

# 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*

Julian Amendolaggine, *Universidad Nacional de La Plata*

Juan Luis Schiavoni, *Universidad Nacional de La Plata*

February 4, 2022

### Abstract

We set up a framework to carry out field experiments—whereby units are grouped into mutually exclusive clusters—that allows for multiple treatments and general forms of intraccluster correlation. We improve upon existing methods by allowing for cluster size heterogeneity, which is typically ignored when designing experiments. Based on data from several existing experiments, we show that ignoring cluster size heterogeneity can severely overestimate power and underestimate minimum detectable effects. We derive formulas for optimal group-level assignment probabilities as well as the power function used to calculate power, sample size, and minimum detectable effects for each treatment effect. We apply our results to the design of partial population experiments for estimating spillover effects and we run a large-scale randomized tax communication campaign in a municipality of Argentina to estimate total and neighborhood spillover effects on property tax compliance and e-billing sign up. We find evidence of spillover effects for clusters with a high proportion of treated units.

JEL CODES: H71 , H26 , H21 , O23.

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

---

\*Corresponding author: Guillermo Cruces, E-mail: [guillermo.cruces@nottingham.ac.uk](mailto:guillermo.cruces@nottingham.ac.uk). The design for this experiment was preregistered with the AEA RCT Registry (RCT ID: **AEARCTR-0006569**). All 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 for the design of randomized controlled trials and the assessment of their impact. The design of such experiments and the identification of spillover effects, however, is a tall order. In this paper, we set up a general framework to carry out experiments of this type, and implement our methods in a large-scale field experiment to estimate spillovers in property tax compliance where the population of interest consists of taxpayers in residential blocks.

We consider the design of experiments in a sample where units are grouped into mutually exclusive clusters (as in, e.g., [Duflo and Saez, 2003](#); [Crépon et al., 2013](#)). We provide an asymptotic distributional approximation and variance formulas to conduct power, sample size and minimum detectable effects calculations for general clustered experimental designs allowing for multiple treatments and general forms of heteroskedasticity and intracluster correlation. We improve upon existing methods by allowing for cluster size heterogeneity, which affects the variance of treatment effect estimators but is typically ignored when designing experiments. Based on data from several existing experiments, we 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. 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.

We implement our methods in a large-scale field experiment 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. In July 2020, this municipality enabled a paperless billing option for taxpayers to receive their monthly bill by e-mail *in lieu* of the regular paper bill. Our campaign consists of sending letters to randomly selected dwellings where we remind neighbors about the electronic option, how to sign up, and we also include information about the status of the account, due dates, and past due debt. We are particularly interested in whether the campaign creates spillover effects on neighbors that live nearby within a treated block but that do not receive a letter.

Randomization took place in two stages. 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. In the second stage, we randomly selected accounts within the last three groups of blocks to receive the treatment letter. We sent approximately 25,000 treatment 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 people targeted by the intervention (treated) relative to pure control blocks) and spillover effects (the behavior of non-targeted neighbors (untreated) within treated blocks relative to pure control blocks) on monthly payments and electronic billing sign-up.

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) and on subscriptions to electronic billing. Our results reveal higher payment rates and e-bill subscriptions of treated and untreated accounts relative to neighbors in pure control blocks where nobody received the communication 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 16% of the reference 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 and precisely estimated. In high-saturation blocks with 80% treated accounts, the effect on payment rates of untreated accounts is about 1 percentage point and statistically significant in the early days of the intervention, losing significance from the due date onward. Intuitively, the early days are precisely when social interactions are more likely to occur. Conversely, social interference is absent in blocks with only 20% treated accounts. In that case, the spillover effect for untreated units oscillates around zero. In addition, blocks with 50% and 80% treated properties appear to respond more strongly than blocks where only 20% was treated. We argue this could be interpreted as some sort of spillover effect. Intuitively, interventions are likely to create interference among treated units within a block when more people receive the letter, nudging some of them to pay the bill.

Our intervention also increases subscriptions to the electronic billing option for treated and untreated accounts, with greater effects in high-saturation blocks, albeit small in absolute value. We first plot the cumulative share of e-billing subscribers over time and show a sharp increase right after the letters started to be delivered. We then run dynamic difference-in-differences regressions comparing subscription rates between each saturation group and pure control blocks, day by day, relative to September 27, 2020 (our baseline date). For treated accounts, we find an immediate

direct effect in the three saturation groups that increases over time. This direct effect is always higher in blocks with 80% treated units, consistent with the presence of spillovers or interference that strengthens the effect of the letter. 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 people switched to the digital bill.

For untreated accounts, the raw data also suggest the presence of spillover effects in subscriptions to e-billing in high-saturation blocks. These effects are precisely estimated but harder to detect, and only significant at the 5-10% level at the beginning of the intervention. The most clear effect arises in blocks with 50% treated accounts with a spillover effect of 0.25 percentage points. The somewhat absence of spillovers in this case might be explained by the fact that the outcome of analysis (e-bill subscriptions) has very low take up, making it harder for interference between neighbors to trigger any meaningful response.

Lastly, for completeness we also estimate the effects of the intervention on backward and forward payments corresponding to other billing periods before and after October 2020. Overall, payment rate levels are low and decrease over time, possibly reflecting the Argentine economic crisis. This is particularly evident in April 2020, when the COVID-19 pandemic hit Argentina and payment rates decreased sharply from 55% to 45%. Despite this general downward trend, when we look at the difference in payment rates between treated and pure control accounts, they reveal a noticeable increase in the pandemic billing periods. Although the October bill presents the highest effect (4.2 percentage points), the letters had some residual effects beyond the intervention month in November and December 2020. Moreover, the letters also induced some neighbors to make backward payments to cancel past due debt, but only for the pandemic bills (April to September 2020).

**Comparison with existing literature.** In terms of the methodological contribution, our results 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, and for heterogeneity in cluster sizes. 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](#)), and thus ignoring this heterogeneity when designing experiments results in overestimating power and underestimating MDEs. 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 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, [Jiang and Imai \(2021\)](#) analyze two-stage completely randomized experiments and provide randomization-based variance estimators and sample size formulas. Our results complement this paper by considering more general assignment mechanisms and by conducting super-population-based large-sample inference in a double array asymptotic framework. This approach allows us to determine the effect of cluster size heterogeneity in the asymptotic distribution of the treatment effect estimators. We also complement the 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.<sup>1</sup> While the latter are randomized controlled trials, they do not address directly the issue of potential spillovers in compliance at the local level. 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 spillovers in mind.

The paper is organized as follows. 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, and the empirical strategy. Section 5 displays the total and neighborhood spillover effects uncovered in the empirical analysis. Section 6 provides some concluding remarks.

---

<sup>1</sup>For previous work in Argentina see [Castro and Scartascini \(2015\)](#) and [Lopez-Luzuriaga and Scartascini \(2019\)](#).

## 2 Why is Accounting for Cluster Heterogeneity Important?

As mentioned, one of our main methodological contributions is to provide variance and power formulas that account for cluster size heterogeneity. 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. 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 conducting two-stage experiments (see Section C for 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,  $sd(n_g) = 15.2$ . We then compare these “adjusted” standard errors and MDEs with the “undajusted” 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,  $\rho$ . The figure shows that the ratio grows rapidly as  $\rho$  increases, 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 the adjusted and unadjusted MDEs and reveals that even for values of  $\rho$  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  $\rho$ .

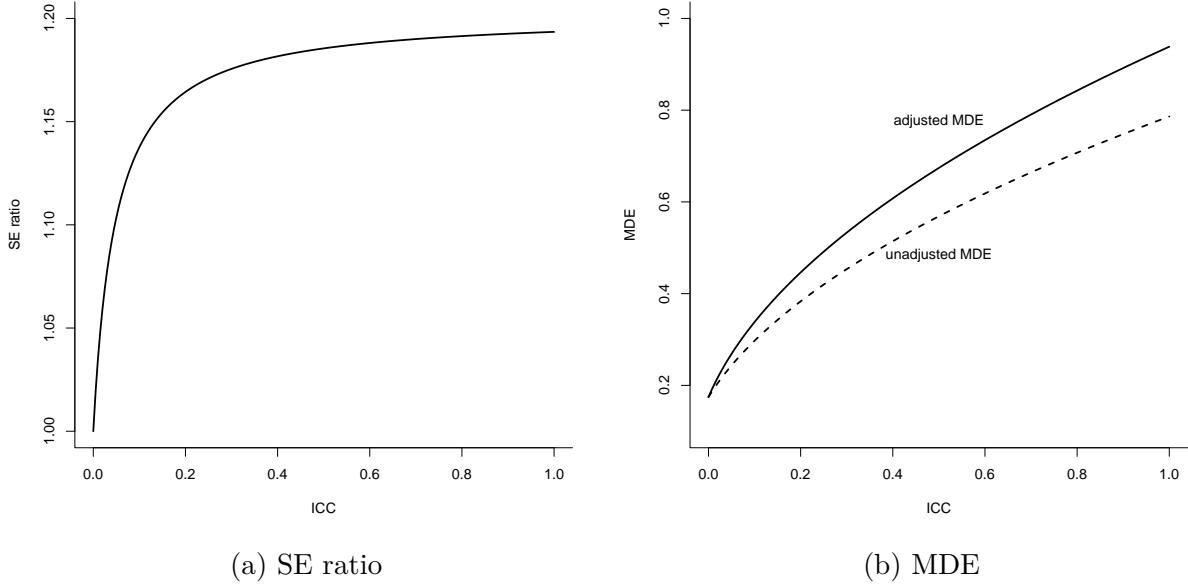
As mentioned above, the underestimation of standard errors and MDEs becomes more severe as the ratio of variance to mean of group sizes increases. For instance, in this illustration, keeping a standard deviation of group sizes of  $sd(n_g) = 15.2$  but reducing average group size to  $\bar{n} = 18$  results in adjusted MDEs that can be between 25% and 30% larger than the unadjusted ones.

