pool within-cluster and pure controls, and estimate a simpler model to measure treatment effects,

$$Y_{ic} \quad = \quad \beta_0 + \beta_1 T_{ic} + \varepsilon_{ic}. \tag{12}$$

This regression returns $\hat{\overline{ITT}} = \hat{\beta}_1$.[28] Power is significantly improved by the larger counterfactual, particularly when τ is high. Theorem 4 characterizes the pooled MDE when the within-cluster controls are included in the counterfactual.

Theorem 4 (MDE with Within-Cluster Controls)**.** *Assume Assumptions 1, 2 and 4. Given RS design (Π, f), the MDE of $\overline{ITT}$ for statistical significance level α and power γ is:*

$$\text{MDE}^T = (t_{1-\gamma} + t_\alpha)\sqrt{\frac{1}{nC}\left(\left(\frac{1+r(n-1)}{\mu(1-\mu)}\right)\tau^2 + \left(\frac{1}{\mu(1-\mu)}\right)\sigma^2\right)}$$

where $r = \eta^2/\mu(1-\mu)$ is the correlation in treatment status between two individuals in the same cluster.

Theorem 4 nests the MDE of this model between the more familiar expressions for the MDE 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.

Corollary 6 (Nesting of MDE)**.** *Let MDE^T_{RS} be the minimum detectable effect for a randomized saturation design with treatment probability μ. Then*

$$\text{MDE}^T_B < \text{MDE}^T_{RS} < \text{MDE}^T_C,$$

where MDE^T_B is the MDE in a blocked design with saturation μ and MDE^T_C is the MDE in a clustered design with share of treatment clusters μ.

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

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

This follows directly from Theorem 4, noting that the blocked design corresponds to $r = 0$ and the clustered design corresponds to $r = 1$.

Corollary 6 provides context for a well-known result. Fixing the treatment probability μ, the MDE 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 MDE is minimized when $\mu(1-\mu)$ is maximized, which occurs at $\mu = 1/2$. As is well known, the optimal design in the absence of spillovers is a blocked study with equal size treatment and control groups.

Proof of Theorem 4. We want to compute $\text{Var}(\hat{\boldsymbol{\beta}})$ for (12). Recall from Section A.1 that $\text{Var}(\hat{\boldsymbol{\beta}}) = A^{-1}BA^{-1}/nC$ where $A = E[x_{ic}x_{ic}']$ and $B = E[X_c'\varepsilon_c\varepsilon_c'X_c]/n$. Therefore, given $x_{ic}' = [\ 1 \quad T_{ic}\ ]$,

$$A = E\begin{bmatrix} 1 & T_{ic} \\ T_{ic} & T_{ic}^2 \end{bmatrix} = \begin{bmatrix} 1 & \mu \\ \mu & \mu \end{bmatrix}$$

and

$$\begin{aligned} B &= \frac{1}{n}E\begin{bmatrix} \left(\sum_{i=1}^n \varepsilon_{ic}\right)^2 & \left(\sum_{i=1}^n \varepsilon_{ic}\right)\left(\sum_{i=1}^n T_{ic}\varepsilon_{ic}\right) \\ \left(\sum_{i=1}^n \varepsilon_{ic}\right)\left(\sum_{i=1}^n T_{ic}\varepsilon_{ic}\right) & \left(\sum_{i=1}^n T_{ic}\varepsilon_{ic}\right)^2 \end{bmatrix} \\ &= \tau^2(n-1)\begin{bmatrix} 1 & \mu \\ \mu & \eta^2+\mu^2 \end{bmatrix} + \left(\tau^2+\sigma^2\right)A \end{aligned}$$

This can be used to compute

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

Using $\eta^2 = r\mu(1-\mu)$, we can express $\text{Var}(\hat{\beta}_1)$ in terms of μ and r.

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

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

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 5 (Linearity). *$\mu(t,p)$ is affine in p for $t \in \{0,1\}$.*

Given Assumption 5, it is natural to estimate:

$$Y_{ic} \quad = \quad \alpha_0 + \alpha_1 T_{ic} + \delta_1 P_c + \delta_2 T_{ic} P_c + \varepsilon_{ic} \tag{13}$$

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, $d\hat{ST}/dp = \hat{\delta}_1 + \hat{\delta}_2$ and $dS\hat{N}T/dp = \hat{\delta}_1$. A test for $dST/dp = dSNT/dp$ is given by the hypothesis test $\delta_2 = 0$.[29]

Theorem 5 characterizes the analytical expression for the MDSE in the affine model, which is proportional to $SE(\hat{\delta_1} + \hat{\delta_2})$ for treated individuals and $SE(\hat{\delta_1})$ for untreated individuals.

