

Guillermo Cruces
Dario Tortarolo
Gonzalo Vazquez-Bare

22/32

Working paper

Design of two-stage experiments with an application to spillovers in tax compliance

Design of Two-Stage 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*

July 13, 2022

Abstract

We set up a framework to conduct experiments for estimating spillover effects when units are grouped into mutually exclusive clusters. We improve upon existing methods by allowing for heteroskedasticity, intra-cluster correlation and cluster size heterogeneity, which are typically ignored when designing experiments. We show that ignoring these factors can severely overestimate power and underestimate minimum detectable effects. We derive formulas for optimal group-level assignment probabilities and the power function used to calculate power, sample size, and minimum detectable effects. We apply our methods to the design of a large-scale randomized communication campaign in a municipality of Argentina to estimate total and neighborhood spillover effects on property tax compliance. Besides the increase in tax compliance of individuals directly targeted with our mailing, we find evidence of spillover effects on untreated individuals in street blocks where a high proportion of taxpayers were notified.

JEL CODES: H71 , H26 , H21 , O23.

KEYWORDS: two-stage designs, partial population experiments, spillovers, randomization, property tax, tax compliance

*We thank Youssef Benzarti, Augustin Bergeron, Javier Birchenall, 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 Amendolagaine and Juan Luis Schiavoni for their invaluable support throughout the project. Corresponding author: Guillermo Cruces, E-mail: guillermo.cruces@nottingham.ac.uk. The design for this experiment was preregistered with the AEA RCT Registry (RCT ID: **AEARCTR-0006569**). All remaining errors are our own.

1 Introduction

There has been a renewed interest in the social interactions behind public policy interventions—in the context of schools, of welfare take-up and of tax compliance, among many others. The presence of interference between units has important consequences and challenges for the design of randomized controlled trials and the assessment of their impact, and the early experimental literature typically considered effects on non-treated units in an ex-post fashion (e.g. [Miguel and Kremer, 2004](#)). In this paper, we set up a general framework to design and carry out experiments of this type. We then employ our methods and design a large scale field experiment to capture relatively elusive spillover effects of communication campaigns on tax compliance.

Our first methodological contribution is to derive an asymptotic distributional approximation and variance formulas to conduct power, sample size and minimum detectable effects calculations for general multi-treatment experimental designs where units are grouped into mutually exclusive clusters (as in, e.g., [Duflo and Saez, 2003](#); [Crépon et al., 2013](#)). We improve upon existing methods by allowing for general forms of heteroskedasticity, intracluster correlation and cluster size heterogeneity, all factors that affect the variance of treatment effect estimators but are typically ignored when designing experiments. To illustrate the importance of this issue empirically, we use data from existing studies to show that the corrected minimum detectable effects (MDEs) can be about 20% and up to 30% larger than the ones that fail to account for cluster heterogeneity.

We consider a double-array asymptotic setting where cluster sizes are allowed, but not required, to grow with the sample size. This allows us to determine the effect of group size heterogeneity on the accuracy of the normal approximation for conducting inference and power calculations. Our analysis nests the commonly analyzed cases with fixed cluster sizes, equally sized clusters, binary treatment and non-clustered experiments, among others. Our results can be straightforwardly implemented after imputing values for outcome variances and intracluster correlations, as in any standard power analysis. We also show that our general formula simplifies to well-known formulas in specific designs (e.g. [Duflo, Glennerster and Kremer, 2007](#)).

Our second methodological contribution is to apply our general results to partial population designs for estimating spillover effects. We show our formulas generalize those of [Hirano and Hahn \(2010\)](#) and [Baird et al. \(2018\)](#) to account for heteroskedasticity, general forms of intra-cluster correlation and cluster size heterogeneity. In addition, we provide a power function to conduct power, sample size, and minimum detectable effects calculations for different treatment effects based on our distributional approximation and variance formulas. Finally, we derive formulas for optimal group-level assignment probabilities.

Lastly, we apply our framework to the design of a large-scale field experiment devised to estimate total and neighborhood 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. Property tax collection fell significantly in the context of the COVID-19 lockdown, and we devised an intervention in October 2020, when mobility restrictions started to ease, to help the Municipality recoup its collection levels. Our campaign consisted of sending personalized letters to randomly selected dwellings with reminders about due taxes, as well as 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 – group sizes, power calculations, etc. – following our framework to maximize the chance of capturing spillover effects of our mailings on neighbors that lived in the same street blocks of treated individuals (i.e., those who received letters from us) but that did not receive a letter.

Randomization took place in two stages. First, we randomly divided our sample of 3,982 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. In the second stage, we randomly selected accounts within the last three groups of blocks to receive the personalized information letter. We sent approximately 25,000 letters between September 28th and October 7th, 2020, corresponding to the October billing period (with due date on the 9th) as well as past due debt (if any). We run saturated regressions that identify total effects (the change in outcomes among treated individuals – i.e., letter recipients – relative to those in pure control blocks) and spillover effects (the behavior of untreated neighbors within treated blocks – blocks with treated individuals – relative to those in pure control blocks) on monthly payments.

We find compelling graphical evidence of total effects and spillover effects on property tax payment rates in the October billing period (the month of the intervention). Our results reveal higher payment rates of treated and untreated accounts relative to neighbors in pure control blocks where no neighbors received the information letter. The regressions show an immediate and statistically significant effect in the payment rate of treated units in the three saturation groups relative to pure control blocks. For blocks with the highest saturation (80% treated accounts), total effects on bill payments emerge (numerically and statistically) on the same day that the letters started to be distributed, reaching a magnitude of about 4.5 percentage points by the due date. This represents a 13.2% increase relative to the payment rate in pure control blocks. We validate the experiment by showing no effects for bill payments in September 2020 (i.e., a billing period before the intervention took place).

Spillover effects are more modest in magnitude but still substantial and precisely estimated, but only in high-saturation street blocks (those with 80% treated accounts). Payment rates of untreated

accounts in those blocks increase by about 1.1 percentage point, a statistically significant effect in the early days of the intervention. Conversely, we do not find evidence of spillover effects in blocks with only 20% or 50% treated accounts, with estimates of spillover effects that oscillate around zero.

Finally, we find some heterogeneous effects along the expected (i.e., pre-registered) dimensions. Spillovers in street blocks with 80% treated individuals are much higher at about 2.6 percentage points (and highly statistically significant) in the blocks above the median of payment rates for 2019, the tax year before our intervention (and the pandemic and the ensuing lockdown), whereas it is not statistically significant in blocks below the median.

Comparison with existing literature. Our paper contributes to a growing literature on experimental design (Hirano and Hahn, 2010; Athey, Eckles and Imbens, 2018; Baird et al., 2018; Bugni, Canay and Shaikh, 2018, 2019; Basse, Feller and Toulis, 2019; Jiang and Imai, 2021; Puelz et al., 2022; Bai, 2022; Viviano, 2022). More specifically, our results are applied to the estimation of spillover effects and generalize those of Hirano and Hahn (2010) 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 particular, cluster size heterogeneity, which is commonly ignored when designing experiments, has two important practical implications. First, when clusters are not equally sized, 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 when designing experiments results in overestimating power and underestimating MDEs, as we illustrate in Section 2. 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 “too heterogeneous” (Carter, Schnepel and Steigerwald, 2017; Djogbenou, MacKinnon and Ørregaard Nielsen, 2019). This fact highlights the importance of analyzing and accounting for the distribution of cluster sizes when designing experiments. Based on recent advances in the econometric literature on inference for clustered data (Hansen and Lee, 2019), our main methodological result provides two statistics that summarize the heterogeneity in the cluster size distribution.

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 and Imai (2021) 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. One advantage of our approach is that it allows us to determine the effect of cluster size heterogeneity in the asymptotic distribution of the treatment effect estimators. As mentioned, we also contribute

to the existing literature by deriving optimal choices of group-level assignment probabilities.

We also contribute to a large empirical literature on property tax compliance and a small but growing literature on spillover effects. There has been a renewed interest on this subject with some recent insightful papers such as [Brockmeyer et al. \(2020\)](#) in Mexico City, [Weigel \(2020\)](#) and [Bergeron, Tourek and Weigel \(2021\)](#) in the Democratic Republic of Congo, [Krause \(2020\)](#) in Haiti, and [Eguino and Schächtele \(2020\)](#) in Argentina, among others.¹ While the latter are randomized controlled trials, they do not address directly the issue of potential spillovers in compliance at the local level. The social interactions behind public policy interventions in tax compliance has remained a relatively elusive issue in this literature. In a recent study of tax professionals as sources of spillovers between taxpayers, [Battaglini et al. \(2019\)](#) highlight that network externalities in compliance behavior has been documented in laboratory experiments. They also discuss more recent studies based on randomized controlled trials that test the importance of spatial proximity. [Rincke and Traxler \(2011\)](#) study enforcement spillovers of TV licensing inspections on untreated households in Austria (see also [Drago, Mengel and Traxler, 2020](#)). In a study of the income tax at the city level in Detroit, [Meiselman \(2018\)](#) fails to find evidence of geographic network effects on neighbors. In the context of firms, [Boning et al. \(2020\)](#) analyze direct and network effects from in-person visits by revenue officers on visited and non-visited firms. Whereas these papers find spillover effects in compliance, their original experiments were not designed to capture these effects. A notable exception is [Pomeranz \(2015\)](#) who shows that increasing enforcement on a randomly-selected group of Chilean firms leads to spillovers up the VAT chain. We build on these pioneering works by designing our intervention with the purpose of capturing spillovers.

The paper is organized as follows. Section 2 provides a brief illustration based on previously published studies of how our methodology can result in better design of partial population experiments. In Section 3, we set up our framework for two-stage experiments and derive the main methodological results. In Section 4, we describe the large-scale randomized communication campaign, the administrative data used in the analysis, the empirical strategy and the results from our empirical analysis. Section 5 provides some concluding remarks.

2 Why is Accounting for Cluster Heterogeneity in Experimental Design Important?

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, we can expect cluster sizes to vary substantially: for instance, electoral precincts, towns, schools or school dis-

¹For previous work in Argentina see [Castro and Scartascini \(2015\)](#) and [Lopez-Luzuriaga and Scartascini \(2019\)](#). [Antinyan and Asatryan \(2019\)](#) present a meta-analysis of nudges in tax compliance.

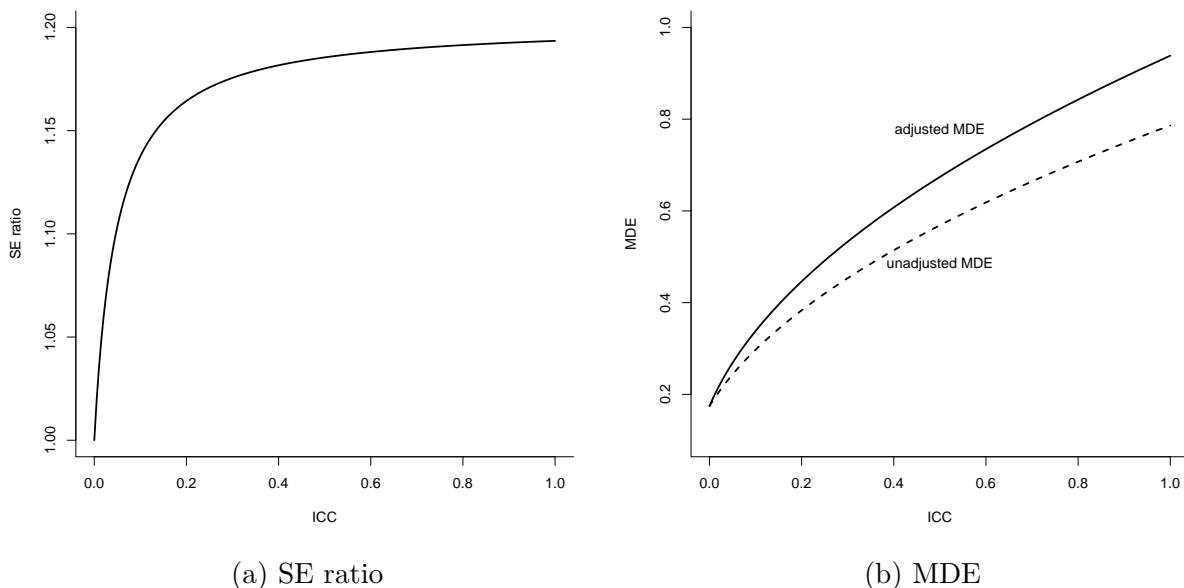
tricts, can have different numbers of voters, population or students. In the application we present below, our tax information campaign reaches city street blocks with a wide range of taxpayers—from 8 in the smallest blocks to 50 in the largest (see Figure 2). When clusters vary in size, the variance 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 the estimators of interest, which in turn results in overestimating power and underestimating MDEs. As we show in Section 3, this problem becomes more serious the larger (i) the ratio of the variance to average cluster size, (ii) the intraclass correlation in outcomes, and (iii) the within-group correlation of treatment assignments.

We illustrate this issue based on data from four published studies: Ichino and Schündeln (2012), Haushofer and Shapiro (2016), Giné and Mansuri (2018) and Imai, Jiang and Malani (2021). These four experiments employ a partial population design where clusters are randomly assigned to different treatment intensities to estimate spillover effects (see Section 3.4 and Section C.1 in the supplemental appendix for further details). Specifically, we use the formulas we derive in Section 3 to calculate standard errors and MDEs accounting for cluster size heterogeneity using the median values of number of groups, $G = 95$, average group size, $\bar{n} = 23.3$, and group size standard deviation, $sd(n_g) = 15.2$, from these four studies. We then compare these adjusted standard errors and MDEs with the unadjusted ones that would be obtained if (incorrectly) ignoring cluster size heterogeneity.

The results from this numerical exercise are shown in Figure 1. Panel (a) shows the ratio of the adjusted to unadjusted standard errors as a function of the intraclass correlation in the outcomes, ρ . The figure shows that the ratio grows rapidly as ρ increases, and stabilizes between 1.15 and 1.2, suggesting that even for moderate levels of intraclass correlation, the adjustment factor due to group size heterogeneity is substantial. Panel (b) shows the adjusted and unadjusted MDEs, and shows that even for values of ρ as low as 0.05, the adjusted MDE can be 10% larger than the unadjusted one, and this difference can grow up to around 20% for larger values of ρ .

The underestimation of standard errors and MDEs becomes more severe as the ratio of variance to mean of group sizes increases. For instance, in this illustration, keeping a standard deviation of group sizes of $sd(n_g) = 15.2$ but reducing average group size from $\bar{n} = 23.3$ to $\bar{n} = 18$ results in adjusted MDEs that can be between 25% and 30% larger than the unadjusted ones. Ignoring this adjustment results in overly optimistic and thus under-powered designs.

Figure 1: Adjusted and unadjusted standard errors and MDEs.