## 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. 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](#)),

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 A2:  $G = 95$ ,  $\bar{n} = 23.3$ ,  $sd(n_g) = 15.2$ .

or households in villages or other geographic administrative units ([Ichino and Schündeln, 2012](#); [Haushofer and Shapiro, 2016](#); [Giné and Mansuri, 2018](#)). In our application, the population of interest consists of taxpayers in residential blocks.

We consider a random 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 start by considering 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  $\mathbf{A}_g = (A_{1g}, \dots, A_{n_g g})$  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 are allowed to be different across clusters, which allows, for example, for stratification at the cluster level.

To define parameters of interest, we introduce the following assumption.

**Assumption 1 (Conditional Moments)** *For all  $i, j$  and  $g$ ,*

- (i)  $\mathbb{E}[Y_{ig}|A_{ig} = a_k] = \mu(a_k)$  for all  $a_k \in \mathcal{A}$ ,
- (ii)  $\mathbb{V}[Y_{ig}|A_{ig} = a_k] = \sigma^2(a_k)$  for all  $a_k \in \mathcal{A}$ ,
- (iii)  $\text{Cov}(Y_{ig}, Y_{jg}|A_{ig} = a_k, A_{jg} = a_l) = c(a_k, a_l)$  for all  $a_k, a_l \in \mathcal{A}$ .

This assumption imposes identical 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.

The parameters of interest are comparisons of conditional averages  $\mathbb{E}[Y_{ig}|A_{ig} = a_k] - \mathbb{E}[Y_{ig}|A_{ig} = a_0] = \mu(a_k) - \mu(a_0)$  for  $k = 1, \dots, K$ . We consider the following saturated regression:

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

where by construction  $\beta_k = \mu(a_k) - \mu(a_0)$  and  $\mathbb{E}[\varepsilon_{ig}|A_{ig}] = 0$ . The vector of treatment effects  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)'$  can be estimated using standard linear regression methods.

We introduce the following restrictions on the treatment assignment mechanism.

### Assumption 2 (Assignment Mechanism)

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

Part (i) of Assumption 2 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 (ii) rules out the case in which some treatment values  $a_k$  have zero probability in the sample.

## 3.2 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  $\hat{\boldsymbol{\beta}}$ .

In what follows, we define  $\sigma^2(a_k) = \mathbb{V}[Y_{ig}|A_{ig} = a_k]$ ,  $c(a_k, a_l) = \text{Cov}(Y_{ig}, Y_{jg}|A_{ig} = a_k, A_{jg} = a_l)$ ,  $\rho(a_k, a_l) = c(a_k, a_l)/(\sigma(a_k)\sigma(a_l))$ ,  $\pi_g(a_k, a_l) = \mathbb{P}_g[A_{ig} = a_k, A_{jg} = a_l]$  for  $i \neq j$  and we let “ $\rightarrow_D$ ” 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 when groups can be large and heterogeneous in size. The commonly analyzed setting with fixed  $n_g$  and/or equally-sized clusters ( $n_g = \bar{n}$  for all  $g$ ) is a particular case of our setup. 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:

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

The proof and the full shape of the covariance matrix (including the off-diagonal elements) are given in supplemental appendix F.3. 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:

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

This formula corresponds to the variance of a difference in means with clustered data.<sup>2</sup> 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.

---

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

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  $\mathbb{P}_g[A_{ig} = a_k, A_{jg} = a_0] = 0$ .

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 intracluster 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. Also notice that conducting inference on linear combinations or smooth functions of the coefficients in  $\hat{\beta}$  is straightforward using the delta method. See Section F 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).<sup>3</sup> 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.

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:

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

which corresponds to Equation (6) in [Duflo, 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

---

<sup>3</sup>Notice 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”.

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 [Duflo, Glennerster and Kremer \(2007\)](#).

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  $\beta_k$  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 intracluster correlations), where  $z_{1-\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile from the standard normal distribution.

### 3.3 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  $t \neq 0$ .

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:

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

As mentioned, the variances and intracluster correlations are the only unknown parameters, whereas the group sizes are known 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 F.1 in the appendix discusses the two most common within-group assignment mechanisms and characterizes these probabilities explicitly.

**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. \(2018\)](#).

**Within-group assignment.** 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  $\xi_g$  and let:

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

so that  $\xi_g$  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  $\mathbb{P}_g[D_{ig} = 1|T_g = t] = p_t$ . See Section F.1 in the appendix for further details.

### 3.4 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 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  $p_t$  depend 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 these choices depend on the parameters that the researcher wants 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. The optimality criterion determines how the variances of all the parameters of interest are aggregated. The literature on optimal design of experiments has proposed several optimality 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).<sup>4</sup> 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.

**Proposition 2** *In the design described in Section 3.3, 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 F.4.

**Constrained Designs.** Often, researchers may need to incorporate different sets of constraints when choosing assignment probabilities. For example, the design may need to account for logistical, 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 set based on budgetary restrictions. 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.

---

<sup>4</sup>Notice that this criterion is different from the one in Baird et al. (2018), who minimize the average standard error.

## 4 A Randomized Tax Communication Campaign

We conducted a partial population experiment to estimate direct and spillover effects of simple communication messages on property tax compliance and electronic billing sign-up. 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 (*Tasa por Servicios Generales*). The experiment consists of a two-level randomized communication campaign where we sent a one-page letter informing neighbors of a new electronic billing option, how to sign up, and we also provided information of the current billing period, past due debt, and how to pay online or in person.<sup>5</sup> <sup>6</sup> The experiment was run on the universe of residential dwellings present in the municipality in 2019.

Randomization took place in two stages—first at the block level and then at the account 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. In the second stage, we randomly selected accounts within the last three groups of blocks to receive the treatment letter.

The timeline of the intervention is displayed in Figure 3. We sent approximately 25,000 treatment letters to account holders who are billed the *Tasa por Servicios Generales*. The letters were delivered between September 28th and October 7th, 2020, corresponding to payments due on October 9th, 2020 (month 10 billing period) as well as past due debt (if any). To study total effects, we measure the change in outcomes among people targeted by the intervention (treated) relative to pure control blocks. To study spillover effects, we analyze the behavior of non-targeted neighbors (untreated) within treated blocks relative to pure control blocks.

### 4.1 Administrative Data

For the empirical analysis, we use two 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

---

<sup>5</sup>Figure A.1 in the appendix provides an anonymous example of the intervention letter. Our letter is indeed very simple and emphasizes action-relevant information, in accordance with De Neve et al. (2021) who show that simplified tax letters are one of the most effective ways to increase tax compliance.

<sup>6</sup>The electronic billing option was officially launched in the municipality in June 2020, and this letter provided a simple reminder for neighbors to sign up to the digital billing option.

payment, days overdue, means of payment (cash or electronic), type of account (residential, retail store, factory), gender of the account holder, age of the account holder, linear front meters of the lot/property, assessed value of the property.

The second database contains the people that signed up to the electronic billing option, which was launched in the municipality in June 2020. 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.

**Baseline data.** For the randomization, power calculations, and simulations, we use baseline data from the year 2019. We make use of three different outcomes:

- `paid_all`: dummy variable equal to 1 if the account paid the twelve bills in 2019,
- `paid_some`: dummy variable equal to 1 if the account paid at least one bill in 2019,
- `paid_six`: dummy variable equal to 1 if the account paid six bills or more in 2019.

The data set is restricted to blocks with size between 8 and 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 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.

## 4.2 Treatment Assignment

Using the notation from Section 3.3, 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 D. 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  $(d, t)$ ;<sup>7</sup> (ii)  $\text{ICC} = 0.1$  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.<sup>8</sup>

<sup>7</sup>This gives a conservative estimate because 0.25 is the upper bound for the variance of a binary variable.

<sup>8</sup>Appendix Figure D.6 plots the power function for each estimator

### 4.3 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  $\varepsilon_{ig}$  to be correlated within blocks and use a cluster-robust variance estimator. In this regression,  $\theta_t$  is interpreted as the spillover effect on untreated units in groups with  $T_g = t$  and  $\tau_t$  is interpreted as the total effect on treated units in groups with  $T_g = t$ .

## 5 Results

### 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.<sup>9</sup> Recall that the due date was October 9th and the letters were delivered between September 28th and October 7th (see Figure 3). Hence, the October billing period constitutes our best shot at identifying immediate spillover effects. We start by showing compelling graphical evidence of the effect of the intervention in Figures 4 to 7 and then we summarize the corresponding point estimates in Table 3.