Theorem 5 (Affine MDSE). *Assume Assumptions 1 and 2 and let (Π, f) be a randomized saturation design with $\kappa \geq 2$ interior saturations. The MDSE of treated individuals for statistical significance level α and power γ is:*

$$\text{MDSE}^T = (t_{1-\gamma} + t_\alpha) \sqrt{\frac{1}{nC} * \left\{(n-1)\,\tau^2 h_1 + (\tau^2 + \sigma^2)\,h_2\right\}}$$

where

$$h_1 = \left(\frac{(\eta^2+\mu^2)^2 - 2\mu(\eta^2+\mu^2)E[P_c^3] + \mu^2 E[P_c^4]}{((\eta^2+\mu^2)^2 - \mu E[P_c^3])^2} \right) \text{ and } h_2 = \left(\frac{\eta^2+\mu^2}{(\eta^2+\mu^2)^2 - \mu E[P_c^3]} \right)$$

An analogous expression characterizes the MDSE of untreated individuals, denoted MDSE^S.

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.[30] 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

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

$$Y_{ic} = \alpha_0 + \alpha_1 T_{ic} + \alpha_2 S_{ic} + \delta_1 P_c + \delta_2 T_{ic} P_c + \varepsilon_{ic}. \tag{14}$$

The intercept δ_2 estimates the spillover effect on untreated individuals at saturation zero. This should be zero, as $SNT(0) = 0$ by definition, so $\alpha_2 = 0$ serves as a hypothesis test for the linearity of the spillover relationship.

[30] 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).

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.[31] With this unperturbed counterfactual in hand, it is possible to correctly estimate the $\overline{\hat{ITT}}$.

Assumption 5 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 5 requires at least two interior saturations, but does not require a pure control group.

Theorem 6 (Consistency with No Control)**.** *Assume 1, 2 and 5, and let* (Π, f) *be a randomized saturation design with* $\kappa \geq 2$ *interior saturations. Then the OLS estimates from* (13) *are consistent estimates of* $\hat{ITT}(p) = \hat{\alpha}_1 + (\hat{\delta}_1 + \hat{\delta}_2)p$ *and* $\hat{SNT}(p) = \hat{\delta}_1 p$*.*

Proof. Given Assumption 5, 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, $\overline{Y}(0, 0)$.

Proof of Theorem 5. We want to compute $\text{Var}(\hat{\boldsymbol{\beta}})$ for (11) when

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

[31] 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.

Recall from Section A.1 that $\text{Var}(\hat{\boldsymbol{\beta}}) = A^{-1}BA^{-1}/nC$ where $A = E[x_{ic}x_{ic}']$ and $B = E[X_c'\varepsilon_c\varepsilon_c'X_c]/n$. Therefore

$$\begin{aligned} A &= \frac{1}{n}\sum_{i=1}^{n} E\begin{bmatrix} 1 & T_{ic} & T_{ic}P_c & S_{ic} & S_{ic}P_c \\ T_{ic} & T_{ic}^2 & T_{ic}^2P_c & T_{ic}S_{ic} & T_{ic}S_{ic}P_c \\ T_{ic}P_c & T_{ic}^2P_c & T_{ic}^2P_c^2 & T_{ic}S_{ic}P_c & T_{ic}S_{ic}P_c^2 \\ S_{ic} & T_{ic}S_{ic} & T_{ic}S_{ic}P_c & S_{ic}^2 & S_{ic}^2P_c \\ S_{ic}P_c & T_{ic}S_{ic}P_c & T_{ic}S_{ic}P_c^2 & S_{ic}^2P_c & S_{ic}^2P_c^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mu & \eta^2+\mu^2 & 1-\mu-\psi & \mu-\eta^2+\mu^2 \\ \mu & \mu & \eta^2+\mu^2 & 0 & 0 \\ \eta^2+\mu^2 & \eta^2+\mu^2 & E[P_c^3] & 0 & 0 \\ 1-\mu-\psi & 0 & 0 & 1-\mu-\psi & \mu-\eta^2+\mu^2 \\ \mu-\eta^2+\mu^2 & 0 & 0 & \mu-\eta^2+\mu^2 & \eta^2+\mu^2-E\left[p^3\right] \end{bmatrix} \end{aligned}$$

$$\begin{aligned} B &= \frac{1}{n}E\left(\begin{bmatrix} (\sum_{i=1}^n \varepsilon_{ic}) \\ (\sum_{i=1}^n T_{ic}\varepsilon_{ic}) \\ (\sum_{i=1}^n T_{ic}P_c\varepsilon_{ic}) \\ (\sum_{i=1}^n S_{ic}\varepsilon_{ic}) \\ (\sum_{i=1}^n S_{ic}P_c\varepsilon_{ic}) \end{bmatrix} * \begin{bmatrix} (\sum_{i=1}^n \varepsilon_{ic}) \\ (\sum_{i=1}^n T_{ic}\varepsilon_{ic}) \\ (\sum_{i=1}^n T_{ic}P_c\varepsilon_{ic}) \\ (\sum_{i=1}^n S_{ic}\varepsilon_{ic}) \\ (\sum_{i=1}^n S_{ic}P_c\varepsilon_{ic}) \end{bmatrix}'\right) \\ &= (n-1)\tau^2 D + \left(\tau^2+\sigma^2\right)A \end{aligned}$$

