Design of Partial Population Experiments
with an Application to Spillovers in Tax Compliance*

Guillermo Cruces, *U. of Nottingham & CEDLAS-UNLP*
Dario Tortarolo, *U. of Nottingham & IFS*
Gonzalo Vazquez-Bare, *UC Santa Barbara*

April 17, 2023

Abstract

We develop a framework to conduct experiments for estimating direct and spillover effects when units are grouped into mutually exclusive clusters. Crucially, our framework accounts for heterogeneous treatment effects across clusters and heterogeneous cluster sizes, which are pervasive in empirical settings but typically ignored in experimental design. We show that failing to account for cluster heterogeneity in experimental design can severely overestimate power and underestimate minimum detectable effects. We study the large-sample behavior of OLS estimators for direct and spillover effects with heterogeneous clusters and use our results to derive simple formulas to calculate power, minimum detectable effects and optimal cluster assignment probabilities. We also set up a potential outcomes framework that justifies interpreting OLS estimands as causal effects. We apply our methods to design a large-scale experiment to estimate the spillover effects of a communication campaign on property tax compliance. We find an increase in tax compliance among individuals directly targeted with our mailing, as well as compliance spillovers on untreated individuals in street blocks with a high proportion of treated taxpayers.

JEL Codes: H71 , H26 , H21 , O23.

Keywords: two-stage designs, partial population experiments, spillovers, randomized controlled trials, cluster experiments, property tax, tax compliance.

*We thank Yuehao Bai, Youssef Benzarti, Augustin Bergeron, Javier Birchenall, Matias Cattaneo, Kelsey Jack, Heather Royer, Doug Steigerwald and Alisa Tazhitdinova for valuable discussions and suggestions, and seminar participants at the 2021 National Tax Association conference, UCSB Applied Microeconomics Lunch, IFS, CEDLAS-UNLP, and the 2022 Advances with Field Experiments conference. We thank Julian Amendolaggine and Juan Luis Schiavoni for their invaluable support throughout the project. Corresponding author: Guillermo Cruces, E-mail: guillermo.cruces@nottingham.ac.uk. This project was reviewed and approved in advance by the Institutional Review Board at the University of Nottingham. The design for this experiment was preregistered in the AEA RCT Registry (RCT ID: **AEARCTR-0006569**). All remaining errors are our own.

1 Introduction

Randomized controlled trials have been extensively used in economics in recent years. A large fraction of these experiments are based on the assumption that the treatment assignment of one unit or subject does not influence the outcomes of others, which is plausible in numerous contexts. However, spillovers between experimental units can be expected in several cases, and this contaminates the estimation of treatment effects. Besides this potential contamination, identifying the presence of spillovers can be of interest in and of itself, as they provide valuable evidence on the nature and magnitude of interactions between subjects.

The presence of spillovers poses significant challenges and has important consequences for the design of randomized controlled trials (RCTs), ranging from the assessment of direct treatment effects to the accurate measurement of indirect effects. While the early experimental literature considered the impact on non-treated units in an ex-post manner (e.g. Miguel and Kremer, 2004), field experiments incorporating spillover effects into their design have become increasingly prevalent in applied research. In settings where units are grouped into independent clusters (such as schools, villages or firms), a common design is the assignment of clusters to different treatment intensities or "saturations", which are then compared to pure control clusters with no treated units (Moffit, 2001; Duflo and Saez, 2003; Hudgens and Halloran, 2008; Hirano and Hahn, 2010; Baird et al., 2018). These partial population experiments enable researchers to cleanly isolate direct treatment effects and to provide direct evidence on spillover effects. In this paper, we provide a framework that deals with the considerable challenges for experimental design posed by the relaxation of the stable unit treatment value assumption and the characteristics of clusters in real world settings. More specifically, our framework allows experimenters to account for cluster size heterogeneity, which is pervasive in empirical settings but overlooked in applied research and in the recent partial population design literature.

Cluster heterogeneity has several important practical implications for the design and analysis of experiments. First, when clusters are heterogeneous in size, variance formulas need an adjustment term that depends on the first and second moments of the cluster size distribution (Cameron and Miller, 2015). Ignoring this heterogeneity generally results in an underestimation of minimum detectable effects (MDEs) and thus an overestimation of statistical power. For instance, a numerical example based on typical cluster sizes from several existing partial population experiments indicates that standard errors and MDEs that impose homogeneous cluster size need to be adjusted by factors ranging from about 7% to about 20% to account for this heterogeneity (see Appendix Section C.2 for details). Failure to do so would result in under-powered studies. Second, cluster heterogeneity can affect the accuracy of the large sample normal approximation, and power calculations based on this approximation may be misleading when cluster sizes are very heterogeneous (Carter, Schnepel and Steigerwald, 2017; Djogbenou, MacKinnon and Ørregaard Nielsen, 2019; Hansen and Lee, 2019).

Third, in the presence of heterogeneous clusters, average treatment effects are likely to exhibit heterogeneity across clusters as well, which complicates the interpretation of frequently analyzed estimands such as differences in means and OLS coefficients.

To address these issues, our first contribution is to derive an asymptotic distributional approximation and variance formulas for OLS estimators of average outcomes and differences in means in a setting where (i) clusters are heterogeneous in size and (ii) the distribution of outcomes (and thus treatment effects) may vary across clusters. We consider a double-array asymptotic setting where cluster sizes are allowed—but not required—to grow with the sample size. We show that, under this type of heterogeneity, OLS estimators of average conditional outcomes estimate a weighted average of outcomes across clusters, where the weights depend on the relative cluster sizes and the within-cluster treatment probabilities.

Our second and main contribution for applied research is to derive a series of simple formulas to perform power calculations in the presence of cluster size heterogeneity. Using the variance of OLS estimators for power calculations in the presence of heterogeneous clusters has a clear practical disadvantage: it requires the experimenter to assign values to parameters that may differ across clusters. Specifically, the variance depends on the deviation of cluster-specific means from the average, which the experimenter may have little or no information about. To circumvent this limitation, we propose a simplified setting where clusters differ in size but not in their outcome distributions. We illustrate how our formulas can be readily implemented to conduct power and minimum detectable effects calculations. We also show how our formulas generalize those available in the existing methodological literature on experimental design (Duflo, Glennerster and Kremer, 2007; Hirano and Hahn, 2010; Baird et al., 2018) by allowing for multiple treatments, cluster size heterogeneity, heteroskedasticity and general forms of intracluster correlation in outcomes and treatments. Since partial population experiments are multi-arm experiments, optimal design requires choosing an optimality criterion that combines the variances of the multiple estimators. We provide a tractable closed form solution to the optimal choice problem under A-optimality—which minimizes the average variance of the estimators of interest—and discuss how alternative optimality criteria may be used in combination with our variance formulas using numerical methods. Our methodology allows researchers to set up a partial population experiment with cluster size heterogeneity with simple sample statistics. Our formulas provide the variances to be minimized and the criterion to minimize them so as to optimally assign clusters to the different (pre-defined) saturations.

Our third contribution is to set up a potential outcomes framework that allows for within-cluster spillovers, heterogeneous effects and varying cluster sizes. We use this framework to analyze the link between difference-in-means or OLS estimators and average potential outcomes. We provide a set of restrictions on the potential outcomes that reconciles the OLS estimands and average direct and spillover effects in partial population experiments.

Lastly, we apply our framework to design and conduct a large-scale field experiment to estimate

direct and spillover effects of a randomized communication campaign on property tax compliance. We conducted the experiment in a large municipality of Argentina where neighbors are required to pay a monthly bill on their real estate, locally known as *Tasa por Servicios Generales* (TSG), which accounts for most of the local own revenues in Argentine municipalities. Our campaign consisted of sending personalized letters to randomly selected dwellings with reminders about due taxes, information about the status of the account, due dates, past due debt and payment methods. While there is ample evidence on the effect of tax reminders on compliance and collection (Antinyan and Asatryan, 2019), our main research objective was to find evidence on relatively elusive spillover effects from information campaigns on tax collection. We designed the experiment based on our methodological results to maximize the likelihood of capturing spillover effects of our mailings on neighbors that live in the same street blocks of treated individuals (i.e., those who received letters from us) but that did not receive a letter. We included three different degrees of saturation, with city street blocks with no letters (pure controls), and blocks with 20%, 50% and 80% of treated individuals. Our results reveal higher payment rates for treated individuals, but also for their untreated neighbors in the same street block (compared to accounts in pure control blocks where no one received the information letter). Spillover effects are lower in magnitude but still substantial and precisely estimated in high-saturation street blocks (those with 80% treated accounts), especially when accounting for expected (pre-registered) heterogeneity in past compliance: payment rates of untreated accounts in high saturation blocks with above median past compliance increased by 2.6 percentage points, compared to direct effects of about 5.1 percentage points.

Comparison with current literature. Our paper contributes to a growing literature on experimental design (Duflo, Glennerster and Kremer, 2007; Bruhn and McKenzie, 2009; Bugni, Canay and Shaikh, 2018, 2019; Bai, 2022) and in particular to the literature on design and analysis of experiments under spillovers or interference (Hirano and Hahn, 2010; Athey, Eckles and Imbens, 2018; Baird et al., 2018; Basse, Feller and Toulis, 2019; Jiang, Imai and Malani, 2022; Puelz et al., 2022; Viviano, 2022; Leung, 2022). More specifically, our results generalize those of Hirano and Hahn (2010), Hudgens and Halloran (2008) and Baird et al. (2018) by allowing for general treatment assignment mechanisms, within-group heteroskedasticity and correlation structures, heterogeneity in cluster sizes and alternative optimality criteria for experimental design.

In related work, Athey, Eckles and Imbens (2018), Basse, Feller and Toulis (2019) and Puelz et al. (2022) derive randomization inference tests for a general class of null hypotheses under interference, and Jiang, Imai and Malani (2022) analyze two-stage completely randomized experiments and provide randomization-based variance estimators and sample size formulas. Our results complement this literature by considering different assignment mechanisms and by conducting super-population-based large-sample (instead of design-based) inference in a double array asymptotic framework. Our approach allows us to determine the effect of cluster size heterogeneity in the asymptotic behavior

of the treatment effect estimators.

We also contribute to a large empirical literature on property taxes and a small but growing empirical literature on spillover effects in tax compliance. On property taxes, recent contributions include Brockmeyer et al. (2020) study of Mexico City, Bergeron, Tourek and Weigel (2021) and Weigel (2020) for the Democratic Republic of Congo, and Krause (2020) for Haiti, among others. The latter two are randomized controlled trials, and in both cases the authors address the presence of spillovers, but in ex-post analysis rather than in the experimental designs. The effect of social interactions in tax compliance interventions has remained a relatively elusive issue in the broader experimental compliance literature. Some notable exceptions are Pomeranz (2015), who detects enforcement spillovers up the VAT chain in Chilean firms, Drago, Mengel and Traxler (2020) who study enforcement spillovers of TV licensing inspections on untreated households in Austria, and Boning et al. (2020) analyze direct and network effects from in-person visits by revenue officers on visited and non-visited firms in the United States (see the review in Pomeranz and Vila-Belda, 2019, for more studies covering spillover effects). In Argentina, a recent study by Carrillo, Castro and Scartascini (2021) finds neighborhood spillover effects from a program that randomly awarded 400 taxpayers with the repair of a sidewalk. Whereas these papers find spillover effects in tax compliance, their original experiments were not designed to capture these effects. We build on these pioneering works with an intervention (and its statistical power) designed with the purpose of capturing spillovers.

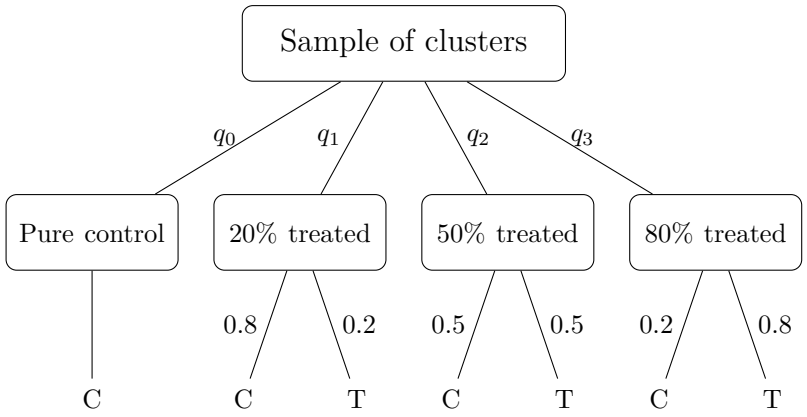
The paper is organized as follows. In Section 2, we set up our framework for partial population experiments and derive the main methodological results. In Section 3, we describe the large-scale randomized communication campaign, the administrative data used in the analysis, the empirical strategy, and evidence of direct and spillover effects. Section 4 provides some concluding remarks.

2 Partial Population Experiments

2.1 Setup

We consider the design of experiments in a sample where units are grouped into mutually exclusive and independent clusters, and where treatment assignments can vary both between and within clusters. We refer to these experiments as *partial population experiments* (Moffit, 2001). Common examples of this type of clustering are students in schools (Miguel and Kremer, 2004; Beuermann et al., 2015), family members in households (Barrera-Osorio et al., 2011; Foos and de Rooij, 2017), job seekers in local labor markets (Crépon et al., 2013), employees in firms or organizations (Duflo and Saez, 2003), or households or voters in villages or other geographic administrative units (Angelucci and De Giorgi, 2009; Ichino and Schündeln, 2012; Haushofer and Shapiro, 2016; Giné and Mansuri, 2018). In our application, a local property tax reminder information campaign, the

Figure 1: A Partial Population Design



population of interest consists of taxpayers in residential city street blocks, and the effect of interest is the impact of the campaign on payments by targeted individuals and the potential spillovers on the non-treated within blocks with different treatment intensities.

Formally, we consider a sample of observations that are divided into mutually independent clusters $g = 1, \ldots, G$, where each cluster g contains n_g observations $i = 1, \ldots, n_g$ and the total sample size is $n = \sum_{g=1}^{G} n_g$. We view cluster sizes as non-random (see Bugni et al., 2022, for a sampling approach in cluster RCTs where cluster sizes are seen as random). In a partial population experiment, clusters are randomly divided into categories or *saturations* denoted by $T_g \in \mathcal{T} = \{0, 1, 2, \ldots, M\}$ where by convention $T_g = 0$ denotes a pure control group and $\mathbb{P}[T_g = t] = q_t \in (0, 1)$. Within each group, a binary treatment D_{ig} is assigned at the individual level with probability $\mathbb{P}[D_{ig} = 1|T_g = t]$ and where $\mathbb{P}[D_{ig} = 0|T_g = 0] = 1$.[1] We also let $\mathbf{D}_g = (D_{1g}, D_{2g}, \ldots, D_{n_g g})'$ be the vector of unit-level treatment assignments in cluster g, $\mathbf{D} = (\mathbf{D}_1', \ldots, \mathbf{D}_G')'$ and $\mathbf{T} = (T_1, \ldots, T_G)'$. Figure 1 provides an example of a partial population design with four saturation levels. Notice that both standard RCTs with independent observations and cluster RCTs are particular cases of partial population experiments, as we show in Examples 1 and 2 below.

The observed outcome of interest for unit i in cluster g is denoted by Y_{ig} and let $\mathbf{Y}_g = (Y_{1g}, \ldots, Y_{n_g g})'$ be the vector of observed outcomes in cluster g. In partial population experiments, the estimands of interest are typically based on comparisons of average outcomes between treated or untreated units in treated clusters to pure control units, $\mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t] - \mathbb{E}[Y_{ig}|T_g = 0]$, pooled across clusters. We take these estimands as given in the first part of the paper since they are the most commonly analyzed in the empirical literature on partial population experiments. We then analyze their causal interpretation within a potential outcomes framework in Section 2.6. Let $\mu_g(d, t) = \mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t]$ be the conditional mean of the outcome in cluster g given

[1] In practice, some saturations may not be feasible for some cluster sizes, for example if $\mathbb{P}[D_{ig} = 1|T_g = t] = 0.5$ but n_g is odd, so the effective proportion of treated in this cluster may be slightly different from $\mathbb{P}[D_{ig} = 1|T_g = t]$. For the sake of brevity, we ignore this issue in the main paper, but we discuss how to adjust the proportions to account for non-integer numbers in Appendix D.1.

assignment (d,t). These magnitudes are often estimated as sample means:

$$\hat{\mu}(d,t) = \frac{\sum_{g=1}^{G} \mathbb{1}(T_g = t) \sum_{i=1}^{n_g} Y_{ig} \mathbb{1}(D_{ig} = d)}{\sum_{g=1}^{G} \mathbb{1}(T_g = t) \sum_{i=1}^{n_g} \mathbb{1}(D_{ig} = d)} = \frac{\sum_g \mathbb{1}_g^t N_g^d \bar{Y}_g^d}{\sum_g \mathbb{1}_g^t N_g^d} \tag{1}$$

where $\mathbb{1}_g^t = \mathbb{1}(T_g = t)$, $N_g^d = \sum_i \mathbb{1}(D_{ig} = d)$ and $\bar{Y}_g^d = \sum_i Y_{ig} \mathbb{1}(D_{ig} = d)/N_g^d$ defined whenever $N_g^d > 0$. These estimators can be computed by running an OLS regression of the outcome on a full set of indicators $(\mathbb{1}(T_g = t, D_{ig} = d))_{(d,t)}$ (without an intercept).

In a setting with homogeneous clusters where $\mu_g(d,t) = \mu(d,t)$ for all g, it is straightforward to show that $\mathbb{E}[\hat{\mu}(d,t)] = \mu(d,t)$, but this is not true in general when clusters are heterogeneous. In the next section we analyze the large-sample behavior of the OLS estimators in a setting with heterogeneous clusters.

2.2 Asymptotic Behavior of OLS Estimators

We now study the asymptotic distribution of the OLS estimators defined in Equation (1) and then apply our results to conduct power and MDE calculations. We start by providing a consistency result and a central limit theorem in a double-array asymptotic setting where both the number of groups and the group sizes grow with the sample size. The goal of letting $n_g \to \infty$ as $n \to \infty$ is to determine how large clusters can be relative to the total sample size to allow for valid inference based on the normal approximation. This type of approximation is more appropriate than the fixed cluster size approach when groups can be large and heterogeneous in size. The settings with bounded cluster sizes and/or equally-sized clusters are nested as a particular case of our analysis. We note that the number of parameters remains fixed in our setup (see Vazquez-Bare, 2022, for an alternative approach in which the number of parameters is allowed to grow with the sample size). We consider the following sampling scheme.

Assumption 1 (Sampling)

(i) $(\mathbf{Y}_g, \mathbf{D}_g, T_g)_{g=1}^G$ *are mutually independent across* g.

(ii) For each g *and for all* $i = 1, \ldots, n_g$, $\mathbb{E}[Y_{ig}^l | D_{ig} = d, T_g = t] = \mu_g^l(d,t)$ *for all* (d,t) *and for all* l *such that* $\mathbb{E}[|Y_{ig}|^l | D_{ig} = d, T_g = t] < \infty$.

(iii) For each g *and for all* $i = 1, \ldots, n_g$, $\mathbb{P}[D_{ig} = d | T_g = t] = p_g(d|t)$ *for all* d *and* t.

Part (i) states that clusters are mutually independent, a standard assumption in the clustering literature. Part (ii) states that average conditional outcomes are the same across units within a cluster. In what follows we define $\mu_g^l(d,t) = \mu_g(d,t)$ for $l = 1$ to reduce notation. Part (iii) states that the treatment probabilities are the same across units within a cluster.

Next, let $\mathbf{D}_{(i)g} = (D_{jg})_{j \neq i}$ denote the vector of observed treatments excluding unit i's and $\mathbf{D}_{(ij)g} = (D_{kg})_{k \neq i,j}$. We introduce the following restriction on conditional outcome moments.

Assumption 2 (Conditional moments) *For all i, j and g,*

(i) $\mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t, \mathbf{D}_{(i)g}] = \mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t]$

(ii) $\mathbb{E}[Y_{ig}Y_{jg}|D_{ig} = d, D_{jg} = d', T_g = t, \mathbf{D}_{(ij)g}] = \mathbb{E}[Y_{ig}Y_{jg}|D_{ig} = d, D_{jg} = d', T_g = t]$.

This assumption is a high-level condition stating that, conditional on own treatment assignment and the cluster-level assignment T_g, the first and second moments of Y_{ig} do not vary with the peers' treatment indicators.[2] Intuitively, this means that the assignment (D_{ig}, T_g) contains all the relevant variation in the outcome moments. In Section 2.6 we show that in a potential outcomes framework, this condition is guaranteed by assuming that peers are exchangeable, so that potential outcomes only depend on the proportion of treated peers, but not their identities.

Finally, we need to restrict cluster heterogeneity to ensure that the estimators are asymptotically normal. We do so in the following way.

Assumption 3 (Cluster heterogeneity and bounded moments)

(i) *For some* $2 \leq r < \infty$, *as* $n \to \infty$, $\max_g n_g^2/n \to 0$ *and* $\left(\sum_g n_g^r\right)^{2/r}/n \leq C < \infty$.

(ii) *For some* $l > r$, $\max_{g,d,t} \mathbb{E}[|Y_{ig}|^l \,|D_{ig} = d, T_g = t] < \infty$.

(iii) $\max_{g,d,t} \left|\mu_g(d,t) - \sum_g n_g \mu_g(d,t)/n\right| \leq \tilde{C} < \infty$.

Condition (i) is taken from Hansen and Lee (2019).[3] The first part in condition (i) ensures that the largest cluster is small relative to the total sample size, so no cluster dominates the sample. The second part of condition (i) is a regularity condition that rules out unbounded r-th moments of the distribution of cluster sizes. As an example, setting $r = 4$ restricts the fourth moment of the cluster size distribution, which rules out heavy tails. In practical terms, this highlights the importance of analyzing the distribution of group sizes when designing an experiment to verify that clusters are small relative to the total sample size. Condition (ii) is a standard regularity condition that ensures that the l-th conditional moment of the outcome is bounded for some $l > r \geq 2$. Part (iii) limits the amount of heterogeneity in average outcomes across clusters.

In what follows, we use "$\to_{\mathbb{P}}$" to denote convergence in probability, "$\to_{\mathcal{D}}$" denote convergence in distribution and define any generic $(2M-1)$-dimensional vector $\mathbf{v}$ as

$$\mathbf{v} = (v(d,t))'_{(d,t)} = (v(0,0), v(0,1), v(1,1), \ldots, v(0,M), v(1,M))'$$

[2] We refer to unit i's "peers" as all the units other than i in the same cluster.

[3] Notice that condition (i) holds automatically when group sizes are seen as fixed in the asymptotic analysis, which corresponds to the case of "many small clusters".

The following theorem characterizes the asymptotic distribution and variance of the OLS estimators in (1).

Theorem 1 *Consider the vector of estimators* $\hat{\boldsymbol{\mu}}_n = (\hat{\mu}(d,t))'_{(d,t)}$ *from* (1) *and define the vector*

$$\boldsymbol{\mu}_n^p = (\mu_n^p(d,t))'_{(d,t)}, \quad \mu_n^p(d,t) = \frac{\sum_g n_g p_g(d|t)\mu_g(d,t)}{\sum_g n_g p_g(d|t)}$$

where $\mu_g(d,t) = \mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t]$. *Let* $\psi_g = (\psi_g(d,t))'_{(d,t)}$ *where*

$$\psi_g(d,t) = \frac{\mathbb{1}_g^t N_g^d(\bar{Y}_g^d - \mu_g(d,t)) + \left(\mathbb{1}_g^t N_g^d - q_t n_g p_g(d|t)\right)(\mu_g(d,t) - \mu_n^p(d,t))}{q_t \sum_g n_g p_g(d|t)/n}$$

and

$$\mathbb{E}[\psi_g] = 0, \quad \Omega_n = \frac{1}{n}\sum_g \mathbb{E}[\psi_g \psi_g'].$$

Suppose that Assumptions 1 to 3 hold, and that:

(i) The minimum eigenvalue of Ω_n *is bounded away from 0,*

(ii) For any (d,t) *such that* $p_g(d|t) > 0$ *for some* g, $\sum_g n_g p_g(d|t)/n \geq c > 0$.

Then $\|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_n^p\| \to_{\mathbb{P}} 0$ *and*

$$\Omega_n^{-1/2}\sqrt{n}\,(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_n^p) = \Omega_n^{-1/2}\frac{1}{\sqrt{n}}\sum_g \psi_g + o_{\mathbb{P}}(1) \to_{\mathcal{D}} \mathcal{N}(\mathbf{0}, I_{2M-1})$$

where I_{2M-1} *is a* $(2M-1)$*-dimensional identity matrix.*

We provide a proof of this result and the explicit formula for the variance matrix Ω_n in Appendix E to conserve space. We briefly note that condition (i) is a standard invertibility condition for the variance matrix, and condition (ii) puts a lower bound on the probabilities that units are observed across the sample for each assignment (d,t), excluding assignments where probabilities are zero by design (since for example $p_g(1|0) = 0$ in pure control groups). There are two main practical implications of Theorem 1. First, each sample mean $\hat{\mu}(d,t)$ estimates a weighted average of cluster-specific means $\mu_g(d,t)$, where the weights depend on the cluster size n_g and the within-cluster probability of treatment $p_g(d|t)$. Second, the distribution of $\hat{\boldsymbol{\mu}}_n$ can be approximated as

$$\hat{\boldsymbol{\mu}}_n \overset{a}{\sim} \mathcal{N}\left(\boldsymbol{\mu}_n^p, \frac{\Omega_n}{n}\right)$$

where the variance matrix Ω_n allows for heterogeneity in cluster sizes and potential outcomes moments, heteroskedasticity, different treatment assignment probabilities across clusters and intra-cluster correlation in both outcomes and unit-level treatment assignments.

2.3 Estimating Differences in Means

The previous section shows that with heterogeneous clusters sample means estimate a weighted average of cluster-specific means. This implies that differences-in-means estimators will generally not recover weighted averages of population differences in means. More precisely, $\hat{\mu}(d,t) - \hat{\mu}(d',t')$ estimates:

$$\mu_n^p(d,t) - \mu_n^p(d',t,) = \frac{\sum_g n_g p_g(d|t)\mu_g(d,t)}{\sum_g n_g p_g(d|t)} - \frac{\sum_g n_g p_g(d'|t')\mu_g(d',t')}{\sum_g n_g p_g(d'|t')}$$

which is not a weighted average of differences in average outcomes because treatment probabilities may differ across clusters. In the particular case in which the treatment probabilities do not vary across clusters, $p_g(d|t) = p(d|t)$ for all g,

$$\mu_n^p(d,t) - \mu_n^p(d',t') = \sum_g \frac{n_g}{n}(\mu_g(d,t) - \mu_g(d',t'))$$

which is a proper weighted average of differences between average outcomes weighted by the relative cluster sizes n_g/n and independent of the treatment probabilities $p(d|t)$. Thus, in settings with heterogeneous clusters, the experimenter may prefer designs in which the within-cluster treatment probabilities do not vary across clusters with the same assignment $T_g = t$.