Figure 4 panel (a) shows the cumulative share of individuals paying the October 2020 bill over time, both for treated units and pure control blocks. The brown dashed line shows the payment rate for pure control units. The blue dashed line corresponds to treated units in group  $T_g = 1$  (blocks with 20% treated). The black dashed line corresponds to treated units in group  $T_g = 2$  (blocks with 50% treated). The red solid line corresponds to treated units in group  $T_g = 3$  (blocks with 80% treated). Panel (b) shows, for each calendar day, the difference between each treated group and the

---

<sup>9</sup>Appendix section B shows balance test regressions that confirm our groups are balanced and comparable.

pure control group (treatment effect coefficients).

Figure 4 reveals a clear direct effect of the intervention on treated accounts. The payment rate of treated units starts 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. Moreover, blocks with 50% and 80% treated properties appear to respond more strongly than blocks where only 20% was treated. This can be interpreted as a spillover effect on the treated, since the intervention could also create interference among treated units strengthening the effect of the letter. For example, neighbors who received the letter and talked to each other might decide to go ahead and pay the bill—a behavior that is more likely to arise when more people are treated in the block.

Likewise, Figure 5 panel (a) shows the cumulative payment rate of the October 2020 bill for untreated units. 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 day, the difference between each untreated group and the pure control group, and therefore captures spillover effects. For comparison, the gray solid line shows the treatment effects for treated units (pooled together from  $T_g = 1, 2, 3$  in Figure 4).

Although smaller in size, Figure 5 reveals a clear spillover effect of the intervention on untreated accounts. Spillover effects mainly arise in high-saturation blocks where 80% of the neighbors were treated. 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.

Figures 6 and 7 present the coefficients and 95% confidence intervals from a saturated regression that estimates, day by day, the difference in payment rates between each treated and untreated group relative to pure control blocks (see equation 6). Figure 6 confirms the previous results, with an immediate and statistically significant effect 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. Figure 7 shows that spillover effects are more modest in magnitude and precisely estimated. In high-saturation blocks with 80% treated accounts, the effect is about 1 percentage point and statistically significant in the early days of the intervention, losing significance from the due date onward.

Finally, Table 3 summarizes the corresponding point estimates for total and spillover effects

reported in figures 6 and 7. Panels A, B, and C display total effects and spillover effects in blocks where 20%, 50%, and 80% were treated, respectively. 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, none of the coefficients are statistically significant or large in magnitude. Columns (2) to (4) show the coefficients and block-clustered standard errors for October 2020 bill payments at three different dates: October 3 (early payments), October 9 (early and on time payments), and October 31 (early, on time, and 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.

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 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 people start to pay their bill, reaching 27.2% in pure control blocks by the due date, making small effects harder to detect. Accordingly, although the spillover effect on untreated units remains unchanged in size, it loses significance by the due date. In contrast, the total effect on treated units increases to 4.3 percentage points, representing 15.8% of the reference payment rate in pure control blocks. After the due date, people continue paying their overdue bill, reaching 34.4% in pure control blocks by the end of the month. Both total and spillover effects remain relatively constant.

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.

## 5.2 Effects on Subscriptions to Electronic Billing

We next analyze the effect of the intervention on subscriptions to electronic billing.<sup>10</sup> We follow a similar structure and start by showing convincing graphical evidence of total and spillover effects (Figures 8 to 11) which are then summarized in Table 4.

Figure 8 panel (a) shows the cumulative share of e-billing subscribers over time, both for treated units and units in pure control blocks. Panel (b) shows, for each calendar day, the difference between each treated group and the pure control group. These raw data clearly show the emergence of treatment effects right after the letters started to be delivered. Note, however, that both figures suffer from differential pre-trends in subscription rates across groups, adding some caution to the interpretation. We believe this is somewhat fortuitous and unrelated to the experiment. Nevertheless, as a way of mitigating this threat, we run a dynamic difference-in-differences comparing

---

<sup>10</sup>The e-billing option was launched by the municipality in June 2020.

subscription rates between each group and pure control blocks, day by day, relative to September 27, 2020 (our baseline date). We present the results in Figure 10. Three 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 treatment groups and the pure control group 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; and, (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.

Figures 9 and 11 reveal a similar story for untreated units and pure control blocks. For comparison, in panel (b) we also add a gray solid line with the effect of pooled treated units. Although less clear than the figures for treated units, the raw data 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 4 summarizes the corresponding diff-in-diffs estimates reported in Figures 6 and 7, with the same structure as Table 3.<sup>11</sup> 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 time. This effect is always 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 people switched to the digital bill.<sup>12</sup> 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 most clear 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.

---

<sup>11</sup>Column (1) validates the experiment by showing a placebo saturated regression that compares subscription rates between each group and the pure control group on September 17, before the intervention began. None of the coefficients are statistically significant or large in magnitude.

<sup>12</sup>Recall that this electronic billing option was launched in June 2020.

### 5.3 Backward and Forward Payments

For completeness, in the last part of our analysis we analyze the effects of the intervention on backward and forward payments corresponding to other billing periods before and after month 10. These results are summarized in Figure 12.

Figure 12 panel (a) shows payment rates in levels for treated units pooled together (black line) and pure control units (gray line), for 24 consecutive monthly bills between January 2019 and December 2020. Panel (b) reports total treatment effects—i.e., the difference between treated and pure control units—and 95% confidence intervals for the 24 billing periods (blue line). Importantly, these payment rates include any payment made until December 2020 and, therefore, capture both timely and overdue payments. The latter are particularly relevant in our context, because the treatment letter included past due balances and can therefore induce neighbors to make backward payments to cancel their debt. To visualize this point and validate our experiment, we estimate the difference in payment rates between treated and control accounts but only considering payments made up to September 27, 2020, a day before letters started to be delivered. We superimpose the coefficients and confidence intervals in panel (b) (red line).

Three important points are worth highlighting from Figure 12: (1) Overall, payment rate levels are low and decrease over time, possibly reflecting the Argentine economic crisis. This is particularly evident in April 2020, when the COVID-19 pandemic hit Argentina and payment rates decreased sharply from 55% to 45%; (2) the difference in payment rates between treated and pure control accounts (blue line in panel b) experiences a noticeable increase in the pandemic billing periods. Although the October bill when the intervention took place presents the highest effect (4.2 percentage points), the letters had some residual effect beyond the intervention month (November and December bills), and also nudged some neighbors to catch up and make backward payments for the pandemic bills (April to September 2020),<sup>13</sup> and (3) placebo total effects (red line), based on payment rates constructed with pre-intervention payments, are precisely estimated and not different from zero. This is reassuring and implies that our sample is balanced and that the effects we estimate are indeed caused by our experiment.

## 6 Conclusion

We provided a general framework to carry out partial population experiments with an application to spillovers in property tax compliance. We first derived an asymptotic approximation and variance formulas to conduct power calculations for general clustered experimental designs allowing for

---

<sup>13</sup>Note that the letters also had a small effect on the billing periods pre-COVID, since the blue line is above the red line (i.e., treated accounts are more likely to cancel their debt relative pure control accounts and to pre-intervention payments), but this effect is not statistically significant.

multiple treatments, general forms of intracluster correlation, and cluster size heterogeneity.

One of the main methodological contributions is to provide variance and power 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 estimated total and neighborhood spillover effects of a randomized communication campaign on property tax compliance in a large municipality of Argentina where neighbors are required to pay a monthly bill on their real estate. We estimate total effects on monthly payments and sign-up to electronic billing, 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 and on subscriptions to electronic billing. Our results reveal higher payment rates and e-bill subscriptions of treated and untreated accounts relative to neighbors in pure control blocks where nobody received the communication letter. For blocks with the highest saturation (80% treated accounts), total effects on payment likelihood emerge (numerically and statistically) on the same day that the letters started to be distributed, reaching a magnitude of about 4.5 percentage points (16% of the reference payment rate in pure control blocks). Spillover effects are more modest in magnitude and precisely estimated. In high-saturation blocks with 80% treated accounts, the effect on payment rates of untreated accounts is about 1 percentage point and statistically significant in the early days of the intervention, which is when social interference is more likely occur. Our intervention also increases subscriptions to the electronic billing option for treated and untreated accounts, with greater effects in high-saturation blocks, albeit small in absolute value.

## References

- 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.
- 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.
- 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 Yannu 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*.
- 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.

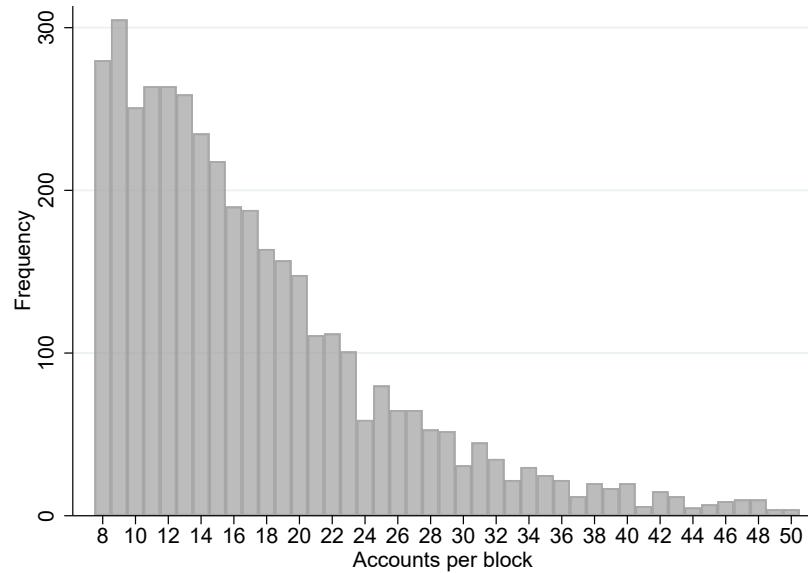
**Hansen, Bruce E., and Seojeong Lee.** 2019. “Asymptotic theory for clustered samples.” *Journal of Econometrics*, 210(2): 268–290.