Notes: Panel (a) shows the ratio of adjusted to unadjusted standard errors as a function of the intraclass correlation (ICC). Panel (b) shows the adjusted (solid line) and unadjusted (dashed line) minimum detectable effects as a function of the intraclass correlation (ICC). Adjusted magnitudes account for group size variability. Unadjusted magnitudes assume no group size variability, i.e. zero variance of group size. Calculations use the median values from Table A4: $G = 95$, $\bar{n} = 23.3$, $sd(n_g) = 15.2$.

3 Design of Two-Stage Experiments

3.1 Setup

In our general setup, we consider the design of experiments in a sample where units are grouped into mutually exclusive clusters within which there might be spillovers in the outcomes of interest. 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 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 tax reminder information campaign, the population of interest consists of taxpayers in residential city street blocks, and the outcome 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 saturations of treated individuals.

We consider a sample of observations that are divided into mutually independent clusters $g = 1, \dots, G$, where each cluster g contains n_g observations $i = 1, \dots, n_g$ and the total sample size is $n = \sum_{g=1}^G n_g$ (which includes the non-clustered setting as the special case in which $n_g = 1$ for all

g). We analyze a general design where the experimenter randomly assigns a multi-valued treatment A_{ig} taking values in a set $\mathcal{A} = \{a_0, a_1, \dots, a_K\}$ where we set a_0 as the baseline treatment status (such as no treatment or a placebo). In our setup, the treatment assignment may vary both within and between clusters, which encompasses (multi-treatment) cluster randomized trials as the special case in which $A_{ig} = A_{jg} = A_g$ for all i and j . The binary treatment case corresponds to $\mathcal{A} = \{0, 1\}$.

The treatment assignments in group g are collected in a vector $\mathbf{A}_g = (A_{1g}, \dots, A_{n_gg})$ characterized by a probability distribution $\pi_g(\mathbf{a}) = \mathbb{P}_g[\mathbf{A}_g = \mathbf{a}]$ for $\mathbf{a} \in \mathcal{A}^{n_g}$. These probabilities can differ across clusters, which allows, for example, for stratification at the cluster level. Similarly, let $\mathbf{Y}_g = (Y_{1g}, \dots, Y_{n_gg})$ be the vector collecting the observed outcomes in group g .

We introduce the following restrictions on the sampling and treatment assignment mechanism.

Assumption 1 (Sampling and Assignment Mechanism)

- (i) $(\mathbf{Y}_g, \mathbf{A}_g)_{g=1}^G$ are mutually independent across g .
- (ii) For each g and for all $i = 1, \dots, n_g$, $\mathbb{P}_g[A_{ig} = a_k] = \pi_g(a_k)$ for $a_k \in \mathcal{A}$.
- (iii) For all $a_k \in \mathcal{A}$, $\sum_{g=1}^G \pi_g(a_k) > 0$.

Part (i) of Assumption 1 states that groups are independent. Part (ii) states that the treatment assignment is identically distributed within each cluster, so that all units within the same group are subject to the same assignment mechanism. Part (iii) rules out the case in which some treatment values a_k have zero probability in the sample.

3.2 Estimands and Estimators of Interest

In multi-treatment experiments, effects are commonly estimated through a saturated regression like the following:

$$Y_{ig} = \alpha + \sum_{k=1}^K \beta_k \mathbb{1}(A_{ig} = a_k) + \varepsilon_{ig} \tag{1}$$

Because the regression is saturated, it follows that the OLS estimator of each coefficient $\hat{\beta}_k$ is a difference in means between each assignment a_k and the baseline assignment a_0 :

$$\hat{\beta}_k = \frac{\sum_g \sum_i Y_{ig} \mathbb{1}(A_{ig} = a_k)}{\sum_g \sum_i \mathbb{1}(A_{ig} = a_k)} - \frac{\sum_g \sum_i Y_{ig} \mathbb{1}(A_{ig} = a_0)}{\sum_g \sum_i \mathbb{1}(A_{ig} = a_0)}.$$

To ensure that these regression coefficients have a causal interpretation, we introduce the following assumption. In what follows, for each i and j , let $\mathbf{A}_{(i)g}$ denote the vector of assignments excluding unit i , and let $\mathbf{A}_{(ij)g}$ be the vector of assignments excluding i and j .

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

$$(i) \mathbb{E}[Y_{ig}|A_{ig} = a_k, \mathbf{A}_{(i)g}] = \mathbb{E}[Y_{ig}|A_{ig} = a_k] = \mu(a_k) \text{ for all } a_k \in \mathcal{A},$$

$$(ii) \mathbb{V}[Y_{ig}|A_{ig} = a_k, \mathbf{A}_{(i)g}] = \mathbb{V}[Y_{ig}|A_{ig} = a_k] = \sigma^2(a_k) \text{ for all } a_k \in \mathcal{A},$$

$$(iii) \mathbb{C}ov(Y_{ig}, Y_{jg}|A_{ig} = a_k, A_{jg} = a_l, \mathbf{A}_{(ij)g}) = \mathbb{C}ov(Y_{ig}, Y_{jg}|A_{ig} = a_k, A_{jg} = a_l) = c(a_k, a_l) \text{ for all } a_k, a_l \in \mathcal{A}.$$

This assumption imposes two restrictions on the first and second conditional moments of the observed outcomes. The first one states that, conditional on own assignment A_{ig} , the other units' assignments do not affect the outcome moments. In other words, the assignment A_{ig} contains all the relevant variation in the outcome moments. In Section 3.4 we show that, in a partial population design, this assumption amounts to assuming that peers are exchangeable, that is, that outcomes depend on the proportion of treated units but not their identities.

The second part of Assumption 2 imposes equal conditional first and second moments across units and clusters, so that these moments do not vary over i and g . In this case, the parameters can be defined in terms of a general population and not on a specific sample. This assumption can be relaxed at the expense of additional notation by switching focus to averages across groups, although this makes the parameters sample-dependent.

Under this assumption, the coefficients $\beta = (\beta_1, \dots, \beta_K)'$ from Equation (6) equal differences in expected outcomes, $\beta_k = \mathbb{E}[Y_{ig}|A_{ig} = a_k] - \mathbb{E}[Y_{ig}|A_{ig} = a_0] = \mu(a_k) - \mu(a_0)$ for $k = 1, \dots, K$ and $\mathbb{E}[\varepsilon_{ig}|A_{ig}] = 0$. In addition, OLS estimators are conditionally unbiased, that is, $\mathbb{E}[\hat{\beta}_k|\mathbf{A}_g] = \beta_k$.

3.3 Asymptotic Distribution and Power Function

In this section we present our main methodological result, which provides an asymptotic approximation to the distribution and the variance of the OLS estimators for the parameters in Equation (1). Let this vector of OLS estimators be $\hat{\beta}$.

In what follows, we define $\sigma^2(a_k) = \mathbb{V}[Y_{ig}|A_{ig} = a_k]$, $c(a_k, a_l) = \mathbb{C}ov(Y_{ig}, Y_{jg}|A_{ig} = a_k, A_{jg} = a_l)$, $\rho(a_k, a_l) = c(a_k, a_l)/(\sigma(a_k)\sigma(a_l))$, $\pi_g(a_k, a_l) = \mathbb{P}_g[A_{ig} = a_k, A_{jg} = a_l]$ for $i \neq j$ and we let “ $\rightarrow_{\mathcal{D}}$ ” denote convergence in distribution. We consider an asymptotic setting in which both the number of groups and the group sizes grow with the sample size. The goal of letting $n_g \rightarrow \infty$ as $n \rightarrow \infty$ is to determine how fast group sizes can grow 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 setting with fixed n_g and/or equally-sized clusters ($n_g = \bar{n}$ for all g) is nested as a particular case of our analysis. The number of parameters remains fixed in our setup (see [Vazquez-Bare, Forthcoming](#),

for an alternative approach in which the number of parameters is allowed to grow with the sample size). The next result follows from applying Theorem 9 in Hansen and Lee (2019) to our setting.

Proposition 1 *Suppose that Assumptions 1 and 2 and the regularity conditions in Assumption 3 in the supplemental appendix hold. If*

$$\max_{g \leq G} \frac{n_g^2}{n} \rightarrow 0, \quad \frac{\sum_{g=1}^G n_g^4}{n^2} \leq C < \infty, \quad (2)$$

then

$$V_n^{-1/2} \sqrt{n}(\hat{\beta} - \beta) \rightarrow_{\mathcal{D}} \mathcal{N}(\mathbf{0}, I_K)$$

where I_K is a K -dimensional identity matrix and:

$$\begin{aligned} V_{n,kk} &= \frac{n\sigma^2(a_k)}{\sum_g n_g \pi_g(a_k)} \left\{ 1 + \rho(a_k, a_k) \frac{\sum_g n_g(n_g - 1)\pi_g(a_k, a_k)}{\sum_g n_g \pi_g(a_k)} \right\} \\ &+ \frac{n\sigma^2(a_0)}{\sum_g n_g \pi_g(a_0)} \left\{ 1 + \rho(a_0, a_0) \frac{\sum_g n_g(n_g - 1)\pi_g(a_0, a_0)}{\sum_g n_g \pi_g(a_0)} \right\} \\ &- 2n\sigma(a_k)\sigma(a_0)\rho(a_k, a_0) \frac{\sum_g n_g(n_g - 1)\pi_g(a_k, a_0)}{\sum_g n_g \pi_g(a_k) \sum_g n_g \pi_g(a_0)} \end{aligned}$$

The proof and the full shape of the covariance matrix (including the off-diagonal elements) are given in supplemental appendix C.4. In terms of practical implementation, the main takeaway from this result is that, provided Condition (2) holds, the variance of each $\hat{\beta}_k$ can be approximated by:

$$\begin{aligned} \mathbb{V}[\hat{\beta}_k] &\approx \frac{V_{n,kk}}{n} = \frac{\sigma^2(a_k)}{\sum_g n_g \pi_g(a_k)} \left\{ 1 + \rho(a_k, a_k) \frac{\sum_g n_g(n_g - 1)\pi_g(a_k, a_k)}{\sum_g n_g \pi_g(a_k)} \right\} \\ &+ \frac{\sigma^2(a_0)}{\sum_g n_g \pi_g(a_0)} \left\{ 1 + \rho(a_0, a_0) \frac{\sum_g n_g(n_g - 1)\pi_g(a_0, a_0)}{\sum_g n_g \pi_g(a_0)} \right\} \\ &- 2\sigma(a_k)\sigma(a_0)\rho(a_k, a_0) \frac{\sum_g n_g(n_g - 1)\pi_g(a_k, a_0)}{\sum_g n_g \pi_g(a_k) \sum_g n_g \pi_g(a_0)}. \end{aligned} \quad (3)$$

This formula corresponds to the variance of a difference in means with clustered data.² More precisely, in the first two terms of the sum, the first components $\sigma^2(a_k)/\sum_g n_g \pi_g(a_k)$ and $\sigma^2(a_0)/\sum_g n_g \pi_g(a_0)$ are the conditional variance of the outcome divided by the expected cell sample size, and the second component is the design effect that accounts for clustering between observations within a group. Notice that the design effect depends on the correlation in outcomes conditional on treatment assignments, $\rho(a_k, a_k)$, the correlation in treatment assignments, captured by $\pi_g(a_k, a_k)$ and $\pi_g(a_0, a_0)$, and the heterogeneity in group sizes. Finally, the third term captures the covariance between the two sample means. This last term is equal to zero whenever $\mathbb{P}_g[A_{ig} = a_k, A_{jg} = a_0] = 0$.

²Heuristically, $\hat{\beta}_k = \bar{Y}_k - \bar{Y}_0$ and thus $\mathbb{V}[\hat{\beta}_k] = \mathbb{V}[\bar{Y}_k - \bar{Y}_0] = \mathbb{V}[\bar{Y}_k] + \mathbb{V}[\bar{Y}_0] - 2\text{Cov}(\bar{Y}_k, \bar{Y}_0)$.

In the above formula, the group sizes n_g are observable and the probabilities $\pi_g(a_k)$, $\pi_g(a_0)$ and $\pi_g(a_k, a_l)$ are determined by the experimental design. Hence, the only unknown terms are the variances $\sigma^2(a_k)$ and $\sigma^2(a_0)$ and intraclass correlations $\rho(a_k, a_k)$, $\rho(a_0, a_0)$ and $\rho(a_k, a_0)$, which can be imputed by the researcher based on a pilot experiment, previous literature or by considering a range of reasonable values, as in standard power analysis. More precisely, based on the distributional approximation and variance formulas in Proposition 1, power, sample size and minimum detectable effects calculations can be conducted for each effect β_k using the power function:

$$\Gamma(\beta_k) \approx 1 - \Phi\left(\frac{\beta_k}{\sqrt{\mathbb{V}[\hat{\beta}_k]}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\beta_k}{\sqrt{\mathbb{V}[\hat{\beta}_k]}} - z_{1-\alpha/2}\right) \quad (4)$$

after imputing the unknown parameters (outcome variances and intraclass correlations), where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile from the standard normal distribution. Also notice that conducting inference and power calculations for linear combinations or smooth functions of the coefficients in $\hat{\beta}$ (see e.g. the pooled and slope effects proposed by Baird et al., 2018) is straightforward using the delta method. See Section C in the appendix for further details.

The approximation in Proposition 1 holds when the sample size is large enough and as long as no cluster is “too large”, as formalized by Condition (2).³ More precisely, the first part of Condition (2) ensures that the largest cluster is small relative to the total sample size, whereas the second part restricts the fourth moment of the distribution of group sizes, which intuitively 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 all clusters are small relative to the total sample size and possibly discard outliers if present. We discuss this further in our empirical analysis.

The following examples show how the general formula in Theorem 1 simplifies to well-known formulas in specific designs.

Example 1 (Non-clustered experiments) *Suppose that all clusters have only one unit, $n_g = 1$. This amounts to analyzing a random sample of individuals as in a standard RCT. Suppose the treatment is assigned independently to each unit with probability $p \in (0, 1)$. In this case, $K = 1$, $A_{ig} \in \{0, 1\}$, $\pi_g(1) = p$, $\pi_g(0) = 1 - p$, and Equation (3) reduces to:*

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

In addition, under homoskedasticity, $\sigma^2(1) = \sigma^2(0) = \sigma^2$ and thus:

$$\mathbb{V}[\hat{\beta}] \approx \frac{\sigma^2}{np(1-p)}$$

³Notice that this condition holds automatically when group sizes are seen as fixed in the asymptotic analysis, which corresponds to the case of “many small groups”.

which corresponds to Equation (6) in [Dufló, Glennerster and Kremer \(2007\)](#).