2.4 Power and MDE Calculations

The distributional approximation and variance formulas in Theorem 1 can be applied to conduct power and minimum detectable effects calculations based on the vector of estimators $\hat{\boldsymbol{\mu}}_n$, subvectors, linear combinations (such as the pooled and slope effects proposed by Baird et al., 2018) or nonlinear functions thereof, applying the delta method when needed. For instance, the difference in means $\beta(d,t) = \mu_n(d,t) - \mu_n(0,0)$, can be estimated as $\hat{\beta}(d,t) = \hat{\mu}(d,t) - \hat{\mu}(0,0)$, and using Theorem 1, the power function for a two-sided hypothesis test of no effect $\mathsf{H_0} : \beta(d,t) = 0$ is:

$$\Gamma(\beta(d,t)) \approx 1 - \Phi\left(\frac{\beta(d,t)}{\sqrt{\mathbb{V}[\hat{\beta}(d,t)]}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\beta(d,t)}{\sqrt{\mathbb{V}[\hat{\beta}(d,t)]}} - z_{1-\alpha/2}\right) \tag{2}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile from the standard normal distribution and where the variance $\mathbb{V}[\hat{\beta}(d,t)] = \mathbb{V}[\hat{\mu}(d,t)] + \mathbb{V}[\hat{\mu}(0,0)] - 2\mathbb{C}\text{ov}(\hat{\mu}(d,t), \hat{\mu}(0,0))$ is approximated using the variance matrix Ω_n after imputing the unknown parameters (such as outcome moments and intracluster correlations) and chosen assignment probabilities.

One practical disadvantage of the variance Ω_n from Theorem 1 for power calculations is that it requires assigning values to parameters that may differ across clusters, and in particular it depends on the deviation of cluster-specific means from the average $\mu_g(d,s) - \mu_n^p(d,t)$, for which the exper-

imenter may have little to no information. This limitation may be circumvented by assuming that outcome moments are homogeneous across clusters, as formalized in the following assumption.

Assumption 4 (Outcome Moments Homogeneity) $\mathbb{E}[Y_{ig}^l|D_{ig} = d, T_g = t] = \mu^l(d,t)$ *and* $\mathbb{E}[Y_{ig}^l Y_{jg}^l|D_{ig} = d, D_{jg} = d', T_g = t] = \tilde{c}^l(d,d',t)$ *for all* g, (d,d',t) *and for the value of* l *defined in Assumption 3.*

As before, we write $\mu^1(d,t) = \mu(d,t)$ to reduce notation. Under this additional assumption we obtain the following result.

Theorem 2 *Consider the vector of estimators* $\hat{\boldsymbol{\mu}}_n = (\hat{\mu}(d,t))'_{(d,t)}$ *from (1) and define the vector*

$$\boldsymbol{\mu} = (\mu(d,t))'_{(d,t)}, \quad \mu(d,t) = \mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t].$$

Let $\psi_g = (\psi_g(d,t))'_{(d,t)}$ *where*

$$\widetilde{\psi}_g(d,t) = \frac{\mathbb{1}_g^t N_g^d(\bar{Y}_g^d - \mu(d,t))}{q_t \sum_g n_g p_g(d|t)/n}, \quad \mathbb{E}[\widetilde{\psi}_g] = 0, \quad \widetilde{\Omega}_n = \frac{1}{n}\sum_g \mathbb{E}[\widetilde{\psi}_g \widetilde{\psi}_g'].$$

Suppose that Assumptions 1 to 4 hold, and that

(i) The minimum eigenvalue of $\widetilde{\Omega}_n$ *is bounded away from 0,*

(ii) For any (d,t) *such that* $p_g(d|t) > 0$ *for some* g, $\sum_g n_g p_g(d|t)/n \geq c > 0$.

Then $\hat{\boldsymbol{\mu}}_n \to_{\mathbb{P}} \boldsymbol{\mu}$ *and*

$$\widetilde{\Omega}_n^{-1/2}\sqrt{n}\,(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) = \widetilde{\Omega}_n^{-1/2}\frac{1}{\sqrt{n}}\sum_g \widetilde{\psi}_g + o_{\mathbb{P}}(1) \to_{\mathcal{D}} \mathcal{N}(\mathbf{0}, I_{2M-1})$$

where I_{2M-1} *is a* $(2M-1)$*-dimensional identity matrix,*

$$\begin{aligned}
\frac{1}{n}\sum_g \mathbb{E}[\widetilde{\psi}_g(d,t)^2] &= \frac{n\sigma^2(d,t)}{q_t \sum_g n_g p_g(d|t)}\left\{1 + \rho(d,t)\frac{\sum_g n_g(n_g-1)p_g(d,d|t)}{\sum_g n_g p_g(d|t)}\right\}, \quad t > 0, \\
\frac{1}{n}\sum_g \mathbb{E}[\widetilde{\psi}_g(0,0)^2] &= \frac{\sigma^2(0,0)}{q_0}\left\{1 + \rho(0,0)\left(\frac{\sum_g n_g^2}{n} - 1\right)\right\}, \\
\frac{1}{n}\sum_g \mathbb{E}[\widetilde{\psi}_g(0,t)\widetilde{\psi}(1,t)] &= n\sigma(0,t)\sigma(1,t)\rho(0,1,t)\frac{\sum_g n_g(n_g-1)p_g(0,1|t)}{\sum_g n_g p_g(0|t)\sum_g n_g p_g(1|t)}, \quad t > 0, \\
\frac{1}{n}\sum_g \mathbb{E}[\widetilde{\psi}_g(d,t)\widetilde{\psi}(d',t')] &= 0, \quad t \neq t'
\end{aligned}$$

and where $\sigma^2(d,t) = \mathbb{V}[Y_{ig}|D_{ig} = d, T_g = t]$, $\rho(d,t) = \mathrm{cor}(Y_{ig}, Y_{ig}|D_{ig} = d, D_{jg} = d, T_g = t)$, $p_g(d,d'|t) = \mathbb{P}[D_{ig} = d, D_{jg} = 1|T_g = t]$, *and* $\rho(0,1,t) = \mathbb{C}\mathrm{ov}(Y_{ig}, Y_{ig}|D_{ig} = 0, D_{jg} = 1, T_g = t)$.

Theorem 2 provides a simplified formula that allows for heterogeneity in cluster sizes and within cluster probabilities, heteroskedasticity across treatment assignments and intracluster correlation in conditional outcomes and treatments, but does not depend on cluster-specific average outcomes like the variance formula in Theorem 1 and can be readily used to conduct power and MDE calculations for direct and spillover effects. For example, one can estimate the effect of assignment (d,t) through

$$\hat{\beta}(d,t) = \hat{\mu}(d,t) - \hat{\mu}(0,0)$$

and approximate the variance as

$$\begin{aligned}\mathbb{V}[\hat{\beta}(d,t)] \approx & \frac{\sigma^2(d,t)}{q_t \sum_g n_g p_g(d|t)} \left\{1 + \rho(d,t) \frac{\sum_g n_g(n_g-1)p_g(d,d|t)}{\sum_g n_g p_g(d|t)}\right\} \\ & + \frac{\sigma^2(0,0)}{nq_0} \left\{1 + \rho(0,0)\left(\frac{\sum_g n_g^2}{n} - 1\right)\right\}\end{aligned} \tag{3}$$

which only depends on the variance and conditional intracluster correlation in outcomes (as in any standard power calculation), the assignment probabilities, which are chosen by the experimenter, and the sample distribution of cluster sizes, which is observable. This variance can be fed into the power formula in Equation (2) to calculate power or MDEs. The following examples show how our general formula simplifies to the ones proposed in the literature under further assumptions.

Example 1 (Standard RCT with a binary treatment) *Suppose that each cluster has only one unit ($n_g = 1$), and there are only two saturations so that each (single-unit) cluster is assigned to treatment or control with probability q and $1-q$ respectively. In this case, $q_t = q$, $q_0 = 1-q$, $\sum_g n_g p_g(1|1) = n$ and thus*

$$\mathbb{V}[\hat{\beta}(1,1)] \approx \frac{\sigma^2(1,1)}{nq} + \frac{\sigma^2(0,0)}{n(1-q)}.$$

In addition, under a homoskedasticity assumption $\sigma^2(1,1) = \sigma^2(0,0) = \sigma^2$ we get:

$$\mathbb{V}[\hat{\beta}(1,1)] \approx \frac{\sigma^2}{nq(1-q)}$$

which is Equation (6) in Duflo, Glennerster and Kremer (2007).

Example 2 (Cluster RCT) *Suppose that clusters are assigned to two saturations $T_g \in \{0,1\}$ and that all units within the same cluster receive the same treatment. In this case, $p_g(1|1) = p_g(1,1,|t) = 1$ and thus*

$$\mathbb{V}[\hat{\beta}(1,1)] \approx \frac{\sigma^2(1,1)}{nq}\left\{1 + \rho(1,1)\frac{\sum_g n_g(n_g-1)}{n}\right\} + \frac{\sigma^2(0,0)}{n(1-q)}\left\{1 + \rho(0,0)\left(\frac{\sum_g n_g^2}{n} - 1\right)\right\}.$$

In addition, suppose that clusters are equally-sized, $n_g = \bar{n}$, and assume a random effects structure

so that $\sigma^2(1,1) = \sigma^2(0,0) = \sigma^2 + \tau^2$ *and* $\rho(1,1) = \rho(0,0) = \tau^2/(\sigma^2+\tau^2)$. *In this case,*

$$\mathbb{V}[\hat{\beta}(1,1)] \approx \frac{1}{q(1-q)} \cdot \frac{\bar{n}\tau^2 + \sigma^2}{G\bar{n}}$$

which is Equation (9) in Duflo, Glennerster and Kremer (2007).

Example 3 (Homoskedastic case with two treatment saturations) *Suppose there is only one treatment intensity and a pure control category, so that* $M = 1$, *as in Duflo and Saez (2003). Let* $q = \mathbb{P}[T_g = 1]$ *and* $p = \mathbb{P}[D_{ig} = 1 | T_g = 1]$. *Assume that* $\sigma^2(d,t) = 1$ *and* $\rho(d,t) = 0$ *for all* (d,t). *In this case, for assignment* $(d,t) = (0,1)$, *Equation (3) simplifies to:*

$$\mathbb{V}[\hat{\beta}(0,1)] \approx \frac{1-pq}{(1-p)q(1-q)}$$

which corresponds to the variance formula in Hirano and Hahn (2010).

Example 4 (Random effects structure with equally-sized clusters) *Suppose that clusters are equally sized,* $n_g = \bar{n}$ *for all* g, *and consider a random effects covariance structure so that* $\sigma^2(d,t) = \sigma^2 + \tau^2$, $\rho(d,t) = \tau^2$ *for all* (d,t). *In addition, suppose that the within-cluster assignment given* $T_g = t$ *sets a fixed number of treated units* $\bar{n}p_t$ *in each cluster, which implies that* $\mathbb{P}[D_{ig} = 1, D_{jg} = 1 | T_g = t] = p_t(\bar{n}p_t - 1)/(\bar{n} - 1)$. *In this case, for assignment* $(1,t)$, *Equation (3) becomes:*

$$\mathbb{V}[\hat{\beta}(1,t)] \approx \frac{\sigma^2+\tau^2}{\bar{n}G}\left\{\bar{n}\rho\left(\frac{1}{q_t} + \frac{1}{q_0}\right) + (1-\rho)\left(\frac{1}{p_t q_t} + \frac{1}{q_0}\right)\right\}$$

which corresponds to Equation (3) in Baird et al. (2018).

2.5 Optimal Design

In partial population designs, the experimenter faces three choices: the number of saturations, M, the probability of each saturation $q_t = \mathbb{P}[T_g = t]$ and the within-cluster treatment probabilities $p_g(d|t)$.

The choice of the number of saturations M depends on the assumptions that the researcher is willing to make about how the average outcomes $\mathbb{E}[Y_{ig} | D_{ig} = d, T_g = t]$ vary as a function of t. If this function is linear (or close to linear), two saturations would be enough to identify the shape of this function, whereas if the function can be approximated by a quadratic function one would need three saturations, and so on. As the number of saturations M increases, the shape of the function that can be identified becomes more flexible at the expense of noisier estimation due to smaller sample sizes in each saturation. We do not discuss the choice of M and the within-cluster probabilities, as these are determined by the parameters that the researcher wants to identify, which in turn depend on the shape of the unknown average outcome function.

Given M and the within-group treatment probabilities, optimally choosing the cluster-level assignment probabilities $\{q_t\}_{t=0}^M$ requires defining an optimality criterion that determines how the variances of all the estimators of interest are aggregated. The literature on optimal design of experiments has proposed several criteria (see e.g. Silvey, 1980; Melas, 2006; Berger and Wong, 2009). We now consider *A-optimality*, which minimizes the trace of the variance-covariance matrix of the treatment effect estimators $(\hat{\beta}(d,t))_{(d,t>0)}$ (or equivalently, the average of the asymptotic variances).[4] The justification of this criterion is that the trace of the variance-covariance matrix can be seen as a measure of the size of the confidence ellipsoid (i.e. the multidimensional confidence interval) for the vector of parameters of interest. One advantage of A-optimality is its tractability: it has a simple closed-form solution in this setting, as shown in the following result.

Theorem 3 *Consider the optimal design problem:*

$$\min_{q_0,q_1,\ldots,q_M} \sum_{t=1}^{M} \left\{ \mathbb{V}[\hat{\beta}(0,t)] + \mathbb{V}[\hat{\beta}(1,t)] \right\}$$

with $q_t > 0$, $\sum_{t=0}^M q_t = 1$ *using the variance formula in Equation* (3). *The optimal assignment probabilities are given by:*

$$q_0^* = \frac{\sqrt{2MB_0}}{\sqrt{2MB_0} + \sum\limits_{t>0} \sqrt{B_t}}, \quad q_t^* = \frac{\sqrt{B_t}}{\sqrt{2MB_0} + \sum\limits_{t>0} \sqrt{B_t}}, \quad t > 0,$$

where

$$B_0 = \frac{\sigma^2(0,0)}{n} \left\{ 1 + \rho(0,0) \left(\frac{\sum_g n_g^2}{n} - 1 \right) \right\}$$

and for $t > 0$

$$\begin{aligned} B_t = &\frac{\sigma^2(1,t)}{\sum_g n_g p_g(1|t)} \left\{ 1 + \rho(1,t) \frac{\sum_g n_g(n_g-1)p_g(1,1|t)}{\sum_g n_g p_g(1|t)} \right\} \\ &+ \frac{\sigma^2(0,t)}{\sum_g n_g p_g(0|t)} \left\{ 1 + \rho(0,t) \frac{\sum_g n_g(n_g-1)p_g(0,0|t)}{\sum_g n_g p_g(0|t)} \right\}. \end{aligned}$$

Theorem 3 complements Theorem 2 and completes our main applied result by providing a tractable optimality criterion for an experimenter that wishes to find optimal assignment probabilities for cluster saturations in the context of a partial population experiment. In a nutshell, Theorem 2 indicates how to compute the variances to be minimized from observable and experimenter-selected parameters, whereas Theorem 3 provides a minimization criterion for these variances with a closed-form solution. A researcher wishing to set up a partial population experiment needs to optimally

[4]Notice that this criterion is different from the one in Baird et al. (2018), who minimize the average standard error. We propose this alternative method as it is more in line with the theoretical literature on experimental design, while also allowing for a simple, closed-form solution to the optimal design problem.

assign clusters to the different (pre-defined) saturations.

Alternative optimality criteria. While A-optimality has the advantage of a simple closed form solution, there are many other optimality criteria that may be preferable in different settings. Optimization problems with these alternative criteria do not have closed form solutions in general, but can be solved numerically based on our variance formula. See Silvey (1980), Melas (2006) and Berger and Wong (2009) for further details and discussions.

It should be noted that researchers may often need to incorporate different sets of constrains (such as logistical, budgetary, political or administrative constrains) when choosing assignment probabilities. These restrictions can be incorporated when choosing q_t, either directly into the optimization problem in Theorem 3 or on a case-specific basis. For example, in the experiment we describe in the next section, the total number of treated units was set by the government agency. We set up a system of equations incorporating this restriction in a "minimax-like" approach to control the variance of the smallest treatment cells (i.e. the noisiest estimators). Section 3.3 and Appendix B.1 provide a detailed step by step description on how we applied Theorems 2 and 3 and these additional constraints in the context of our tax compliance information campaign.

2.6 Potential Outcomes Framework

In this section we introduce a potential outcomes framework to study the causal interpretation of the OLS estimands discussed in the previous sections. Let $Y_{ig}(d, \mathbf{d}_g, t)$ denote unit i's (random) potential outcomes where d denotes own treatment, $\mathbf{d}_g \in \{0,1\}^{n_g-1}$ is a vector denoting unit i's peers' treatments and t denotes the cluster-level assignment. To be able to compare outcomes across clusters, our first assumption is an exclusion restriction stating that the cluster level assignment t does not directly affect potential outcomes. This assumption is likely to hold when the cluster level assignment variable T_g is an external randomization device drawn by the experimenter.

Assumption 5 (Exclusion restriction) $Y_{ig}(d, \mathbf{d}_g, t) = Y_{ig}(d, \mathbf{d}_g)$ *for all* $(d, \mathbf{d}_g, t)$.

While this assumption is required to identify treatment effects using variation across clusters, to our knowledge we are the first to make it explicit. In this setup, $Y_{ig}(1, \mathbf{d}_g) - Y_{ig}(0, \mathbf{d}_g)$ is the direct effect of the treatment on unit i in cluster g, $Y_{ig}(0, \mathbf{d}_g) - Y_{ig}(0, \tilde{\mathbf{d}}_g)$ is the spillover effect on an untreated unit and $Y_{ig}(1, \mathbf{d}_g) - Y_{ig}(1, \tilde{\mathbf{d}}_g)$ is the spillover effect on a treated unit. This potential outcomes structure allows for within-cluster spillovers, an assumption often known as *stratified interference* (Hudgens and Halloran, 2008). The observed outcome of interest for unit i in cluster g is denoted by $Y_{ig} = \sum_{d,\mathbf{d}_g} Y_{ig}(d, \mathbf{d}_g) \mathbb{1}(\mathbf{D}_g = (d, \mathbf{d}_g))$.

Next, we assume that the vector of treatment assignments $(\mathbf{D}_g, T_g)$ is independent of the vector of potential outcomes, which is guaranteed by random assignment of the treatment.

Assumption 6 (Independence) $(Y_{ig}(d, \mathbf{d}_g))_{(d,\mathbf{d}_g)} \perp\!\!\!\perp (\mathbf{D}_g, T_g)$.

Finally, we assume that peers are exchangeable, so that potential outcomes depend on the proportion of treated peers, but not on their identities. This assumption reduces the dimensionality of potential outcomes and is ubiquitous when analyzing spillovers (see Vazquez-Bare, 2022, and references therein for further discussion). In what follows let $\mathbf{1}_g$ be an $(n_g - 1)$-dimensional column vector of ones.

Assumption 7 (Exchangeability) *For all* $\mathbf{d}_g$, $Y_{ig}(d, \mathbf{d}_g) = Y_{ig}(d, s_g)$ *where* $s_g = \mathbf{1}_g'\mathbf{d}_g/(n_g - 1)$ *is the proportion of unit* i*'s treated peers.*

The following result links moments of observed outcomes to average potential outcomes.

Lemma 1 *Let* $N_g^1 = \sum_{i=1}^{n_g} D_{ig}$ *be the total number of treated units in cluster* g*. Suppose that Assumptions 5 to 7 hold and that* N_g^1 *is nonrandom conditional on* T_g*, with* $\mathbb{P}[N_g^1 = n_g p_g(d|t)|T_g = t] = 1$*. Then Assumption 2 holds and*

$$\mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t] = \mathbb{E}[Y_{ig}(d, s_g(d|t))], \quad s_g(d|t) = \frac{n_g p_g(d|t) - d}{n_g - 1}.$$

Lemma 1 provides conditions under which the conditional mean $\mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t]$ equals an average potential outcome for a given proportion of treated peers. Importantly, this result shows that the proportion of treated peers depends on the cluster size n_g. Thus, when cluster sizes are heterogeneous, average potential outcomes may vary across clusters, even within clusters with the same assignment $T_g = t$. To see this, consider the following example. Suppose there are two cluster sizes, $n_g = 6$ and $n_g = 10$, and consider clusters with $p_g(1|t) = 0.5$ so that half the units are assigned to treatment. In clusters with $n_g = 6$, the total number of treated units will be 3 and thus the proportion of treated peers is $3/5 = 0.6$ for untreated units and $2/5 = 0.4$ for treated units. On the other hand, in clusters with $n_g = 10$ there will be 5 treated units and thus the proportion of treated peers is $5/9 \approx 0.55$ for untreated units and $4/9 \approx 0.44$ for treated units. Hence, average potential outcomes may be different across units in clusters with different sizes, even under the same treatment saturation T_g. To be able to apply the variance and power formulas from Theorem 2 while interpreting OLS estimands as average potential outcomes, we introduce the following assumption.

Assumption 8 (Potential Outcomes Homogeneity) *For* l *as defined in Assumption 3,*

(i) $\mathbb{E}[Y_{ig}^l(d, s)] = \tilde{\mu}^l(d, s)$ *for all* g *and* (d, s).

(ii) For each (d, t) *there exists an* $s(d|t)$ *such that* $\max_g \left|\tilde{\mu}^l(d, s_g(d|t)) - \tilde{\mu}^l(d, s(d|t))\right| = 0$ *where* $s_g(d|t) = (n_g p_g(d|t) - d)/(n_g - 1)$.

Part (i) states that, for a given (d, s), potential outcome moments do not vary across clusters. Notice that this condition is not enough to eliminate heterogeneity in average potential outcomes, since as mentioned previously, when cluster sizes vary, average potential outcomes are evaluated at different values of s across differently-sized clusters with the same assignment. For this reason, part (ii) ensures that $\tilde{\mu}^l(d, s)$ is invariant to the perturbations in s generated by the variation in cluster sizes. This condition requires $\tilde{\mu}^l(d, s)$ to be locally flat in a window around $s(d|t)$ for each (d, t). While this second condition may be unlikely to hold exactly in practice, it can be a reasonable approximation when $s_g(d|t)$ shows little variation across g for each (d, t) (which happens for example when clusters are not very small) and/or the function $\tilde{\mu}^l(d, s)$ does not change abruptly under small perturbations in s. This condition can also be relaxed to hold approximately in the sense that $\max_g |\tilde{\mu}(d, s_g(d|t)) - \tilde{\mu}(d, s(d|t))| < \varepsilon$ for some small $\varepsilon > 0$. Under Assumptions 5 to 8, Lemma 1 implies that Assumption 4 holds and thus the results in Theorem 2 can be applied to conduct power calculations for direct and spillover effects while linking OLS estimands to average potential outcomes.

3 A Randomized Tax Communication Campaign

3.1 Background

As discussed in the introduction, there is a large body of evidence on nudges and tax compliance (Antinyan and Asatryan, 2019), but there is relatively scant evidence on the social interactions behind these interventions. We designed and implemented a public policy intervention based on the framework presented in the previous section to illustrate its potential to capture the presence of social effects in tax compliance – establishing credible evidence on these effects was our second research question.

Our randomized controlled trial was designed as a partial population experiment with the purpose of estimating the direct and spillover effects of a personalized communication campaign on property tax compliance. The intervention took place in a large municipality of Argentina where neighbors are billed and required to pay a municipal property tax on a monthly basis (the *Tasa por Servicios Generales*). The experiment consisted of a two-level randomized communication campaign where we sent a one-page personalized letter with information on the current billing period, past due debt, and how to pay online or in person.[5]

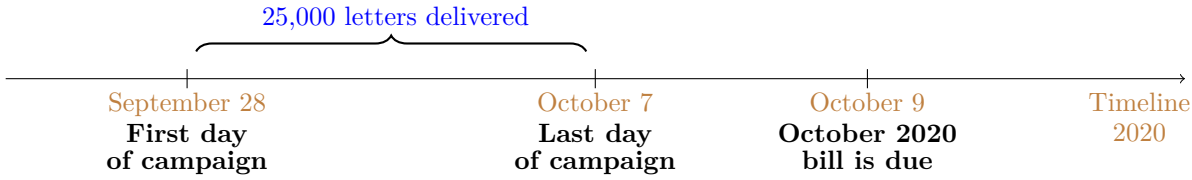
Randomization took place in two stages—first at the city street block level, and then at the taxpayer account (i.e., property) level. In the first stage, we randomly divided our sample of 3,982

[5]Figure A.1 in the appendix provides an anonymized example of the intervention letter. Our simple design emphasized action-relevant information, in accordance with De Neve et al. (2021) who show that simplified tax letters are an effective way to increases tax compliance.

blocks (clusters) into four groups with different intensity of treatment: (1) pure control blocks where no accounts were notified, (2) blocks with 20% of the accounts treated, (3) blocks with 50% of the accounts treated, and (4) blocks with 80% of the accounts treated. These different treatment intensities were designed to capture whether spillovers depend on the saturation of our information campaign at the city street block level (namely, low, medium and high saturation levels).[6] In the second stage, we randomly selected accounts within the last three groups of blocks to receive the treatment letter. The experiment was run on the universe of residential dwellings present in the municipality in 2019.

The timeline of the intervention is displayed below. We sent approximately 25,000 treatment letters to account holders who are billed the *Tasa por Servicios Generales*. The letters were delivered between September 28th and October 7th, 2020, corresponding to payments due on October 9th, 2020 as well as past due debt (if any). Direct effects of the campaign are captured by the difference in outcomes among individuals targeted by the intervention (treated) compared to those in pure control blocks. To study spillover or indirect effects, we compare the payment behavior of non-targeted neighbors within treated blocks (untreated) relative to those in pure control blocks.

Figure 2: Timeline of the randomized communication campaign



3.2 Administrative Data

For the empirical analysis, we use a combination of administrative databases provided by the revenue agency of the municipality where the experiment took place. The main database is constructed from the monthly bills issued to account holders between January 2018 and December 2020. The unit of observation is an account (*cuenta*), which coincides with a dwelling unit. The data contain the following billing details and demographic characteristics of the account holder (*titular*): account number (unique ID), address, block number, name of locality (neighborhood), year and month of the bill (12 bills per year), monthly fee (in pesos), paid fee (amount in pesos), due date, date of payment, days overdue, means of payment (cash or electronic), type of account (residential, retail store, factory), gender of the account holder, age of the account holder, linear front meters of the lot/property, assessed value of the property.

The municipality required us to target city street blocks with 8 to 50 accounts. Figure 3 shows

[6] The choice $p_1 = 20\%$, $p_2 = 50\%$ and $p_3 = 80\%$ attempts to balance parsimony with flexibility to detect nonlinearities in total and spillover effects without having to estimate too many parameters. See Section 2.5 for further discussion.

the distribution of accounts per block. Table 2 shows some descriptive statistics for the year 2019. Our sample size consists of 68,808 accounts distributed in 3,982 blocks. The frequency of payments is highly polarized. About 45 percent of the accounts paid the twelve 2019 monthly bills and about 35 percent did not pay any bill at all.[7] We call these two core groups *always payers* and *never payers*, respectively. The proportion of always payers is relatively low (45 percent) and, therefore, leaves room for potential behavioral responses from non-compliant and partially-compliant neighbors, and this was compounded by the context of the pandemic, during which lockdown measures reduced payments even from highly compliant individuals.