**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.** forthcoming. “Causal Inference with Interference and Noncompliance in Two-Stage Randomized Experiments.” *Journal of the American Statistical Association*.
- 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.
- 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*.
- 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.

# 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: Timeline of the randomized communication campaign

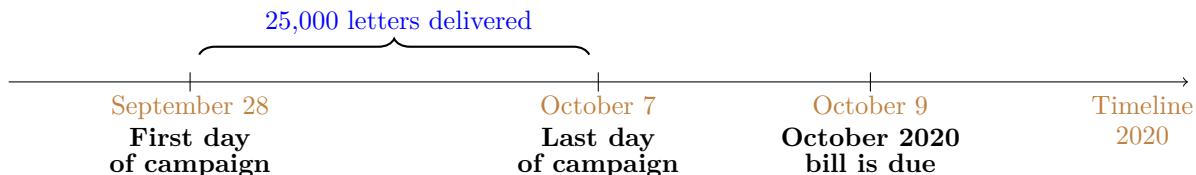
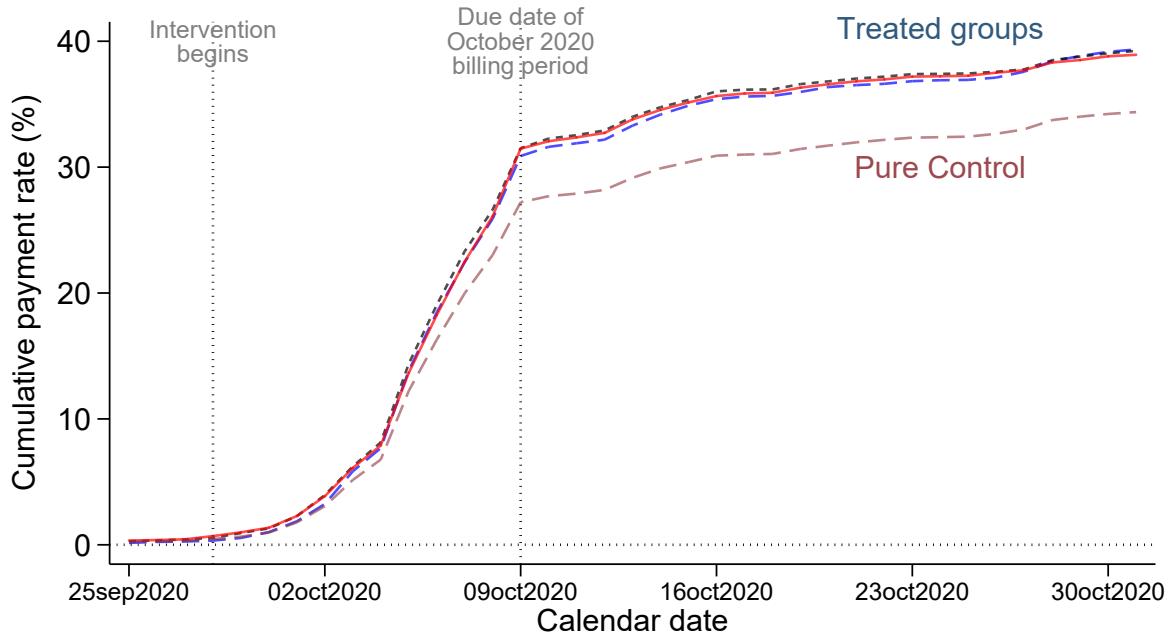
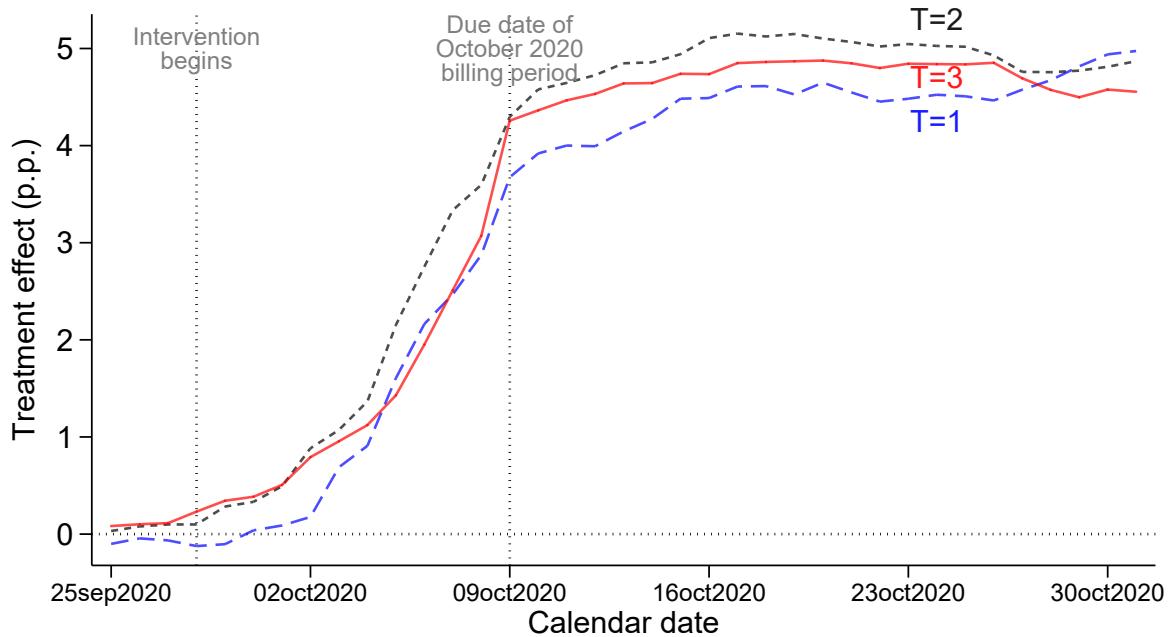


Figure 4: Payment rate of the October 2020 bill: Treated groups

(a) Payment rates in levels



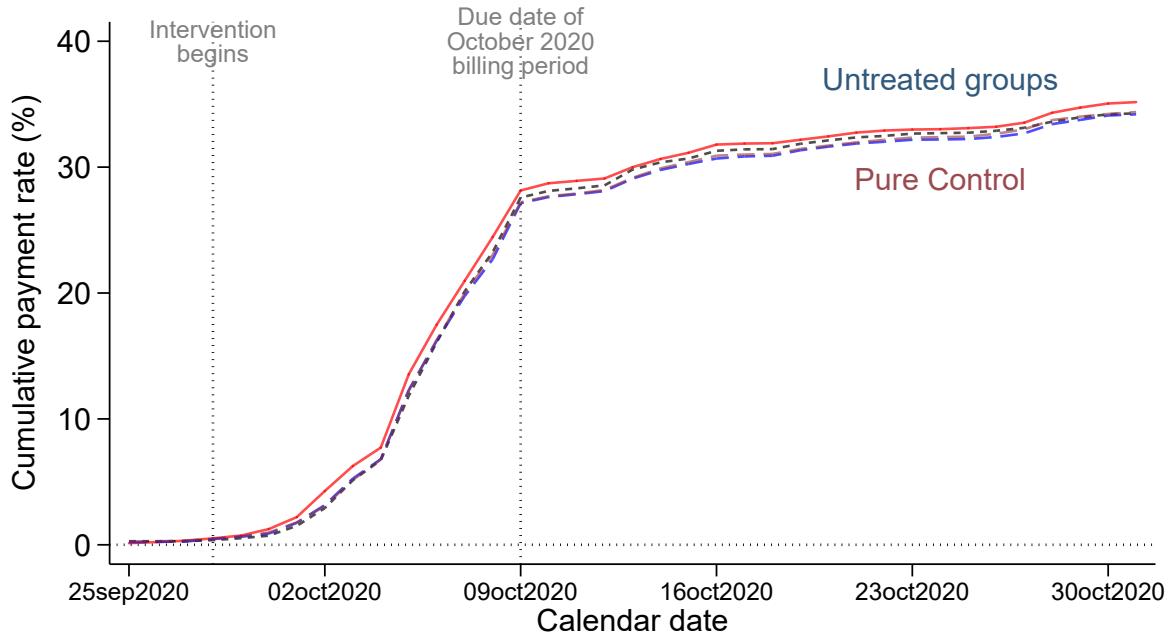
(b) Difference relative to pure control group