Example 2 (Clustered randomized experiments) Suppose that clusters are assigned to a binary treatment with probability $\lambda \in (0, 1)$ and that all units within a cluster receive the same treatment, $A_{ig} = A_g \in \{0, 1\}$, which implies $K = 1$ and $\pi_g(a_1, a_0) = 0$. In addition, suppose that all clusters are equally sized so that $n_g = \bar{n}$ for all g . Then, Equation (3) reduces to:

$$\mathbb{V}[\hat{\beta}] \approx \frac{\sigma^2(1)}{G\bar{n}\lambda} [1 + \rho(1)(\bar{n} - 1)] + \frac{\sigma^2(0)}{G\bar{n}(1 - \lambda)} [1 + \rho(0)(\bar{n} - 1)].$$

In addition, assume a random effects structure so that $\sigma^2(1) = \sigma^2(0) = \sigma^2 + \tau^2$ and $\rho(1) = \rho(0) = \tau^2/(\sigma^2 + \tau^2)$. In this case,

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

which corresponds to Equation (9) in [Dufló, Glennerster and Kremer \(2007\)](#).

3.4 Partial Population Designs

We now apply Proposition 1 to partial population designs for estimating spillover effects. In a partial population design, groups are randomly divided into categories denoted by $T_g \in \mathcal{T} = \{0, 1, 2, \dots, M\}$ where by convention $T_g = 0$ denotes a pure control group and $\mathbb{P}[T_g = t] = q_t$. Within each group, treatment is assigned at the individual level with a probability that depends on the value of T_g , $\mathbb{P}_g[D_{ig} = 1|T_g = t] = p_g(t)$ and where $\mathbb{P}_g[D_{ig} = 0|T_g = 0] = 1$. Thus, in this case $A_{ig} = (D_{ig}, T_g)$ and $\pi_g(d, t) = p_g(t)^d(1 - p_g(t))^{1-d}q_t$. In addition, $\mathbb{P}_g[A_{ig} = (d, t), A_{jg} = (0, 0)] = 0$ for any $t \neq 0$.

In this setting, Assumption 2 requires that outcome moments do not vary with peers' assignments, conditional on own assignment, for example $\mathbb{E}[Y_{ig}|A_{ig} = a_k, \mathbf{A}_{(i)g}] = \mathbb{E}[Y_{ig}|A_{ig} = a_k]$. In a partial population experiment, this assumption reduces to $\mathbb{E}[Y_{ig}|D_{ig} = d, T_g, \mathbf{D}_{(i)g}] = \mathbb{E}[Y_{ig}|D_{ig} = d, T_g = t]$. Since $T_g = t$ determines the proportion of treated units in the group, this requirement amounts to assuming that, given the proportion of treated units determined by T_g , the identities of the treated peers do not affect the outcome. In such cases, it is usually said that peers are exchangeable. This assumption is commonly invoked in the spillovers literature (see [Vazquez-Bare, Forthcoming](#), and references therein for further discussion).

Denote the assignment $(D_{ig}, T_g) = (d, t)$ by “ dt ” and the assignment $(D_{ig}, T_g) = (0, 0)$ by “0”. Applying Proposition 1 to this case, under Condition (2) the variance of each treatment effect

estimator $\hat{\beta}_{dt}$ can be approximated by:

$$\begin{aligned} \mathbb{V}[\hat{\beta}_{dt}] \approx & \frac{\sigma^2(dt, dt)}{q_t \sum_g n_g p_g(t)^d (1 - p_g(t))^{1-d}} \left\{ 1 + \rho(dt, dt) \frac{\sum_g n_g (n_g - 1) \mathbb{P}_g[D_{ig} = d, D_{jg} = d | T_g = t]}{\sum_g n_g p_g(t)^d (1 - p_g(t))^{1-d}} \right\} \\ & + \frac{\sigma^2(0, 0)}{n q_0} \left\{ 1 + \rho(0, 0) \left(\frac{\sum_g n_g^2}{n} - 1 \right) \right\}. \end{aligned} \tag{5}$$

As mentioned, the variances and intra-cluster correlations are the only unknown parameters, whereas the group sizes are observed in the sample and the probabilities q_t , $p_g(t)$ and $\mathbb{P}_g[D_{ig} = d, D_{jg} = d | T_g = t]$ are chosen by the researcher. Section C.2 in the appendix discusses the two most common within-group assignment mechanisms, namely, fixed margins and Bernoulli trials, and characterizes these probabilities explicitly.

Next, we show how our general formula in Equation (5) simplifies to the formulas proposed in the literature under further assumptions.

Example 3 (Homoskedastic case with two treatment intensities) *Suppose there is only one treatment intensity and a pure control category, so that $M = 1$ and $A_{ig} \in \{(0, 0), (0, 1), (1, 1)\}$, 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 $\sigma^2(a_k) = 1$ and $\rho(a_k, a_l) = 0$ for $k, l = 0, 1$. In this case, for assignment $(d, t) = (0, 1)$, Equation (5) simplifies to:*

$$\mathbb{V}[\hat{\beta}_{01}] \approx \frac{1 - pq}{(1 - p)q(1 - q)}$$

which corresponds to the variance formula in [Hirano and Hahn \(2010\)](#).

Example 4 (Random effects structure with equally-sized groups) *Consider the case in which groups are equally sized, $n_g = \bar{n}$ for all g , and a random effects covariance structure so that $\sigma^2(a_k) = \sigma^2 + \tau^2$, $\rho(a_k, a_l) = \tau^2$ for all k, l . In addition, suppose that the within-group assignment given $T_g = t$ sets a fixed number of treated units $\bar{n}p_t$ in each group, 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 (5) becomes:*

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

which corresponds to Equation (3) in [Baird et al. \(2018\)](#).

While these two examples provide useful guidance for practitioners on how to design experiments to estimate spillovers, they make restrictive assumptions that may result in under-powered designs, as illustrated in Section 2. Equation (5) generalizes these cases by allowing for general forms of heteroskedasticity, intra-cluster correlation and non-homogeneous clusters.⁴

⁴A practical issue that arises when allowing for heterogeneously-sized clusters is that it may not be possible to assign

3.5 Design of Partial Population Experiments

Our results can be used to optimally choose assignment probabilities. To see the intuition, suppose for simplicity that the probabilities $\mathbb{P}_g[D_{ig} = d|T_g = t]$ and $\mathbb{P}_g[D_{ig} = d, D_{jg} = d|T_g = t]$ do not vary over g . When designing a partial population experiment, the researcher needs to specify (i) the number of treatment intensities M , (ii) the group-level assignment probabilities $\{q_t\}_{t=0}^M$ where $q_t = \mathbb{P}[T_g = t]$ and (iii) the within-group treatment probabilities $\{p_t\}_{t=0}^M$ and $\{\mathbb{P}[D_{ig} = d, D_{jg} = d|T_g = t]\}_{t=0}^M$ where $p_t = \mathbb{P}[D_{ig} = 1|T_g = t]$.

The choices of M and of the within-group treatment probabilities p_t are closely related, and depend on the structure of spillover effects that the researcher wants to be able to estimate. A larger M allows for more “granularity” which may give a more complete assessment of spillovers, at the expense of complicating estimation by introducing more parameters. This issue is discussed in [Vazquez-Bare \(Forthcoming\)](#) in a setting with equally-sized groups. Given a value of M , the choice of within-treatment probabilities p_t depends on the researcher’s prior on the treatment intensities that generate spillovers. For instance, suppose $M = 3$ (i.e. one pure control and three treatment intensities). If the researcher believes that spillovers only materialize when the treatment intensity is large, a possible choice would be $p_1 = 50\%$, $p_2 = 70\%$ and $p_3 = 90\%$. On the other hand, the choice $p_1 = 20\%$, $p_2 = 50\%$ and $p_3 = 80\%$ may be more useful when the researcher does not have a clear prior on the structure of spillovers and may therefore prefer a more uniform distribution of treatment intensities. We do not discuss optimal choices of M and p_t in this paper, as they depend on the parameters that the researcher wishes to identify, which in turn depend on an unknown function (the outcome response function).

We now discuss the choice of $\{q_t\}_{t=0}^M$ given M and the within-group treatment probabilities. Optimally choosing this set of probabilities requires defining an optimality criterion that determines how the variances of all the parameters 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 focus on *A-optimality*, which minimizes the trace of the variance-covariance matrix of the estimators (or equivalently, the average of the asymptotic variances).⁵ 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. While other criteria are possible, A-optimality has the advantage of having a simple closed-form solution in this setting, as shown in the following proposition.

the exact desired number of units to treatment. We propose a method to deal with this issue in [Section C.2](#) of the appendix.

⁵Notice 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 optimality problem in [Proposition 2](#).

Proposition 2 *In the design described in Section 3.4, consider the optimal design problem:*

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

with $q_t > 0$, $\sum_{t=0}^M q_t = 1$ using the variance formula in Equation (5). The optimal assignment probabilities are given by:

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

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 $t > 0$

$$B_t = \frac{\sigma^2(1t, 1t)}{\sum_g n_g p_g(t)} \left\{ 1 + \rho(1t, 1t) \frac{\sum_g n_g (n_g - 1) \mathbb{P}_g[D_{ig} = 1, D_{jg} = 1 | T_g = t]}{\sum_g n_g p_g(t)} \right\} \\ + \frac{\sigma^2(0t, 0t)}{\sum_g n_g (1 - p_g(t))} \left\{ 1 + \rho(0t, 0t) \frac{\sum_g n_g (n_g - 1) \mathbb{P}_g[D_{ig} = 0, D_{jg} = 0 | T_g = t]}{\sum_g n_g (1 - p_g(t))} \right\}.$$

The proof is given in supplemental appendix C.5.

Constrained Designs. Researchers may often need to incorporate different sets of practical constraints when choosing assignment probabilities. For example, the design may need to account for logistical, budgetary, political, administrative or other types of constraints that restrict the total number of units that receive treatment. These restrictions can be incorporated when choosing q_t . In the next section, we describe the design of our partial population experiment where the total number of treated units was in part given by a budgetary restriction. To choose the assignment probabilities, we set up a system of equations incorporating this restriction and ensuring that the variance of the smallest treatment cells are equal, to ensure a certain level of precision for the “hardest” treatment effect to estimate.

4 A Randomized Property Tax Communication Campaign

4.1 Background and Experimental Design

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 maximize the statistical power 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.⁶

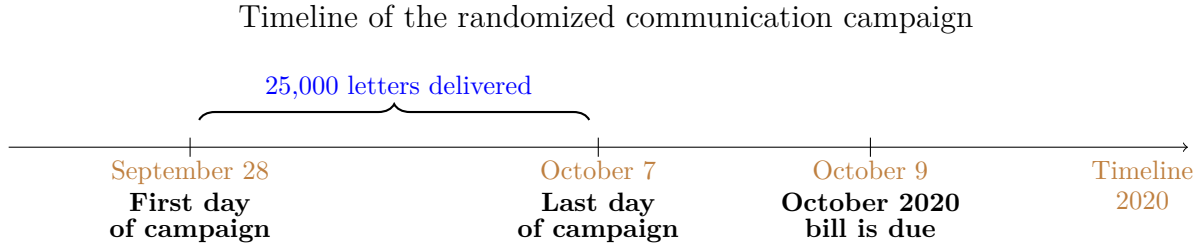
Randomization took place in two stages—first at the city street block level, and then at the taxpayer account (i.e., property) level. In the first stage, we randomly divided our sample of 3,982 blocks (clusters) into four groups with different intensity of treatment: (1) pure control blocks where no 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).⁷ 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,

⁶Figure 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 increase tax compliance.

⁷As explained in Subsection 3.5, the choice $p_1 = 20\%$, $p_2 = 50\%$ and $p_3 = 80\%$ may be useful when the researcher does not have a clear prior on the structure of spillovers and may therefore prefer a more uniform distribution of treatment intensities. Our choice attempts to balance parsimony with flexibility to detect nonlinearities in total and spillover effects without having to estimate too many parameters. Ultimately, the choice of within-treatment probabilities p_t depends on the researcher’s prior on the treatment intensities that generate spillovers. The optimal choice of p_t depends on the parameters that the researcher wishes to identify and, thus, it is not addressed in our paper.

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.



4.2 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.

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.

The municipality required us to target city street blocks with 8 to 50 accounts. Figure 2 shows 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.⁸ 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

⁸For the full distribution, see Figure A.5.

payments even from highly compliant individuals.

4.3 Treatment Assignment

Using the notation from Section 3.4, let n_g indicate the number of units (accounts - *cuentas*) per group (block - *cuadra*) with $g = 1, \dots, G$ and let $N = \sum_g n_g$ be the total sample size. The group-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 $D_{ig} \in \{0, 1\}$. We have that:

$$\mathbb{P}[D_{ig} = 1 | T_g = t] = p_t = \begin{cases} 0 & \text{if } t = 0 \\ 0.2 & \text{if } t = 1 \\ 0.5 & \text{if } t = 2 \\ 0.8 & \text{if } t = 3. \end{cases}$$

Choice of q_t . The expected number of treated units/letters sent is

$$n_1 = n(0.2q_1 + 0.5q_2 + 0.8q_3)$$

On the other hand, since the assignments $T_g = 1$ and $T_g = 3$ are symmetric, we set $q_1 = q_3$. If the goal is to send L letters, the choice of q_t should satisfy:

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

Finally, to ensure that the variances of the estimators are similar across assignments, we set:

$$q_2 = Rq_3$$

where R depends on the intraclass correlation and the variance of the outcomes. We use the results in Proposition 1 to approximate the variances and obtain the ratio R . We provide further details in Appendix B.1. We were able to send $L = 25,061$ letters and determined the sample sizes 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

Power and MDE. Finally, we use the power function formula (4) to conduct power calculations for each estimator using the following parameters: (i) $\sigma^2(d, t) = 0.25$ for all (d, t) ;⁹ (ii) ICC = 0.1 which is close to (but larger than) the estimated intraclass correlation of the baseline outcome; (iii) 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.¹⁰

4.4 Estimation

Given an outcome Y_{ig} , our goal is to estimate $\beta_{0t} = \mathbb{E}[Y_{ig}|D_{ig} = 0, T_g = t] - \mathbb{E}[Y_{ig}|D_{ig} = 0, T_g = 0]$ for $t = 1, 2, 3$, which can be seen as spillover effects on untreated units in groups with $T_g = t$ compared to pure controls, and $\beta_{1t} = \mathbb{E}[Y_{ig}|D_{ig} = 1, T_g = t] - \mathbb{E}[Y_{ig}|D_{ig} = 0, T_g = 0]$ which are total effects on treated units in groups with $T_g = t$ compared to pure controls.

We jointly estimate the parameters of interest through the following saturated OLS regression:

$$Y_{ig} = \alpha + \sum_{t=1}^3 \beta_{0t} \mathbb{1}(T_g = t)(1 - D_{ig}) + \sum_{t=1}^3 \beta_{1t} \mathbb{1}(T_g = t)D_{ig} + \varepsilon_{ig} \quad (6)$$

where we allow ε_{ig} to be correlated within blocks and use a cluster-robust variance estimator. In this regression, θ_t is interpreted as the spillover effect on untreated units in groups with $T_g = t$ and τ_t is interpreted as the total effect on treated units in groups with $T_g = t$.

⁹This gives a conservative estimate because 0.25 is the upper bound for the variance of a binary variable.