Baseline data. For the randomization, power calculations, and simulations, we use baseline data from the year 2019. We rely on three different pre-treatment outcomes: (i) an indicator equal to 1 if the account paid the twelve monthly bills of 2019, (ii) an indicator equal to 1 if the account paid at least one bill in 2019, and (iii) an indicator equal to 1 if the account paid six bills or more in 2019.

3.3 Experimental Design

Following the notation from Section 2, let n_g indicate the number of units (accounts - *cuentas*) per group (block - *cuadra*) with $g = 1, \ldots, G$ and let $n = \sum_g n_g$ be the total sample size. The cluster-level (block) treatment indicator is denoted by $T_g \in \{0, 1, 2, 3\}$ with distribution $\mathbb{P}[T_g = t] = q_t$ for $t = 0, 1, 2, 3$ where $T_g = 0$ indicates the pure control group, $T_g = 1$ indicates the groups with 20% treated, $T_g = 2$ indicates groups with 50% treated, and $T_g = 3$ indicates groups with 80% treated. The unit-level (account) treatment indicator is $D_{ig} \in \{0, 1\}$.

The municipality requested that the total number of letters sent be around 25,000 for budgetary reasons, and logistics implied that we set $L = 25,061$. To incorporate this constraint into the choice of the saturation probabilities q_t, we set up a system of equations as follows. The expected number of treated units is $n_1 = n(0.2q_1 + 0.5q_2 + 0.8q_3)$. Since the assignments $T_g = 1$ and $T_g = 3$ can be seen as symmetric, we set $q_1 = q_3$. Finally, we add an equation that ensures that the variance of the effect at 50% saturation is equal to the variance for the "small" cells (untreated units in 80% groups and treated units in 20% groups), so that $\mathbb{V}[\hat{\beta}(d, 2)] = \mathbb{V}[\hat{\beta}(0, 3)] = \mathbb{V}[\hat{\beta}(1, 1)]$. This gives a fourth equation of the form $q_2 = Rq_3$ where R depends on the intracluster correlation and the

[7] For the full distribution, see Figure A.5.

variance of the outcomes. Our system of equations is therefore:

$$\begin{aligned} 1 &= q_0 + q_1 + q_2 + q_3 \\ L &= n(0.2q_1 + 0.5q_2 + 0.8q_3) \\ q_1 &= q_3 \\ q2 &= Rq_3. \end{aligned}$$

We use the results in Theorem 2 to approximate the variances and calculate the ratio R. We provide further details in Appendix B.1. After finding the cluster assignment probabilities, our resulting sample sizes are shown in Table 1.

Table 1: Sample sizes

		Blocks	Control Obs	Treated Obs
$T_g = 0$	Pure control	1,102	19,103	0
$T_g = 1$	20% treated	1,099	15,060	3,853
$T_g = 2$	50% treated	680	5,905	5,897
$T_g = 3$	80% treated	1,100	3,677	15,311
Total		3,981	43,745	25,061

We jointly estimate the direct and spillover effects through the following saturated OLS regression:

$$Y_{ig} = \alpha + \sum_{t=1}^{3} \beta(0,t)\mathbb{1}(T_g = t)(1 - D_{ig}) + \sum_{t=1}^{3} \beta(1,t)\mathbb{1}(T_g = t)D_{ig} + \varepsilon_{ig} \tag{4}$$

where we allow ε_{ig} to be correlated within blocks and use a cluster-robust variance estimator. We use the power function formula (2) to calculate MDEs for each estimator using the following parameters: (i) $\sigma^2(d,t) = 0.25$ (the upper bound for binary variables) (ii) $\rho(d,t) = 0.1$ which is close to (but larger than) the estimated intracluster correlation of the outcome in our baseline data; (iii) the sample and group sizes given by the baseline data. The power calculations give minimum detectable effects between 2.6 and 3.3 percentage points.[8]

[8]Appendix Figure B.13 plots the power function for each estimator

3.4 Empirical Results

3.4.1 Total and Spillover Effects on the October 2020 bill

We begin the analysis by estimating total and neighborhood spillover effects on timely payments of the October 2020 property tax bill.[9] The due date was October 9th and the letters were delivered between September 28th and October 7th. We start by showing compelling graphical evidence of the effect of the intervention in Figures 4 to 6 and then we summarize the corresponding point estimates in Tables 3 and 4.

Figure 4 panel (a) shows the cumulative share of individuals paying the October 2020 bill over time, both for *treated* units and pure control blocks. The brown dashed line shows the payment rate for pure control units. The blue dashed line corresponds to treated units in group $T_g = 1$ (blocks with 20% treated). The black dashed line corresponds to treated units in group $T_g = 2$ (blocks with 50% treated). The red solid line corresponds to treated units in group $T_g = 3$ (blocks with 80% treated). Panel (b) shows, for each calendar day, the difference between each treated group and the pure control group (i.e., the treatment effect coefficients). Similarly, Figure 5 shows the analog but for *untreated* units and pure control blocks. Panel (b) thus captures spillover effects.[10]

Figure 4 reveals a clear positive direct effect of the intervention on tax compliance of treated accounts. The payment rate of treated units started to diverge from the pure control group as soon as the intervention began, reaching the maximum effect exactly by the due date of the current billing period, and staying relatively constant afterwards.[11]

Although smaller in size, Figure 5 reveals a clear spillover effect of the intervention on untreated accounts. Spillover effects mainly arise in high-saturation blocks where 80% of the neighbors were treated, and, to a lesser extent, for blocks where 50% of units were treated. The payment rate of untreated units starts to diverge from the pure control group right after the intervention began, reaching the maximum effect by the due date of the current billing period, and declining slightly afterwards. Conversely, social interference seems to be absent in blocks with only 20% treated accounts, where the spillover effect for untreated units oscillates around the zero line.

Figure 6 presents the coefficients and 95% confidence intervals from a saturated regression that estimates, day by day, the difference in payment rates between each treated and each untreated group relative to pure control blocks in which no accounts were treated (see equation 4). The top panel shows the effect on treated (left) and untreated (right) units in blocks with 80% treated, and the middle and bottom panels display the analog results for blocks with 50% and 20% treated

[9] Appendix section A.2 presents the results from balance test regressions. These results confirm that our groups are balanced and comparable.

[10] For comparison, the gray solid line shows the treatment effect for treated units (pooled together from $T_g = 1, 2, 3$ in Figure 4).

[11] Appendix A.5 shows that untreated blocks are not (indirectly) affected by adjacent treated blocks and thus provide a valid counterfactual for our analysis.

units.[12] The estimates displayed in the left panels of Figure 6 indicate an immediate and statistically significant increase in the payment rate of treated units in the three saturation groups relative to pure control blocks. Note that for the highest saturation group with 80% treated units, the effect emerges (numerically and statistically) on the same day that the letters started to be distributed, reaching a magnitude of about 4.5 percentage points. The right panels of Figure 6 show that spillover effects are more modest in magnitude and precisely estimated. In high-saturation blocks with 80% treated accounts, payment rates increase by about 1.1 percentage point and the effect is statistically significant in the early days of the intervention, losing significance from the due date onward. In all the cases, both total and spillover effects remain relatively constant after the due date (October 9th).

Table 3 summarizes the corresponding point estimates for total and spillover effects reported in Figure 6. Panels A, B, and C display total effects and spillover effects in blocks where 80%, 50%, and 20% were treated, respectively. The omitted category comprises accounts in blocks where no accounts were treated. To validate our experiment, column (1) shows a placebo saturated regression using timely payments of the September 2020 property tax bill as the dependent variable (i.e., a billing period before the intervention took place). Reassuringly, these coefficients are small in magnitude and none is statistically significant at standard levels.[13] Columns (2) to (3) show the coefficients and block-clustered standard errors for October 2020 bill payments at two different dates: October 3 (early payments) and October 31 (includes overdue payments). To benchmark our estimates, in the last row we report the average payment rate in pure control blocks at each of these dates (i.e., the constant of each regression).

From Table 3, we can see that in the early stage of the intervention, high-saturation blocks with 80% treated accounts present a statistically significant total and spillover effect of about 1.1 percentage point. This effect is relatively large in magnitude if we consider that by this date, only 5.2% of neighbors in pure controls block had paid their October 2020 bill. Naturally, as time goes by more individuals start to pay their bill, reaching 34.4% in pure control blocks by the end of the month, making small effects harder to detect. Accordingly, although the spillover effect on untreated units remains unchanged in size, it loses statistical significance. In contrast, the total effect on treated units increases to 4.5 percentage points, which represents 13.2% of the payment rate in pure control blocks.

In sum, our property tax experiment uncovers both total and spillover effects by estimating a higher payment rate of treated and untreated accounts relative to neighbors in pure control blocks where nobody received the communication letter. In both cases, effects are larger in high-saturation blocks, albeit short-lived for spillovers when considering the full sample.

[12] These point estimates coincide with those reported in panels (b) of Figures 4 and 5.

[13] Figure A.4 in the appendix presents the analog of Figure 6 for the pre-treatment September 2020 bill. Reassuringly, the evidence indicates a zero pre-treatment effect on payment rates between each treated and untreated group relative to pure control blocks.

3.4.2 Heterogeneous Effects

The results from the full experimental sample presented in the previous section unearthed modest spillover effects only in the high saturation group and only in the early days of the intervention. However, as discussed in our experiment's pre-analysis plan, it is highly likely that our treatment effects could vary along a fundamental dimension, namely pre-treatment tax compliance behavior. The relevance of this dimension of heterogeneity was anticipated and pre-registered in the experiment's pre-analysis plan.

In this section, we study heterogeneous effects along this dimension. To do so, we divide the sample in blocks that exhibited average compliance (i.e., payments) above and below the median compliance in 2019. We define past compliance by computing the average number of payments of the twelve monthly bills for 2019 in each block. We use this measure to divide our sample in two groups – those above and those below the median block average payment rate.[14]

The logic of this heterogeneity analysis goes as follows. A large fraction of neighbors that typically paid their bills stopped doing so during the pandemic in the first few months of 2020. This decrease in compliance was stronger in blocks that had higher compliance in 2019. Hence, we argue that such a core group of "good compliers" is more likely to be nudged to pay by our intervention, and where spillover effects are more likely to show up.[15]

This additional evidence is presented in Table 4, which is analogous to Table 3 but presents two sets of results—below and above median 2019 compliance. The direct effects at the end of the first month are generally larger but not substantially different: for blocks with 80%, 50% and 20% saturation, direct effects are about 5.1, 5.7 and 4.4 percentage points for street blocks above the median average compliance in 2019, compared to about 4.1, 4.8 and 5.4 for those below.[16]

The division of the sample in these two groups shows a much starker contrast for indirect or spillover effects. As in the main analysis in Table 3, there is a spillover effect in early payments for the 80% saturation group but only for city blocks above median compliance in 2019. This effect is relatively large (1.58 percentage points, larger in fact than the direct effect of 1.06). There is also a significant spillover effect for the 20% saturation group, but it is relatively small and it dissipates when looking at the end-of-month effects. For those in above median 2019 compliance city blocks in the 80% saturation group, the end-of-month spillover effect is much larger: 2.56 percentage points,

[14] The distribution of the 68,806 accounts by the number of bills paid in 2019 is bi-modal, with a core group of neighbors not paying any bill (35%) and another group paying all of them (45%). Panel (a) of Figure A.5 shows the individual-level distribution. Panel (b) shows the block-level distribution with the corresponding moments used to divide our sample.

[15] Figure A.6 suggests that 2018 and 2019 are comparable in terms of compliance, but compliance decreased substantially in 2020 because of the pandemic—the sharp fall corresponds to the lockdown measures put in place. Figure A.7 shows that payment rates in 2020 decreased more in blocks with higher compliance in 2019. In contrast, 2018 and 2019 show similar levels of compliance.

[16] The differences are relatively small for early payments, and not significant for the placebo September 2020 bill.

about half of the direct effect in the same group (5.09 percentage points).

The daily direct and indirect effect of our campaign for the group with 80% of individuals treated in street blocks above and below median compliance in 2019 is illustrated in Figure 7, which makes the pattern in Table 4 all the more apparent.[17]

To sum up, the mild spillover effect reported in the previous section is much stronger and driven by individuals living in blocks with high compliance in 2019, as predicted and registered in our pre-analysis plan. The effect is only present in blocks where 80% of the accounts were treated, where spillovers were more likely to emerge.

3.4.3 Other Margins

Subscriptions to electronic billing. We find evidence that our tax communication campaign also increases the subscriptions to receive an electronic bill by e-mail.[18] These effects are greater in high-saturation blocks, albeit small in absolute value. Appendix Section A.3 presents convincing graphical evidence of total and spillover effects (Figure A.8) which are then summarized in Table A2, although spillover effects in this outcome are much more tenuous.

Backward and forward payments. We also find that the effects of our letters are not solely concentrated on the October 2020 billing period (the bill targeted by our intervention). Section A.4 presents convincing graphical evidence that the letters also increased the payment rates in subsequent billing periods. Perhaps more striking, we also show that some neighbors made backward payments to cancel past-due debt from previous billing periods. This is especially prominent after April 2020 when the COVID-19 lockdown measures were established in Argentina (See Figure A.9).

4 Conclusion

We provide a general framework to analyze and design partial population experiments with an application to spillovers in property tax compliance. We derive an asymptotic approximation and variance formulas to conduct power calculations for general clustered experimental designs allowing for multiple treatments, general forms of intracluster correlation, and cluster heterogeneity. To incorporate cluster heterogeneity into the experimental design, we consider a double-array asymptotic setting where both the number of clusters and the cluster sizes grow with the sample size, which nests the commonly analyzed case with fixed cluster size and/or equally sized clusters. We

[17] Table 4 confirms that spillover effects are driven by blocks with baseline compliance above the median in high saturation blocks (80% treated). Spillover effects are more muted and insignificant in medium (50% treated) and low (20% treated) saturation blocks, however. Reassuringly, the first two columns also show no effects for the pre-intervention bill of September 2020 either above or below the median.

[18] Note that nudging individuals to sign up to e-billing was an explicit content of the letter (see Figure A.1).

then apply our results to conduct power and MDE calculations in partial population experiments to derive formulas for optimal group-level assignment probabilities. Our formulas and design are easy to adapt to other experimental settings.

In our application, we estimate total and neighborhood spillover effects of a randomized communication campaign on property tax compliance in a large municipality of Argentina where neighbors must pay a monthly bill on their real estate. We estimate total effects on monthly payments, and also analyze whether the campaign creates spillover effects on neighbors that live nearby within a treated block but that do not receive a letter. We find compelling evidence of total effects and spillover effects on property tax payment rates. Our results reveal higher payment rates of treated and untreated accounts relative to neighbors in pure control blocks where nobody received the communication letter. We find that these indirect or spillover effects are much stronger in city street blocks that exhibited a higher degree of tax compliance in the pre-treatment period. This application showcases the usefulness of our methodological framework for designing partial population experiments.

References

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.

Antinyan, Armenak, and Zareh Asatryan. 2019. "Nudging for tax compliance: A meta-analysis." ZEW - Leibniz Centre for European Economic Research ZEW Discussion Papers 19-055.

Athey, Susan, Dean Eckles, and Guido W. Imbens. 2018. "Exact P-values for Network Interference." *Journal of the American Statistical Association*, 113(521): 230–240.

Baird, Sarah, Aislinn Bohren, Craig McIntosh, and Berk Özler. 2018. "Optimal Design of Experiments in the Presence of Interference." *The Review of Economics and Statistics*, 100(5): 844–860.

Bai, Yuehao. 2022. "Optimality of Matched-Pair Designs in Randomized Controlled Trials." *American Economic Review*, 112(12): 3911–3940.

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

Basse, G W, A Feller, and P Toulis. 2019. "Randomization tests of causal effects under interference." *Biometrika*, 106(2): 487–494.

Berger, Martijn P.F., and Weng-Kee Wong. 2009. *An Introduction to Optimal Designs for Social and Biomedical Research.* Wiley.

Bergeron, Augustin, Gabriel Tourek, and Jonathan Weigel. 2021. "The State Capacity Ceiling on Tax Rates: Evidence from Randomized Tax Abatements in the DRC." *Mimeo.*

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

Boning, William C., John Guyton, Ronald Hodge, and Joel Slemrod. 2020. "Heard it through the grapevine: The direct and network effects of a tax enforcement field experiment on firms." *Journal of Public Economics*, 190(C).

Brockmeyer, A, A Estefan, K Ramirez Arras, and J.C. Suarez Serrato. 2020. "Taxing Property in Developing Countries: Theory and Evidence from Mexico." *IFS Working Paper.*

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4): 200–232.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh. 2018. "Inference under Covariate-Adaptive Randomization." *Journal of the American Statistical Association*, 113(524): 1784–1796.

Bugni, Federico A, Ivan A. Canay, and Azeem M. Shaikh. 2019. "Inference under Covariate-Adaptive Randomization with Multiple Treatments." *Quantitative Economics*, 10(4): 1747–1785.

Bugni, Federico A., Ivan A. Canay, Azeem M. Shaikh, and Max Tabord-Meehan. 2022. "Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes." *arXiv:2204.08356*.

Cameron, Adrian Colin, and Douglas L Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, 50(2): 317–372.

Carrillo, Paul E., Edgar Castro, and Carlos Scartascini. 2021. "Public good provision and property tax compliance: Evidence from a natural experiment." *Journal of Public Economics*, 198: 104422.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2017. "Asymptotic Behavior of a t-Test Robust to Cluster Heterogeneity." *The Review of Economics and Statistics*, 99(4): 698–709.

Crépon, 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.

De Neve, Jan-Emmanuel, Clément Imbert, Johannes Spinnewijn, Teodora Tsankova, and Maarten Luts. 2021. "How to Improve Tax Compliance? Evidence from Population-Wide Experiments in Belgium." *Journal of Political Economy*, 129(5): 1425–1463.

Djogbenou, Antoine A., James G. MacKinnon, and Morten Ørregaard Nielsen. 2019. "Asymptotic theory and wild bootstrap inference with clustered errors." *Journal of Econometrics*, 212(2): 393–412.

Drago, Francesco, Friederike Mengel, and Christian Traxler. 2020. "Compliance Behavior in Networks: Evidence from a Field Experiment." *American Economic Journal: Applied Economics*, 12(2): 96–133.

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." In *Handbook of Development Economics*. Vol. 4 of *Handbook of Development Economics*, , ed. T. Paul Schultz and John A. Strauss, 3895–3962. Elsevier.

Foos, Florian, and Eline A. de Rooij. 2017. "All in the Family: Partisan Disagreement and Electoral Mobilization in Intimate Networks—A Spillover Experiment." *American Journal of Political Science*, 61(2): 289–304.

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

Hansen, Bruce E., and Seojeong Lee. 2019. "Asymptotic theory for clustered samples." *Journal of Econometrics*, 210(2): 268–290.

Haushofer, Johannes, and Jeremy Shapiro. 2016. "The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya." *The Quarterly Journal of Economics*, 131(4): 1973–2042.

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

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

Ichino, Nahomi, and Matthias Schündeln. 2012. "Deterring or Displacing Electoral Irregularities? Spillover Effects of Observers in a Randomized Field Experiment in Ghana." *The Journal of Politics*, 74(1): 292–307.

Imai, Kosuke, Zhichao Jiang, and Anup Malani. 2021. "Causal Inference With Interference and Noncompliance in Two-Stage Randomized Experiments." *Journal of the American Statistical Association*, 116(534): 632–644.

Jiang, Zhichao, Kosuke Imai, and Anup Malani. 2022. "Statistical Inference and Power Analysis for Direct and Spillover Effects in Two-Stage Randomized Experiments." *Biometrics*.

Krause, Benjamin. 2020. "Balancing Purse and Peace: Tax Collection, Public Goods and Protests." *Mimeo*.

Leung, Michael P. 2022. "Rate-optimal cluster-randomized designs for spatial interference." *The Annals of Statistics*, 50(5): 3064 – 3087.

Melas, Viatcheslav B. 2006. *Functional Approach to Optimal Experimental Design.* Springer New York.

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

Moffit, Robert. 2001. "Policy Interventions, Low-level Equilibria and Social Interactions." In *Social Dynamics*. , ed. Stephen N. Durlauf and Peyton Young, 45–82. MIT Press.

Pomeranz, Dina. 2015. "No Taxation without Information : Deterrence and Self-Enforcement in the Value Added Tax." *The American Economic Review*, 105(8): 2539–2569.

Pomeranz, Dina, and José Vila-Belda. 2019. "Taking State-Capacity Research to the Field: Insights from Collaborations with Tax Authorities." *Annual Review of Economics*, 11(1): 755–781.

Puelz, David, Guillaume Basse, Avi Feller, and Panos Toulis. 2022. "A graph-theoretic approach to randomization tests of causal effects under general interference." *Journal of the Royal Statistical Society: Series B*, 84(1): 174–204.

Silvey, Samuel D. 1980. *Optimal Design: An Introduction to the Theory for Parameter Estimation.* Springer Netherlands.

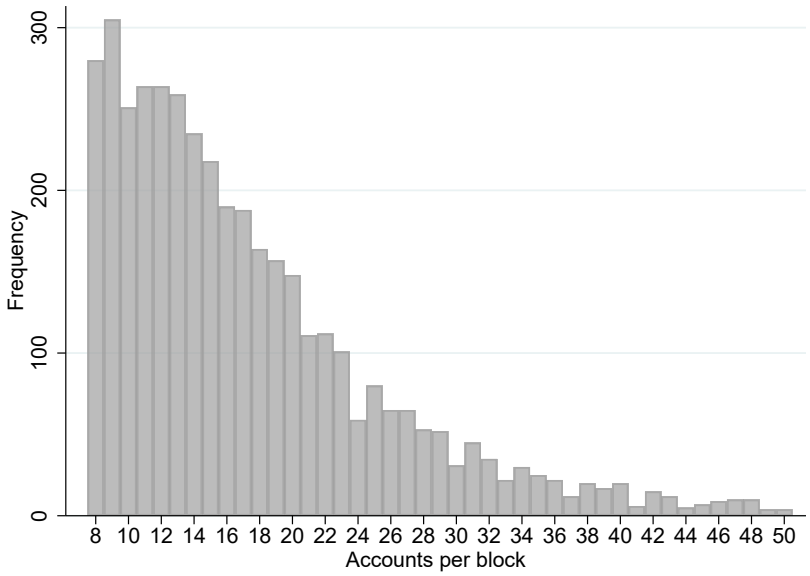
Vazquez-Bare, Gonzalo. 2022. "Identification and Estimation of Spillover Effects in Randomized Experiments." *Journal of Econometrics*.

Viviano, Davide. 2022. "Policy design in experiments with unknown interference." *working paper*.

Weigel, Jonathan L. 2020. "The Participation Dividend of Taxation: How Citizens in Congo Engage More with the State When it Tries to Tax Them." *The Quarterly Journal of Economics*, 135(4): 1849–1903.

5 Figures and Tables

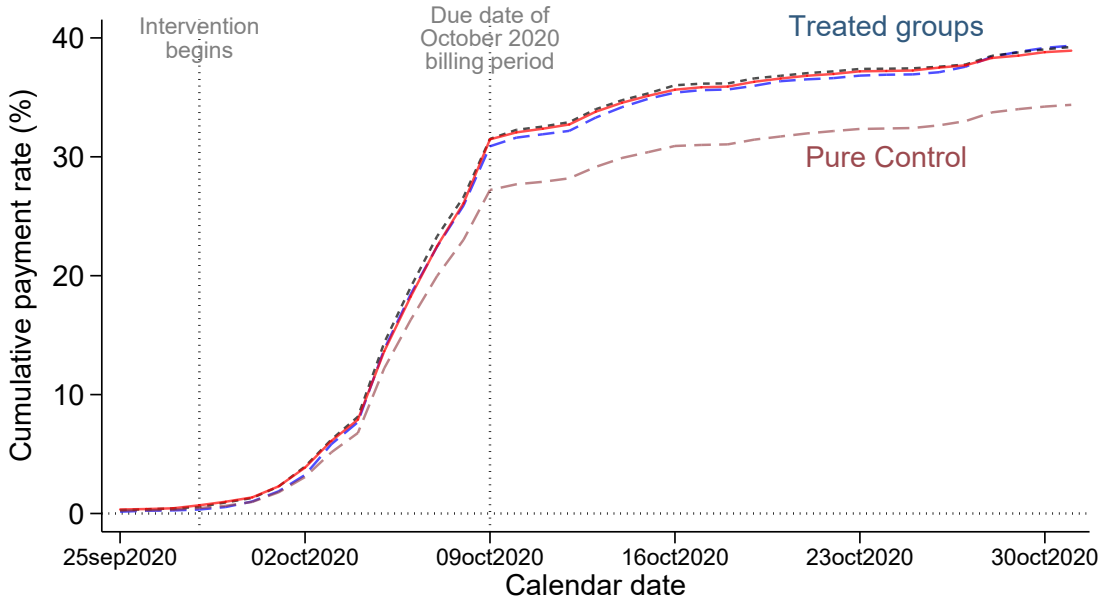
Figure 3: Distribution of accounts per block



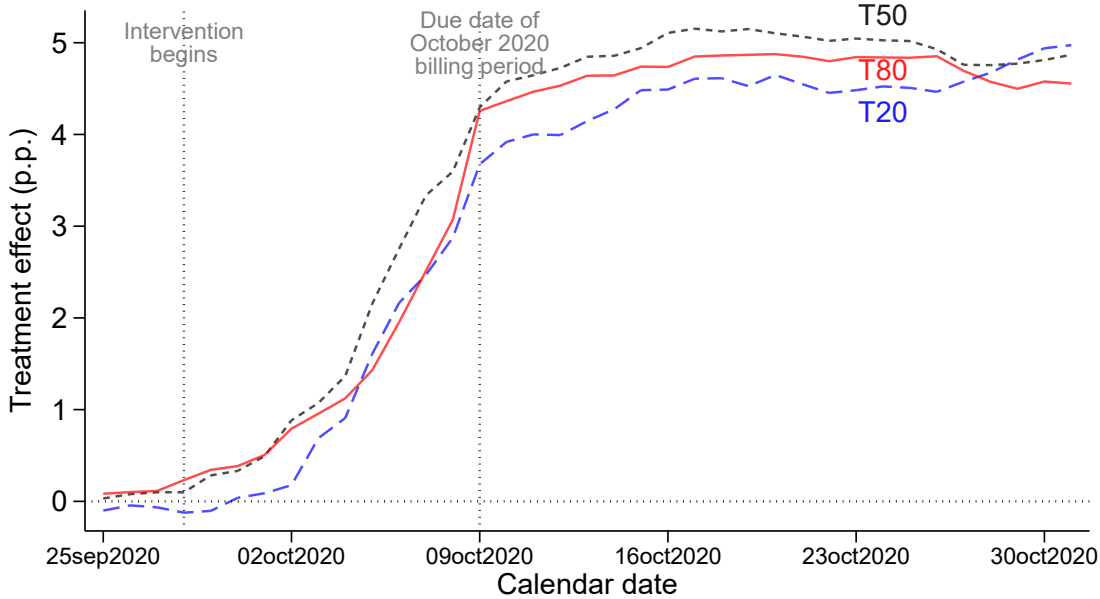
Notes: This figure shows the distribution of accounts per block using data from the year 2019. We use these data to design the experiment. Our sample size consists of 68,808 accounts distributed in 3,982 blocks.