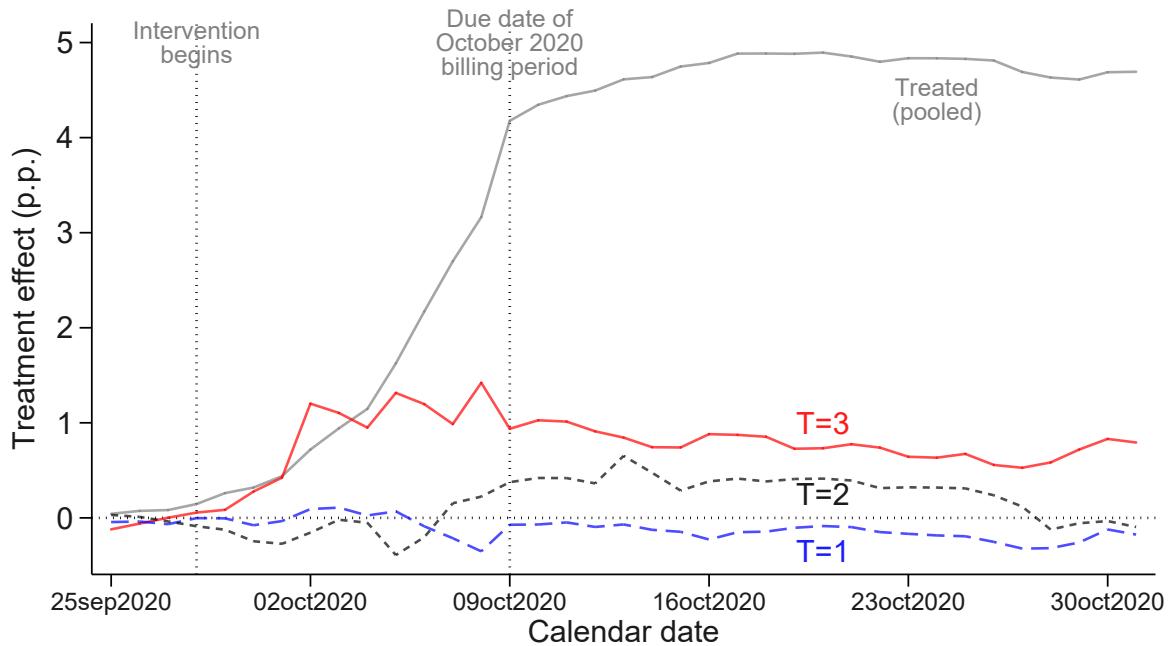
*Notes:* These figures show the effect of the intervention on payments of the October 2020 bill for treated groups. Panel (a) shows the cumulative share of individuals paying the October 2020 bill over time. The brown dashed line shows the payment rate for pure control units. The blue dashed line corresponds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

Figure 5: Payment rate of the October 2020 bill: Untreated groups

(a) Payment rates in levels

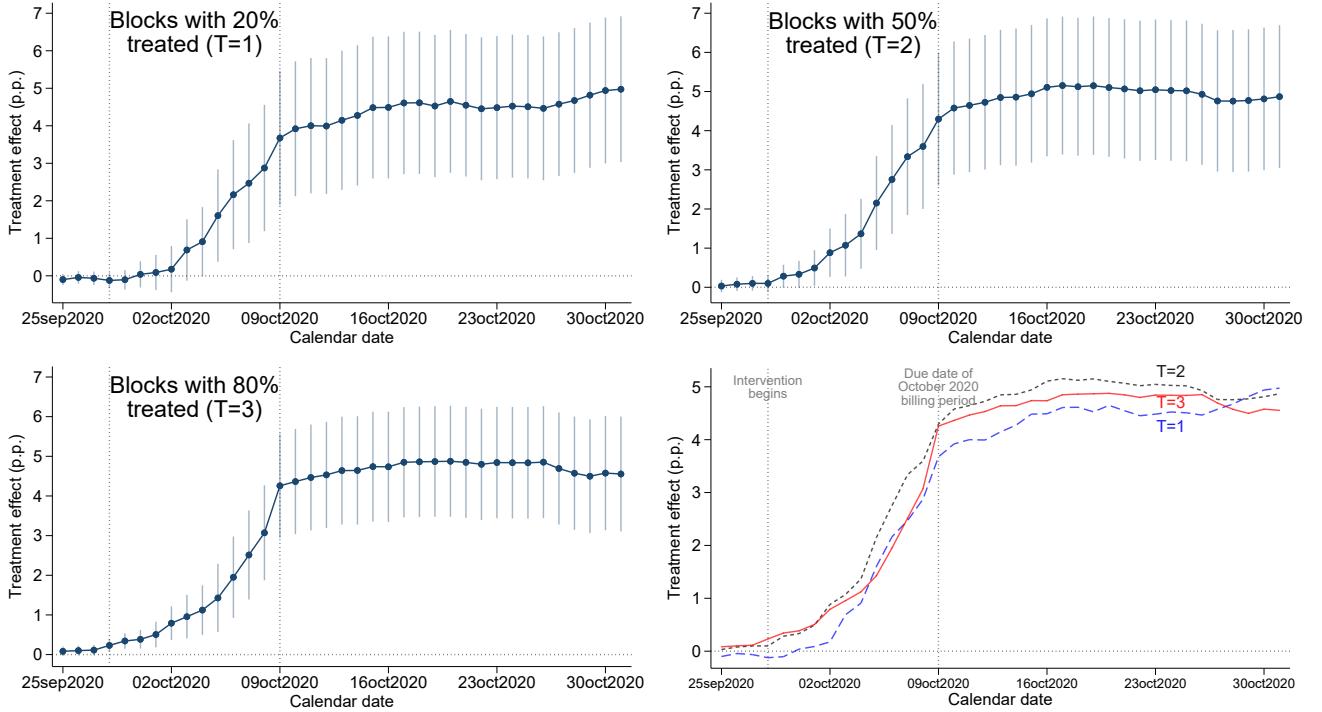


(b) Difference relative to pure control group



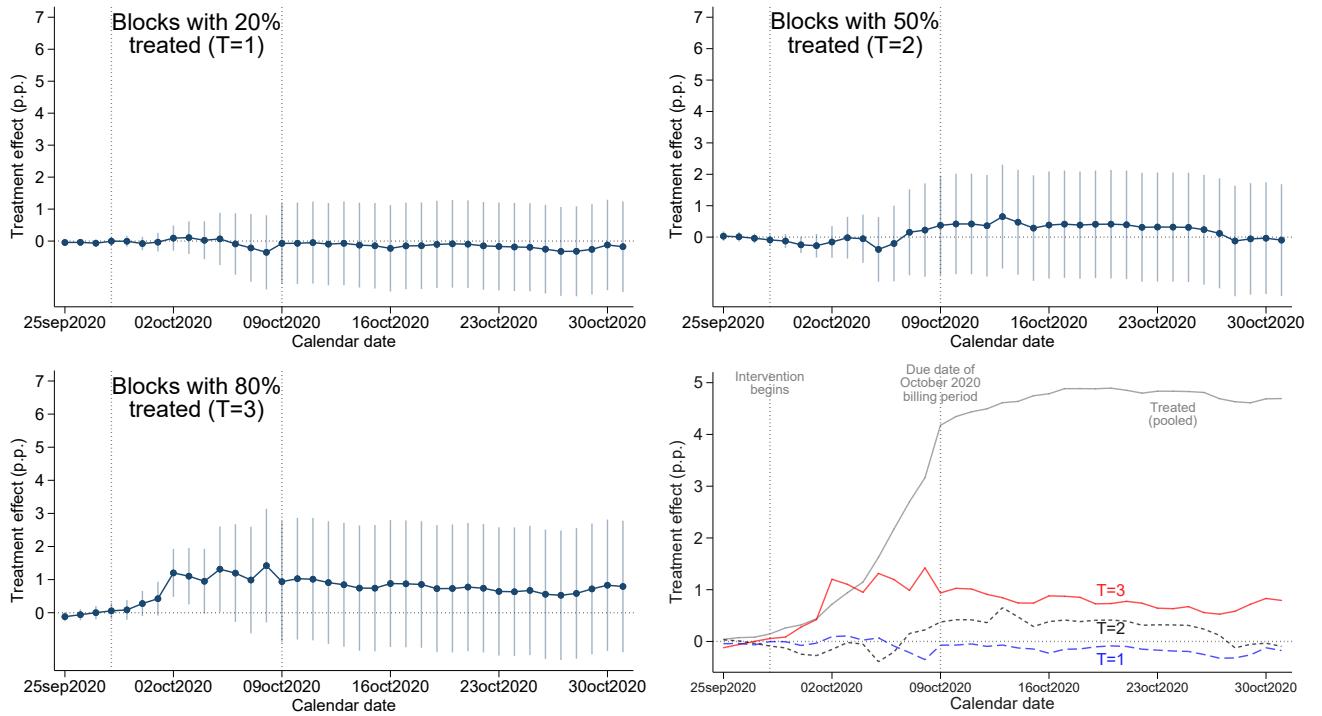
*Notes:* These figures show the effect of the intervention on payments of the October 2020 bill for untreated groups. Panel (a) shows the cumulative share of individuals paying the October 2020 bill over time. The brown dashed line shows the payment rate for pure control units. The blue dashed line corresponds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

Figure 6: Difference between treated and pure control groups (Oct'20 bill)



*Notes:* These figures show the coefficients and 95% confidence intervals from a saturated regression that computes, at each calendar date, the payment rate difference between each treated group and the pure control group. The top left panel corresponds to treated units in group  $T_g = 1$  (blocks with 20% treated). The top right panel corresponds to treated units in group  $T_g = 2$  (blocks with 50% treated). The bottom left panel corresponds to treated units in group  $T_g = 3$  (blocks with 80% treated). The bottom right panel superimposes the point estimates from the previous three panels (same as Figure 4 panel (b)). Standard errors are clustered by block. The first vertical bar denotes the start of the intervention. The due date for the October 2020 bill was October 9th and is indicated with another vertical bar. The letters were delivered between September 28th and October 7th.