¹⁰Appendix Figure B.10 plots the power function for each estimator

4.5 Property Tax Information Campaign: Empirical Results

4.5.1 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.¹¹ 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 3 to 5 and then we summarize the corresponding point estimates in Tables 3 and 4.

Figure 3 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 4 shows the analog but for *untreated* units and pure control blocks. Panel (b) thus captures spillover effects.¹²

Figure 3 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.

Although smaller in size, Figure 4 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 5 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 6). 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

¹¹Appendix section A.2 presents the results from balance test regressions. These results confirm that our groups are balanced and comparable.

¹²For comparison, the gray solid line shows the treatment effect for treated units (pooled together from $T_g = 1, 2, 3$ in Figure 3).

units.¹³ The estimates displayed in the left panels of Figure 5 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 5 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 5. 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.¹⁴ 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.

¹³These point estimates coincide with those reported in panels (b) of Figures 3 and 4.

¹⁴Figure A.4 in the appendix presents the analog of Figure 5 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.

4.5.2 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.¹⁵

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.¹⁶

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.¹⁷

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,

¹⁵The 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.

¹⁶Figure 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.

¹⁷The 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 6, which makes the pattern in Table 4 all the more apparent.¹⁸

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.

4.5.3 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.¹⁹ 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).

5 Conclusion

We provide a general framework to carry out partial population experiments with an application to spillovers in property tax compliance. The estimation of spillovers and other indirect effects must be built into the experimental design and not as an afterthought. Yet, how to incorporate these aspects is not obvious. We derive an asymptotic approximation and variance formulas to conduct power calculations for general clustered experimental designs allowing for multiple treatments, general forms of intra-cluster correlation, and cluster size heterogeneity.

One of the main methodological contributions of the analysis is to provide variance and power

¹⁸Table 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.

¹⁹Note that nudging individuals to sign up to e-billing was an explicit content of the letter (see Figure A.1).

formulas that account for cluster size heterogeneity, which is typically ignored when designing experiments. When clusters vary in size, the variance of treatment effect estimators contains an adjustment factor that depends on the average and the variance of cluster size. Ignoring this adjustment factor underestimates the variance of the estimators of interest, which in turn results in overestimating power and underestimating MDEs. We illustrate this issue based on data from four published studies conducting two-stage experiments. The corrected MDEs can be about 20% and up to 30% larger than the ones that fail to account for 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 then apply our results to the design of partial population experiments for estimating spillover effects and use our results 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 graphical 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.

The results have clear implications for the design of partial population experiments. Spillovers and other indirect effects must be accounted for and incorporated not only in terms of being registered in a pre-analysis plan, but also and most importantly in terms of accounting for them in power calculations.

In terms of the design of tax collection policies, we find evidence of interactions in tax payment behavior. While our results do not point out to substantial savings in communication costs by relying on word of mouth and indirect effects, they still point out that there is some degree of interactions between taxpayers that could be taken into account for the design of optimal communication and tax enforcement policies.

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.” *working paper*.
- 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.
- Battaglini, Marco, Luigi Guiso, Chiara Lacava, and Eleonora Patacchini.** 2019. “Tax Professionals: Tax-Evasion Facilitators or Information Hubs?” C.E.P.R. Discussion Papers CEPR Discussion Papers 13656.
- 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.”
- 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.

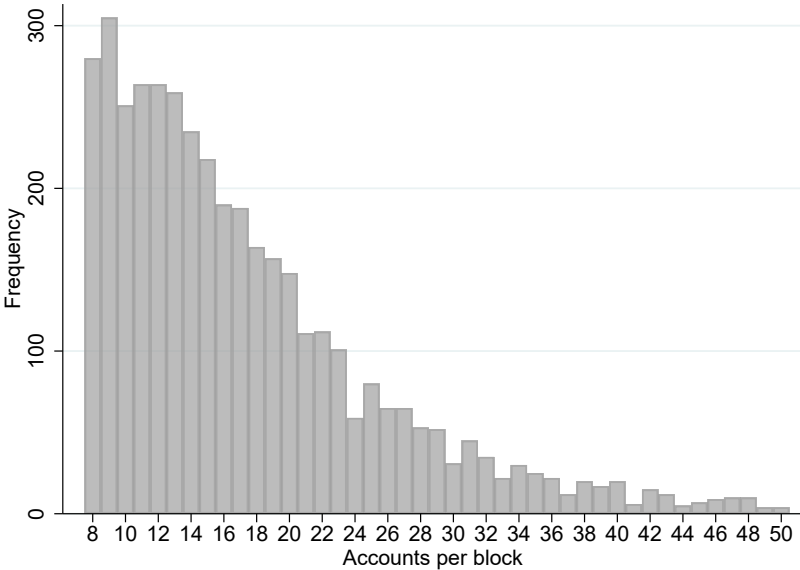
- Boning, William C., John Guyton, Ronald Hodge, and Joel Slemrod.** 2020. “Heard it through the grapevine: The direct and network effects of a tax enforcement field experiment on firms.” *Journal of Public Economics*, 190(C).
- Brockmeyer, A, A Estefan, K Ramirez Arras, and J.C. Suarez Serrato.** 2020. “Taxing Property in Developing Countries: Theory and Evidence from Mexico.” *IFS Working Paper*.
- Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh.** 2018. “Inference under Covariate-Adaptive Randomization.” *Journal of the American Statistical Association*, 0(0): 1–13.
- 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.
- Cameron, Adrian Colin, and Douglas L Miller.** 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, 50(2): 317–372.
- 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.
- Castro, Lucio, and Carlos Scartascini.** 2015. “Tax compliance and enforcement in the pampas evidence from a field experiment.” *Journal of Economic Behavior & Organization*, 116: 65–82.
- 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.
- Eguino, Huáscar, and Simeon Schächtele.** 2020. “A playground for tax compliance? Testing fiscal exchange in an RCT in Argentina.” IDB Working Paper Series.
- 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.
- 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, Zichao, and Kosuke Imai.** 2021. “Statistical Inference and Power Analysis for Direct and Spillover Effects in Two-Stage Randomized Experiments.” *working paper*.
- Krause, Benjamin.** 2020. “Balancing Purse and Peace: Tax Collection, Public Goods and Protests.”
- Lopez-Luzuriaga, Andrea, and Carlos Scartascini.** 2019. “Compliance spillovers across taxes: The role of penalties and detection.” *Journal of Economic Behavior & Organization*, 164: 518–534.

- Meiselman, Ben.** 2018. “Ghostbusting in Detroit: Evidence on nonfilers from a controlled field experiment.” *Journal of Public Economics*, 158(C): 180–193.
- 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.
- Pomeranz, Dina.** 2015. “No Taxation without Information : Deterrence and Self-Enforcement in the Value Added Tax.” *The American Economic Review*, 105(8): 2539–2569.
- 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.
- Rincke, Johannes, and Christian Traxler.** 2011. “Enforcement Spillovers.” *The Review of Economics and Statistics*, 93(4): 1224–1234.
- Silvey, Samuel D.** 1980. *Optimal Design: An Introduction to the Theory for Parameter Estimation*. Springer Netherlands.
- Vazquez-Bare, Gonzalo.** Forthcoming. “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.

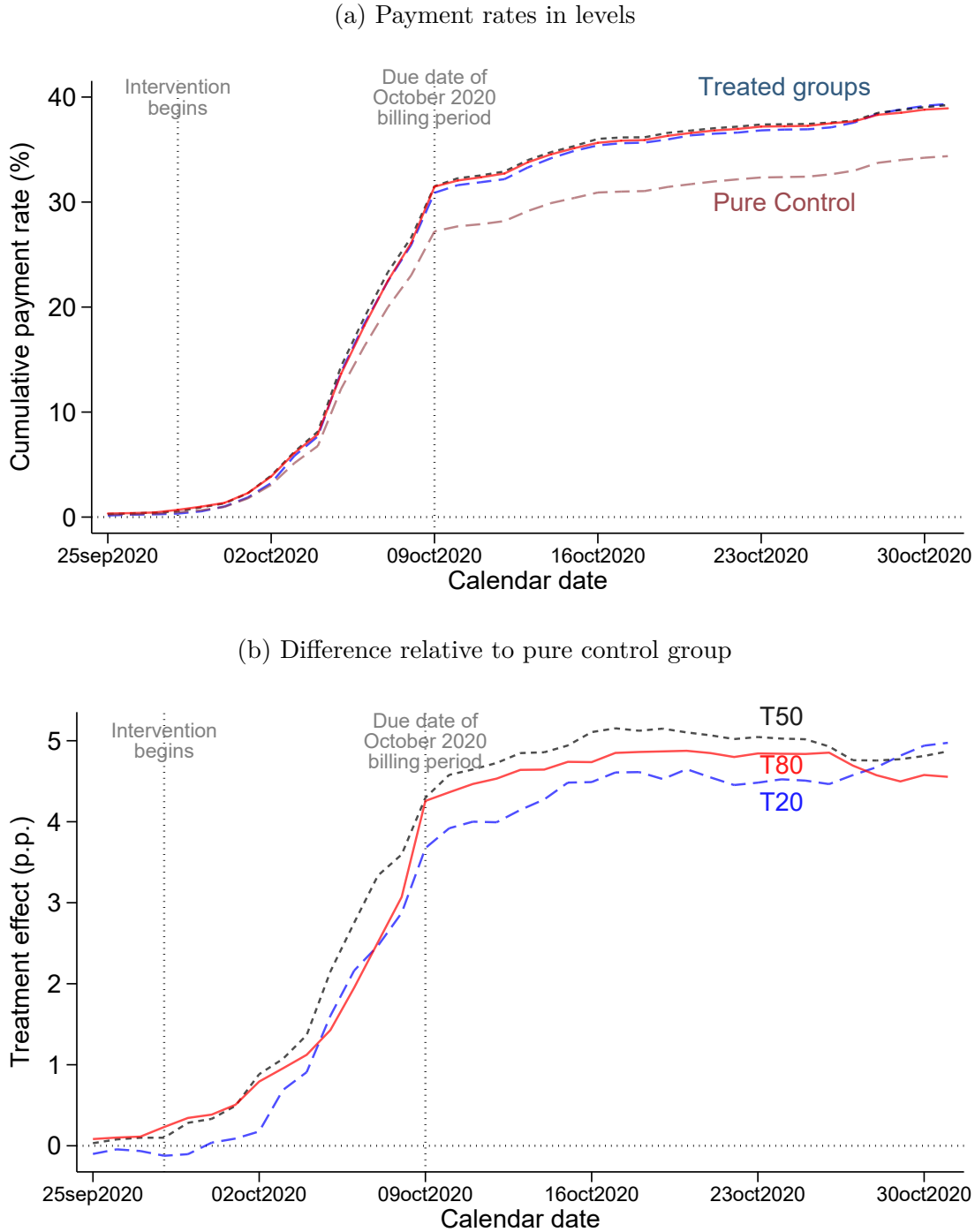
6 Figures and Tables

Figure 2: Distribution of accounts per block



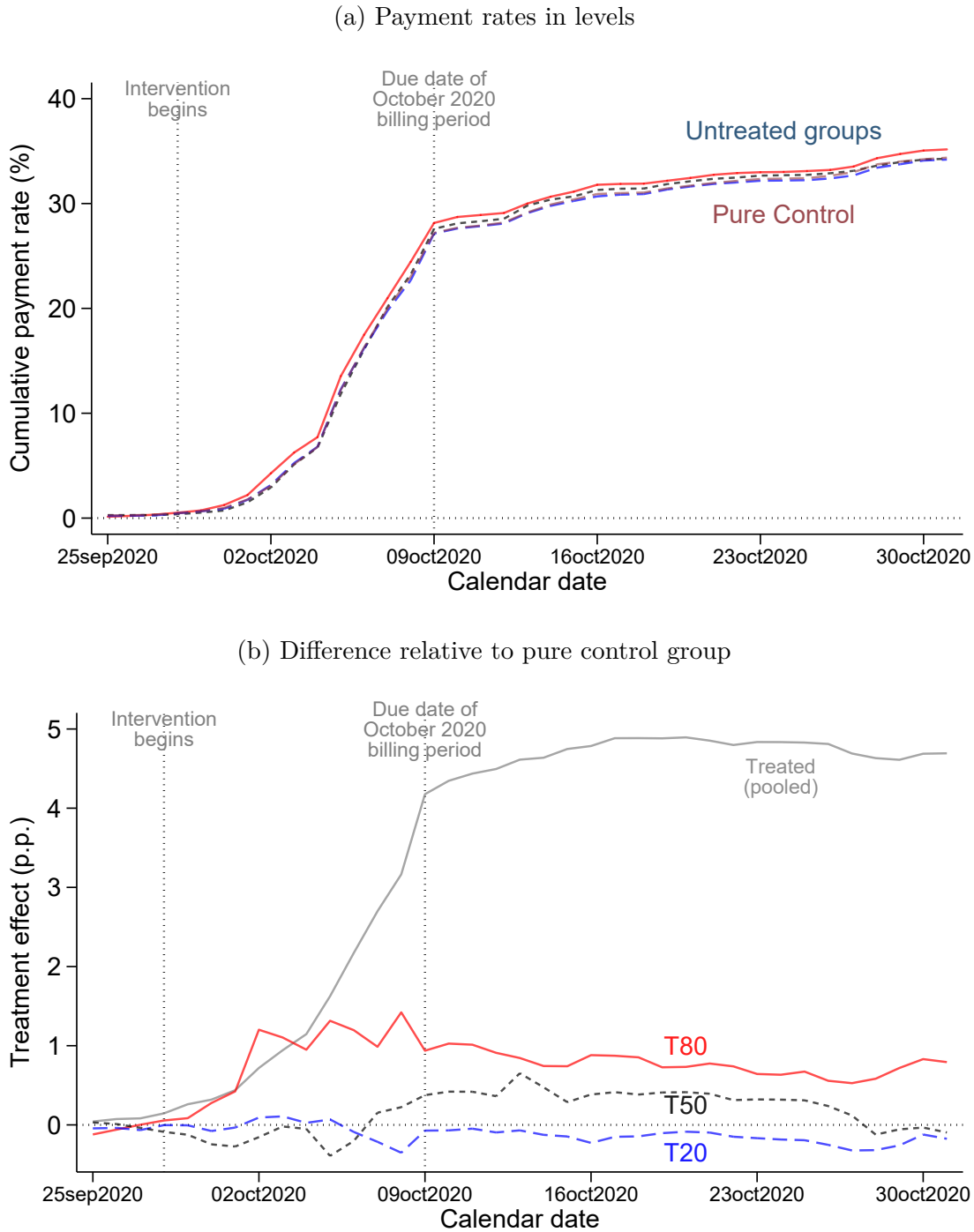
Notes: This figure shows the distribution of accounts per block using data from the year 2019. We use these data to design the experiment. Our sample size consists of 68,808 accounts distributed in 3,982 blocks.