Figure 4: Payment rates: Treated groups vs Pure control blocks

(a) Payment rates in levels



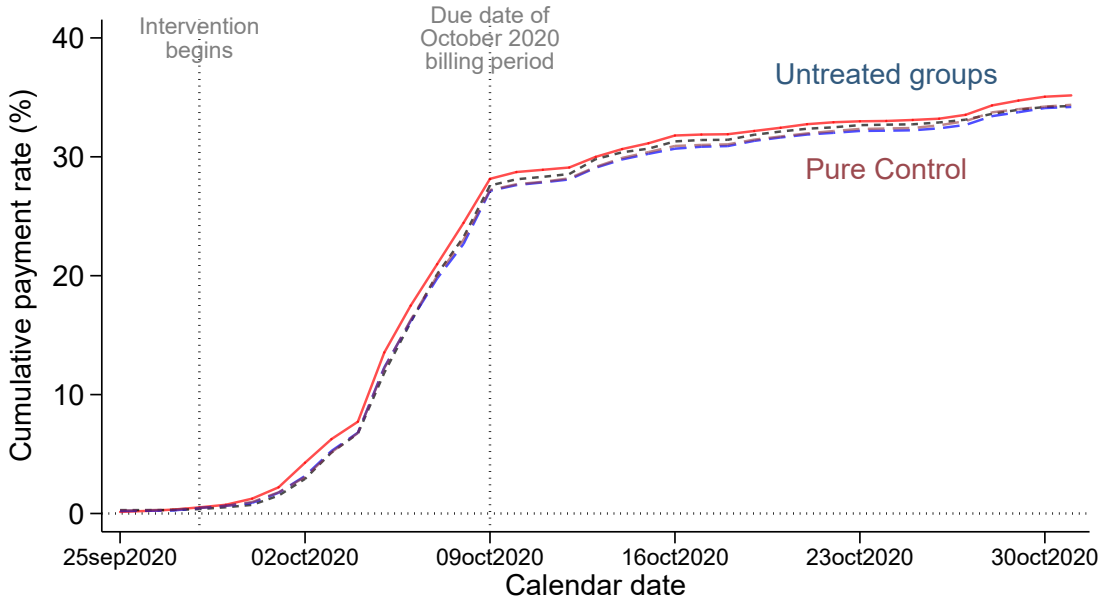
(b) Difference relative to pure control group



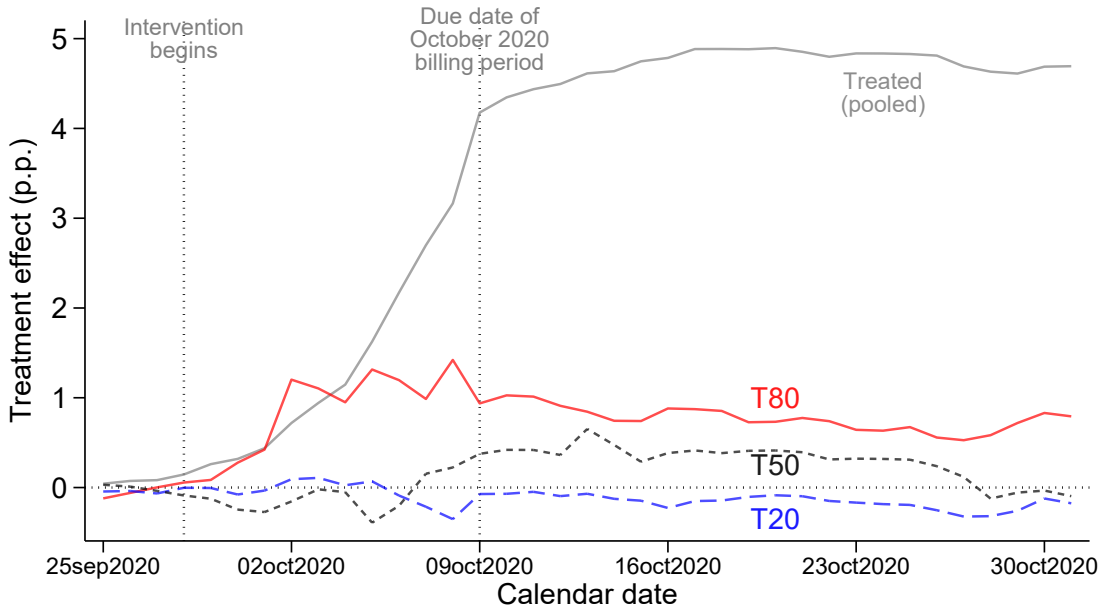
Notes: These figures show the effect of the intervention on payments of the October 2020 bill for treated groups. Panel (a) shows the cumulative share of individuals paying the October 2020 bill over time. The brown dashed line shows the payment rate for pure control units. The blue dashed line corresponds to treated units in group $T_g = 1$ (blocks with 20% treated). The black dashed line corresponds to treated units in group $T_g = 2$ (blocks with 50% treated). The red solid line corresponds to treated units in group $T_g = 3$ (blocks with 80% treated). Panel (b) shows, for each calendar date, the difference between each treated group and the pure control group (treatment effect coefficients). The letters were delivered between September 28th and October 7th. The first vertical bar denotes the start of the intervention. The due date was October 9th and is indicated with another vertical bar.

Figure 5: Payment rates: Untreated groups vs Pure control blocks

(a) Payment rates in levels

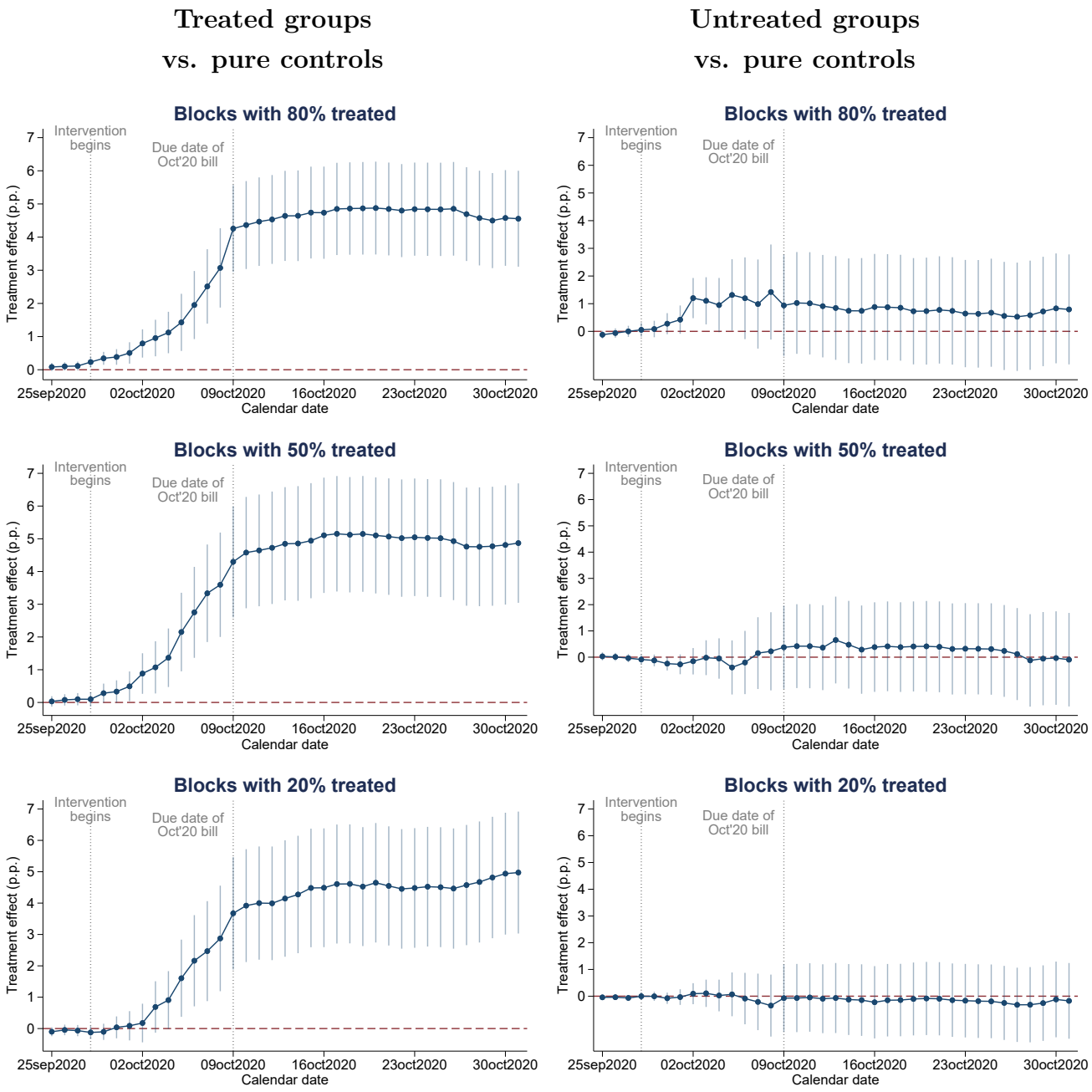


(b) Difference relative to pure control group



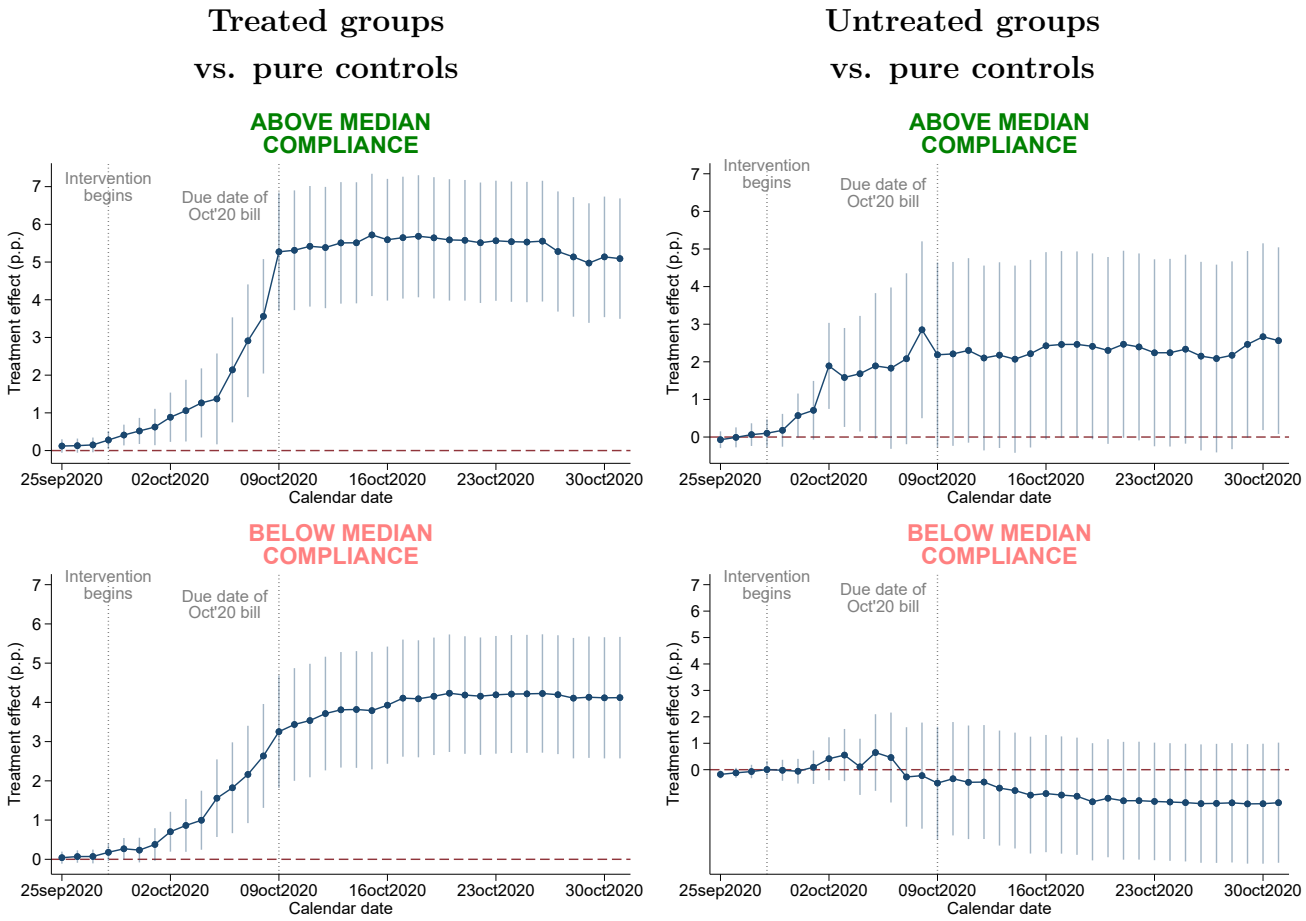
Notes: These figures show the effect of the intervention on payments of the October 2020 bill for untreated groups. Panel (a) shows the cumulative share of individuals paying the October 2020 bill over time. The brown dashed line shows the payment rate for pure control units. The blue dashed line corresponds to untreated units in group $T_g = 1$ (blocks with 20% treated). The black dashed line corresponds to untreated units in group $T_g = 2$ (blocks with 50% treated). The red solid line corresponds to untreated units in group $T_g = 3$ (blocks with 80% treated). Panel (b) shows, for each calendar date, the difference between each untreated group and the pure control group (treatment effect coefficients). For comparison, the gray solid line shows the treatment effects for treated units (pooled from $T_g = 1, 2, 3$). The letters were delivered between September 28th and October 7th. The first vertical bar denotes the start of the intervention. The due date was October 9th and is indicated with another vertical bar.

Figure 6: Direct effects on treated accounts and spillover effects on untreated accounts



Notes: These figures show the coefficients and 95% confidence intervals from a saturated regression that computes, at each calendar day, the payment rate difference between each treated and untreated group relative to the pure control group (i.e., blocks where no accounts were treated). The top panel shows the effect on treated (left) and untreated (right) units in blocks with 80% treated ($T_g = 3$). The middle panel shows the effect on treated (left) and untreated (right) units in blocks with 50% treated ($T_g = 2$). The bottom panel shows the effect on treated (left) and untreated (right) units in blocks with 20% treated ($T_g = 3$). These point estimates coincide with those reported in panel (b) of Figures 4 and 5. Standard errors are clustered by block. The first vertical bar denotes the start of the intervention. The due date for the October 2020 bill was October 9th and is indicated with another vertical bar. The letters were delivered between September 28th and October 7th.

Figure 7: Heterogeneity of total and spillover effects on property tax payments in blocks below and above median compliance in 2019. Blocks with 80% treated.



Notes: These figures show the coefficients and 95% confidence intervals from a saturated regression that computes, at each calendar day, the payment rate difference between treated and untreated groups relative to the pure control group (i.e., blocks where no accounts were treated). We focus the attention to blocks where 80% of the units were treated. The top figures show the effect on treated (left) and untreated (right) units in blocks with baseline compliance above the median. The bottom figures repeat this in blocks with baseline compliance below the median. We define compliance as the share of bills paid by block in 2019. The median compliance is 0.56 (see Figure A.5). Standard errors are clustered by block. The first vertical bar shows the due date for the September 2020 bill. This corresponds to a bill issued and due for payment before our intervention began, thus serving as a placebo. The second vertical bar indicates the start of the intervention. The letters were delivered between September 28th and October 7th.

Table 2: Descriptive statistics in 2019 (baseline year)

	Blocks	Obs	Mean	SD	ICC
Paid the twelve bills in 2019	3,981	68,808	0.449	0.497	0.062
Paid at least one bill in 2019	3,981	68,808	0.650	0.477	0.071
Paid six bills or more in 2019	3,981	68,808	0.572	0.495	0.073

Notes: This table shows descriptive statistics about the frequency of payments in 2019. This is the baseline year we used for the randomization, power calculations, and simulations. The data set is restricted to blocks with size between 8 and 50 accounts. Figure 3 shows the distribution of accounts per block. Our sample size consists of 68,808 accounts distributed in 3,982 blocks. The frequency of payments is very polarized. About 45 percent of the accounts paid the twelve bills and about 35 percent did not pay any bill. We call these two core groups *always payers* and *never payers*, respectively. The perfect compliance rate of 45 percent is presumably low and, therefore, leaves room for potential behavioral responses from non-compliant and partially-compliant neighbors.

Table 3: Total and spillover effects on property tax payments

Dependent variable:	**Placebo bill:**	**Intervention bill:**	
Pr(pay the bill)	Sep'20	Early	By Oct 31
	(1)	(2)	(3)
A. Blocks with 80% treated			
Treated	0.12	0.96***	4.55***
	(0.69)	(0.28)	(0.74)
Untreated	-0.30	1.10**	0.79
	(0.95)	(0.43)	(1.01)
B. Blocks with 50% treated			
Treated	0.76	1.07***	4.87***
	(0.88)	(0.41)	(0.93)
Untreated	0.26	-0.02	-0.10
	(0.88)	(0.34)	(0.91)
C. Blocks with 20% treated			
Treated	0.85	0.69*	4.97***
	(0.93)	(0.42)	(0.99)
Untreated	0.07	0.11	-0.18
	(0.68)	(0.26)	(0.72)
Payment Rate of Pure Control	29.70	5.15	34.37
Observations	68,806	68,806	68,806
Number of clusters (blocks)	3,981	3,981	3,981

Notes: This table shows the results from saturated OLS regressions (equation 4 in the text). Each column corresponds to a separate regression. The omitted category corresponds to blocks where no accounts were treated (pure control). Panel A shows the results for blocks where 80% were treated, panel B for blocks with 50% treated, and panel C for blocks with 20% treated. The dependent variable in each column is: (1) an indicator for paying the September 2020 bill by September 15th (pre intervention); (2) an indicator for paying the October 2020 bill by October 3rd (early payments); (3) an indicator for paying the October 2020 bill by October 31st (includes early, on time, and overdue payments). The first column corresponds to a pre-intervention bill and considers payments made before the letters were delivered (placebo). The estimates correspond exactly to the numbers shown in Figure (6). The letters were delivered between September 28th and October 7th. The due date for the October 2020 bill was October 9th. The row *Payment Rate of Pure Control* displays the constant of each regression, corresponding to the average payment rate in blocks with no treated units). Standard errors clustered by blocks are reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 4: Heterogeneity of total and spillover effects on property tax payments in blocks below and above median compliance in 2019

	Placebo bill:		**Intervention bill:**			
	Sep'20		Early		By Oct 31	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
	(1)	(2)	(3)	(4)	(5)	(6)
A. Blocks with 80% treated						
Treated	0.10	0.28	0.86**	1.06**	4.12***	5.09***
	(0.73)	(0.81)	(0.34)	(0.42)	(0.79)	(0.81)
Untreated	-1.55	0.78	0.55	1.58**	-1.25	2.56**
	(1.09)	(1.24)	(0.50)	(0.67)	(1.16)	(1.27)
B. Blocks with 50% treated						
Treated	1.54	0.69	1.24**	1.02	4.81***	5.67***
	(0.99)	(1.12)	(0.50)	(0.62)	(1.07)	(1.08)
Untreated	0.81	0.36	0.10	-0.03	1.34	-0.76
	(0.94)	(1.15)	(0.43)	(0.50)	(1.00)	(1.14)
C. Blocks with 20% treated						
Treated	1.32	0.27	0.85*	0.52	5.41***	4.40***
	(1.11)	(1.24)	(0.52)	(0.63)	(1.21)	(1.27)
Untreated	0.27	-0.32	0.68**	-0.42	0.61	-1.09
	(0.72)	(0.80)	(0.33)	(0.38)	(0.77)	(0.82)
Payment Rate of Pure Control	20.05	38.19	3.63	6.49	23.53	43.91
Observations	32,361	36,445	32,361	36,445	32,361	36,445
Number of clusters (blocks)	2,013	1,968	2,013	1,968	2,013	1,968

Notes: This table shows the results from saturated OLS regressions (equation (4) in the text) in which we break the main results from Table (3) for blocks below and above median compliance in 2019. We define compliance as the share of bills paid by block in 2019 with median value of 0.56 (see Figure A.5). The dependent variable in each column is: (1) and (2) an indicator for paying the September 2020 bill by September 15th (pre intervention); (3) and (4) an indicator for paying the October 2020 bill by October 3rd (early payments); (5) and (6) an indicator for paying the October 2020 bill by October 31st (includes early, on time, and overdue payments). The letters were delivered between September 28th and October 7th. The due date for the October 2020 bill was October 9th. The row *Payment Rate of Pure Control* displays the constant of each regression, corresponding to the average payment rate in blocks with no treated units). Standard errors clustered by blocks are reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Supplementary Materials for:

"Design of Two-Stage Experiments with an Application to Spillovers in Tax Compliance"

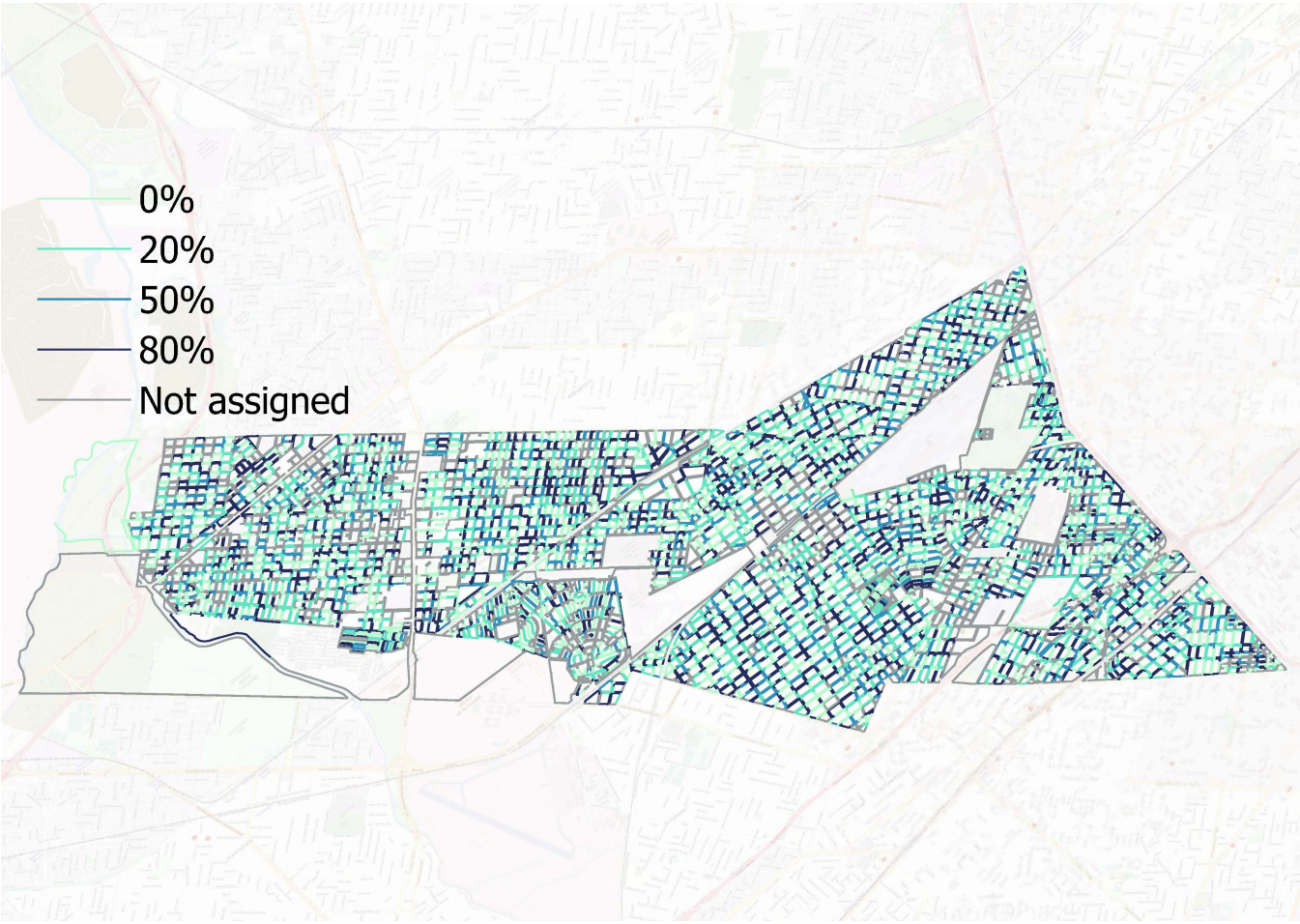
A Additional Material and Results

A.1 Additional Material

Figure A.1: Example of the intervention letter

Notes: This figure shows an anonymized example of the letters sent during the intervention between September 28th and October 7th, 2020. The headline reads: *"Your municipal taxes are now available on the electronic bill."* The information below the headline contains the name of the account holder, the address, and the account number. The main text of the letter reads: *"We would like to tell you that now in Tres de Febrero your municipal General Service Fee (TSG) bill is 100% digital. In other words, paper is no longer used. You can access it and pay for it from your cell phone or computer. In this way, we take care of each other by reducing circulation and we also take care of the environment. It is a difficult situation and we appreciate the effort you are making to keep up with your taxes, because that translates directly into constructions and services that do not stop in your neighborhood. We inform you of the status of your account and show you how easy it is:"* The table below this text shows the account number, the amount due in the October 2020 billing period, the amount of past due debt from previous months of 2020, and the amount of past due date from earlier years. The large box below the table explains: (1) how to sign up for the electronic billing, and (2) how to pay the bill and the different means of payment (online or in person). Finally, below the box, the text reads: *"In case of doubts, contact us at reclamos.mistasas@tresdefebrero.gov.ar. If this letter arrived by mistake at your address, inform us in that same email. Many thanks!"*

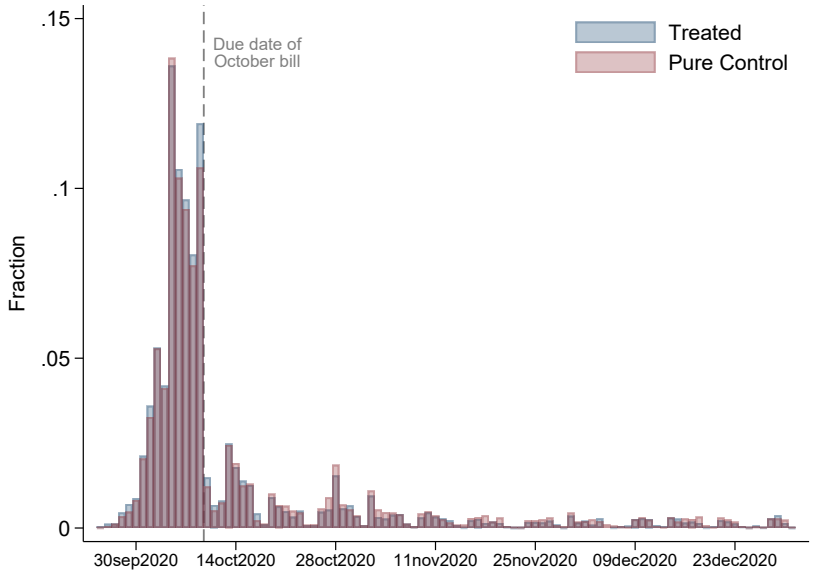
Figure A.2: Map of the municipality with the experimental design



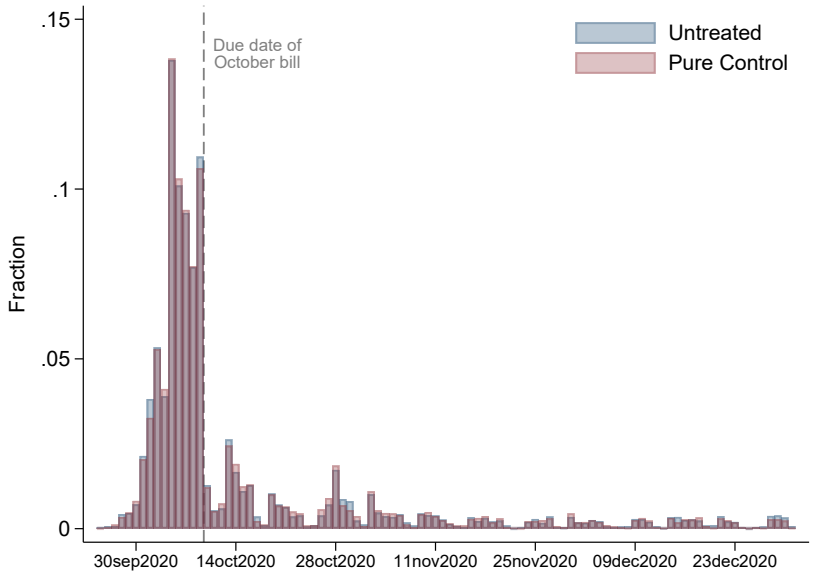
Notes: This figure shows a map of the municipality where the 2-level randomized communication campaign took place. We highlight the group-level assignment of blocks (*cuadras*) with different colors: pure control blocks with 0% treated (light green), blocks with 20% treated accounts (green), blocks with 50% treated (blue), and blocks with 80% treated (dark blue). We use gray for blocks that were not part of the experiment (e.g., industrial or commercial blocks).