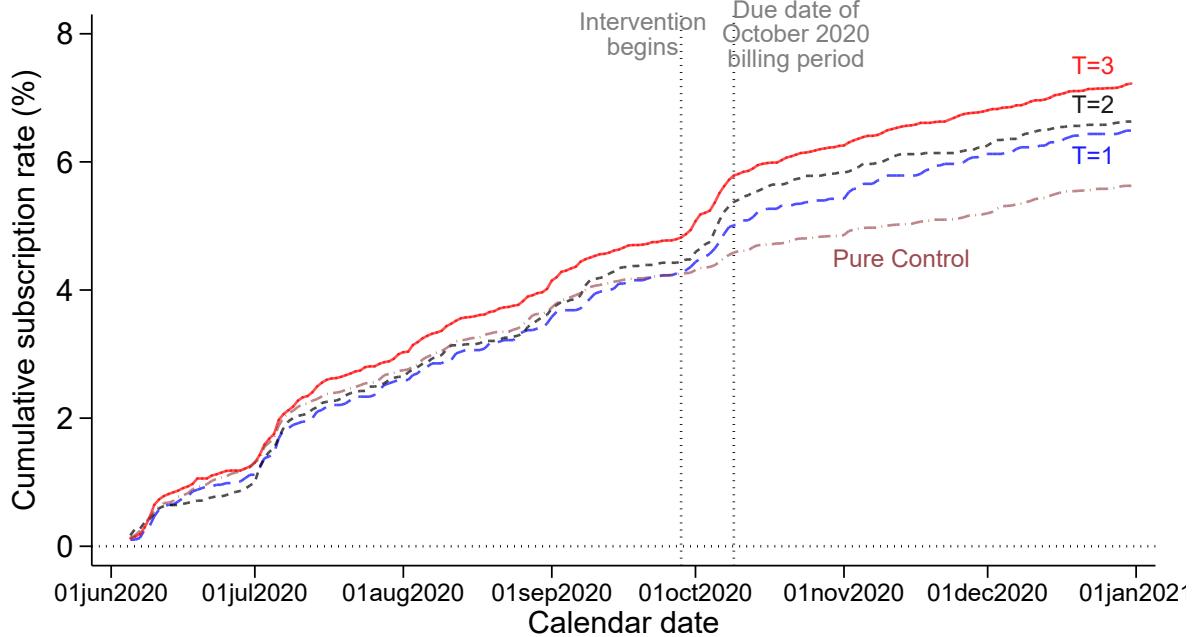
Figure 7: Difference between untreated and pure control groups (Oct'20 bill)



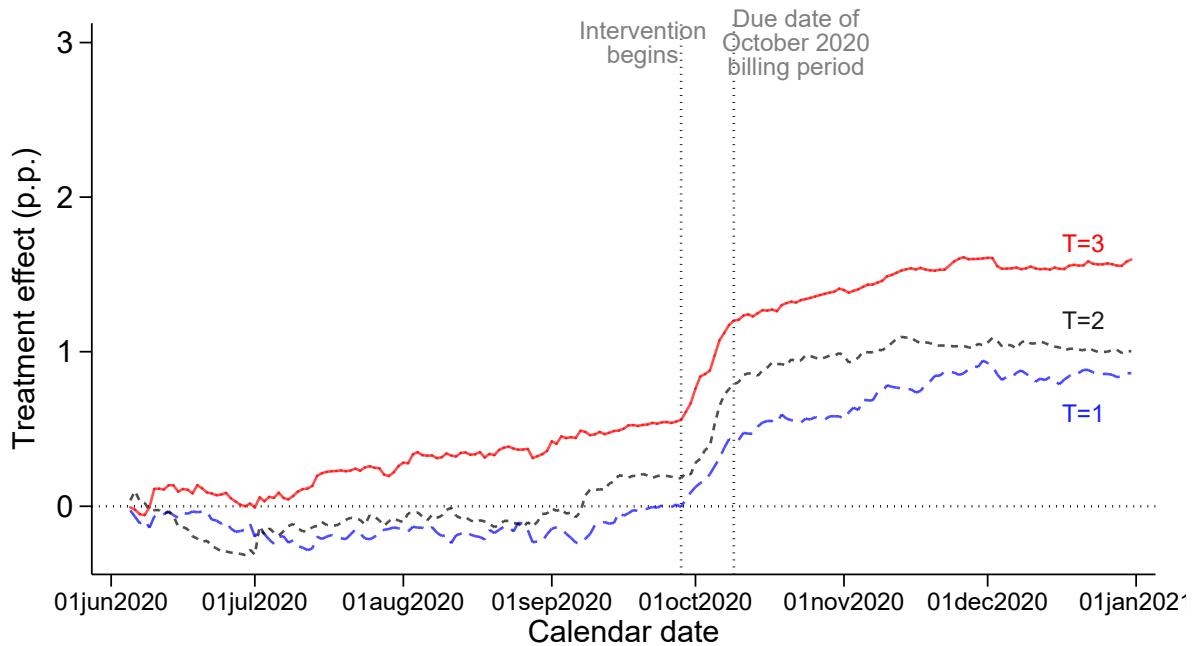
*Notes:* These figures show the coefficients and 95% confidence intervals from a saturated regression that computes, at each calendar date, the payment rate difference between each untreated group and the pure control group. The top left panel corresponds to untreated units in group  $T_g = 1$  (blocks with 20% treated). The top right panel corresponds to untreated units in group  $T_g = 2$  (blocks with 50% treated). The bottom left panel corresponds to untreated units in group  $T_g = 3$  (blocks with 80% treated). The bottom right panel superimposes the point estimates from the previous three panels (same as Figure 5 panel (b)). 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 8: Subscription to electronic billing: Treated groups

(a) Subscription rates in levels



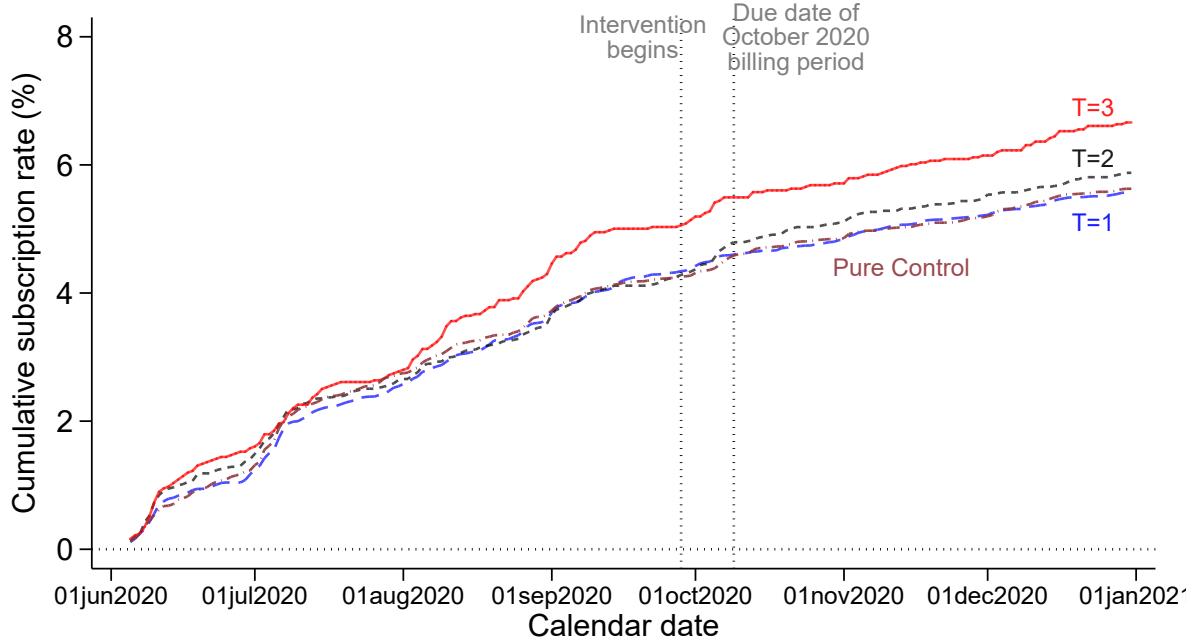
(b) Difference relative to pure control group