Figure 3: Payment rates: Treated groups vs Pure control blocks



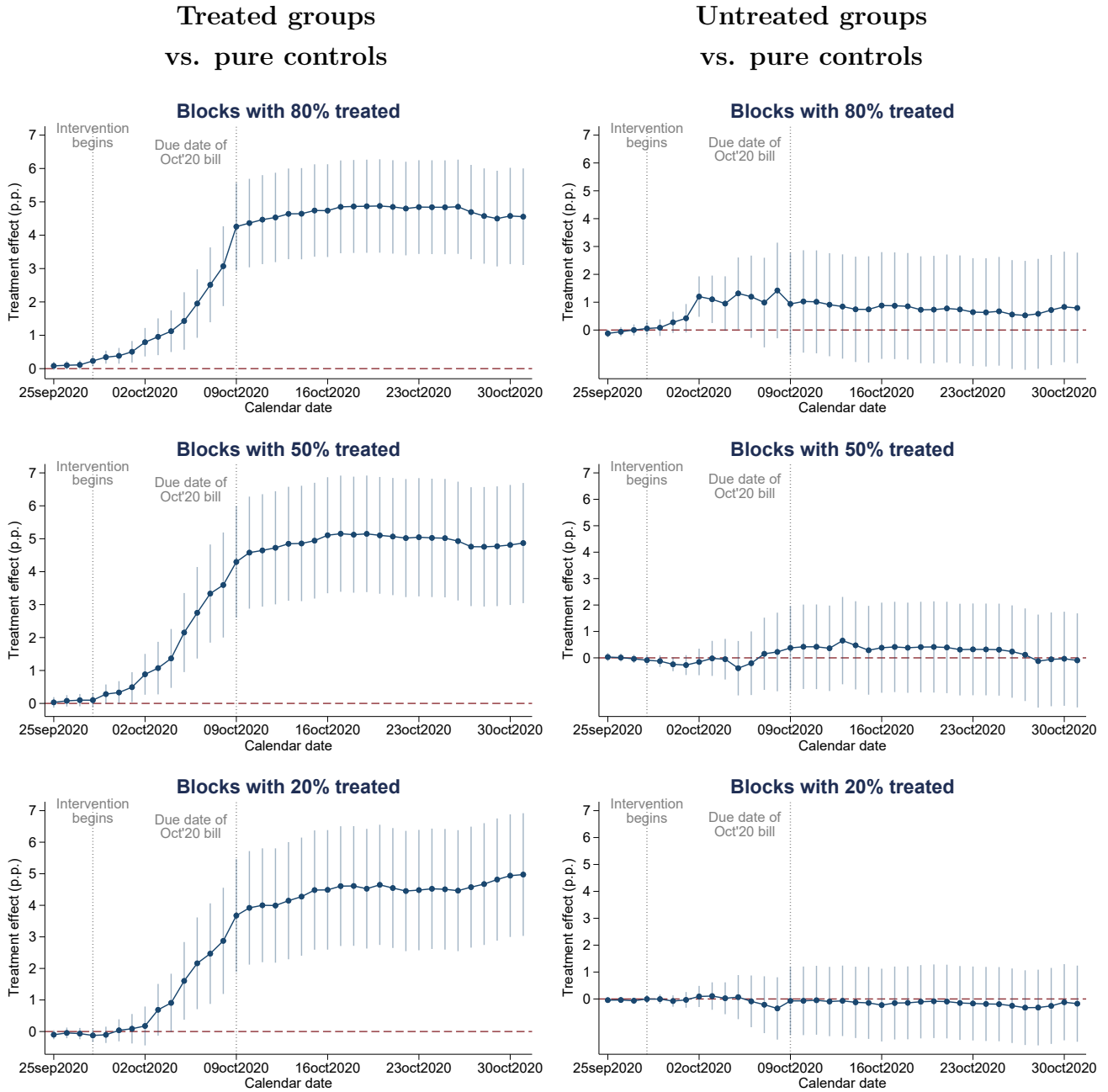
Notes: These figures show the effect of the intervention on payments of the October 2020 bill for treated groups. Panel (a) shows the cumulative share of individuals paying the October 2020 bill over time. The brown dashed line shows the payment rate for pure control units. The blue dashed line corresponds 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 4: Payment rates: Untreated groups vs Pure control blocks



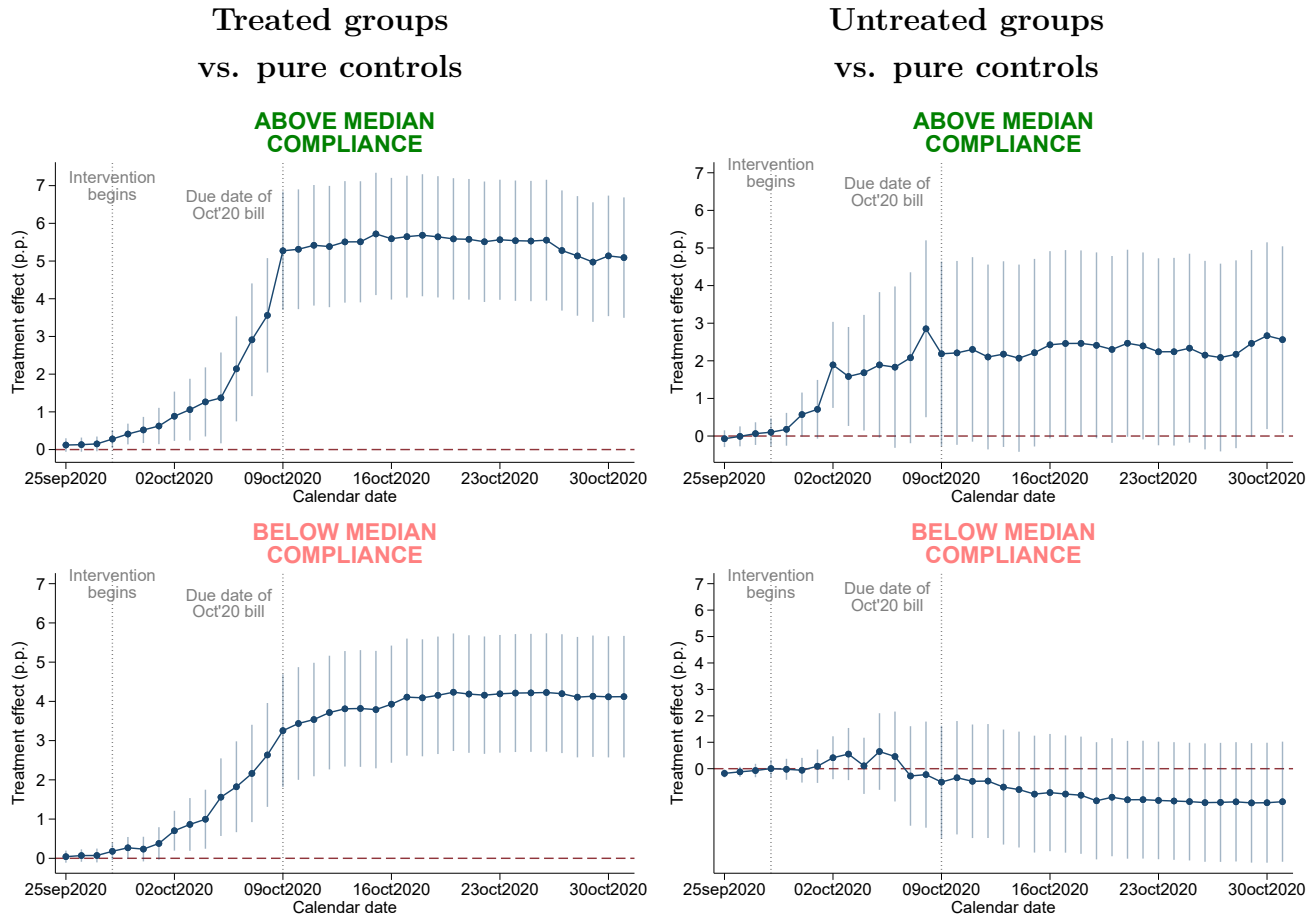
Notes: These figures show the effect of the intervention on payments of the October 2020 bill for untreated groups. Panel (a) shows the cumulative share of individuals paying the October 2020 bill over time. The brown dashed line shows the payment rate for pure control units. The blue dashed line corresponds 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 5: 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 3 and 4. 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 6: 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 2 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: Pr(pay the bill)	Placebo bill:	Intervention bill:	
	Sep'20 (1)	Early (2)	By Oct 31 (3)
<i>A. Blocks with 80% treated</i>			
Treated	0.12 (0.69)	0.96*** (0.28)	4.55*** (0.74)
Untreated	-0.30 (0.95)	1.10** (0.43)	0.79 (1.01)
<i>B. Blocks with 50% treated</i>			
Treated	0.76 (0.88)	1.07*** (0.41)	4.87*** (0.93)
Untreated	0.26 (0.88)	-0.02 (0.34)	-0.10 (0.91)
<i>C. Blocks with 20% treated</i>			
Treated	0.85 (0.93)	0.69* (0.42)	4.97*** (0.99)
Untreated	0.07 (0.68)	0.11 (0.26)	-0.18 (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 6 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 (5). 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 (1)	Above Median (2)	Below Median (3)	Above Median (4)	Below Median (5)	Above Median (6)
A. Blocks with 80% treated						
Treated	0.10 (0.73)	0.28 (0.81)	0.86** (0.34)	1.06** (0.42)	4.12*** (0.79)	5.09*** (0.81)
Untreated	-1.55 (1.09)	0.78 (1.24)	0.55 (0.50)	1.58** (0.67)	-1.25 (1.16)	2.56** (1.27)
B. Blocks with 50% treated						
Treated	1.54 (0.99)	0.69 (1.12)	1.24** (0.50)	1.02 (0.62)	4.81*** (1.07)	5.67*** (1.08)
Untreated	0.81 (0.94)	0.36 (1.15)	0.10 (0.43)	-0.03 (0.50)	1.34 (1.00)	-0.76 (1.14)
C. Blocks with 20% treated						
Treated	1.32 (1.11)	0.27 (1.24)	0.85* (0.52)	0.52 (0.63)	5.41*** (1.21)	4.40*** (1.27)
Untreated	0.27 (0.72)	-0.32 (0.80)	0.68** (0.33)	-0.42 (0.38)	0.61 (0.77)	-1.09 (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 (6) 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

Tus impuestos municipales ahora vienen en la BOLETA DIGITAL

ID: XXXXX

TITULAR:
DIRECCIÓN: CAP. MADARIAGA N° LOCALIDAD: 11 de Septiembre
C.P.: 1657
PARTIDA: XXXXXX/7

Te queremos contar que ahora en Tres de Febrero tu boleta municipal de la Tasa por Servicios Generales (TSG) es 100% digital. O sea, ya no se usa más el papel. Podés acceder a ella y pagarla desde el celular o la computadora. De esta manera, nos cuidamos entre todos al reducir la circulación y también cuidamos el medio ambiente. Es una situación difícil y te agradecemos el esfuerzo que estás haciendo para estar al día con tus impuestos, porque eso se transforma directamente en obras y servicios que no paran en tu barrio. Te informamos el estado de tu cuenta y te mostramos lo fácil que es:

PARTIDA: XXXXX/7	
Cuota 10 vencimiento 10 de octubre 2020:	347,29
Deuda año en curso*:	1.702,58
Deuda años anteriores*:	289,54

* Al 15/09/2020

¿CÓMO PAGAR?

Ingresando a tasas.tresdefebrero.gov.ar completá los datos:

DESCARGÁ O PAGÁ TU BOLETA

Tasa o Servicio a Pagar

Servicios generales

Identificación o

Ingresar / Ingresar

🔍 BUSCAR BOLETA

✉️ RECIBIR LA BOLETA POR MAIL

CLICKEÁ ESTE BOTÓN

📧 y recibirla todos los meses en tu mail.

También podés entrar a miboleta.tresdefebrero.gov.ar

1) Podés pagar **ONLINE** con

- mercado pago** → En el momento desde nuestra web.
- PagoMis Cuéntas** **LRN** → Obteniendo el código de pago electrónico para pagar desde la plataforma de tu banco o cajero automático.

2) Podés pagar en **EFFECTIVO** en

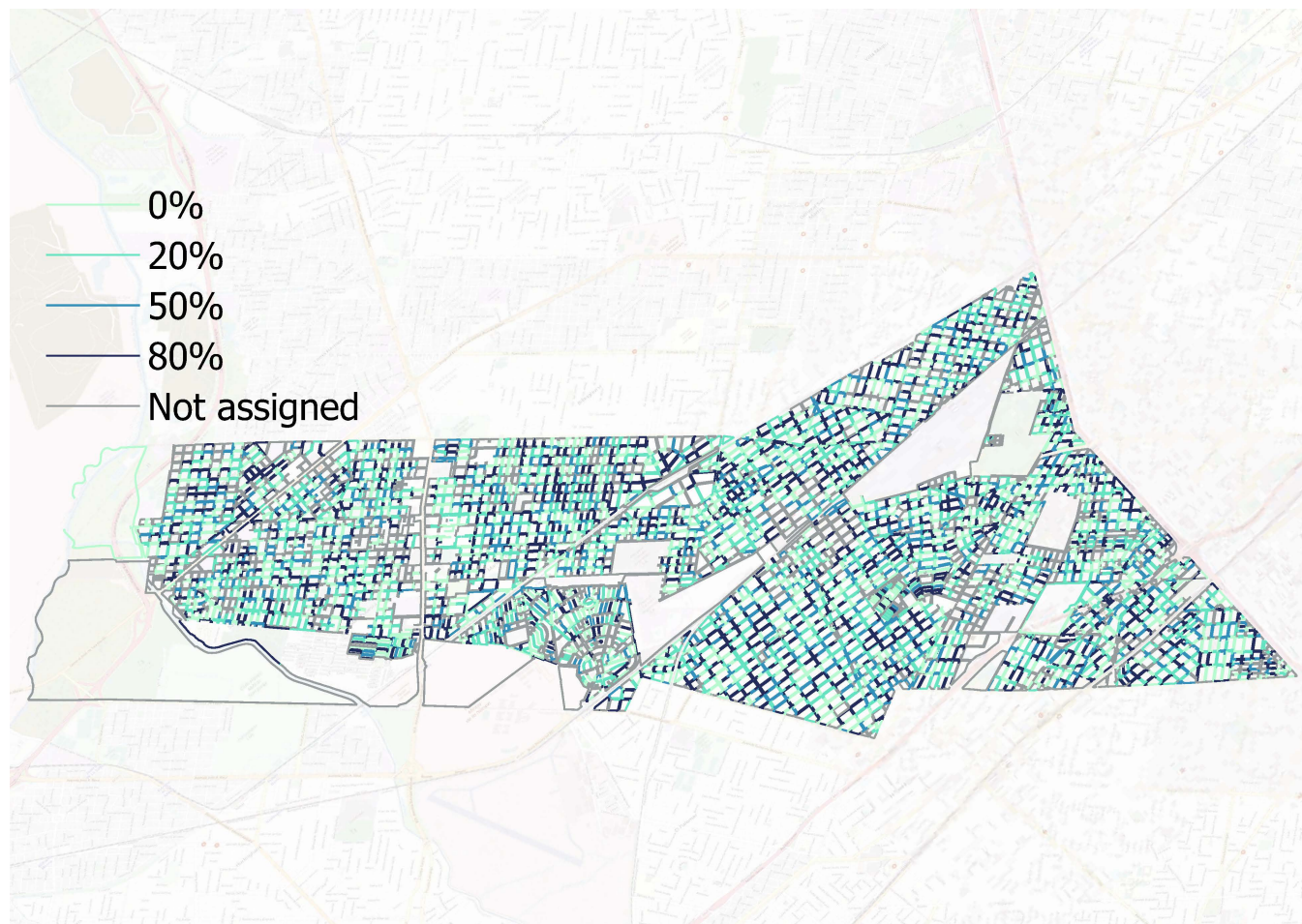
- rapipago** → DESCARGALA o levá tu NÚMERO DE PARTIDA.