where

$$D = \begin{bmatrix} 1 & \mu & E[P_c^2] & 1-\mu-\psi & \mu-E[P_c^2] \\ \mu & E[P_c^2] & E[P_c^3] & \mu-E[P_c^2] & E[P_c^2]-E\left[P_c^3\right] \\ E[P_c^2] & E[P_c^3] & E[P_c^4] & E[P_c^2]-E\left[P_c^3\right] & E[P_c^3]-E\left[P_c^4\right] \\ 1-\mu-\psi & \mu-E[P_c^2] & E[P_c^2]-E\left[P_c^3\right] & 1-2\mu+E[P_c^2]-\psi & \mu-2E[P_c^2]+E\left[P_c^3\right] \\ \mu-E[P_c^2] & E[P_c^2]-E\left[P_c^3\right] & E[P_c^3]-E\left[P_c^4\right] & \mu-2E[P_c^2]+E\left[P_c^3\right] & E[P_c^2]-2E\left[P_c^3\right]+E\left[P_c^4\right] \end{bmatrix}$$

Using mathematica to compute $\text{Var}(\hat{\delta}) = \frac{1}{nC} * A^{-1}BA^{-1}$ and taking the diagonal entries yields the result. The MDSE^T is a function of $SE(\hat{\delta}_3)$, while the MDSE^S is a function of $SE(\hat{\delta}_4)$.

TABLE A1. Robustness check using cross-EA variation in treatment intensity

Dependent Variable:	Terms Enrolled		Average Test Score		Ever Married		Ever Pregnant	
	(1)	(2)	(3)	(4)	(7)	(8)	(9)	(10)
CCT	0.119	0.126	0.022	0.008	0.000	-0.023	0.040	-0.011
	(0.0431)***	(0.085)	(0.00896)**	(0.015)	(0.023)	(0.044)	(0.026)	(0.041)
UCT	0.059	0.052	0.005	-0.019	-0.064	-0.090	-0.057	-0.114
	(0.050)	(0.111)	(0.013)	(0.018)	(0.0269)**	(0.0484)*	(0.0240)**	(0.0508)**
Within CCT EA Control	0.013	0.016	0.021	0.023	0.010	0.011	0.008	0.008
	(0.047)	(0.047)	(0.014)	(0.0134)*	(0.023)	(0.023)	(0.026)	(0.025)
Within UCT EA Control	-0.100	-0.095	-0.020	-0.015	0.000	0.002	-0.021	-0.020
	(0.074)	(0.077)	(0.023)	(0.023)	(0.034)	(0.036)	(0.029)	(0.031)
# of treated EAs within 3 km	-0.021	-0.020	-0.005	-0.002	0.005	0.005	0.004	0.003
	(0.018)	(0.020)	(0.005)	(0.006)	(0.009)	(0.012)	(0.009)	(0.011)
# of treated EAs between 3 & 6 km	0.010	0.019	0.001	0.006	-0.004	-0.002	-0.005	-0.003
	(0.013)	(0.016)	(0.003)	(0.004)	(0.006)	(0.007)	(0.006)	(0.008)
# of total EAs within 3 km	0.012	0.011	0.006	0.004	-0.003	-0.004	0.001	0.002
	(0.012)	(0.013)	(0.00281)**	(0.004)	(0.006)	(0.007)	(0.006)	(0.007)
# of total EAs between 3 & 6 km	-0.004	-0.008	-0.002	-0.004	0.000	-0.001	0.004	0.002
	(0.007)	(0.008)	(0.002)	(0.00216)*	(0.003)	(0.004)	(0.003)	(0.004)
Treated individual * # of treated EAs within 3 kilometers		0.003		0.005		-0.004		-0.007
		(0.021)		(0.004)		(0.011)		(0.012)
Treated individual * # of treated EAs between 3 and 6 kilometers		0.001		-0.006		0.013		0.009
		(0.040)		(0.008)		(0.022)		(0.023)
Treated individual * # of total EAs within 3 kilometers		-0.029		-0.014		-0.007		-0.004
		(0.026)		(0.00467)***		(0.015)		(0.015)
Treated individual * # of total EAs between3 and 6 kilometers		0.012		0.007		0.004		0.007
		(0.014)		(0.00276)**		(0.008)		(0.007)
Observations	2,579	2,579	2,612	2,612	2,649	2,649	2,650	2,650
R-squared	0.098	0.098	0.418	0.42	0.144	0.144	0.199	0.2

Regressions are OLS models using Round 3 data with robust standard errors clustered at the EA level. All regressions are weighted with both sampling and saturation weights to make the results representative of the target population in the study EAs. Baseline values of the following variables are included as controls in the regression analyses: age dummies, strata dummies, household asset index, highest grade attended, and an indicator for ever had sex. Parameter estimates statistically different than zero at 99 percent (***), 95 percent (**), and 90 percent (*) confidence.