Figure A.3: Distribution of payment date for treated, untreated, and pure control (October 2020 billing period)

(a) Treated vs. Pure Control

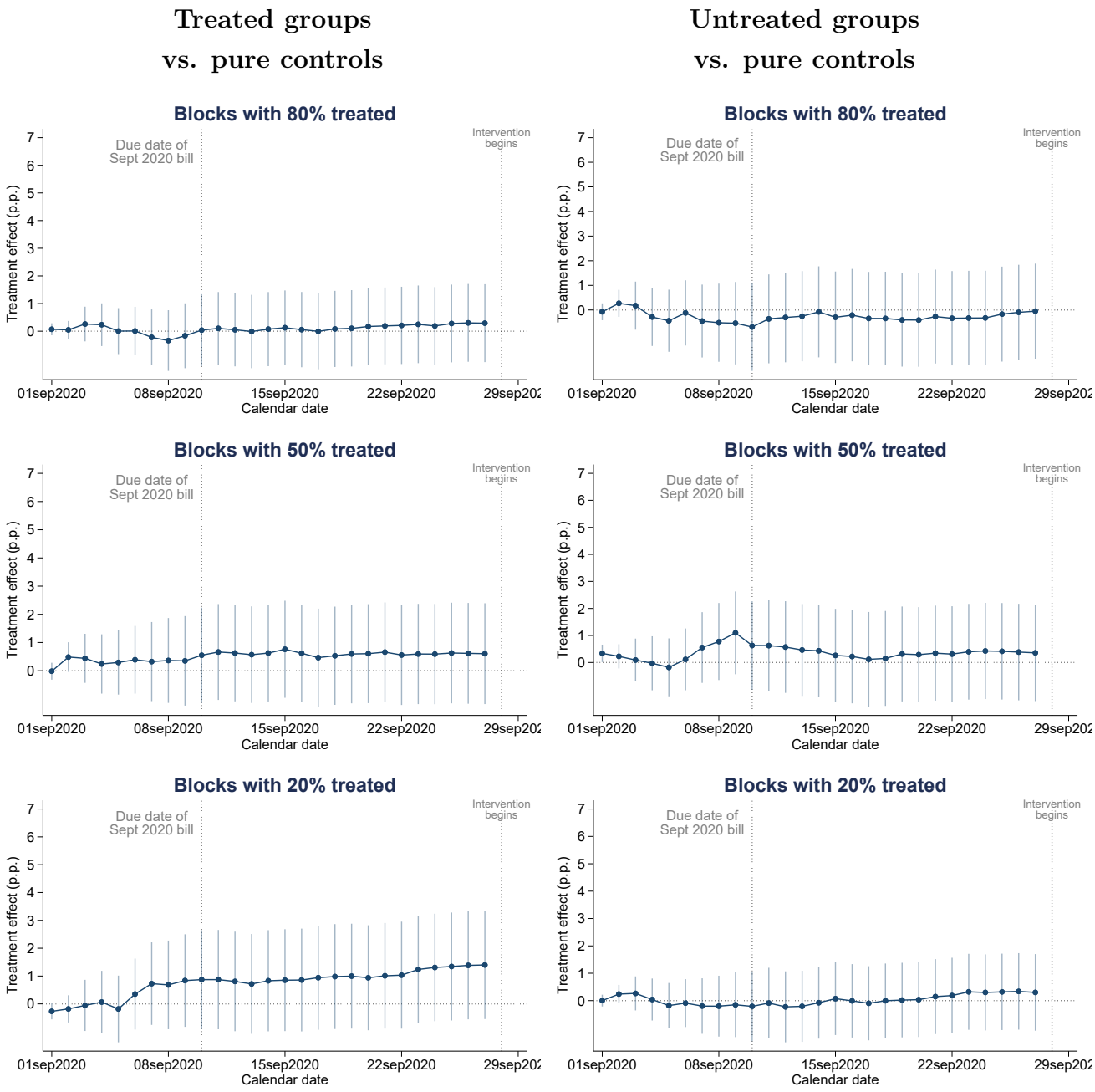


(b) Untreated vs. Pure Control



Notes: These figures show the fraction of individuals paying the October 2020 bill before and after the due date (October 9th, 2020). Panel (a) shows the distribution of payments for treated units (in blue) relative to pure control units (in red). We pool together treated units from $T_g = 1, 2, 3$. Panel (b) shows the distribution of payments for untreated units (in blue) relative to pure control units (in red). We pool together untreated units from $T_g = 1, 2, 3$. The area of each histogram integrates to one. A larger bar in a particular date means that the payment frequency of the corresponding group is higher than the other group.

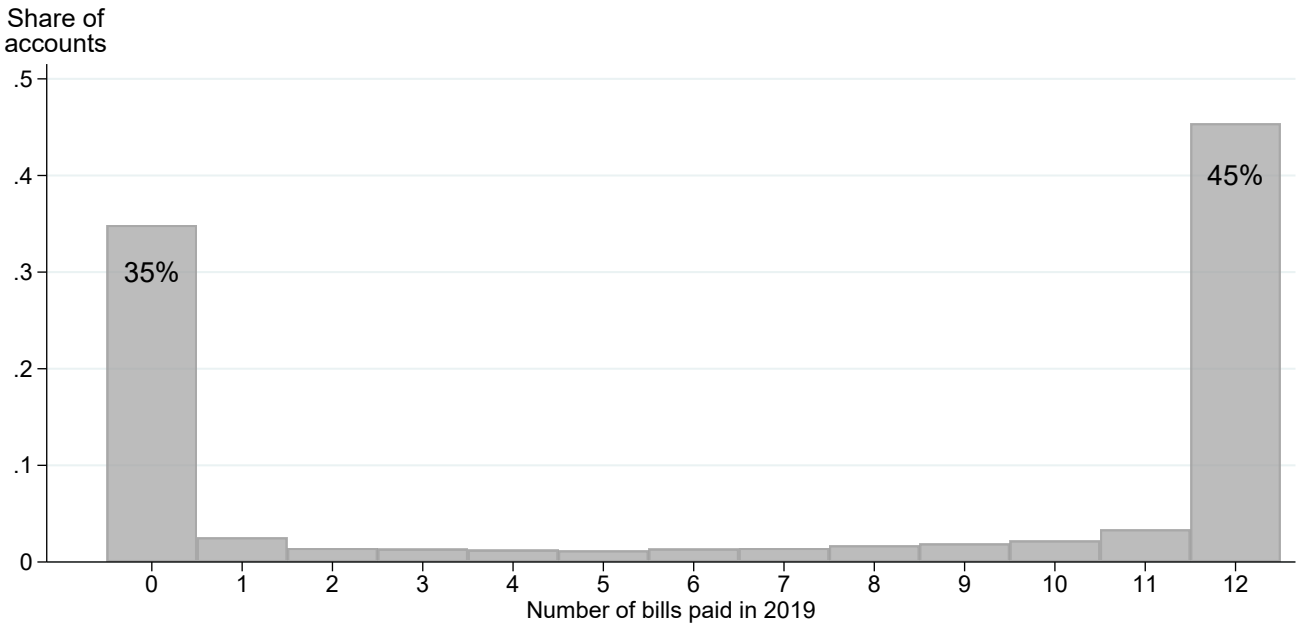
Figure A.4: Placebo. Direct and spillover effects for the pre-intervention Sep'20 bill



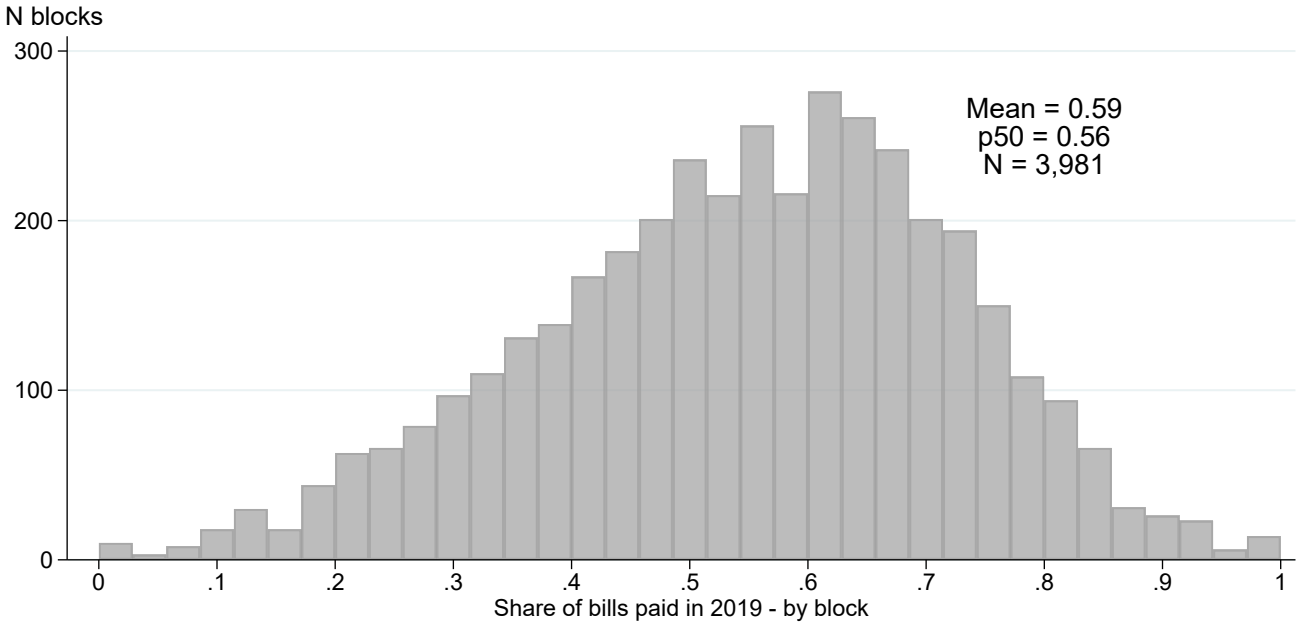
Notes: These figures show the coefficients and 95% confidence intervals from a saturated regression that computes, at each calendar day, the payment rate difference between each treated and untreated group relative to the pure control group (i.e., blocks where no accounts were treated). The top panel shows the effect on treated (left) and untreated (right) units in blocks with 80% treated ($T_g = 3$). The middle panel shows the effect on treated (left) and untreated (right) units in blocks with 50% treated ($T_g = 2$). The bottom panel shows the effect on treated (left) and untreated (right) units in blocks with 20% treated ($T_g = 3$). Standard errors are clustered by block. The first vertical bar shows the due date for the September 2020 bill. This corresponds to a bill issued and due for payment before our intervention began, thus serving as a placebo. The second vertical bar indicates the start of the intervention. The letters were delivered between September 28th and October 7th.

Figure A.5: Distribution of bill payments in 2019 for individuals and blocks

(a) **Number of monthly bills paid in 2019 (by individuals)**



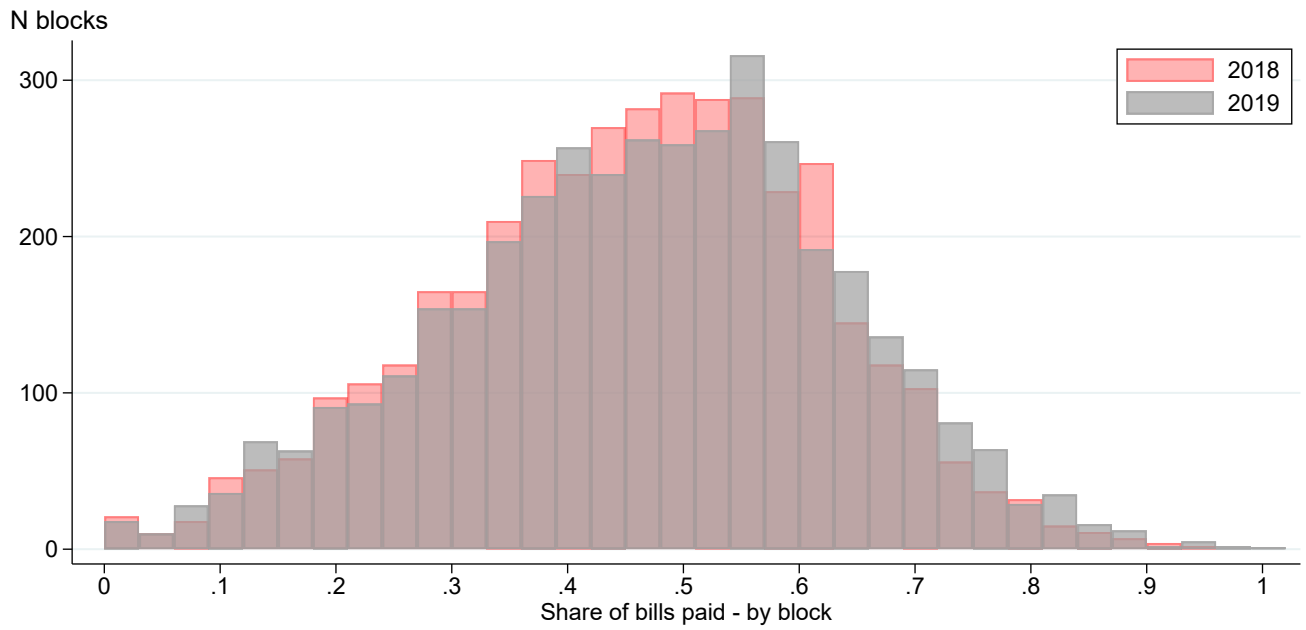
(b) **Share of bills paid in 2019 (by blocks)**



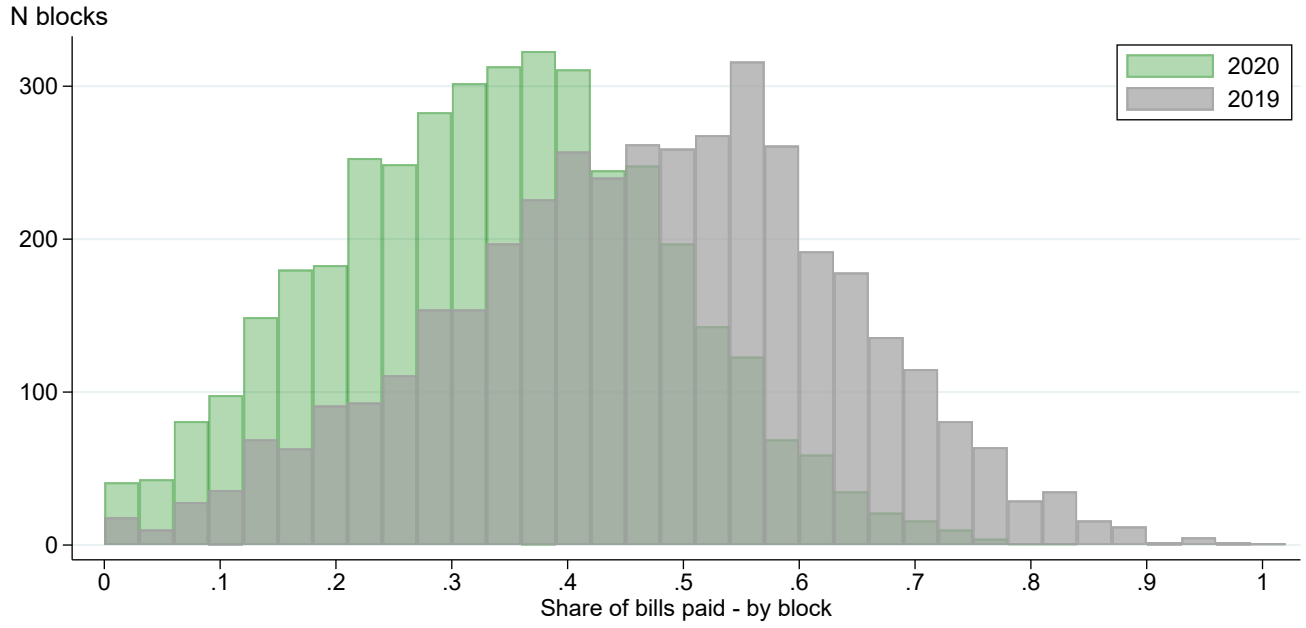
Notes: Panel (a) shows the distribution of the 68,806 accounts by the number of bills paid in 2019. The distribution is bi-modal with a core group of neighbors not paying any bill (35%) and another group paying all of them (45%). Panel (b) uses the information from panel (a) to compute the share of total bills paid in 2019 for each block. We use this measure of block-level compliance for the heterogeneity analysis, to split our sample into blocks below and above the median of 0.56 (see Table 4). These two figures and values look very similar for the year 2018.

Figure A.6: Compliance in the first nine months of 2018, 2019, and 2020

(a) **2018 vs 2019**

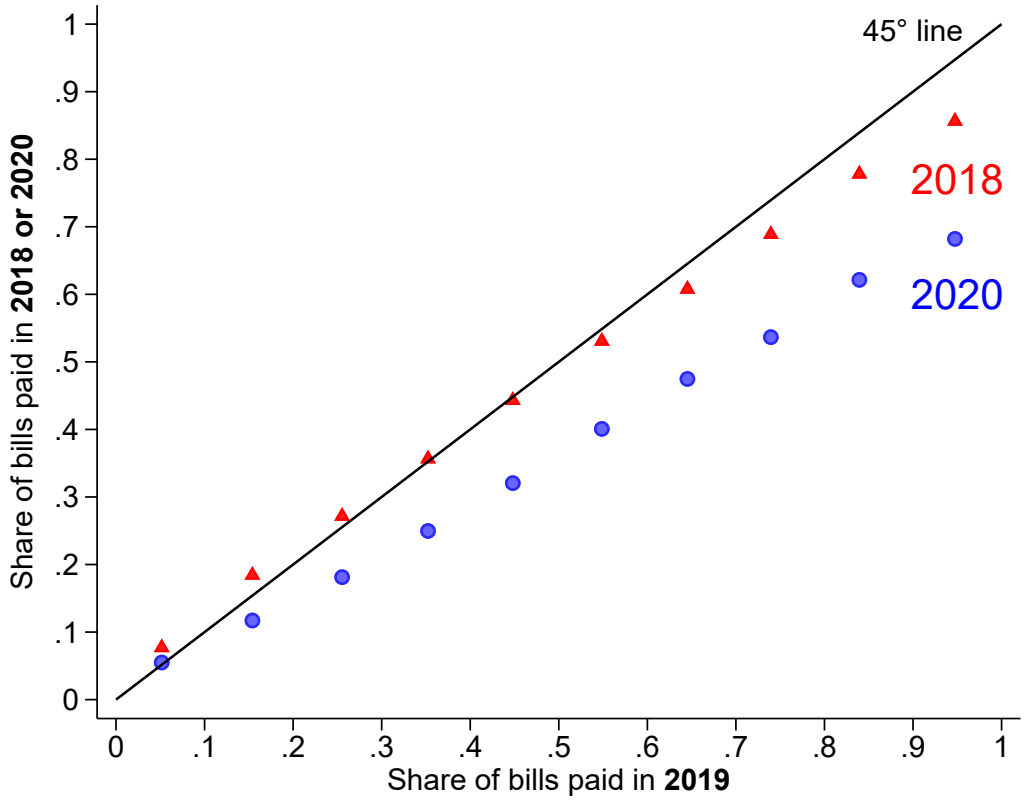


(b) **2019 vs 2020**



Notes: These figures show compliance in the first 9 billing periods of the year. For each block we compute the share of total bills paid out of 9. Panel (a) compares 2018 and 2019 and panel (b) compares 2019 and 2020. We restrict the analysis to the first 9 bills because our intervention takes place in October. To make it comparable, the numerator excludes overdue payments (i.e., payments made after the due date of each month). The figure suggests that 2018 and 2019 are comparable in terms of compliance and that compliance decreases substantially in 2020 because of the pandemic.

Figure A.7: Payment rates in 2020 decreased more in blocks with higher compliance in 2019



Notes: This figure compares compliance in 2018 or 2020 (vertical axis) relative to 2019 (horizontal axis) at the block level. To that end, we split the sample of blocks into ten evenly-spaced groups using the share of payments in 2019 (horizontal axis). For each bin, we then compute the average share of payments in 2018, 2019, and 2020. The red triangles compare 2018 against 2019 and the blue circles compare 2020 against 2019. The 45° line corresponds to the situation where compliance remains unchanged over time. The figure suggests that the drop in compliance in 2020 highlighted in Figure A.6 is more prominent for higher levels of baseline compliance. That is, blocks that had high compliance in 2019 are those where the payment rate decreased the most in the first nine months of 2020. In contrast, 2018 and 2019 display similar levels of compliance. This stylized fact suggests that blocks with high compliance in 2019 (and low compliance in 2020) are more likely to be nudged by our intervention and, thus, where spillovers are more likely to manifest.

A.2 Balance checks

We run balance test checks to verify the comparability of the treated, untreated, and pure control groups in terms of demographic and account-related characteristics in 2019. We jointly estimate the parameters of interest through the following saturated OLS regression:

$$X_{ig} = \alpha + \sum_{t=1}^{3} \theta_t \mathbb{1}(T_g = t)(1 - D_{ig}) + \sum_{t=1}^{3} \tau_t \mathbb{1}(T_g = t) D_{ig} + \varepsilon_{ig} \tag{5}$$

where X_{ig} is one of the account holder or dwelling characteristics contained in our baseline data. We allow ε_{ig} to be correlated within blocks and use a cluster-robust variance estimator. In this regression, θ_t captures the average difference of X_{ig} of untreated units in groups with $T_g = t$ relative to the pure control group and τ_t captures the average difference of X_{ig} of treated units in groups with $T_g = t$ relative to the pure control group. The results are reported in Table A1 and reassuringly confirm that our groups are highly balanced. The null effect on timely payments (i.e., excluding past-due payments) of the September 2020 bill—the bill prior to our intervention— sheds further light on the balance between groups (see Figure A.4).

Table A1: Balance test saturated regressions

	Property Value	Front Metres	House type	Tenant Male	Tenant Age	Bill amount	N Bills paid 2019	Digital payment
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Blocks with 80% treated:								
Treated	0.01	−8.27	−0.00	−0.00	−0.14	2.81	0.05	−0.00
	(0.02)	(17.77)	(0.00)	(0.01)	(0.40)	(7.81)	(0.09)	(0.01)
Untreated	0.00	−1.76	0.00	0.00	−0.53	6.27	−0.06	−0.00
	(0.02)	(20.70)	(0.01)	(0.01)	(0.53)	(12.95)	(0.12)	(0.01)
B. Blocks with 50% treated:								
Treated	0.01	12.65	−0.00	−0.00	−0.47	1.16	0.03	0.00
	(0.02)	(20.38)	(0.01)	(0.01)	(0.50)	(9.21)	(0.11)	(0.01)
Untreated	0.01	25.30	−0.00	−0.00	−0.42	1.88	0.02	0.01
	(0.02)	(20.66)	(0.01)	(0.01)	(0.48)	(9.66)	(0.11)	(0.01)
C. Blocks with 20% treated:								
Treated	0.02	32.57*	−0.01	0.01	0.10	5.94	0.07	−0.01
	(0.02)	(16.79)	(0.01)	(0.01)	(0.54)	(9.55)	(0.12)	(0.01)
Untreated	0.02	19.14	−0.01	−0.01	0.12	1.32	0.00	0.00
	(0.02)	(14.05)	(0.00)	(0.01)	(0.40)	(7.77)	(0.09)	(0.01)
Mean Pure Control	13.64	841.50	0.91	0.62	19.15	368.66	6.71	0.35
Observations	64,932	68,808	68,808	46,419	52,714	68,808	68,808	38,112
Number of clusters	3,979	3,981	3,981	3,973	3,976	3,981	3,981	3,968

Notes: This table shows balance test regressions to formally test for differences in observable characteristics between the treatment and control groups. Each column corresponds to a separate regression (equation (5) in the text). The dependent variables in each column are: (1) the log of assessed property value; (2) the front metres of the property; (3) an indicator for the property being a house versus a house with a store; (4) whether the tenant is male; (5) a proxy for the tenant's age (first two digits of the ID); (6) the amount paid in the bill corresponding to December 2019 (including zeroes); (7) the number of bills paid in 2019 (the maximum is 12); (8) for those who paid, whether they did so digitally. The row *Mean Pure Control* displays the constant of each regression, corresponding to the average of the dependent variable for accounts in blocks with no treated units ($T_g = 0$). Missing/non-missing indicators for the dependent variables with missing observations (columns 1, 4, 5 and 8) are also balanced between groups (results not reported). Standard errors clustered by blocks are reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01

A.3 Effects on Subscriptions to Electronic Billing

The communication campaign also included information about how to sign up for electronic billing, a system introduced in June 2020. We briefly analyze the effect of our mailing on subscription to this service.

We rely on a database that contains the individuals that signed up to the electronic billing option. This database goes through December 2020 and contains the account number, date of subscription, and email address. This source is linked with the main data through the unique account identifier.

We analyze the effect of the intervention on subscriptions to electronic billing. We present convincing graphical evidence that the tax communication campaign increased the subscriptions to receive an electronic bill by e-mail. These effects are greater in high-saturation blocks, albeit small in absolute value.

The results are summarized in Figure A.8, which follows a similar structure as Figure 6 but for e-bill subscriptions. We run dynamic difference-in-differences comparing subscription rates between each treated and each untreated group relative to pure control blocks, day by day (fixing September 27, 2020 as the baseline date).