Por dudas comunicate con nosotros a reclamos.mistasas@tresdefebrero.gov.ar
Si esta carta llegó por error a tu domicilio, informanos en ese mismo correo electrónico

¡Muchas gracias!

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!”

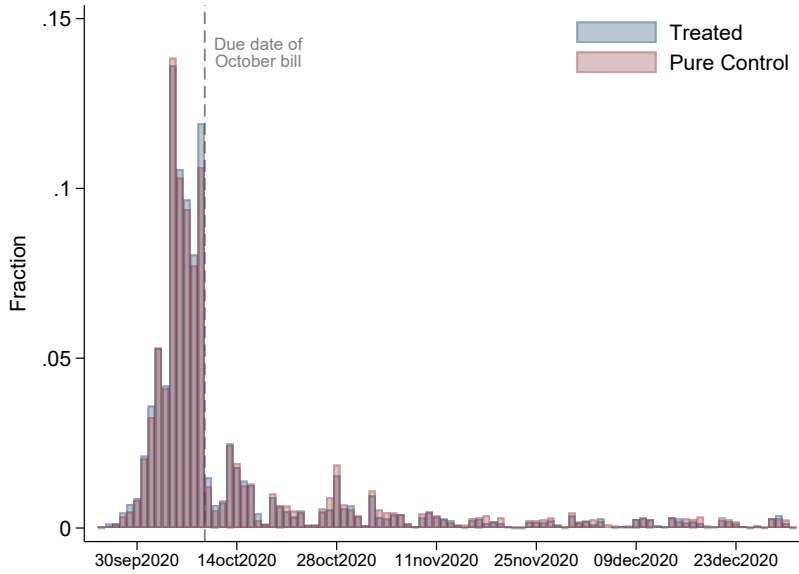
Figure A.2: Map of the municipality with the experimental design



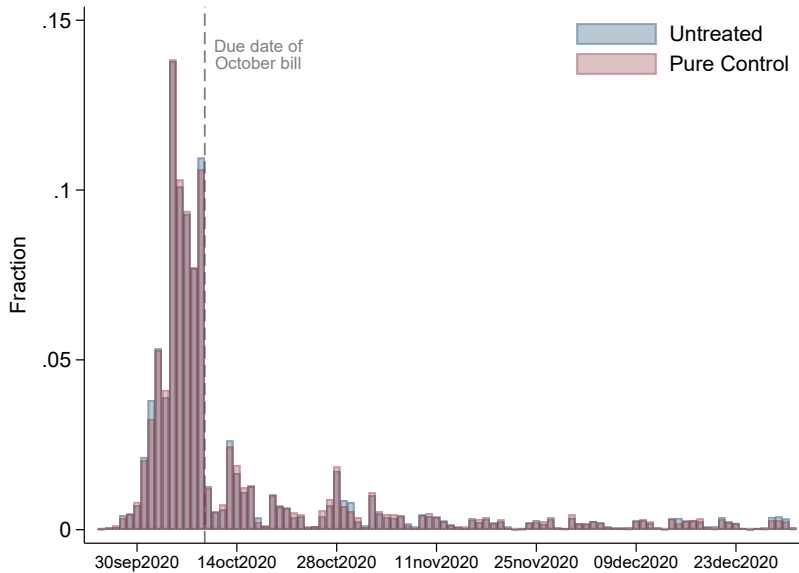
Notes: This figure shows a map of the municipality where the 2-level randomized communication campaign took place. We highlight the group-level assignment of blocks (*cuadras*) with different colors: pure control blocks with 0% treated (light green), blocks with 20% treated accounts (green), blocks with 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

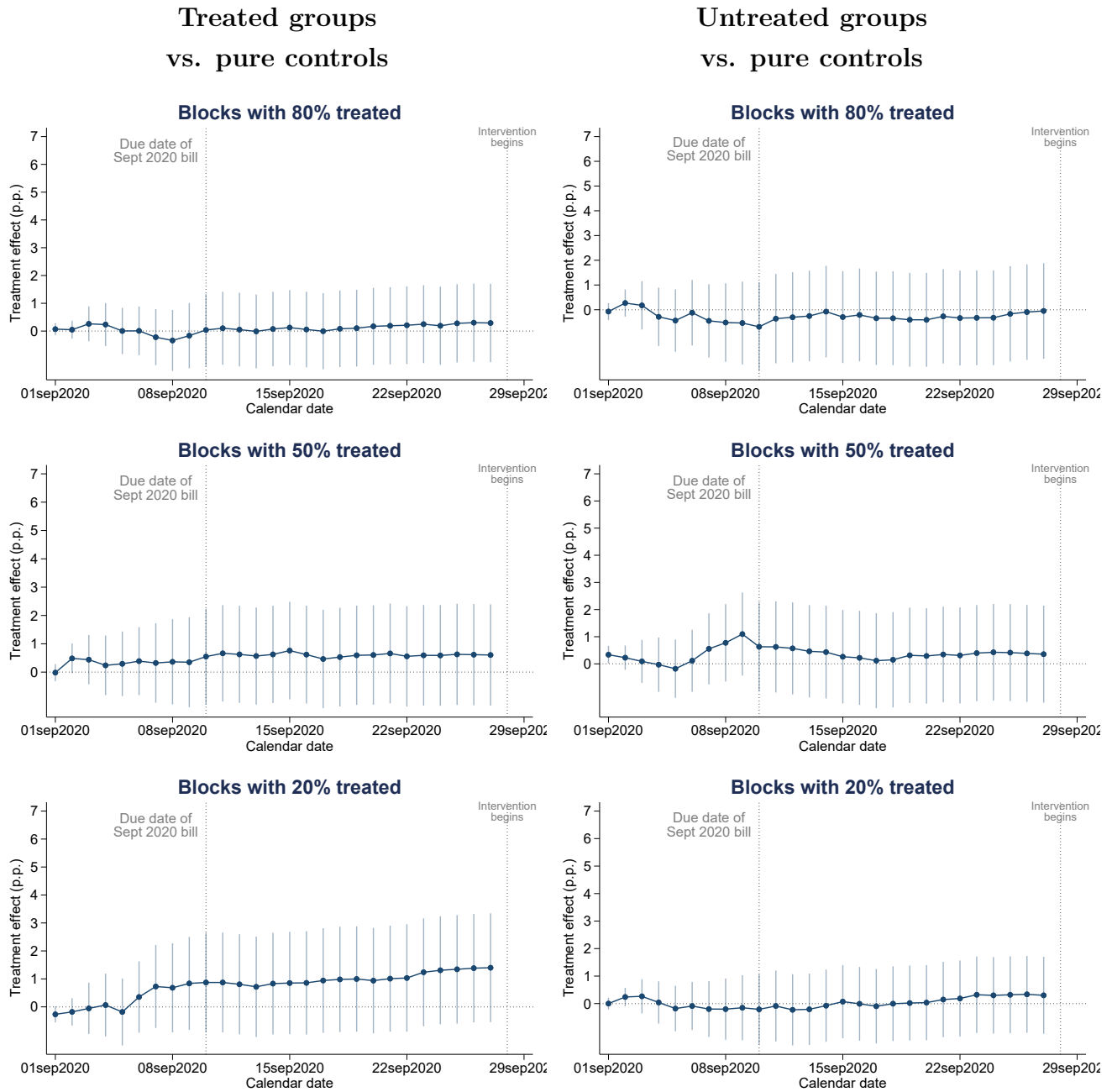


(b) Untreated vs. Pure Control



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.

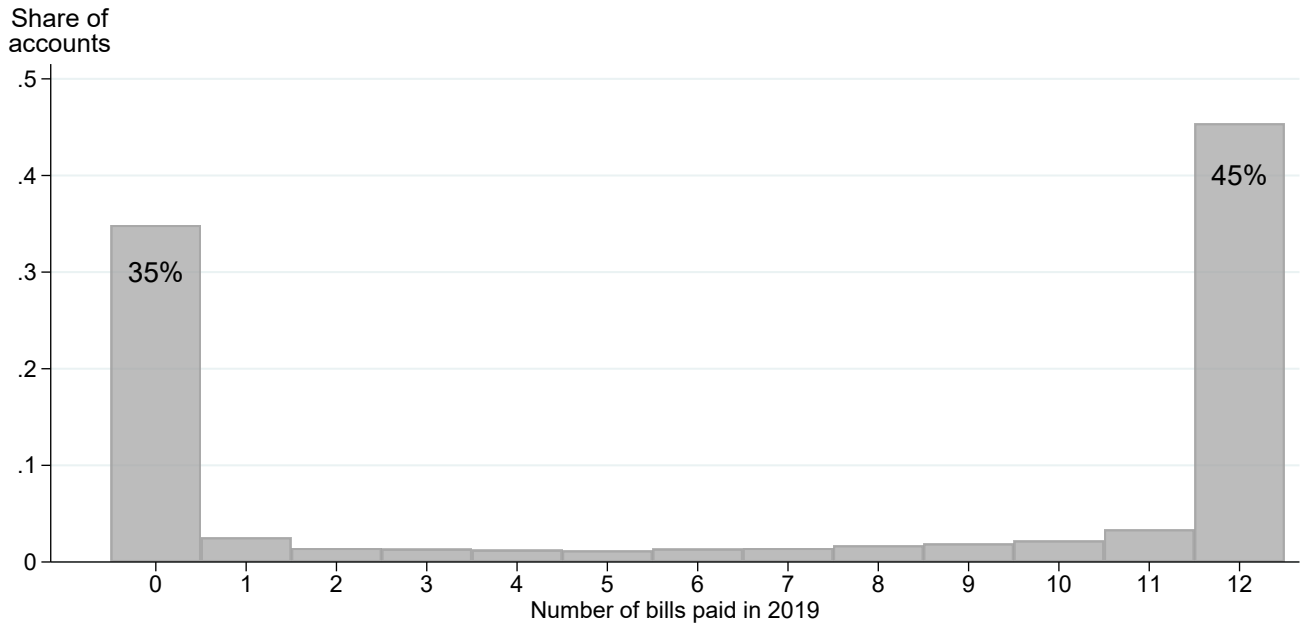
Figure A.4: Placebo. Direct and spillover effects for the pre-intervention Sep'20 bill



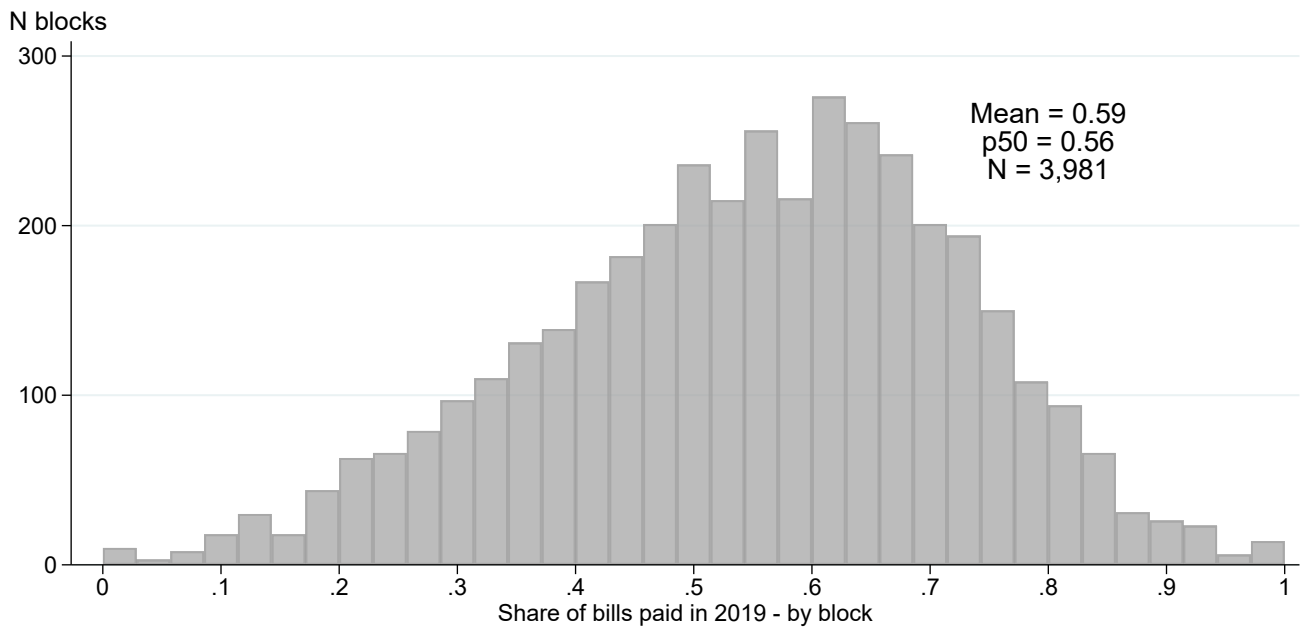
Notes: These figures show the coefficients and 95% confidence intervals from a saturated regression that computes, at each calendar day, the payment rate difference between each treated and untreated group relative to the pure control group (i.e., blocks where no accounts were treated). The top panel 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)