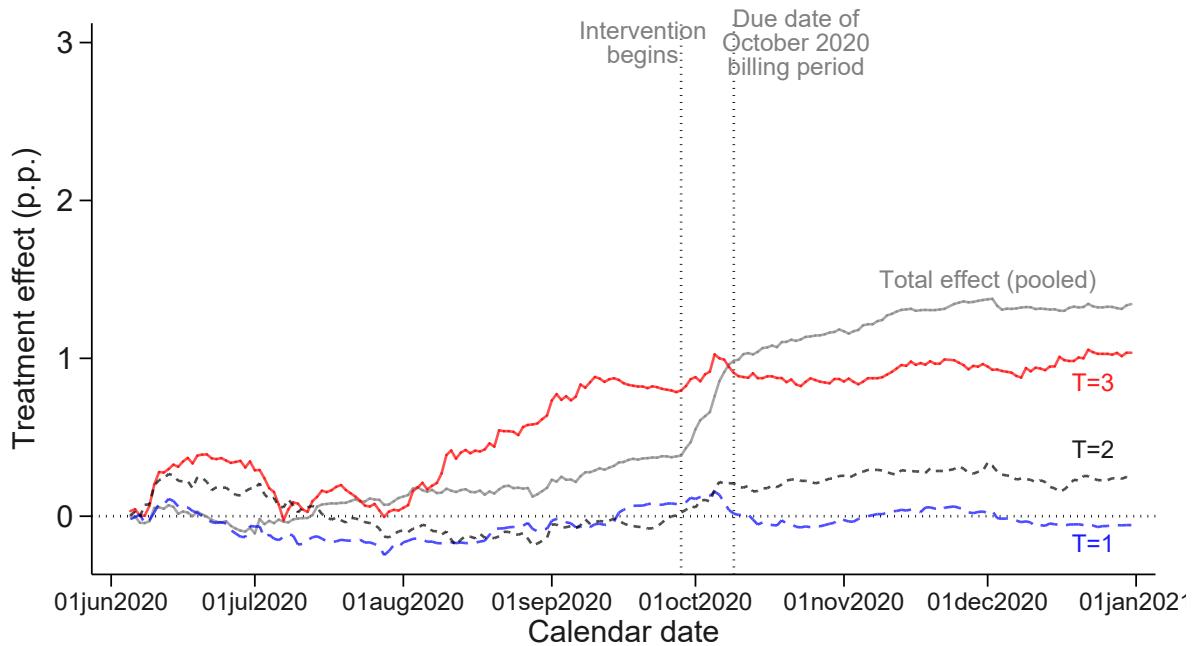
*Notes:* These figures show the effect of the intervention on subscriptions to electronic billing for treated groups. Panel (a) shows the cumulative share of e-billing subscribers over time. The brown dashed line shows the subscription rate for pure control units. The blue dashed line corresponds to 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. The e-billing option was launched by the municipality in June 2020. 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 9: Subscription to electronic billing: Untreated groups

(a) Subscription rates in levels

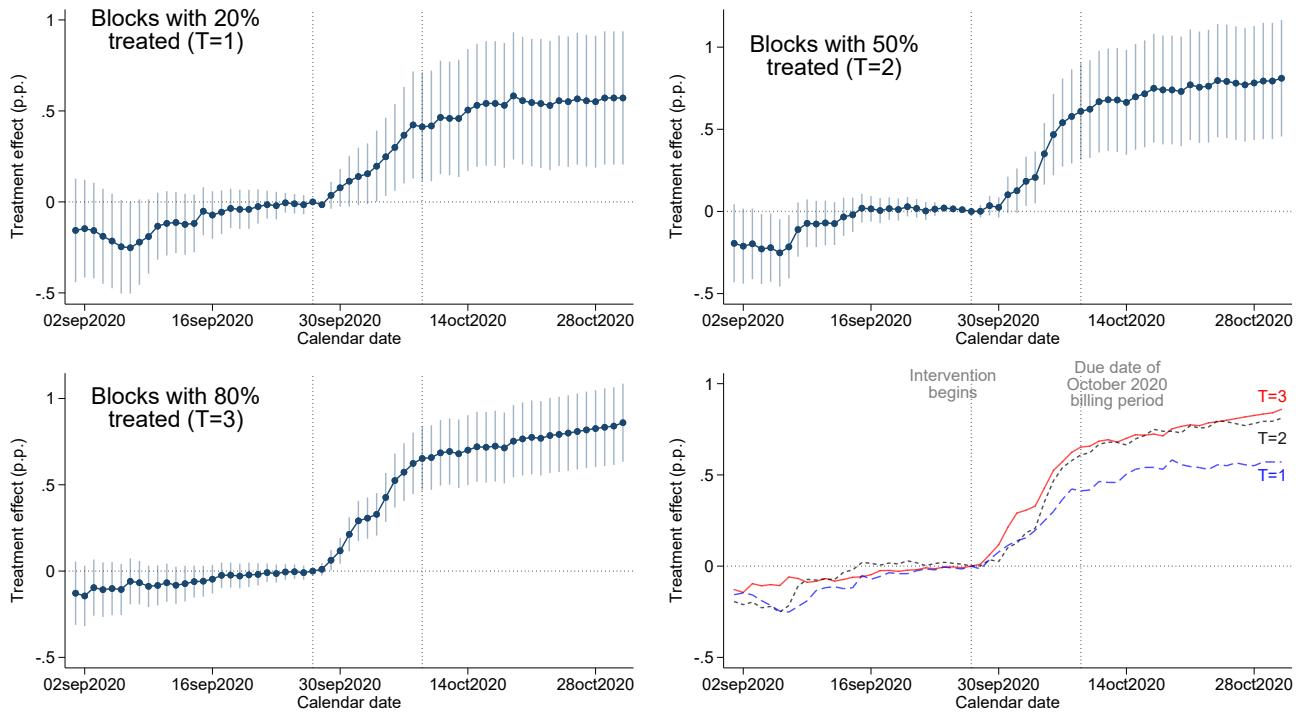


(b) Difference relative to pure control group



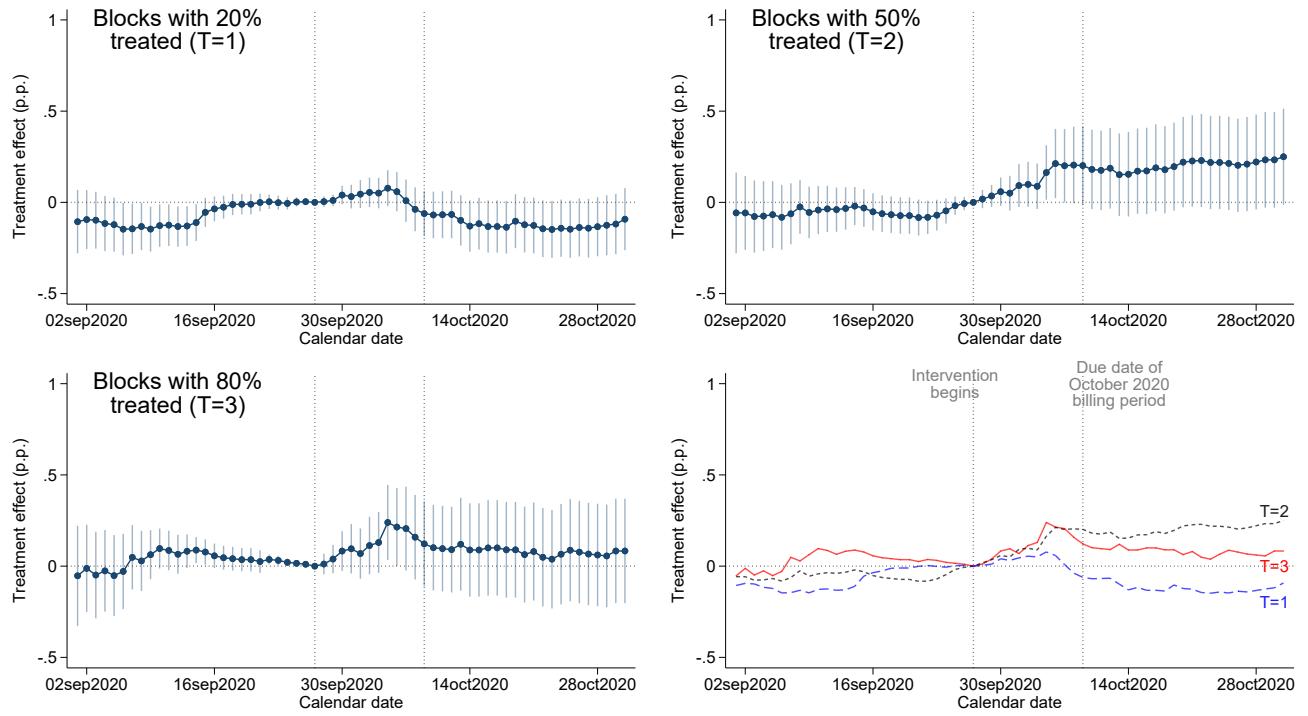
*Notes:* These figures show the effect of the intervention on subscriptions to electronic billing for untreated groups. Panel (a) shows the cumulative share of e-billing subscribers over time. The brown dashed line shows the subscription rate for pure control units. The blue dashed line corresponds to 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. For comparison, the gray solid line shows the effect for pooled treated units ( $T_g = 1, 2, 3$ ) relative to the pure control group. The e-billing option was launched by the municipality in June 2020. 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 10: Diff-in-diffs for treated groups (subscriptions to e-billing)



*Notes:* These figures show the coefficients and 95% confidence intervals from dynamic difference-in-differences regressions. All the coefficients are estimated relative to September 27th, 2020 (baseline date). The top left panel compares treated units in group  $T_g = 1$  (blocks with 20% treated) relative to pure control units. The top right panel compares treated units in group  $T_g = 2$  (blocks with 50% treated) relative to pure control units. The bottom left panel compares treated units in group  $T_g = 3$  (blocks with 80% treated) relative to pure control units. The bottom right panel superimposes the point estimates from the previous three panels. Standard errors are clustered by block. The first vertical bar denotes the start of the intervention. The due date was October 9th and is indicated with another vertical bar. The letters were delivered between September 28th and October 7th.