Four important points are worth highlighting: (1) trends are generally parallel, as we estimate no significant differences between the treatment and control groups prior to the intervention; (2) the difference in subscription rates between treated accounts and pure control blocks experiences a noticeable break at the time we started sending letters, which is reassuring and implies that the effects we estimate are indeed caused by our experiment; (3) total effects are greater in high-saturation blocks with 50% and 80% treated units relative to low-saturation blocks where only 20% received the letter. As happened with payment rates, this could be interpreted as a spillover effect, whereby the intervention creates interference between treated units strengthening the effect of the letter; and (4) although less clear than the left-hand-side panels for treated units, the right-hand-side panels of Figure A.8 also suggest the presence of spillover effects in subscriptions to e-billing for untreated accounts in high-saturation blocks. As was the case with payment rates, these effects are harder to detect. They are precisely estimated but only significant at the 5% level at the beginning of the intervention.

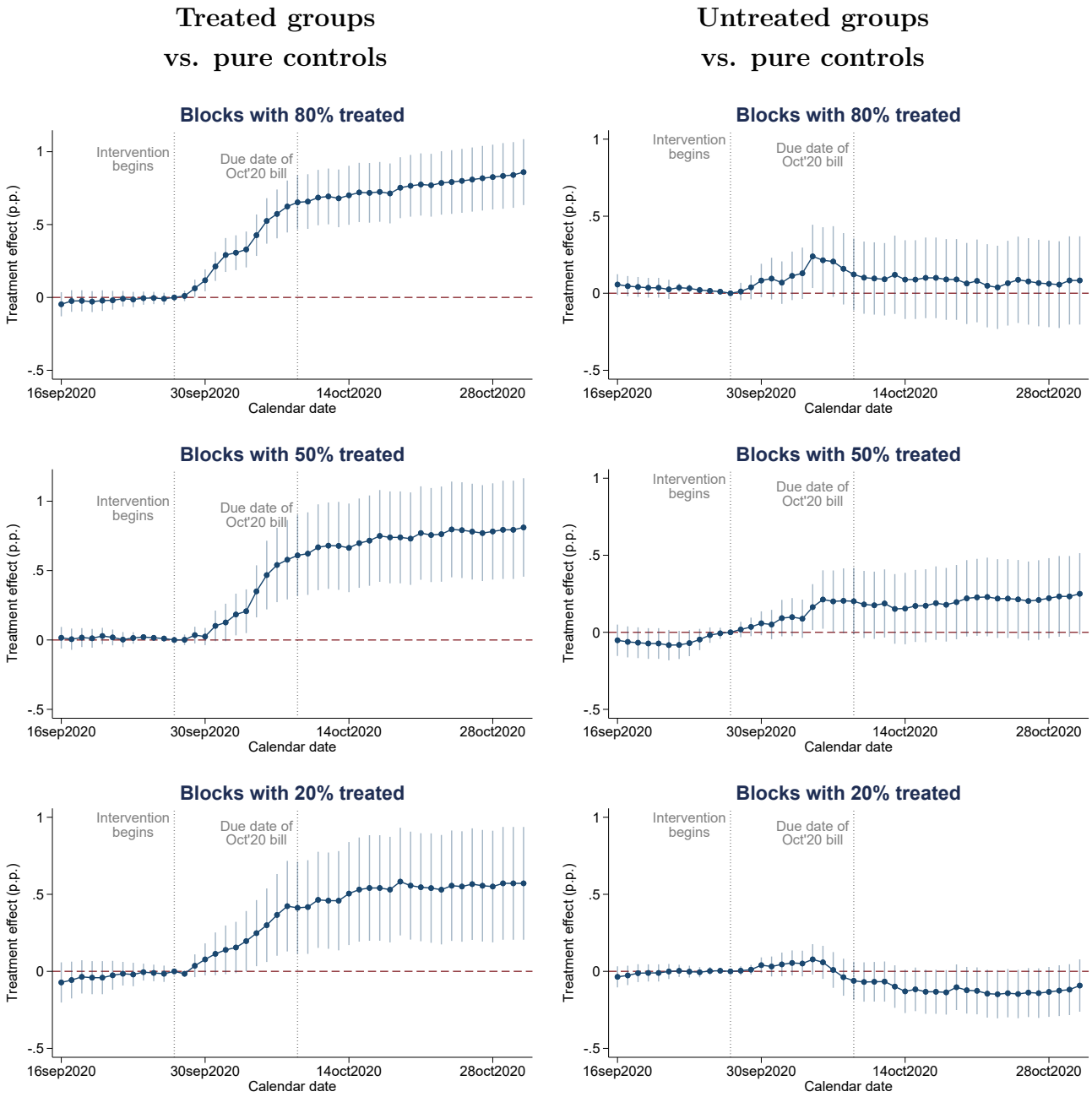
Lastly, Table A2 summarizes the corresponding diff-in-diffs estimates reported in Figures A.8, with the same structure as Table 3.[1] To benchmark our estimates, in the last row we report the share of e-bill subscribers in pure control blocks on September 27 (our baseline date). For treated accounts, the table shows an immediate effect in the three saturation groups that increases over

[1]Column (1) validates the experiment by showing a placebo saturated regression that compares subscription rates between each group and the pure control group on September 17, before the intervention began. None of the coefficients are statistically significant or large in magnitude.

time. This effect is higher in blocks with 80% treated units, consistent with interference that strengthens the effect. In such blocks, the total effect reaches 0.86 percentage points by the end of October. Although, this represents about 20% of the baseline 4.25% share of e-bill subscribers, we find it striking that so few individuals switched to the digital bill. In the case of untreated accounts, spillover effects on subscription rates are smaller and therefore much harder to detect than in the analysis of payment rates. The clearest effect arises in blocks with 50% treated accounts with a spillover effect of 0.25 percentage points, significant at the 10% level. The somewhat absence of spillovers in this case can be explained by the fact that the outcome of analysis (subscription rate) has very low take up, making it harder for interference between neighbors to emerge.

In sum, we find that our tax communication campaign also generates total effects and spillover effects among neighbors in subscriptions to electronic billing. These effects are greater in high-saturation blocks, albeit small in absolute value.

Figure A.8: Direct effects on treated accounts and spillover effects on untreated accounts (subscriptions to e-billing). Difference in differences



Notes: These figures show the coefficients and 95% confidence intervals from dynamic difference-in-differences regressions where the outcome of interest is a dummy equal to one if the account is subscribed to an electronic bill. All the coefficients are estimated with respect to September 27th, 2020 (baseline date) and relative to the pure control group (i.e., blocks where no accounts were treated). The top panel shows the effect on treated (left) and untreated (right) units in blocks with 80% treated ($T_g = 3$). The middle panel shows the effect on treated (left) and untreated (right) units in blocks with 50% treated ($T_g = 2$). The bottom panel shows the effect on treated (left) and untreated (right) units in blocks with 20% treated ($T_g = 1$). Standard errors are clustered by block. The first vertical bar denotes the start of the intervention. The due date for the October 2020 bill was October 9th and is indicated with another vertical bar. The letters were delivered between September 28th and October 7th.

Table A2: Total effects and spillover effects for subscriptions to e-billing

Dependent variable:	**Placebo:**	**Intervention:**	
Pr(subscribe to e-bill)	By Sep 20	Early	By Oct 31
	(1)	(2)	(3)
A. Blocks with 80% treated			
Treated	-0.02	0.31***	0.86***
	(0.04)	(0.06)	(0.12)
Untreated	0.04	0.11	0.08
	(0.03)	(0.08)	(0.15)
B. Blocks with 50% treated			
Treated	0.03	0.18**	0.81***
	(0.03)	(0.08)	(0.18)
Untreated	-0.07	0.10	0.25*
	(0.05)	(0.06)	(0.13)
C. Blocks with 20% treated			
Treated	-0.04	0.15*	0.57***
	(0.05)	(0.08)	(0.19)
Untreated	-0.01	0.05	-0.09
	(0.03)	(0.04)	(0.09)
Mean of Pure Control at baseline	4.25	4.25	4.25
Observations	137,612	137,612	137,612
Number of clusters (blocks)	3,981	3,981	3,981

Notes: This table shows the results from a saturated dynamic difference-in-differences regression where the dependent variable is an indicator for subscribing to electronic billing. The regression computes the outcome difference between each of the treated and untreated groups relative to the pure control group for each calendar date relative to September 27th, 2020 (baseline date). The estimates correspond exactly to the numbers shown in Figure (A.8). Column (1) shows the results for e-bill subscriptions made before the letters were delivered (placebo); Column (2) shows the results for early subscriptions right after the letters started to be delivered (by October 3); Column (3) shows the results for subscriptions made up to the end of October 2020. The letters were delivered between September 28 and October 7. The due date for the October 2020 bill was October 9th. The row *Mean of Pure Control* displays the constant of the regression, corresponding to the average susbcription rate for units in blocks with no treated units on September 27, 2020. Standard errors clustered by blocks are reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01

A.4 Timing of Payments and Due Bills

For completeness, we analyze the effects of the intervention on backward and forward payments corresponding to billing periods before and after month 10, the month of our intervention. These results are summarized in Figure A.9.

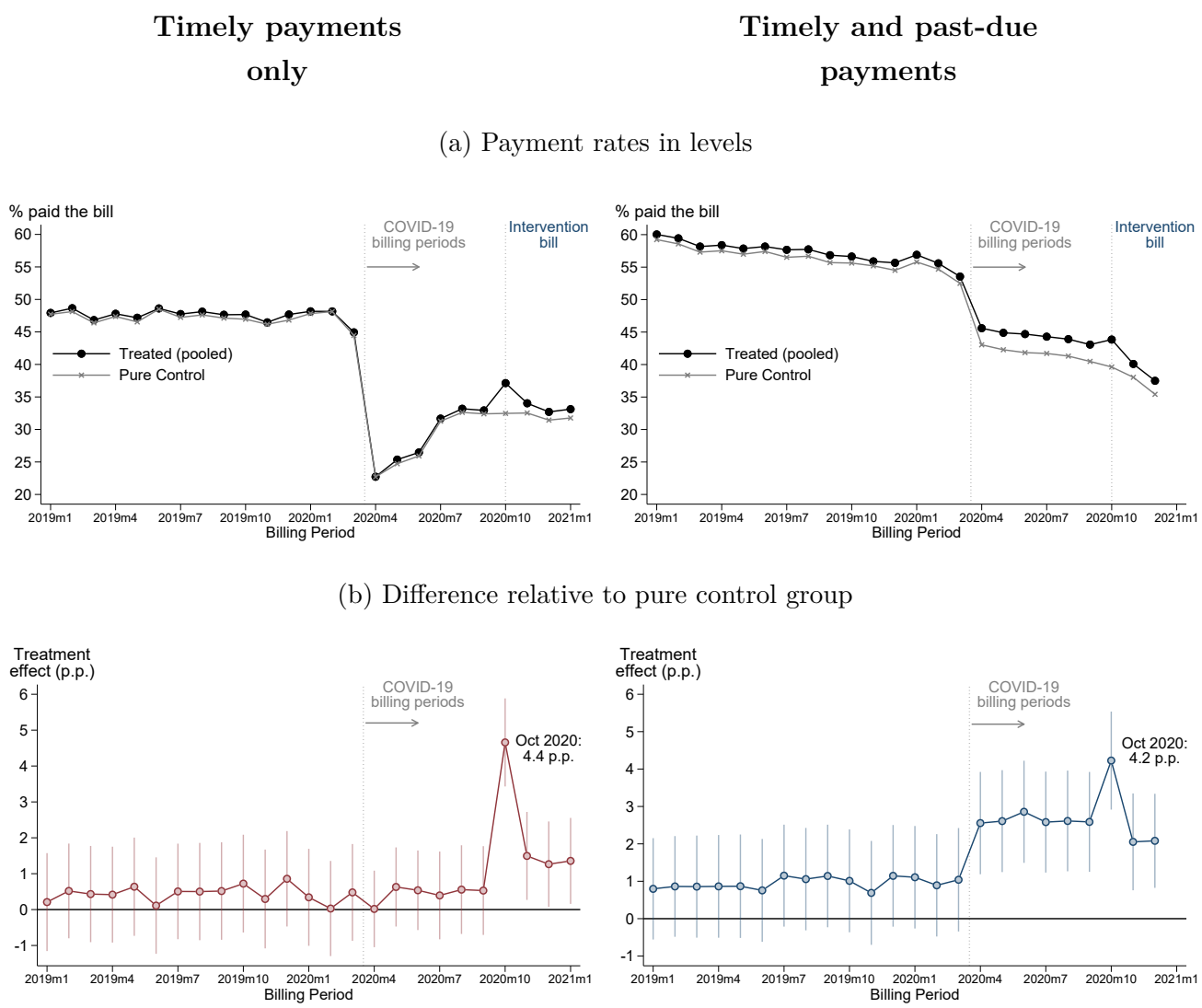
Intuitively, neighbors can pay their property tax bill at any time before or after the due date and, hence, payments from previous billing periods can also be affected by our intervention.[2] To illustrate this, the left panels of Figure A.9 only consider timely payments, defined as bills paid before the 27th of the corresponding month. We set any payment made after the 27th as unpaid in our data. Hence, pre-intervention bills mechanically exclude any past-due payment triggered by our intervention. In contrast, the right panels of Figure A.9 consider timely as well as past-due payments made until December 2020 and, thus, capture backward payments triggered by our intervention (e.g., individuals that decide to pay the October 2020 bill as well as previous unpaid bills after receiving the letter).

The top figures show payment rates in levels for treated units (black line) and pure control units (gray line), for 24 consecutive monthly bills between January 2019 and December 2020. Treated units are pooled from groups $T_g = 1, 2, 3$. The bottom figures report total treatment effects—i.e., the difference between treated and pure control units—and 95% confidence intervals for the 24 billing periods. The first vertical bar denotes the start of the COVID-19 pandemic in Argentina and the second vertical bar flags the October'20 bill targeted by our intervention.

Four important points are worth noting: (1) Overall, payment rate levels are low. The top left panel shows that about 48% of households pay their bill before the 27th of each month. This share is relatively constant until March 2020 when the COVID-19 pandemic hit Argentina and payment rates decreased sharply to 23%; (2) a similar pattern emerges when we consider timely and past-due payments. The reason why levels are higher and decrease over time is that as time goes by it is more likely that individuals cancel unpaid bills; (3) placebo direct effects (red line), based on payment rates constructed with timely payments only, are precisely estimated and not different from zero for the 21 pre-intervention bills. For the October 2020 bill, however, timely payments are 4.4 p.p. higher in treated units relative to control blocks. This is reassuring and implies that our sample is balanced and that the effects we estimate are indeed caused by our experiment; and (4) when we account for past-due payments, the blue line shows that our intervention nudged some individuals to catch up with unpaid bills. The difference in payment rates between treated and pure control accounts experiences a noticeable increase in the pandemic billing periods from April 2020 onward. Although the October bill when the intervention took place presents the highest effect (4.2 p.p.), the letters also had some residual positive effect in November and December too.

[2]The treatment letter included past due balances and could therefore induce neighbors to make backward payments to cancel debt.

Figure A.9: Total effects on pre- and post-intervention bills



Notes: These figures show the effect of the communication campaign on payment rates of pre- and post-intervention bills. The left panels only consider timely payments, defined as bills paid before the 27th of the corresponding month (i.e., any payment made after the 27th is considered unpaid). Hence, pre-intervention bills mechanically exclude any past-due payment triggered by our intervention. The right panels consider timely as well as past-due payments made until December 2020 and, thus, capture backward payments triggered by our intervention (e.g., individuals that after receiving the letter pay the October 2020 bill as well as previous unpaid bills). The top figures show payment rates in levels for treated units (black line) and pure control units (gray line), for 24 consecutive monthly bills between January 2019 and December 2020. Treated units are pooled from groups $T_g = 1, 2, 3$. The bottom figures report total treatment effects—i.e., the difference between treated and pure control units—and 95% confidence intervals for the 24 billing periods. The letters were delivered between September 28th and October 7th. The vertical bar denotes the start of the COVID-19 pandemic in Argentina. Each coefficient is estimated in separate regressions. Standard errors are clustered at the block level. The red line shows no difference on timely payments for pre-intervention bills. In contrast, when we account for past-due payments, the blue line shows that our intervention nudged some individuals to catch up with unpaid bills from April 2020 onwards.

A.5 Are Untreated Blocks Affected by the Intervention?

A crucial aspect of partial population experiments is the unit within which the experimenter will test the presence of spillovers. In some settings, these are relatively straightforward to establish: electoral precincts for political outcomes, towns for regional policies, schools or school districts for educational interventions. In our application, we aim to measure information spillovers among taxpayers. Discussions with municipal tax authorities and with taxpayers, as well as the context of our intervention, led us to select city street blocks as the relevant clusters for potential information spillovers about tax reminders and deadlines and their effects on tax compliance. Specifically, the campaign was motivated by the sharp drop in compliance in April 2020 induced by the severe lockdown imposed in the Greater Buenos Aires area in Argentina during the COVID pandemic in a context where most payments were made in person (see Figure A.9). The lockdown was strongly enforced and as a result citizens' mobility was severely limited, which justifies the choice of the city street block—a relatively small cluster—as the relevant unit for information spillover, since it reflects the limited physical interactions generated by the lockdown. A further justification is the city's street layout, which consists mainly of relatively homogeneous straight streets with orthogonal intersections in square/rectangular city blocks (see Figure A.2).

A potential concern with this setup is that the city street block may not be the relevant unit to capture information spillovers. The random assignment process and the city's physical layout imply that taxpayers in pure control street blocks (i.e., blocks where no one received a tax reminder) were still adjacent and/or surrounded by blocks with treated taxpayers, as shown by inspection of the map in Figure A.2. Interference between adjacent blocks is possible, and this would induce a downward bias in our results, since individuals in pure control (untreated) blocks would be affected by the information campaign via spillovers from adjacent (treated) blocks. Our empirical setup allows for an auxiliary test to rule out this concern, and establish that units in pure control blocks indeed provide a valid counterfactual in our analysis.[3]

To test the robustness of untreated blocks as pure controls, we leverage our experimental assignment process which implies that the "intensity" of treatment in the surrounding blocks is random by definition. Pure control units are by chance surrounded by blocks with varying degrees of treatment intensity (0%, 20%, 50%, or 80%), and thus by a random number of treated taxpayers. If there is interference between treated and untreated blocks, we should observe that pure control payment rates increase with the exposure of untreated blocks to treated blocks.

We construct our measure of the potential exposure of a street-block to the intervention in two steps. First, we use GIS software to calculate a buffer of 100, 200, and 300 meters around the centroid of each street-block (see the three figures in the top panel of Figure A.10), given the

[3]This is an auxiliary analysis in the sense that while it exploits features of our experimental design, it does not correspond to our pre-registered analysis and only represents an ex-post robustness check.

typical street block length of 100 meters. Second, for each street-block and radius, we calculate the share of properties receiving a letter (treated) relative to all the properties in the buffer zone. The three figures in the middle panel of Figure A.10 display the distribution of the share of treated units around pure control blocks.

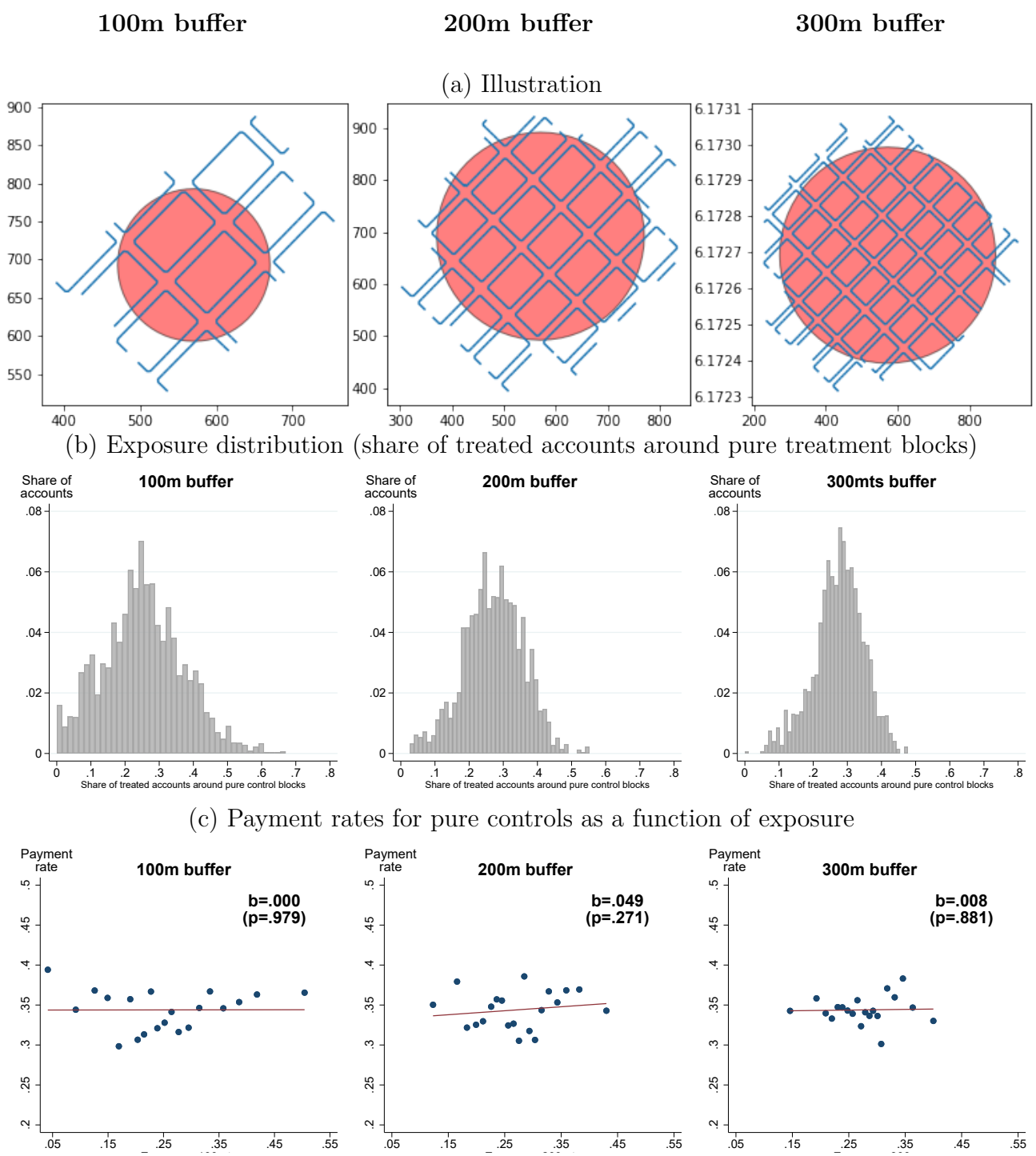
With this exposure measure at hand, we test whether payment rates of the October 2020 bill in pure control blocks increase with the exposure to the proportion of treated units in surrounding blocks. The figures in the bottom panel of Figure A.10 present parametric and non-parametric evidence of this relationship. Each panel shows a binned scatterplot of payment rates of the October 2020 bill (y-axis) by equally-sized bins of exposure to treated units within the buffer zone (x-axis). Reassuringly, the relationship is flat, and it is robust to increasing the size of the buffer zone to 200 and 300 meters. This is confirmed by the small linear regression coefficients and large p-values reported in these figures.

Our main results indicated that we only found spillover effects in our main research design for high saturation blocks with high previous compliance, as illustrated by the results in Figure 7 and Table 4. We conduct a similar analysis with the exposure measure for the 100 meters buffer in Figure A.11. The parametric and non-parametric results presented there confirm a flat gradient for untreated blocks with both high and low compliance in 2019, further confirming that untreated blocks were not affected by the intervention even when considering this relevant dimension of heterogeneity.

Finally, for completeness, we also study the relationship between payment rates and exposure to adjacent treated blocks in blocks where 80% of the units were treated, again for the 100 meters buffer. The results of this exercise are reported in Figure A.12. The left panel corresponds to the October 2020 bill affected by our intervention, whereas the middle and right panels correspond to pre-intervention bills of July and August 2020. In all these cases, the relationship between exposure and payment rates is flat and statistically not significant for both the pure control blocks (with blue dots and blue linear fit) and the 80% saturation blocks (with red triangles and a red linear fit). Interestingly, the vertical distance between the red and blue linear fit in the left panel captures the treatment effect of our experiment, which is clearly uniform in the exposure measure.

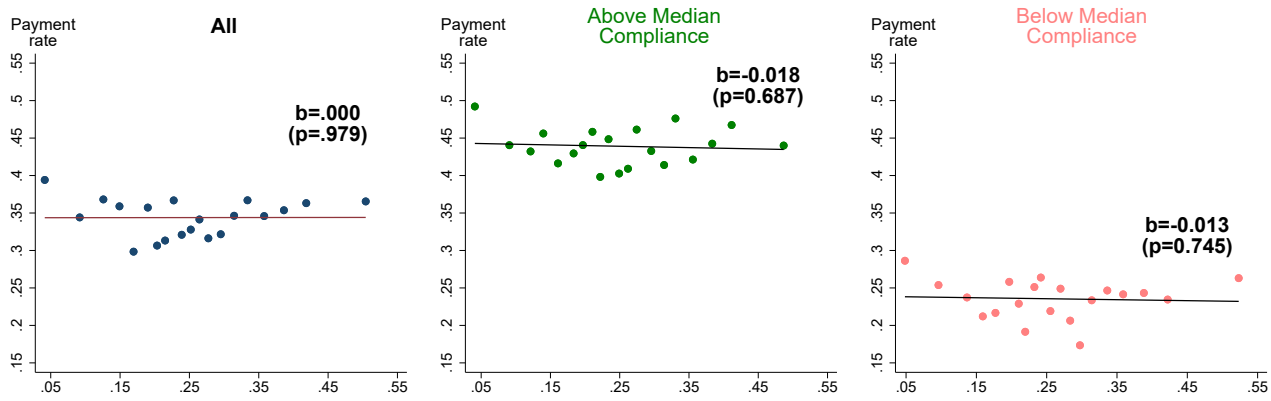
Taken together, the results from the exercise in this section indicate that pure control blocks were not affected by adjacent treated blocks, and thus provide a valid counterfactual for the analysis. In more general terms, information spillovers do not seem to have happened at a higher degree of aggregation than the city street block. When combined with the presence of information spillovers documented in the main body of the paper, the city street block seems to have been the relevant level of information dissemination for this campaign.