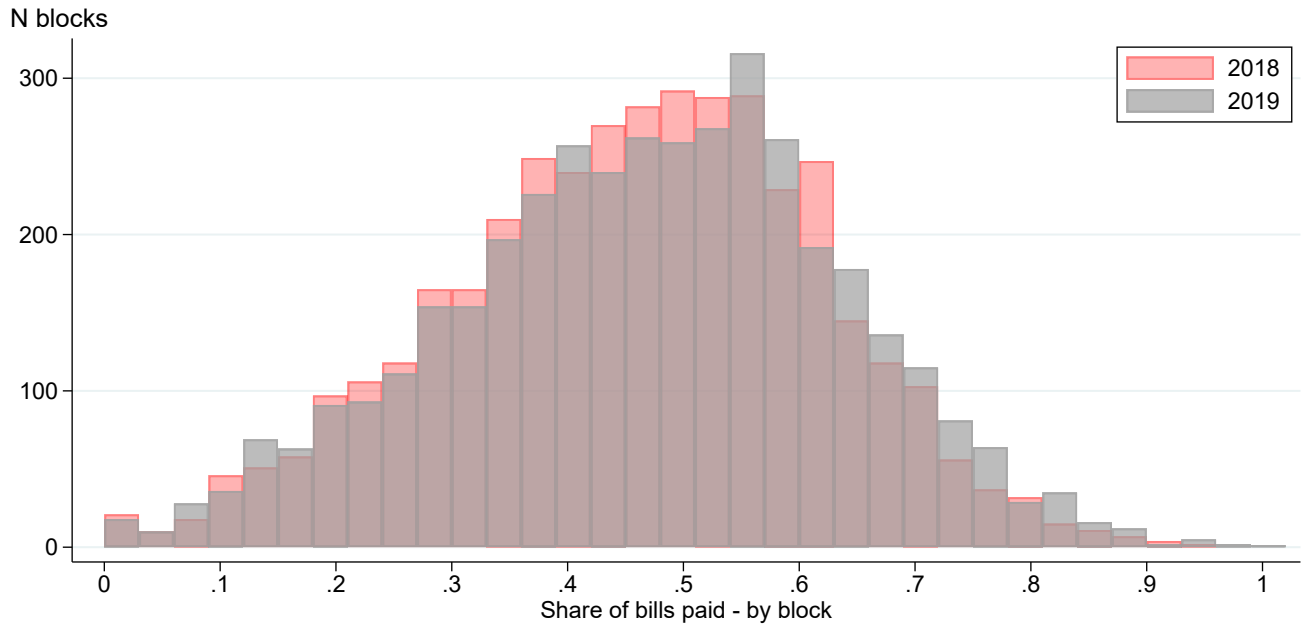
(b) Share of bills paid in 2019 (by blocks)



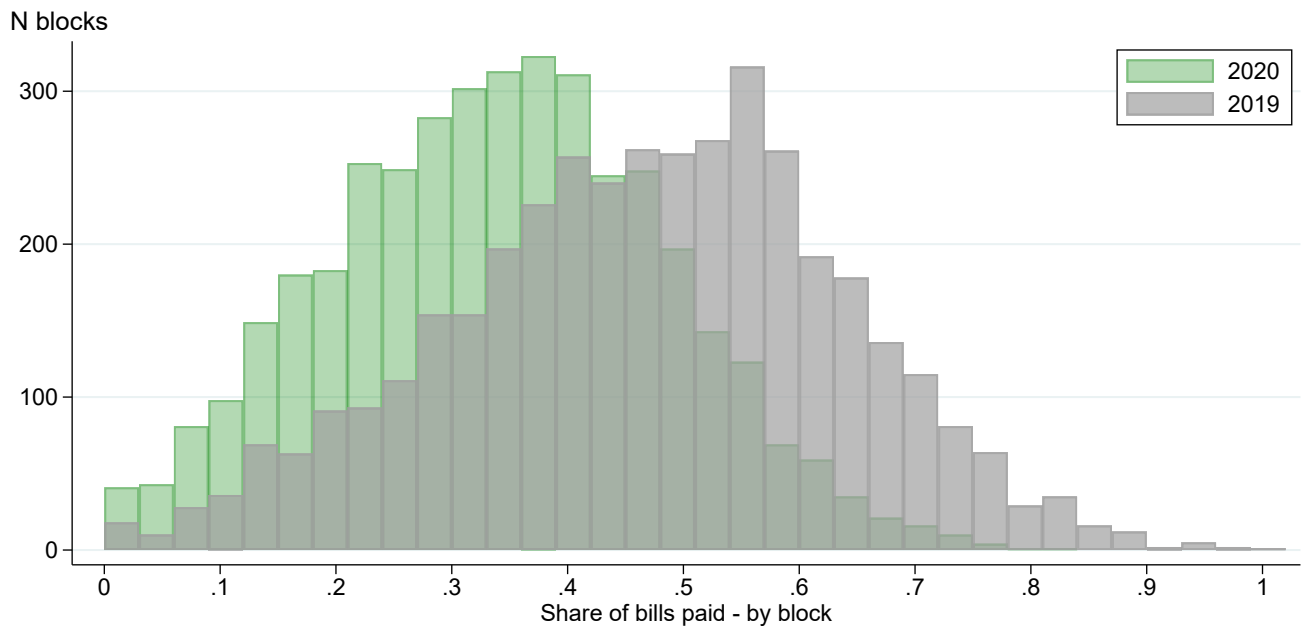
Notes: Panel (a) shows the distribution of the 68,806 accounts by the number of bills paid in 2019. The distribution is bi-modal with a core group of neighbors not paying any bill (35%) and another group paying all of them (45%). Panel (b) uses the information from panel (a) to compute the share of total 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

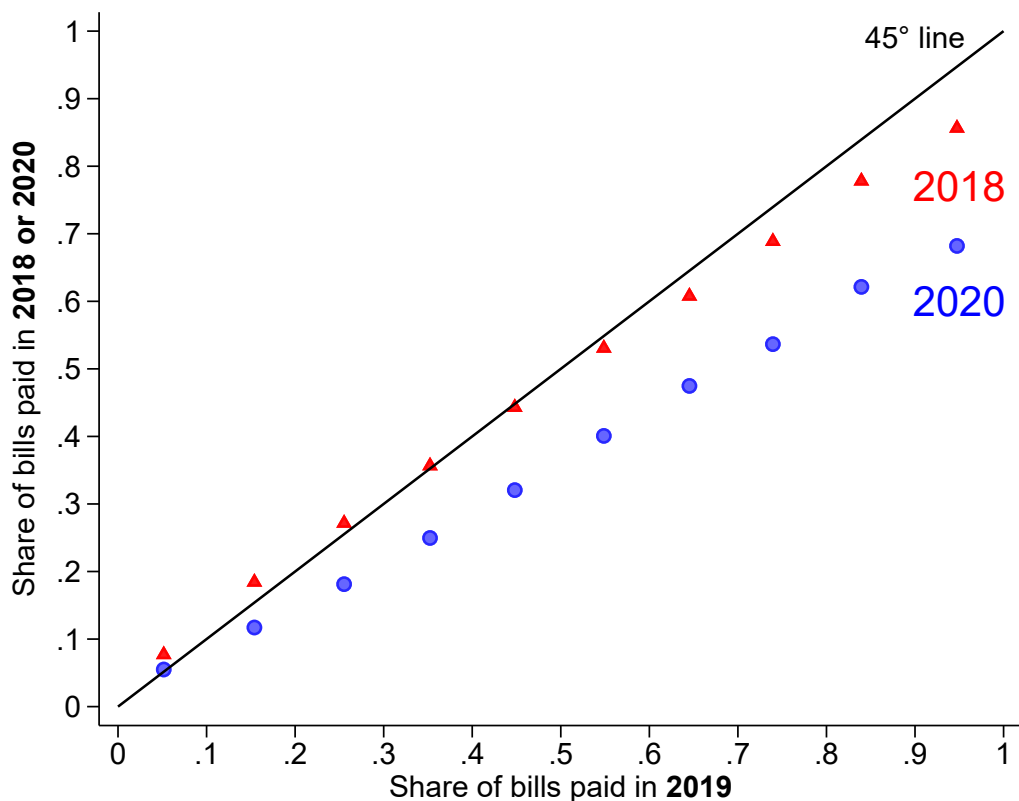


(b) 2019 vs 2020



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:

$$X_{ig} = \alpha + \sum_{t=1}^3 \theta_t \mathbb{1}(T_g = t)(1 - D_{ig}) + \sum_{t=1}^3 \tau_t \mathbb{1}(T_g = t)D_{ig} + \varepsilon_{ig} \quad (7)$$

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 (0.02)	-8.27 (17.77)	-0.00 (0.00)	-0.00 (0.01)	-0.14 (0.40)	2.81 (7.81)	0.05 (0.09)	-0.00 (0.01)
Untreated	0.00 (0.02)	-1.76 (20.70)	0.00 (0.01)	0.00 (0.01)	-0.53 (0.53)	6.27 (12.95)	-0.06 (0.12)	-0.00 (0.01)
B. Blocks with 50% treated:								
Treated	0.01 (0.02)	12.65 (20.38)	-0.00 (0.01)	-0.00 (0.01)	-0.47 (0.50)	1.16 (9.21)	0.03 (0.11)	0.00 (0.01)
Untreated	0.01 (0.02)	25.30 (20.66)	-0.00 (0.01)	-0.00 (0.01)	-0.42 (0.48)	1.88 (9.66)	0.02 (0.11)	0.01 (0.01)
C. Blocks with 20% treated:								
Treated	0.02 (0.02)	32.57* (16.79)	-0.01 (0.01)	0.01 (0.01)	0.10 (0.54)	5.94 (9.55)	0.07 (0.12)	-0.01 (0.01)
Untreated	0.02 (0.02)	19.14 (14.05)	-0.01 (0.00)	-0.01 (0.01)	0.12 (0.40)	1.32 (7.77)	0.00 (0.09)	0.00 (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 (7) 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 5 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.

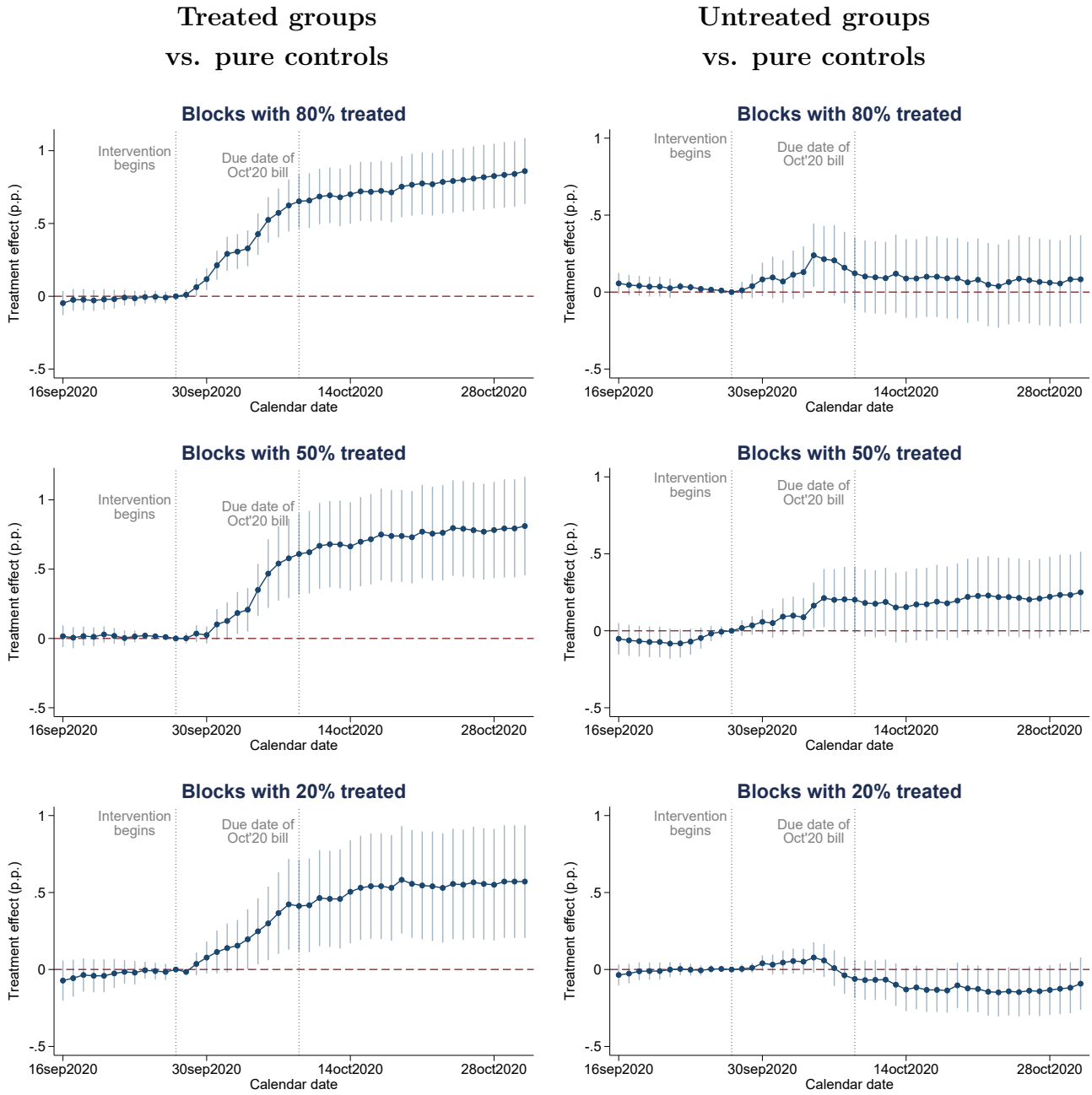
Lastly, Table A2 summarizes the corresponding diff-in-diffs estimates reported in Figures A.8, with the same structure as Table 3.¹ To benchmark our estimates, in the last row we report the share of e-bill subscribers in pure control blocks on September 27 (our baseline date). For treated accounts, the table shows an immediate effect in the three saturation groups that increases over

¹Column (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: Pr(subscribe to e-bill)	Placebo:	Intervention:	
	By Sep 20 (1)	Early (2)	By Oct 31 (3)
<i>A. Blocks with 80% treated</i>			
Treated	-0.02 (0.04)	0.31*** (0.06)	0.86*** (0.12)
Untreated	0.04 (0.03)	0.11 (0.08)	0.08 (0.15)
<i>B. Blocks with 50% treated</i>			
Treated	0.03 (0.03)	0.18** (0.08)	0.81*** (0.18)
Untreated	-0.07 (0.05)	0.10 (0.06)	0.25* (0.13)
<i>C. Blocks with 20% treated</i>			
Treated	-0.04 (0.05)	0.15* (0.08)	0.57*** (0.19)
Untreated	-0.01 (0.03)	0.05 (0.04)	-0.09 (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 subscription 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.