Figure A.10: Robustness of untreated blocks as pure controls



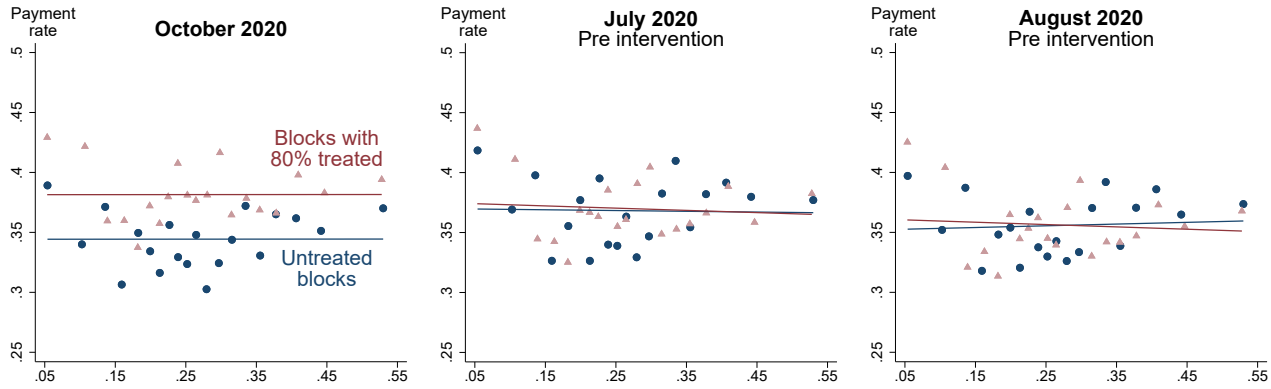
Notes: The top three panels illustrate the way we compute buffer zones around the centroid of each street-block using GIS tools in our data. We consider radiuses of 100 meters (left panel), 200 meters (middle panel), and 300 meters (right panel). The middle panel three figures show the distribution of accounts in pure control street-blocks according to their exposure to treated accounts. The bottom three panels show binned scatterplots of payment rates of the October 2020 bill (y-axis) in pure control blocks and their exposure to treated units within the buffer zone (x-axis). The x-axis is grouped into equally-sized bins. The coefficient and p-value of each regression are also reported in each panel. The regressions flexibly control for a cubic polynomial of the number of properties in the buffer zone. This variable is highly correlated with payment rates and its omission leads to omitted variable bias.

Figure A.11: Payment rates and exposure of untreated blocks above and below median 2019 compliance, 100 meters buffer



Notes: This figure shows binned scatterplots of payment rates (y-axis) in pure control blocks by equally-sized bins of exposure to treated units within a buffer zone of 100 meters (x-axis). The left panel replicates the bottom left panel of Figure A.10. The middle and right panels split pure control blocks into blocks with above- and below-median compliance defined in 2019, respectively. The regressions flexibly control for a cubic polynomial of the number of properties in the buffer zone. This variable is highly correlated with payment rates and its omission leads to omitted variable bias.

Figure A.12: Payment rates and exposure of untreated blocks and blocks with 80% treated units, 100 meters buffer



Notes: This figure shows binned scatterplots of payment rates (y-axis) by equally-sized bins of exposure to treated units within a buffer zone of 100 meters (x-axis). The left panel shows the gradient in both untreated blocks (blue dots) and blocks with 80% treated units (red triangles) for the October 2020 bill (the one affected by the intervention). The middle and right panels correspond to the pre-intervention bills of July and August 2020, respectively. The regressions flexibly control for a cubic polynomial of the number of properties in the buffer zone. This variable is highly correlated with payment rates and its omission leads to omitted variable bias.

B Experimental Design: Additional Material

B.1 Choice of q_tnd power calculations

For simplicity, we assume that the assignment probabilities are the same across groups and that treatment is assigned independently within groups. The "hardest" effect to estimate correspond to the assignments $(d,t) = (1,1)$, i.e. treated in 20% groups, and $(d,t) = (0,3)$, i.e. controls in 80% groups. To ensure the variance of these estimators is similar to the variance of the $(d,t) = (0,2)$ estimator, and using that $q_1 = q_3$, we need:

$$\frac{\sigma^2(0,3)}{0.2q_3}\left\{1 + 0.2\rho_{03,03}\left(\frac{\bar{n}_2}{\bar{n}} - 1\right)\right\} = \frac{\sigma^2(0,2)}{0.5q_2}\left\{1 + 0.5\rho_{02,02}\left(\frac{\bar{n}_2}{\bar{n}} - 1\right)\right\}.$$

where $\bar{n}_2 = \sum_g n_g^2$. We will assume that all the variances are the same, $\sigma^2(0,3) \approx \sigma^2(0,2) = \sigma^2$ and that all the intraclass correlations are the same and equal to 0.1, which is slightly larger than the one estimated for the baseline data. After some simplifications we have that:

$$q_2\left\{1 + 0.02\left(\frac{\bar{n}_2}{\bar{n}} - 1\right)\right\} = 0.4q_3\left\{1 + 0.05\left(\frac{\bar{n}_2}{\bar{n}} - 1\right)\right\}.$$

Using the sample sizes from the baseline data and setting $L = 25,000$ gives the assignment probabilities shown below:

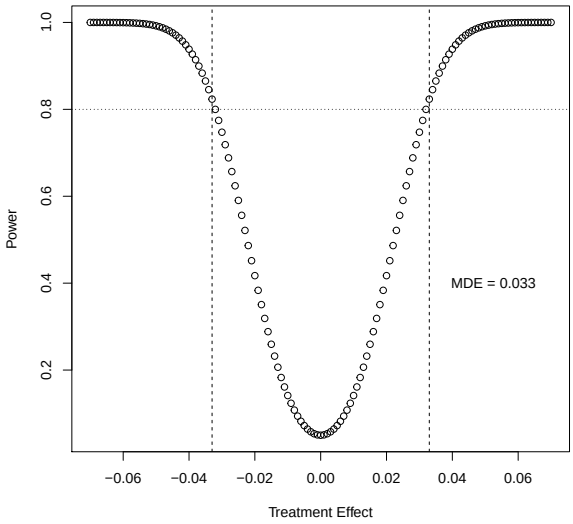
$$\{q_0, q_1, q_2, q_3\} = \{0.273, 0.282, 0.162, 0.282\}$$

The final sample sizes depicted in Table 1 respond to logistical and other practical considerations. For power calculations, Figure B.13 plots the power function for each estimator, using the following parameters:

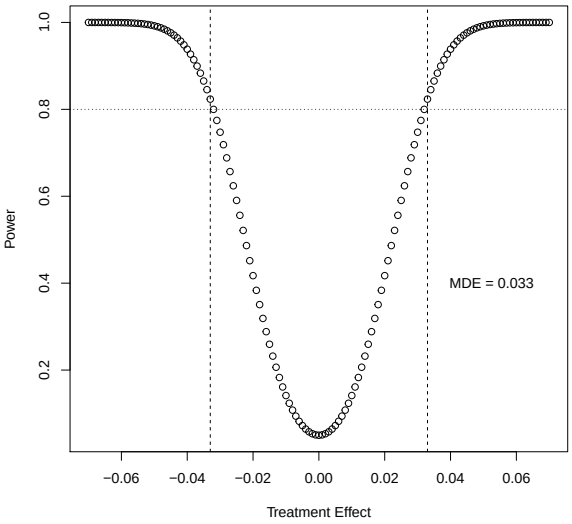
- $\sigma^2(d,t) = 0.25$ for all (d,t). This gives a conservative estimate because 0.25 is the upper bound for the variance of a binary variable.
- ICC $= 0.1$ which is close to (but larger than) the estimated intraclass correlation of the baseline outcome.
- The sample and group sizes given by the baseline data.

The power calculations give a minimum detectable effect between 2.6 and 3.3 percentage points.

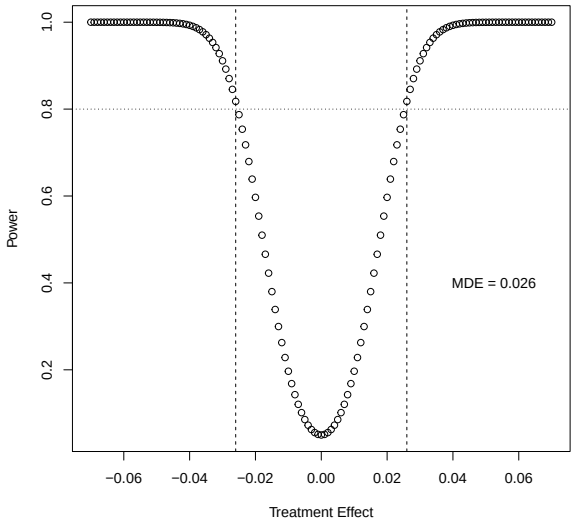
Figure B.13: Power functions



(a) $(d,t) = (0,3)$ or $(d,t) = (1,1)$



(b) $(d,t) = (0,2)$ or $(d,t) = (1,2)$



(c) $(d,t) = (1,3)$ or $(d,t) = (0,1)$

Due to logistical restrictions, our final sample sizes had to be adjusted. We report our effective sample sizes in Table 1 in the main paper. It is important to clarify that, given our large sample size, this adjustment had a negligible effect on power and MDEs.

B.2 Simulations for power calculations

We conduct a simulation study to confirm our analytical power calculations. We assume $(T_1, T_2, \ldots, T_G)$ are iid with distribution: $\mathbb{P}[T_g = t] = q_t$ and the variable is constructed as:

$$T_g = \mathbb{1}(q_0 < U_g \leq q0 + q_1) + 2\mathbb{1}(q_0 + q_1 < U_g \leq q_0 + q_1 + q_2) + 3\mathbb{1}(U_g > q_0 + q_1 + q_2)$$

with $U_g \sim \mathsf{Uniform}(0,1)$. The individual treatment indicator is assigned according to the rule:

$$D_{ig} = \mathbb{1}(U_{ig}^1 \leq 0.2)\mathbb{1}(T_g = 1) + \mathbb{1}(U_{ig}^2 \leq 0.5)\mathbb{1}(T_g = 2) + \mathbb{1}(U_{ig}^3 \leq 0.8)\mathbb{1}(T_g = 3)$$

where $U_{ig}^k \sim \mathsf{Uniform}(0,1)$ for $k = 1, 2, 3$, independent of each other.

We construct seven potential outcomes $Y_{ig}(d,t)$ for $d = 0, 1$ and $t = 0, 1, 2, 3$. Based on the baseline June 2019 outcome Y_{ig}^{base}, the potential outcomes are constructed in the following way:

$$\begin{aligned} Y_{ig}(0,0) &= Y_{ig}^{base} \\ Y_{ig}(d,t) &= \mathbb{1}(U_{dt} \leq c_{dt})(1 - Y_{ig}(0,0)) + \mathbb{1}(\tilde{U}_{dt} \leq c_{dt} + k)Y_{ig}(0,0) \end{aligned}$$

for $(d,t) \neq (0,0)$, where U_{dt} and $\tilde{U}_{dt}$ are independent uniforms. According to this model,

$$\begin{aligned} \mathbb{E}[Y_{ig}(0,0)] &= \mu_0 \\ \mathbb{E}[Y_{ig}(d,t)] &= c_{dt} + \mu_0 k \\ \mathbb{C}\mathrm{ov}(Y_{ig}(0,0), Y_{ig}(d,t)) &= k\mu_0(1-\mu_0) \end{aligned}$$

Therefore, we can set:

$$c_{0t} = \theta_t + \mu_0(1-k), \quad c_{1t} = \tau_t + \mu_0(1-k)$$

and

$$k = \frac{\rho}{\mu_0(1-\mu_0)}$$

where ρ is some specified level for the covariance.

Finally, we set $\mu_0 = \bar{Y}^{base} \approx 0.568$ and $\rho = 0.2$. A value of $\rho = 0.2$ implies a correlation between $Y_{ig}(0,0)$ and $Y_{ig}(d,t)$ between 0.6 and 0.8. The implied intraclass correlation for all potential outcomes is approximately $\mathsf{ICC} = 0.05$.

In each simulation, we use the baseline outcome from June 2019 as the potential outcome for pure controls, and construct the remaining potential outcomes adding the corresponding direct or spillover effects. See the appendix for details. The results are shown in Table A3. The last parameter is set to zero to simulate the probability of type I error.

Table A3: Simulation results

	True value	Prob(reject)
θ_1	0.021	0.812
θ_2	0.026	0.798
θ_3	0.027	0.791
τ_1	0.028	0.801
τ_2	0.026	0.800
τ_3	0.000	0.045

The simulation results are in line with the analytical calculations in the previous section, with slightly lower MDEs because some statistics such as the ICC are in fact lower in the sample. The last row in the table confirms that the probability of incorrectly rejecting the null of no effect is around 5%, as expected.

C Why is Cluster Heterogeneity Important?

C.1 A Numerical Illustration

One of our main methodological contributions is to provide variance and power formulas that account for cluster size heterogeneity. In field experiments with clustered designs, cluster sizes typically vary substantially. For instance, electoral precincts, towns, schools or school districts, typically have different numbers of voters, population or students. In our main application, our tax information campaign reaches city street blocks with a wide range of number of taxpayers—from 8 in the smallest blocks to 50 in the largest (see Figure 3). When clusters vary in size, the variance of treatment effect estimators requires an adjustment factor that depends on the average and the variance of cluster size. Ignoring this adjustment factor underestimates the variance of treatment effect estimators, which in turn results in overestimating power and underestimating MDEs. As we show in Section 2, this problem becomes more serious the larger (i) the ratio of the variance to average cluster size, (ii) the intracluster correlation in outcomes, and (iii) the intracluster correlation of treatment assignments.

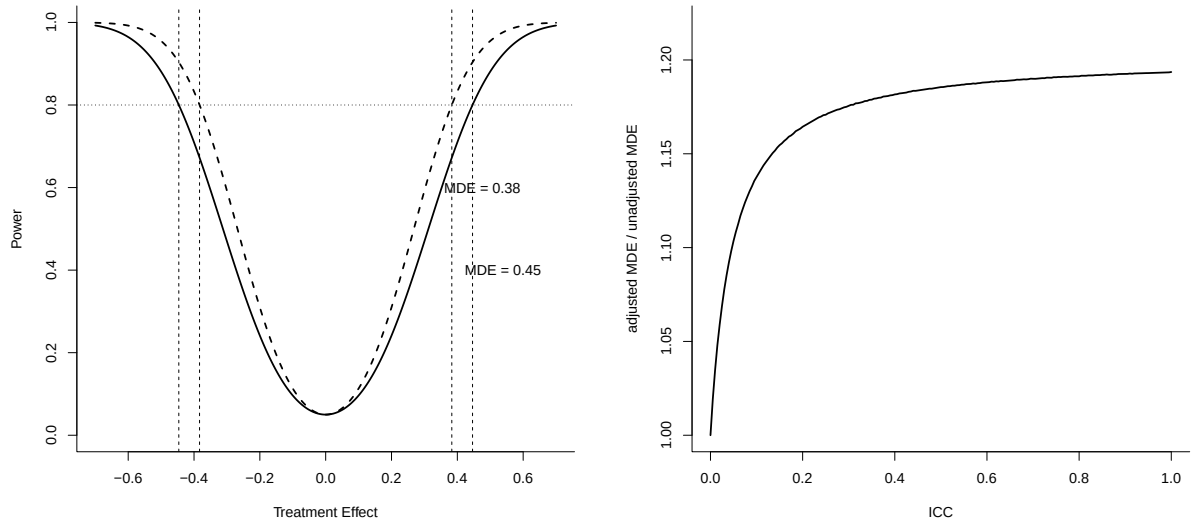
We illustrate this issue in a hypothetical partial population experiment with $G = 95$ clusters with an average cluster size of $\bar{n} = 23.3$ and a standard deviation of cluster size of $sd(n_g) = 15.2$ and an intracluster correlation in potential outcomes of ICC $= 0.2$. These are the median values from four studies in the empirical literature using partial population experiments (see Section C.2 below for further details). Specifically, we use the formulas we derive in Section 2 to calculate power and MDEs accounting for cluster size heterogeneity and compare them to the unadjusted ones that would be obtained if incorrectly ignoring cluster size heterogeneity.

The results from this numerical exercise are shown in Figure C.14. Panel (a) shows the power functions and corresponding MDEs. In this case, accounting for cluster size heterogeneity results in an MDE that is about 18% larger than the unadjusted one. Panel (b) shows the ratio of the adjusted to the unadjusted MDEs as a function of the intracluster correlation in outcomes, and shows that even for low values of intracluster correlation, the adjusted MDE can be substantially larger than the unadjusted one, with this difference increasing with larger intracluster correlations. This example illustrates that, when designing a partial population experiment, ignoring cluster size heterogeneity typically results in power and MDE calculations that are overly optimistic.

C.2 Details of Numerical Illustration

Table A4 summarizes the distribution of group sizes in four published studies employing partial population designs: Giné and Mansuri (2018), Haushofer and Shapiro (2016), Ichino and Schündeln (2012) and Imai,

Figure C.14: Adjusted and unadjusted power functions and MDEs.



(a) Power functions

(b) Ratio of adjusted to unadjusted MDE

Notes: Panel (a) shows the unadjusted (dashed line) and adjusted (solid line) power functions and their corresponding MDEs at 80% power. Panel (b) shows the ratio of the adjusted to the unadjusted MDEs as a function of the intracluster correlation (ICC) in outcomes. Adjusted magnitudes account for cluster size variability. Unadjusted magnitudes assume no group size variability, i.e. zero variance of cluster size. Calculations use the following values: $G = 95$, $\bar{n} = 23.3$, $sd(n_g) = 15.2$, ICC $= 0.2$, the median values from Table A4.

Jiang and Malani (2021).

For our numerical illustration, we calculate the estimators standard errors and minimum detectable effects based on our formulas from Section 2 using the cluster size distribution of these four studies. We refer to these magnitudes as "adjusted" standard errors and MDEs, since they are adjusted for group size variation. For comparison, we also calculate the "unadjusted" standard errors and MDEs using average cluster size and assuming that the variance of group size is equal to zero, that is, ignoring cluster size heterogeneity. To make the results comparable, we use as a benchmark the design in our application to tax compliance, which has four saturations: $p_0 = 0$, $p_1 = 0.2$, $p_2 = 0.5$, $p_3 = 0.8$. We compute the optimal probabilities $\{q_0, q_1, q_2, q_3\}$ using Theorem 3. We assume for simplicity that outcomes are homoskedastic with $\sigma^2(dt, dt) = 1$ for all d, t so that effects are measured in standard deviations, and consider four values for the intraclass correlation, $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The parameter of interest is the spillover effect on untreated units in groups with 80% treated.

The results are shown in Table A5. When the intraclass correlation is low ($\rho = 0.1$), accounting for group size heterogeneity increases standard errors and MDEs between 6.8% and 14.5%. The problem worsens for larger intraclass correlations. When $\rho = 0.5$, adjusted standard errors and MDEs are between 8.3% and 19.6% larger, and between 8.5% and 20.2% larger when $\rho = 0.8$.

Figure C.14 plots the ratio of adjusted to unadjusted standard errors and the adjusted and unadjusted MDEs as a function of the intraclass correlation using the median values from Table A4. The ratio of MDEs

Table A4: Sample sizes in existing literature

	Sample size	No. of groups	Ave. group size	Sd. group size
Giné and Mansuri (2018)	2,736	67	39.4	16.7
Haushofer and Shapiro (2016)	1,440	123	23.4	14.8
Ichino and Schündeln (2012)	868	39	22.3	9.6
Imai, Jiang and Malani (2021)	10,030	434	23.1	15.5
Mean	3,769	165.8	27.05	14.2
Median	2,088	95	23.3	15.2

Table A5: Numerical results

	Standard error			**MDE**		
	Adj.	Unadj.	Ratio	Adj.	Unadj.	Ratio
$\rho = 0.1$						
GM	0.1262	0.1181	1.0687	0.3536	0.3308	1.0689
HS	0.1053	0.0932	1.1307	0.2951	0.2610	1.1307
IS	0.1768	0.1667	1.0608	0.4954	0.4670	1.0608
IJM	0.0569	0.0497	1.1453	0.1595	0.1393	1.1450
$\rho = 0.2$						
GM	0.1695	0.1573	1.0773	0.4749	0.4408	1.0774
HS	0.1389	0.1201	1.1561	0.3892	0.3367	1.1559
IS	0.2300	0.2142	1.0736	0.6445	0.6003	1.0736
IJM	0.0752	0.0640	1.1737	0.2106	0.1794	1.1739
$\rho = 0.5$						
GM	0.2593	0.2393	1.0835	0.7265	0.6705	1.0835
HS	0.2098	0.1783	1.1761	0.5877	0.4997	1.1761
IS	0.3437	0.3171	1.0840	0.9630	0.8884	1.0840
IJM	0.1136	0.0950	1.1961	0.3183	0.2661	1.1962
$\rho = 0.8$						
GM	0.3252	0.2997	1.0851	0.9112	0.8397	1.0851
HS	0.2622	0.2218	1.1818	0.7345	0.6215	1.1818
IS	0.4284	0.3941	1.0869	1.2002	1.1042	1.0869
IJM	0.1420	0.1181	1.2024	0.3979	0.3309	1.2025

increases rapidly for values of ρ, and stabilizes between 1.15 and 1.2, suggesting that even for moderate intraclass correlations, the adjustment factor due to group size heterogeneity may be substantial. Panel (b) shows how the difference between adjusted and unadjusted MDEs becomes larger as the intraclass correlation grows.

D Supplemental Econometric Appendix

D.1 Within-Group Assignment Mechanisms

D.1.1 Fixed Margins

The within-group treatment is often assigned by choosing a fixed number of treated units within each group. Given $T_g = t$, suppose the researcher wants to assign a proportion p_t of, or a total of $n_g p_t$, units to treatment. Assigning exactly $n_g p_t$ units to treatment is not possible when $n_g p_t$ is not an integer. We propose the following procedure to deal with this issue. Define a binary random variable ξ_g and let:

$$N_g^1 = \lfloor n_g p_t \rfloor + \xi_g \mathbb{1}(n_g p_t \notin \mathbb{N}).$$

so that ξ_g plays the role of an adjusting factor that randomly rounds the number of treated up or down. Suppose that, given $T_g = t$, the probability that $\xi_g = 1$ is:

$$\mathbb{P}_g[\xi_g = 1 | T_g = t] = \begin{cases} 0 & \text{if } n_g p_t \in \mathbb{N} \\ n_g p_t - \lfloor n_g p_t \rfloor & \text{if } n_g p_t \notin \mathbb{N}. \end{cases}$$

This implies that, given $T_g = t$, the expected number of treated units in group g is $n_g p_t$ and that $\mathbb{P}_g[D_{ig} = 1 | T_g = t] = p_t$. This implies that, given $T_g = t$, the expected number of treated units in group g is $n_g p_t$ and that $\mathbb{P}_g[D_{ig} = 1 | T_g = t] = p_t$. More precisely,

$$\begin{aligned} \mathbb{E}[N_g^1 | T_g = t] &= \lfloor n_g p_t \rfloor + \mathbb{E}[\xi_g | T_g = t] \mathbb{1}(n_g p_t \notin \mathbb{N}) \\ &= \lfloor n_g p_t \rfloor + (n_g p_t - \lfloor n_g p_t \rfloor) \mathbb{1}(n_g p_t \notin \mathbb{N}) \\ &= n_g p_t \end{aligned}$$

using that $\lfloor n_g p_t \rfloor = n_g p_t$ when $n_g p_t \in \mathbb{N}$. It follows that:

$$\mathbb{E}\left[\frac{N_g^1}{n_g} \middle| T_g = t \right] = \mathbb{P}_g[D_{ig} = 1 | T_g = t] = p_t$$

which doesn't vary across groups conditional on $T_g = t$. On the other hand, defining $N_g^0 = n_g - N_g^1$, we have that:

$$\mathbb{E}\left[\frac{N_g^0}{n_g} \middle| T_g = t \right] = \mathbb{P}_g[D_{ig} = 0 | T_g = t] = 1 - p_t.$$

Next, for this assignment mechanism,

$$\begin{aligned}\mathbb{P}_g[D_{ig}=1, D_{jg}=1|T_g=t] &= \mathbb{E}\left[\frac{N_g^1}{n_g}\left(\frac{N_g^1-1}{n_g-1}\right)\middle| T_g=t\right] \\ &= \frac{\mathbb{E}[(N_g^1)^2|T_g=t]-\mathbb{E}[N_g^1|T_g=t]}{n_g(n_g-1)}\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}[(N_g^1)^2|T_g=t] &= \mathbb{E}[(\lfloor n_g p_t\rfloor + \xi_g \mathbb{1}(n_g p_t \notin \mathbb{N}))^2|T_g=t] \\ &= n_g^2 p_t^2 \mathbb{1}(n_g p_t \in \mathbb{N}) \\ &+ \left((\lfloor n_g p_t\rfloor+1)^2 \mathbb{P}_g[\xi_g=1|T_g=t] + \lfloor n_g p_t\rfloor^2 \mathbb{P}_g[\xi_g=0|T_g=t]\right)\mathbb{1}(n_g p_t \notin \mathbb{N}) \\ &= n_g^2 p_t^2 \mathbb{1}(n_g p_t \in \mathbb{N}) \\ &+ \left((\lfloor n_g p_t\rfloor+1)^2 (n_g p_t - \lfloor n_g p_t\rfloor) + \lfloor n_g p_t\rfloor^2 (1-n_g p_t - \lfloor n_g p_t\rfloor)\right)\mathbb{1}(n_g p_t \notin \mathbb{N}).\end{aligned}$$

Similarly,

$$\mathbb{P}_g[D_{ig}=0, D_{jg}=0|T_g=t] = \frac{\mathbb{E}[(N_g^0)^2|T_g=t]-\mathbb{E}[N_g^0|T_g=t]}{n_g(n_g-1)}$$

where

$$\mathbb{E}[(N_g^0)^2|T_g=t] = \mathbb{E}[(n_g-N_g^1)^2|T_g=t] = n_g^2 + \mathbb{E}[(N_g^1)^2|T_g=t] - 2n_g^2 p_t$$

Notice that even if $\mathbb{P}_g[D_{ig}=d|T_g=t]$ does not change across g, the joint probabilities do. Nevertheless, these terms can be calculated for any sample using the chosen probabilities p_t and the group sizes $\{n_g\}_{g=1}^G$.

D.1.2 Bernoulli Trials

Alternatively, the within-group treatment may be assigned to each unit independently as a "coin flip" with probability p_t. Under this mechanism, independence between treatment indicators implies that:

$$\begin{aligned}\mathbb{P}_g[D_{ig}=1|T_g=t] &= \mathbb{P}[D_{ig}=1|T_g=t] = p_t \\ \mathbb{P}_g[D_{ig}=d, D_{jg}=d|T_g=t] &= \mathbb{P}[D_{ig}=d|T_g=t]^2.\end{aligned}$$

which do not vary over g. It follows that:

$$\frac{\sum_g n_g(n_g-1)\mathbb{P}_g[D_{ig}=d, D_{jg}=d|T_g=t]}{\sum_g n_g \mathbb{P}_g[D_{ig}=d|T_g=t]} = p_t^d(1-p_t)^{1-d}\left(\frac{\sum_g n_g^2}{n}-1\right)$$

Then the variances are approximated by:

$$\mathbb{V}[\hat{\beta}_{0t}] \approx \frac{\sigma^2(0t)}{nq_t(1-p_t)}\left\{1+\rho_{0t}(1-p_t)\left(\frac{\sum_g n_g^2}{n}-1\right)\right\} + \frac{\sigma^2(00)}{nq_0}\left\{1+\rho_{00}\left(\frac{\sum_g n_g^2}{n}-1\right)\right\}$$

and

$$\mathbb{V}[\hat{\beta}_{1t}] \approx \frac{\sigma^2(1t)}{nq_tp_t}\left\{1+\rho_{1t}p_t\left(\frac{\sum_g n_g^2}{n}-1\right)\right\}+\frac{\sigma^2(00)}{nq_0}\left\{1+\rho_{00}\left(\frac{\sum_g n_g^2}{n}-1\right)\right\}.$$

E Proofs

E.1 Setup and Definitions

Following the notation in the paper, consider clusters $g = 1, \ldots, G$ with cluster size n_g, units $i = 1, \ldots, n_g$ and total sample size $n = \sum_g n_g$. The cluster-level treatment assignment is $T_g \in \{0, \ldots, M\}$ with $\mathbb{P}[T_g = t] = q_t$, and the individual-level treatment indicator D_{ig} with $\mathbb{P}[D_{ig} = d | T_g = t] = p_g(d|t)$. Let $A_{ig} = (T_g, D_{ig})$, $\mathbf{A}_g = (D_{1g}, \ldots, D_{n_g g}, T_g)'$ and $\mathbf{A} = (\mathbf{A}_1', \ldots, \mathbf{A}_G')'$.

Within each cluster, the total number of units receiving treatment $D_{ig} = d$ is $N_g^d = \sum_i \mathbb{1}(D_{ig} = d)$, and conditional on $N_g^d > 0$, the within-cluster average outcome under $D_{ig} = d$ is $\bar{Y}_{ig}^d = \sum_{i=1}^{n_g} Y_{ig} \mathbb{1}(D_{ig} = d)/N_g^d$.

Letting $\mathbb{1}_{ig}^{dt} = \mathbb{1}(D_{ig} = d, T_g = t)$ and $= (\mathbb{1}_{ig}^{dt})_{(d,t)}'$, the vector of OLS estimators for the sample means is:

$$\hat{\boldsymbol{\mu}}_n = \left(\sum_g \mathbb{1}_g' \mathbb{1}_g \right)^{-1} \sum_g \mathbb{1}_g' \mathbf{Y}_g = (\mathbf{N})^{-1} \sum_g \mathbb{1}_g' \mathbf{Y}_g$$

where $\mathbf{N} = diag(N(d,t))_{(d,t)}$ is a diagonal matrix with entries $N(d,t) = \sum_g \mathbb{1}_g^t N_g^d$ and where $\mathbb{1}_g^t = \mathbb{1}(T_g = t)$.

Also define $\mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t] = \mu_g(d,t)$, $\mathbb{V}[Y_{ig}|D_{ig} = d, T_g = t] = \sigma_g^2(d,t)$, $\mathbb{C}\text{ov}(Y_{ig}, Y_{jg}|D_{ig} = d, D_{jg} = d', T_g = t) = c_g(d,d',t)$ with $\mathbb{C}\text{ov}(Y_{ig}, Y_{jg}|D_{ig} = d, D_{jg} = d, T_g = t) = c_g(d,t)$ for brevity and similarly $\rho_g(d,d',t) = c_g(d,d',t)/(\sigma_g(d,t)\sigma_g(d',t))$ and $\rho_g(d,t) = \rho_g(d,d,t)$. Finally, let $p_g(d,d'|t) = \mathbb{P}[D_{ig} = d, D_{jg} = d'|T_g = t]$.

E.2 Auxiliary Results

E.2.1 Convergence of Sample Sizes

$$\frac{\mathbf{N}}{n} \times \mathbb{E}\left[\frac{\mathbf{N}}{n}\right]^{-1} \to_{\mathbb{P}} I_{2M-1}.$$

where $\mathbb{E}[\mathbf{N}/n] = diag\left(q_t \sum_g n_g \pi_g(d|t)/n\right)_{(d,t)}$.

Proof. For any (d,t),

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{n}\sum_g \mathbb{1}_g^t N_g^d\right] &= \frac{1}{n^2}\sum_g \mathbb{V}[\mathbb{1}_g^t N_g^d] = \frac{1}{n^2}\sum_g \left\{\mathbb{V}\left[\mathbb{1}_g^t \mathbb{E}[N_g^d|T_g]\right] + \mathbb{E}\left[\mathbb{1}_g^t \mathbb{V}[N_g^d|T_g]\right]\right\} \\
&= \frac{1}{n^2}\sum_g \left\{n_g^2 p_g(d,t)^2 q_t(1-q_t) + q_t n_g p_g(d|t)(1-p_g(d|t)) + q_t n_g(n_g-1)\left(p_g(d,d|t) - p_g(d|t)^2\right)\right\} \\
&= q_t(1-q_t)\sum_g \frac{n_g^2}{n^2} p_g(d,t)^2 + q_t \sum_g \frac{n_g}{n^2} p_g(d|t)(1-p_g(d|t)) + q_t \sum_g \frac{n_g(n_g-1)}{n^2}\left(p_g(d,d|t) - p_g(d|t)^2\right) \\
&= O\left(\frac{\sum_g n_g^2}{n^2}\right) = o(1).
\end{aligned}$$

since $\sum_g n_g^2/n^2 \leq \max_g n_g/n \to 0$. Therefore, by Markov's inequality

$$\begin{aligned}
\mathbb{P}\left[\left\|\frac{\mathbf{N}}{n} \times \mathbb{E}\left[\frac{\mathbf{N}}{n}\right]^{-1} - I_{2M-1}\right\| > \varepsilon^2\right] &= \mathbb{P}\left[\sum_{d,t}\left(\frac{N(d,t)/n}{\mathbb{E}[N(d,t)/n]} - 1\right)^2 > \varepsilon^2\right] \\
&\leq \sum_{d,t} \mathbb{P}\left[\left|\frac{1}{n}\sum_g \mathbb{1}_g^t N_g^d - \mathbb{E}\left[\frac{1}{n}\sum_g \mathbb{1}_g^t N_g^d\right]\right| > \frac{\varepsilon}{\sqrt{2M-1}}\frac{\mathbb{E}[N(d,t)]}{n}\right] \\
&\leq \frac{(2M-1)^2}{\varepsilon^2}\sum_{d,t}\left(\frac{n}{q_t \sum_g n_g p_g(d|t)}\right)^2 \mathbb{V}\left[\frac{1}{n}\sum_g \mathbb{1}_g^t N_g^d\right] \\
&\leq \frac{(2M-1)^3}{\varepsilon^2}\cdot\frac{1}{c}\cdot\max_{d,t}\mathbb{V}\left[\frac{1}{n}\sum_g \mathbb{1}_g^t N_g^d\right] \to 0
\end{aligned}$$

using that $\sum_g n_g p_g(d|t)/n$ is bounded below.

E.2.2 Moments of $\bar{Y}_g^d$

$$\mathbb{1}_g^t \mathbb{E}[\bar{Y}_g^d | T_g = t, \mathbf{D}_g] = \mathbb{1}_g^t \mu_g(d,t),$$

$$\begin{aligned}
\mathbb{1}_g^t \mathbb{V}[\bar{Y}_g^d | T_g = t, \mathbf{D}_g] &= \frac{\mathbb{1}_g^t}{(N_g^d)^2}\left\{\sum_i \mathbb{V}[Y_{ig}|T_g = t, D_{ig} = d]D_{ig} + 2\sum_i\sum_{j>i} D_{ig}D_{jg}\mathbb{C}\text{ov}(Y_{ig}, Y_{jg}|D_{ig}=1, D_{jg}=1)\right\} \\
&= \frac{\mathbb{1}_g^t}{N_g^d}\sigma_g^2(d,t) + 2c_g(d,t)\mathbb{1}_g^t \sum_i\sum_{j>i}\frac{D_{ig}D_{jg}}{(N_g^d)^2}
\end{aligned}$$

and

$$\mathbb{E}\left[\mathbb{1}_g^t (N_g^d)^2 \mathbb{V}[\bar{Y}_g^d | T_g = t, \mathbf{D}_g]\right] = \sigma_g^2(d,t) q_t n_g p_g(d|t) + c_g(d,t) n_g(n_g - 1) q_t p_g(d,d|t)$$

Proof. By direct calculation, letting $\mathbb{1}_{ig}^d = \mathbb{1}(D_{ig} = d)$,

$$\begin{aligned}
\mathbb{1}_g^t \mathbb{E}[\bar{Y}_g^d | T_g = t, \mathbf{D}_g] &= \mathbb{1}_g^t \mathbb{E}\left[\frac{1}{N_g^d}\sum_i Y_{ig}\mathbb{1}_{ig}^d \middle| T_g = t, \mathbf{D}_g\right] = \frac{\mathbb{1}_g^t}{N_g^d}\sum_i \mathbb{E}[Y_{ig}|T_g = t, \mathbf{D}_g]\mathbb{1}_{ig}^d \\
&= \mathbb{1}_g^t \mu_g(d,t)
\end{aligned}$$

where the last equality follows from Assumption 2. Similarly,

$$\begin{aligned}
\mathbb{1}_g^t \mathbb{V}[\bar{Y}_g^d | T_g = t, \mathbf{D}_g] &= \frac{\mathbb{1}_g^t}{(N_g^d)^2}\left\{\sum_i \mathbb{V}[Y_{ig}|T_g = t, D_{ig} = d]D_{ig} + 2\sum_i\sum_{j>i} D_{ig}D_{jg}\mathbb{C}\text{ov}(Y_{ig}, Y_{jg}|D_{ig}=1, D_{jg}=1)\right\} \\
&= \frac{\mathbb{1}_g^t}{N_g^d}\sigma_g^2(d,t) + 2\mathbb{1}_g^t c_g(d,t) \sum_i\sum_{j>i}\frac{D_{ig}D_{jg}}{(N_g^d)^2}
\end{aligned}$$

and the third expression follows from taking expectation.

E.2.3 Bound on $|\mu_g(d,t) - \mu_n^p(d,t)|$

$$|\mu_g(d,t) - \mu_n^p(d,t)| \leq \max_g |\mu_g(d,t) - \mu_n(d,t)| \left(1 + \frac{1}{c}\right)$$

Proof. First,

$$\begin{aligned}
|\mu_n(d,t) - \mu_n^p(d,t)| &= \left| \sum_g \frac{n_g}{n} \mu_g(d,t) - \sum_g \frac{n_g p_g(d|t)}{n p_g(d|t)} \mu_g(d,t) \right| \\
&= \left| \sum_g \frac{n_g}{n} \mu_g(d,t) \left(1 - \frac{p_g(d|t)}{\bar{p}_n(d|t)}\right) \right| \\
&= \left| \sum_g \frac{n_g}{n} (\mu_g(d,t) - \mu_n(d,t)) \left(\frac{\bar{p}_n(d|t) - p_g(d|t)}{\bar{p}_n(d|t)}\right) \right| \\
&\leq \frac{1}{c} \left| \sum_g \frac{n_g}{n} (\mu_g(d,t) - \mu_n(d,t)) \left(\bar{p}_n(d|t) - p_g(d|t)\right) \right|
\end{aligned}$$

Thus,

$$\begin{aligned}
|\mu_g(d,t) - \mu_n^p(d,t)| &\leq |\mu_g(d,t) - \mu_n(d,t)| + |\mu_n(d,t) - \mu_n^p(d,t)| \\
&\leq \max_g |\mu_g(d,t) - \mu_n(d,t)| + \frac{1}{c} \max_g |\mu_g(d,t) - \mu_n(d,t)| \max_g |\bar{p}_n(d|t) - p_g(d|t)| \\
&= \max_g |\mu_g(d,t) - \mu_n(d,t)| \left(1 + \frac{1}{c} \max_g |\bar{p}_n(d|t) - p_g(d|t)|\right) \\
&\leq \max_g |\mu_g(d,t) - \mu_n(d,t)| \left(1 + \frac{1}{c}\right).
\end{aligned}$$

E.3 Proof of Lemma 1

Let $S_{ig} = (N_g^1 - D_{ig})/(n_g - 1)$ be the observed proportion of treated peers for unit i. Then

$$\begin{aligned}
\mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t] &= \sum_s \mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t, S_{ig} = s] \mathbb{P}[S_{ig} = s|D_{ig} = d, T_g = s] \\
&= \sum_s \mathbb{E}[Y_{ig}(d,s)|D_{ig} = d, T_g = t, S_{ig} = s] \mathbb{P}[S_{ig} = s|D_{ig} = d, T_g = s] \\
&= \sum_s \mathbb{E}[Y_{ig}(d,s)] \mathbb{P}[S_{ig} = s|D_{ig} = d, T_g = s] \\
&= \sum_s \mathbb{E}[Y_{ig}(d,s)] \mathbb{P}[N_g^1 = d + s(n_g - 1)|D_{ig} = d, T_g = s] \\
&= \mathbb{E}[Y_{ig}(d, s_g(d|t))]
\end{aligned}$$

where the first equality follows from the law of iterated expectations, the second equality follows from the definition of potential outcomes under Assumptions 5 and 7, the third equality uses the independence in Assumption 6, the fourth equality follows from the definition of S_{ig} and the fifth equality follows from the

fact that $N_g^1 = n_g p_g(d|t)$. Finally, let $\mathcal{D}(t)$ denote the set of possible values for $\mathbf{D}_{(i)g}$ given $T_g = t$. Then,

$$\begin{aligned}
\mathbb{E}[Y_{ig}|D_{id} = d, T_g = t, \mathbf{D}_{(i)g}] &= \sum_{\mathbf{d}_g \in \mathcal{D}(t)} \mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t, \mathbf{D}_{(i)g} = \mathbf{d}_g]\mathbb{1}(\mathbf{D}_{(i)g} = \mathbf{d}_g) \\
&= \sum_{\mathbf{d}_g \in \mathcal{D}(t)} \mathbb{E}[Y_{ig}(d, (\mathbf{d}_g'\mathbf{1}_g - d)/(n_g - 1))]\mathbb{1}(\mathbf{D}_{(i)g} = \mathbf{d}_g) \\
&= \mathbb{E}[Y_{ig}(d, (n_g p_g(d|t)_g - d)/(n_g - 1))]
\end{aligned}$$

where the last equality uses the fact that $\mathbf{d}_g'\mathbf{1}_g = n_g p_g(d|t)$ for any $\mathbf{d}_g \in \mathcal{D}(t)$. An analogous argument gives the result for the second moments. □

E.4 Proof of Theorem 1

For any (d, t),

$$\begin{aligned}
\hat{\mu}(d,t) - \mu_n^p(d,t) &= \hat{\mu}(d,t) - \frac{\sum_g \mathbb{1}_g^t N_g^d \mu_g(d,t)}{\sum_g \mathbb{1}_g^t N_g^d} + \frac{\sum_g \mathbb{1}_g^t N_g^d \mu_g(d,t)}{\sum_g \mathbb{1}_g^t N_g^d} - \mu_n^p(d,t) \\
&= \frac{\sum_g \mathbb{1}_g^t N_g^d (\bar{Y}_g^d - \mu_g(d,t))}{N(d,t)} \\
&+ \frac{\sum_g (\mathbb{1}_g^t N_g^d - q_t n_g p_g(d|t))(\mu_g(d,t) - \mu_n^p(d,t))}{N(d,t)} \\
&+ \frac{\sum_g q_t n_g p_g(d|t)(\mu_g(d,t) - \mu_n^p(d,t))}{N(d,t)} \\
&= \frac{\sum_g \mathbb{1}_g^t N_g^d (\bar{Y}_g^d - \mu_g(d,t))}{N(d,t)} + \frac{\sum_g (\mathbb{1}_g^t N_g^d - q_t n_g p_g(d|t))(\mu_g(d,t) - \mu_n^p(d,t))}{N(d,t)}
\end{aligned}$$

where the last equality uses that

$$\sum_g q_t n_g p_g(d|t)(\mu_g(d,t) - \mu_n^p(d,t)) = q_t \left(\sum_g n_g p_g(d|t)\mu_g(d,t) - \mu_n^p(d,t) \sum_g n_g p_g(d|t) \right) = 0.$$

Next,

$$\begin{aligned}
\hat{\mu}(d,t) - \mu_n^p(d,t) &= \frac{\sum_g \mathbb{1}_g^t N_g^d (\bar{Y}_g^d - \mu_g(d,t))}{N(d,t)} + \frac{\sum_g (\mathbb{1}_g^t N_g^d - q_t n_g p_g(d|t))(\mu_g(d,t) - \mu_n^p(d,t))}{N(d,t)} \\
&= \frac{\mathbb{E}[N(d,t)]}{N(d,t)} \cdot \frac{1}{n} \sum_g \frac{\mathbb{1}_g^t N_g^d (\bar{Y}_g^d - \mu_g(d,t)) + (\mathbb{1}_g^t N_g^d - q_t n_g p_g(d|t))(\mu_g(d,t) - \mu_n^p(d,t))}{q_t \sum_g n_g p_g(d|t)/n} \\
&= \frac{\mathbb{E}[N(d,t)]}{N(d,t)} \cdot \frac{1}{n} \sum_g \psi_g(d,t)
\end{aligned}$$

where

$$\psi_g(d,t) = \frac{\mathbb{1}_g^t N_g^d (\bar{Y}_g^d - \mu_g(d,t)) + (\mathbb{1}_g^t N_g^d - q_t n_g p_g(d|t))(\mu_g(d,t) - \mu_n^p(d,t))}{q_t \bar{p}_n(d|t)}, \quad \mathbb{E}[\psi_g(d,t)] = 0$$

with $\bar{p}_n(d|t) = \sum_g n_g p_g(d|t)/n$ and

$$\begin{aligned}
\mathbb{V}[\psi_g(d,t)] &= \frac{1}{q_t^2 \bar{p}_n(d|t)^2} \left\{ \mathbb{E}\left[\mathbb{1}_g^t (N_g^d)^2 \mathbb{V}[\bar{Y}_g^d | T_g = t, \mathbf{D}_g] \right] + (\mu_g(d,t) - \mu_n^p(d,t))^2 \mathbb{V}[\mathbb{1}_g^t N_g^d] \right\} \\
&+ \frac{2}{q_t^2 \bar{p}_n(d|t)^2} (\mu_g(d,t) - \mu_n^p(d,t)) \mathbb{C}\text{ov}(\mathbb{1}_g^t N_g^d (\bar{Y}_g^d - \mu_g(d,t)), \mathbb{1}_g^t N_g^d) \\
&= \frac{1}{q_t^2 \bar{p}_n(d|t)^2} \left\{ \sigma_g^2(d,t) q_t n_g p_g(d|t) + c_g(d,t) n_g (n_g - 1) q_t p_g(d,d|t) \right\} \\
&+ \frac{(\mu_g(d,t) - \mu_n^p(d,t))^2}{q_t^2 \bar{p}_n(d|t)^2} \left\{ q_t(1-q_t) n_g^2 p_g(d|t)^2 + q_t n_g p_g(d|t)(1 - p_g(d|t)) \right\} \\
&+ \frac{(\mu_g(d,t) - \mu_n^p(d,t))^2}{q_t^2 \bar{p}_n(d|t)^2} \left\{ q_t n_g (n_g - 1) \left(p_g(d,d|t) - p_g(d|t)^2 \right) \right\}.
\end{aligned}$$

From this,

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{n} \sum_g \psi_g(d,t) \right] &= \frac{1}{n^2} \sum_g \mathbb{V}[\psi_g(d,t)] \\
&= \frac{1}{q_t^2 \bar{p}_n(d|t)^2} \sum_g \frac{n_g}{n^2} \sigma_g^2(d,t) q_t p_g(d|t) \\
&+ \frac{1}{q_t^2 \bar{p}_n(d|t)^2} \sum_g \frac{n_g(n_g-1)}{n^2} c_g(d,t) q_t p_g(d,d|t) \\
&+ \frac{q_t(1-q_t)}{q_t^2 \bar{p}_n(d|t)^2} \sum_g \frac{n_g^2}{n^2} p_g(d|t)^2 (\mu_g(d,t) - \mu_n^p(d,t))^2 \\
&+ \frac{q_t}{q_t^2 \bar{p}_n(d|t)^2} \sum_g \frac{n_g}{n^2} p_g(d|t)(1 - p_g(d|t)) (\mu_g(d,t) - \mu_n^p(d,t))^2 \\
&+ \frac{q_t}{q_t^2 \bar{p}_n(d|t)^2} \sum_g \frac{n_g(n_g-1)}{n^2} \left(p_g(d,d|t) - p_g(d|t)^2 \right) (\mu_g(d,t) - \mu_n^p(d,t))^2 \\
&= O\left(\frac{\sum_g n_g^2}{n^2} \right) = o(1)
\end{aligned}$$

since $\sigma_g^2(d,t)$ and $|\mu_g(d,t) - \mu_n(d,t)|$ are bounded from above, $\bar{p}_n(d|t)$ is bounded from below and $\max_g n_g/n \to 0$. This implies that

$$|\hat{\mu}(d,t) - \mu_n^p(d,t)| \to_{\mathbb{P}} 0$$

for all (d,t), which gives the consistency result. Next, stack the elements $\psi_g(d,t)$ in a vector ψ_g and note that

$$\Omega_n = \mathbb{V}\left[\frac{1}{\sqrt{n}} \sum_g \psi_g \right] = \frac{1}{n} \sum_g \mathbb{E}[\psi_g \psi_g']$$

where

$$\frac{1}{n}\sum_g \mathbb{E}[\psi_g(d,t)^2] = \frac{n}{q_t\left(\sum_g n_g p_g(d|t)\right)^2}\sum_g \left\{ n_g \sigma_g^2(d,t) p_g(d|t)\left(1+\rho_g(d,t)(n_g-1)\frac{p_g(d,d|t)}{p_g(d|t)}\right)\right.$$
$$\left. + (\mu_g(d,t)-\mu_n^p(d,t))^2\left(n_g(1-q_t)p_g(d|t)^2 + p_g(d|t) + (n_g-1)\mathbb{C}\text{ov}(D_{ig},D_{jg}|T_g=t)\right)\right\},$$

$$\frac{1}{n}\sum_g \mathbb{E}[\psi_g(d,t)\psi_g(d',t)] = \frac{n\sum_g c_g(0,1,t) n_g p_g(0,1|t)}{q_t\left(\sum_g n_g p_g(d|t)\right)\left(\sum_g n_g p_g(d'|t)\right)}$$
$$+ \frac{n\sum_g (\mu_g(d,t)-\mu_n^p(d,t))(\mu_g(d',t)-\mu_n^p(d',t))\mathbb{C}\text{ov}(\mathbb{1}_g^t N_g^1, \mathbb{1}_g^t N_g^0)}{q_t\left(\sum_g n_g p_g(d|t)\right)\left(\sum_g n_g p_g(d'|t)\right)},$$

$$\frac{1}{n}\sum_g \mathbb{E}[\psi_g(d,t)\psi_g(d',t')] = -\frac{n\sum_g n_g^2 p_g(d|t)p_g(d'|t')(\mu_g(d,t)-\mu_n^p(d,t))(\mu_g(d',t')-\mu_n^p(d',t'))}{\left(\sum_g n_g p_g(d|t)\right)\left(\sum_g n_g p_g(d'|t')\right)}.$$

and this variance matrix is invertible because its minimum eigenvalue is bounded below by assumption. Finally, write

$$\frac{1}{n}\sum_g \psi_g(d,t) = \frac{1}{n}\sum_g\sum_i \psi_{ig}(d,t)$$

where

$$\psi_{ig}(d,t) = \frac{1}{q_t \bar{p}_n(d|t)}\left\{ \mathbb{1}_g^t \mathbb{1}_{ig}^d (Y_{ig}-\mu_g(d,t)) + \left(\frac{\mathbb{1}_g^t N_g^d}{n_g} - q_t p_g(d|t)\right)(\mu_g(d,t)-\mu_n^p(d,t))\right\}.$$

Then we have that for $s > r \geq 2$,

$$\mathbb{E}\left[|\psi_{ig}(d,t)|^s\right] \leq \frac{1}{q_t^s c^s}\left(\mathbb{E}\left[|Y_{ig}-\mu_g(d,t)|^s\right]^{1/s} + |\mu_g(d,t)-\mu_n^p(d,t)|^{1/s}\right) < \infty$$

uniformly over i, g, d, t since as moments are uniformly bounded and using Minkowski's inequality. Thus

$$\max_{i,g}\mathbb{E}[\|\psi_{ig}\|^s] \leq (2M-1)^{s/2}\max_{i,g,d,t}\mathbb{E}[|\psi_{ig}(d,t)|^2] < \infty$$

and by Theorem 2 in Hansen and Lee (2019),

$$\Omega_n^{-1/2}\frac{1}{\sqrt{n}}\sum_g \psi_g \to_{\mathcal{D}} \mathcal{N}(0, I_{(2M-1)}).$$

To complete the proof, notice that by previous calculations:

$$\Omega_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_n^p) = \mathbf{N}^{-1}\mathbb{E}[\mathbf{N}]\Omega_n^{-1/2}\frac{1}{\sqrt{n}}\sum_g \psi_g = \Omega_n^{-1/2}\frac{1}{\sqrt{n}}\sum_g \psi_g + o_{\mathbb{P}}(1)$$

as required. □

E.5 Proof of Theorem 2

This proof follows from the proof of Theorem 1 setting $\mu_g(d,t) = \mu_n^p(d,t) = \mu(d,t)$ throughout. □

E.6 Proof of Theorem 3

Based on Equation (3), the minimization problem is equivalent to:

$$\min_{q_0,q_1,\ldots,q_M} \sum_{t=1}^{M} \frac{B_t}{q_t} + \frac{2MB_0}{q_0} = f(q_0, q_1, \ldots, q_M)$$

subject to $q_t > 0$, $\sum_t q_t = 1$ where B_0 and B_t are defined in the proposition. The first-order condition for each q_t, $t > 0$ are given by:

$$\frac{\partial f}{\partial q_t} = -\frac{B_t}{q_t^2} + \frac{2MB_0}{q_0^2} = 0 \quad \Longleftrightarrow \quad q_t^* = \sqrt{\frac{B_t}{2MB_0}} q_0^*$$

Since $\sum_{t>0} q_t = 1 - q_0$, this gives:

$$1 - q_0^* = q_0^* \sum_{t>0} \sqrt{\frac{B_t}{2MB_0}}$$

and thus:

$$q_0^* = \frac{\sqrt{2MB_0}}{\sqrt{2MB_0} + \sqrt{\sum_{t>0} B_t}}, \quad q_t^* = \frac{\sqrt{B_t}}{\sqrt{2MB_0} + \sqrt{\sum_{t>0} B_t}}, \quad t > 0.$$

On the other hand, the second-order conditions are given by:

$$\frac{\partial^2 f}{\partial q_t^2} = \frac{2B_t}{q_t^3} + \frac{2MB_0}{q_0^3}, \quad \frac{\partial^2 f}{\partial q_t \partial q_l} = \frac{2MB_0}{q_0^3}$$

and therefore the Hessian matrix $\mathbf{H}$ can be written as:

$$\mathbf{H} = \operatorname{diag}\left(\frac{2B_1}{q_1^3}, \ldots, \frac{2B_M}{q_M^3}\right) + \left(\frac{2MB_0}{q_0^3}\right) \mathbf{1}_M \mathbf{1}_M'$$

where $\mathbf{1}_M$ is an $M \times 1$ vector of ones. Thus, for any non-zero $M \times 1$ vector $\mathbf{v}$,

$$\mathbf{v}'\mathbf{H}\mathbf{v} = \sum_{t=1}^{M} \frac{2B_1 z_t^2}{q_1^3} + \left(\frac{2MB_0}{q_0^3}\right) \mathbf{v}'\mathbf{1}_M\mathbf{1}_M'\mathbf{v} = \sum_{t=1}^{M} \frac{2B_t z_t^2}{q_1^3} + \left(\frac{2MB_0}{q_0^3}\right) \left(\sum_{t=1}^{M} z_t\right)^2 > 0$$

using that $B_t > 0$ for all t so the Hessian is positive definite as required. □