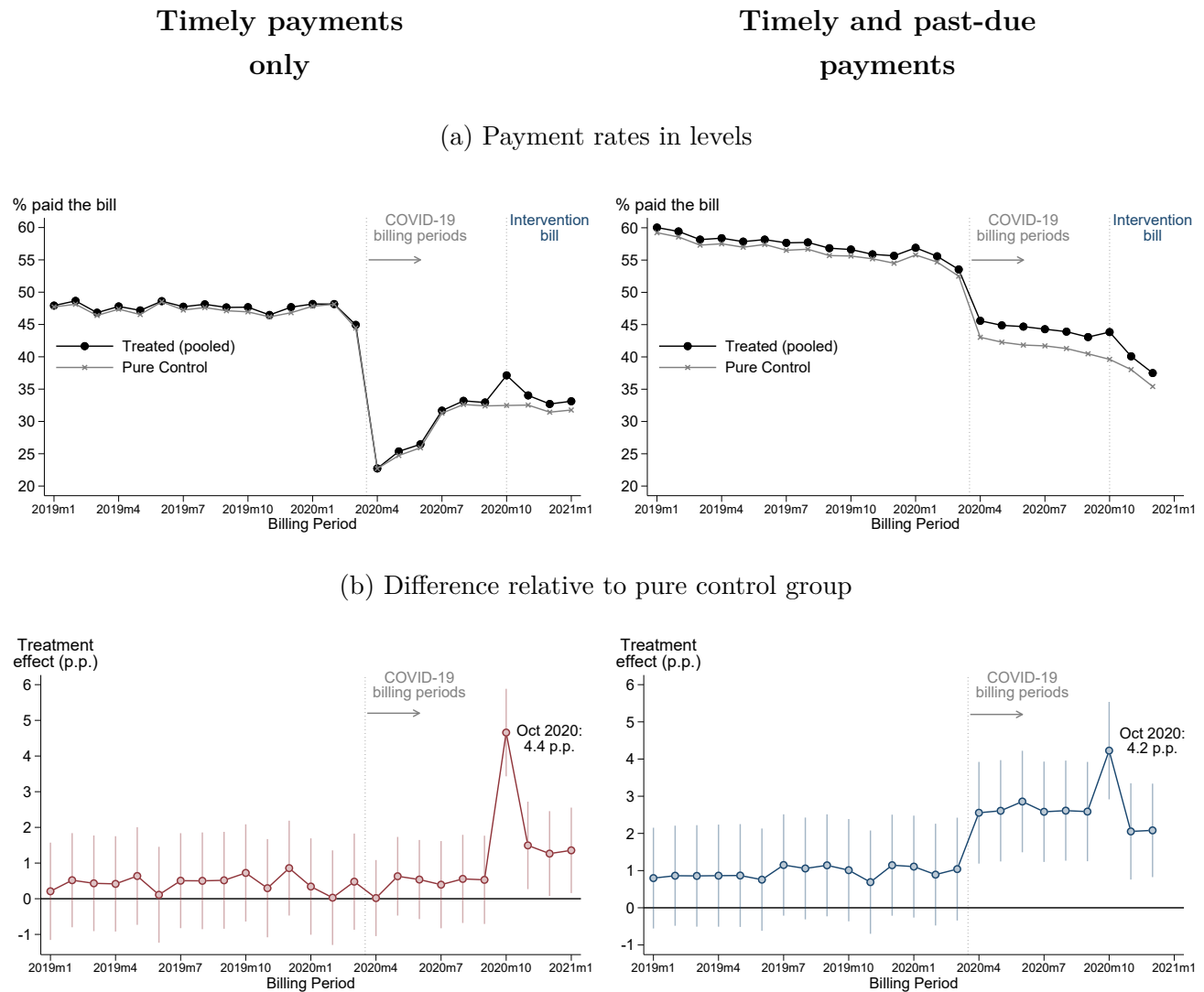
Intuitively, neighbors can pay their property tax bill at any time before or after the due date and, hence, payments from previous billing periods can also be affected by our intervention.² To illustrate this, the left panels of Figure A.9 only consider timely payments, defined as bills paid before the 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.

²The 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.

B Experimental Design: Additional Material

B.1 Choice of q_t and 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:

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

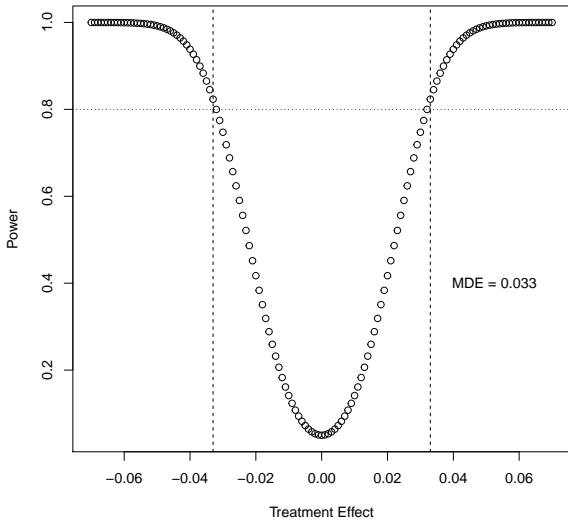
The final sample sizes depicted in Table 1 respond to logistical and other practical considerations. For power calculations, Figure B.10 plots the power function for each estimator, using the following parameters:

- $\sigma^2(d, t) = 0.25$ 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.

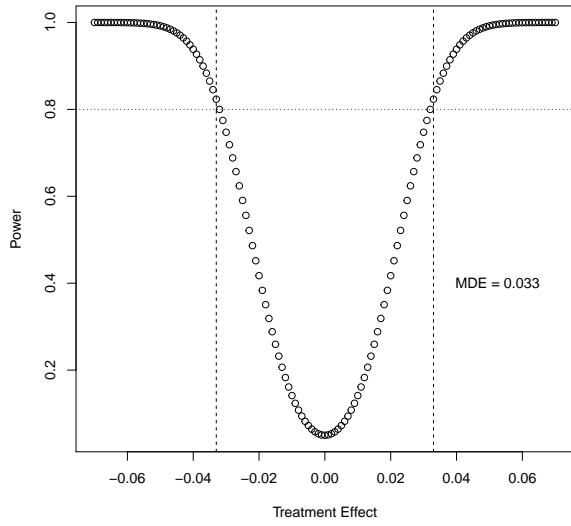
The power calculations give a minimum detectable effect between 2.6 and 3.3 percentage points.

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.

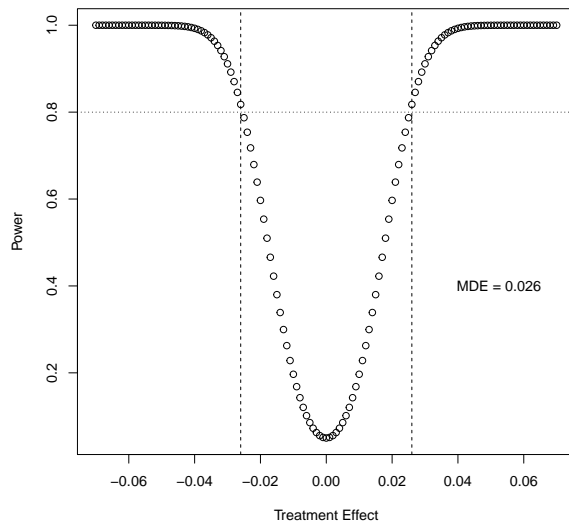
Figure B.10: Power functions



(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 $(d, t) = (0, 1)$

B.2 Simulations for power calculations

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

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

with $U_g \sim \text{Uniform}(0, 1)$. 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 $U_{ig}^k \sim \text{Uniform}(0, 1)$ 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,

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

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 $\text{ICC} = 0.05$.

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.

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.

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

C Supplemental Econometric Appendix

C.1 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, Jiang and Malani \(2021\)](#).

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

For our numerical illustration, we calculate the estimators standard errors and minimum detectable effects based on our formulas from Section 3 using the group 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 group 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 consider a design with four saturations, $p_0 = 0$, $p_1 = 0.2$, $p_2 = 0.5$, $p_3 = 0.8$, and calculate optimal probabilities $\{q_0, q_1, q_2, q_3\}$ based on Proposition 2. 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 three values for the intraclass correlation, $\rho \in \{0.1, 0.5, 0.8\}$. 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 $\rho = 0.8$.

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
$\rho = 0.5$						
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

Figure 1 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 for the group size distribution. The ratio of standard errors 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.

C.2 Within-Group Assignment Mechanisms

C.2.1 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

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

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 $\{n_g\}_{g=1}^G$.

C.2.2 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

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

C.3 Definitions and Regularity Conditions for Main Results

Let $\mathbb{1}_{ig}(a_k) = \mathbb{1}(A_{ig} = a_k)$ and consider the regression:

$$Y_{ig} = \sum_{k=0}^K \theta_k \mathbb{1}_{ig}(a_k) + u_{ig} = \mathbb{1}'_g \boldsymbol{\theta} + u_{ig}$$

where by construction $\theta_k = \mathbb{E}[Y_{ig}|A_{ig} = a_k]$ and $\mathbb{E}[u_{ig}|\mathbf{A}_g] = 0$. The OLS estimator is given by:

$$\hat{\boldsymbol{\theta}} = \left(\sum_g \mathbb{1}'_g \mathbb{1}_g \right)^{-1} \sum_g \mathbb{1}'_g \mathbf{y}_g$$

where $\mathbf{y}_g = (Y_{1g}, Y_{2g}, \dots, Y_{n_gg})'$. Next, let $\mathbf{u}_g = (u_{1g}, u_{2g}, \dots, u_{n_gg})'$ and define:

$$\Omega_n = \frac{1}{n} \sum_g \mathbb{E}[\mathbb{1}'_g \mathbf{u}_g \mathbf{u}'_g \mathbb{1}_g]$$

$$W_n = \frac{1}{n} \sum_g \mathbb{E}[\mathbb{1}'_g \mathbb{1}_g]$$

$$\tilde{V}_n = W_n^{-1} \Omega_n W_n^{-1}.$$

We introduce the following regularity conditions. In what follows, let $\lambda_{\min}(Q)$ denote the minimum eigenvalue of matrix Q .

Assumption 3 (Regularity conditions) *The following conditions hold.*

1. There exists a constant \tilde{C} such that $\lambda_{\min}(W_n) \geq \tilde{C} > 0$.
2. There exists a constant λ such that $\lambda_{\min}(\Omega_n) \geq \lambda > 0$.
3. $\sup_{i,g} \mathbb{E}[|Y_{ig}|^{10}] < \infty$.

Parts 1 and 2 above ensure that covariance matrices are well-defined. Part 1 ensures that there are no empty assignment cells. Because W_n is diagonal, Part 1 is equivalent to $\min_{a_k} \sum_g n_g \pi_g(a_k)/n \geq \tilde{C} > 0$. If the assignment probabilities are equal across groups, this condition reduces to $\min_{a_k} \pi(a_k) \geq \tilde{C}$, so that all assignment probabilities are bounded away from zero. To get intuition on Part 2, in the case with only two treatments $A_{ig} \in \{0, 1\}$ and homoskedasticity, this requirement reduces to assuming that the intraclass correlation satisfies $|\rho| < 1$. More generally, this condition restricts the amount of intraclass correlation to ensure that all the elements in the covariance matrix are well-defined. Finally, Part 3 imposes bounded moments of the outcome. Notice that this assumption is automatically satisfied when the outcome itself is bounded (e.g. binary), as is the case for most of our outcomes.

C.4 Proof of Proposition 1

We verify the assumptions for Theorem 9 in Hansen and Lee (2019). First, by direct calculation, the matrices defined in Section C.3 are:

$$W_{n,kk} = \frac{\sum_g n_g \pi_g(a_k)}{n}, \quad W_{n,kl} = 0, \quad k \neq l$$

and

$$\begin{aligned} \Omega_{n,kk} &= \sigma^2(a_k) \frac{\sum_g n_g \pi_g(a_k)}{n} + c(a_k, a_k) \frac{\sum_g n_g (n_g - 1) \pi_g(a_k, a_k)}{n}, \\ \Omega_{n,kl} &= c(a_k, a_l) \frac{\sum_g n_g (n_g - 1) \pi_g(a_k, a_l)}{n}, \quad k \neq l. \end{aligned}$$

As a result,

$$\begin{aligned} \tilde{V}_{n,kk} &= \frac{n\sigma^2(a_k)}{\sum_g n_g \pi_g(a_k)} \left\{ 1 + \rho(a_k, a_k) \frac{\sum_g n_g (n_g - 1) \pi_g(a_k, a_k)}{\sum_g n_g \pi_g(a_k)} \right\} \\ \tilde{V}_{n,kl} &= n\sigma(a_k)\sigma(a_l)\rho(a_k, a_l) \frac{\sum_g n_g (n_g - 1) \pi_g(a_k, a_l)}{\sum_g n_g \pi_g(a_k) \sum_g n_g \pi_g(a_l)}, \quad k \neq l. \end{aligned}$$

Under Assumptions 2, 1 and 3, and Condition (2), the conditions for Theorem 9 in Hansen and Lee (2019) hold and thus for any sequence of full-rank $(K + 1) \times J$ matrices R_n ,

$$(R_n' \tilde{V}_n R_n)^{-1/2} R_n' \sqrt{n}(\hat{\theta} - \theta) \rightarrow_{\mathcal{D}} \mathcal{N}(\mathbf{0}, I_J).$$

Finally, letting:

$$R'_n = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & & \\ -1 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

we obtain:

$$R'_n \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}, \quad R_n \boldsymbol{\theta} = \boldsymbol{\beta}, \quad V_n = R'_n \tilde{V}_n R_n$$

where:

$$\begin{aligned} V_{n,kk} &= \frac{n\sigma^2(a_k)}{\sum_g n_g \pi_g(a_k)} \left\{ 1 + \rho(a_k, a_k) \frac{\sum_g n_g(n_g - 1) \pi_g(a_k, a_k)}{\sum_g n_g \pi_g(a_k)} \right\} \\ &+ \frac{n\sigma^2(a_0)}{\sum_g n_g \pi_g(a_0)} \left\{ 1 + \rho(a_0, a_0) \frac{\sum_g n_g(n_g - 1) \pi_g(a_0, a_0)}{\sum_g n_g \pi_g(a_0)} \right\} \\ &- 2n\sigma(a_k)\sigma(a_0)\rho(a_k, a_0) \frac{\sum_g n_g(n_g - 1) \pi_g(a_k, a_0)}{\sum_g n_g \pi_g(a_k) \sum_g n_g \pi_g(a_0)} \end{aligned}$$

and

$$\begin{aligned} V_{n,kl} &= \frac{n\sigma^2(a_0)}{\sum_g n_g \pi_g(a_0)} \left\{ 1 + \rho(a_0, a_0) \frac{\sum_g n_g(n_g - 1) \pi_g(a_0, a_0)}{\sum_g n_g \pi_g(a_0)} \right\} \\ &+ n\sigma(a_k)\sigma(a_l)\rho(a_k, a_l) \frac{\sum_g n_g(n_g - 1) \pi_g(a_k, a_l)}{\sum_g n_g \pi_g(a_k) \sum_g n_g \pi_g(a_l)} \\ &- n\sigma(a_k)\sigma(a_0)\rho(a_k, a_0) \frac{\sum_g n_g(n_g - 1) \pi_g(a_k, a_0)}{\sum_g n_g \pi_g(a_k) \sum_g n_g \pi_g(a_0)} \\ &- n\sigma(a_l)\sigma(a_0)\rho(a_l, a_0) \frac{\sum_g n_g(n_g - 1) \pi_g(a_l, a_0)}{\sum_g n_g \pi_g(a_l) \sum_g n_g \pi_g(a_0)} \end{aligned}$$

which completes the proof. \square

C.5 Proof of Proposition 2

Based on Equation (5), the minimization problem is equivalent to:

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

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:

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

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:

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

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

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

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