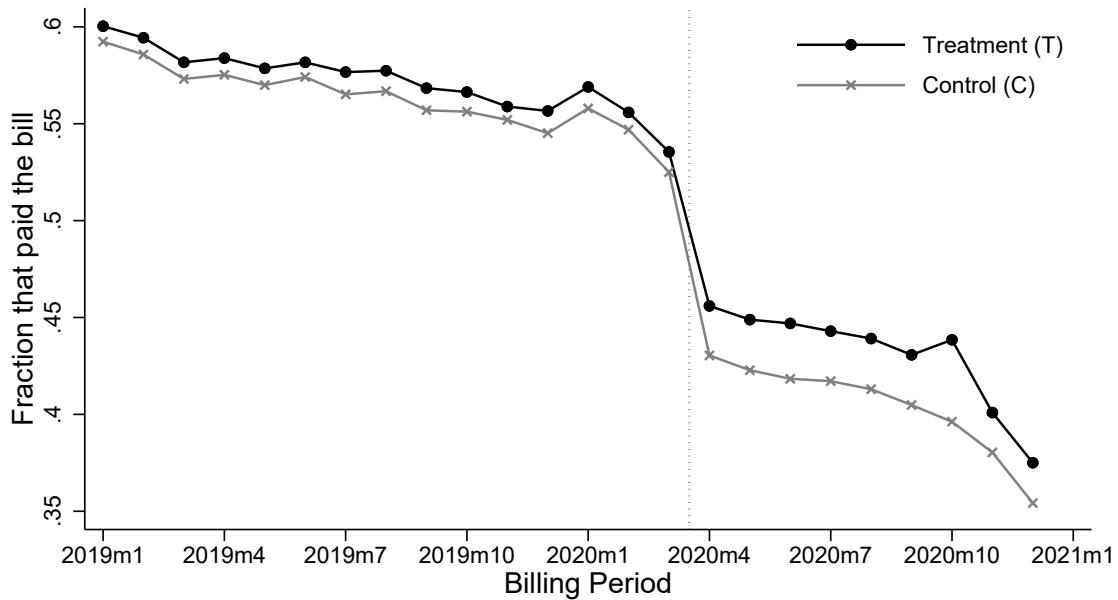
Figure 11: Diff-in-diffs for untreated groups (subscriptions to e-billing)



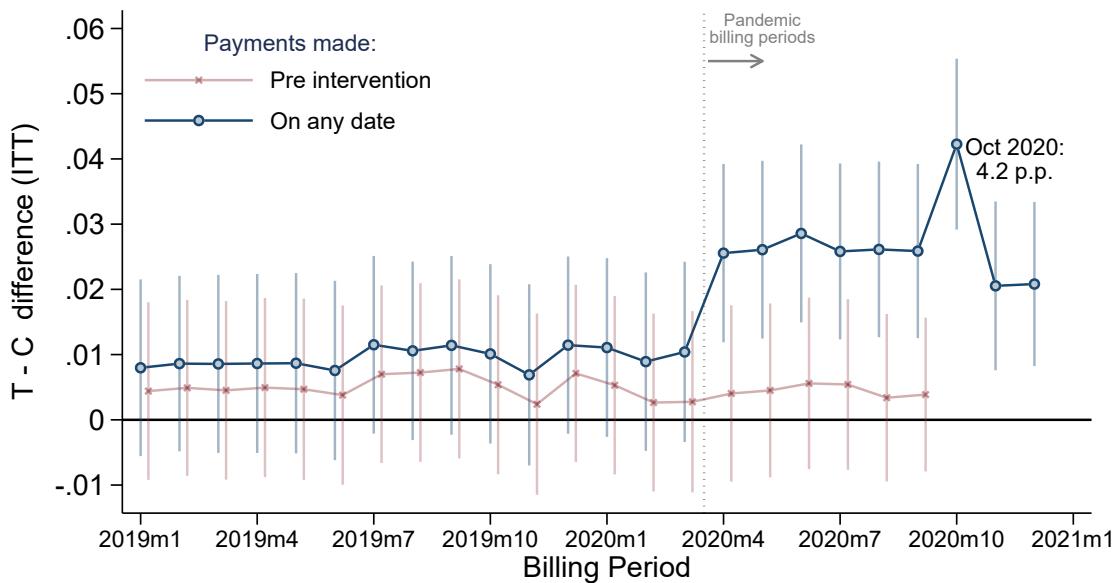
*Notes:* These figures show the coefficients and 95% confidence intervals from dynamic difference-in-differences regressions. All the coefficients are estimated relative to September 27th, 2020 (baseline date). The top left panel compares untreated units in group  $T_g = 1$  (blocks with 20% treated) relative to pure control units. The top right panel compares untreated units in group  $T_g = 2$  (blocks with 50% treated) relative to pure control units. The bottom left panel compares untreated units in group  $T_g = 3$  (blocks with 80% treated) relative to pure control units. The bottom right panel superimposes the point estimates from the previous three panels. Standard errors are clustered by block. The first vertical bar denotes the start of the intervention. The due date was October 9th and is indicated with another vertical bar. The letters were delivered between September 28th and October 7th.

Figure 12: Total effects on pre- and post-intervention bills

(a) Payment rates in levels



(b) Difference relative to pure control group



*Notes:* These figures show the effect of the communication campaign on payment rates of pre- and post-intervention bills. Panel (a) shows 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  $T_g = 1, 2, 3$ . Panel (b) reports total treatment effects—i.e., the difference between treated and pure control units—and 95% confidence intervals for the 24 billing periods (blue line). These payment rates include any payment made until December 2020 and, therefore, capture both timely and overdue payments. As a placebo exercise, the red line estimates the difference in payment rates between treated and control accounts but only considering payments made up to September 27, 2020, a day before letters started to be delivered. 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 on a separate regression. Standard errors are clustered at the block level.

Table 2: Descriptive statistics

|           | Blocks | Obs   | Mean  | SD    | ICC   |
|-----------|--------|-------|-------|-------|-------|
| paid_all  | 3982   | 68808 | 0.449 | 0.497 | 0.062 |
| paid_some | 3982   | 68808 | 0.650 | 0.477 | 0.071 |
| paid_six  | 3982   | 68808 | 0.572 | 0.495 | 0.073 |

Table 3: Total effects and spillover effects for municipal bill payments

| <b>Dependent variable:</b><br>Pr(pay the bill)            | <b>Placebo bill:</b> |                     | <b>Intervention bill:</b> |                     |
|---|----------------------|---------------------|---------------------------|---------------------|
|   | Sep'20 bill<br>(1)   | Early<br>(2)        | By due date<br>(3)        | By 31 Oct<br>(4)    |
| <b>A. Blocks with 20% treated (<math>T_g = 1</math>):</b> |                      |                     |                           |                     |
| Treated   | 0.009<br>(0.009)     | 0.007*<br>(0.004)   | 0.037***<br>(0.009)       | 0.050***<br>(0.010) |
| Untreated   | 0.001<br>(0.007)     | 0.001<br>(0.003)    | -0.001<br>(0.006)         | -0.002<br>(0.007)   |
| <b>B. Blocks with 50% treated (<math>T_g = 2</math>):</b> |                      |                     |                           |                     |
| Treated   | 0.008<br>(0.009)     | 0.011***<br>(0.004) | 0.043***<br>(0.009)       | 0.049***<br>(0.009) |
| Untreated   | 0.003<br>(0.009)     | -0.000<br>(0.003)   | 0.004<br>(0.008)          | -0.001<br>(0.009)   |
| <b>C. Blocks with 80% treated (<math>T_g = 3</math>):</b> |                      |                     |                           |                     |
| Treated   | 0.001<br>(0.007)     | 0.010***<br>(0.003) | 0.043***<br>(0.007)       | 0.046***<br>(0.007) |
| Untreated   | -0.003<br>(0.009)    | 0.011**<br>(0.004)  | 0.009<br>(0.009)          | 0.008<br>(0.010)    |
| Mean of Pure Control at baseline ( $T_g = 0$ )            | 0.297                | 0.052               | 0.272                     | 0.344               |
| Observations  | 68,806               | 68,806              | 68,806                    | 68,806              |
| Number of clusters  | 3,981                | 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 dependent variables in each column are: (1) an indicator for paying the September 2020 bill by 15 September (pre intervention); (2) an indicator for paying the October 2020 bill by 3 October (early payments); (3) an indicator for paying the October 2020 bill by 9 October (on time payments); (4) an indicator for paying the October 2020 bill by 31 October (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 Figures (6) and (7). 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 each regression, corresponding to the average payment rate for units in blocks with no treated units ( $T_g = 0$ ). Standard errors clustered by blocks are reported in parentheses. \* p<0.10, \*\* p<0.05, \*\*\* p<0.01

Table 4: Total effects and spillover effects for subscriptions to e-billing

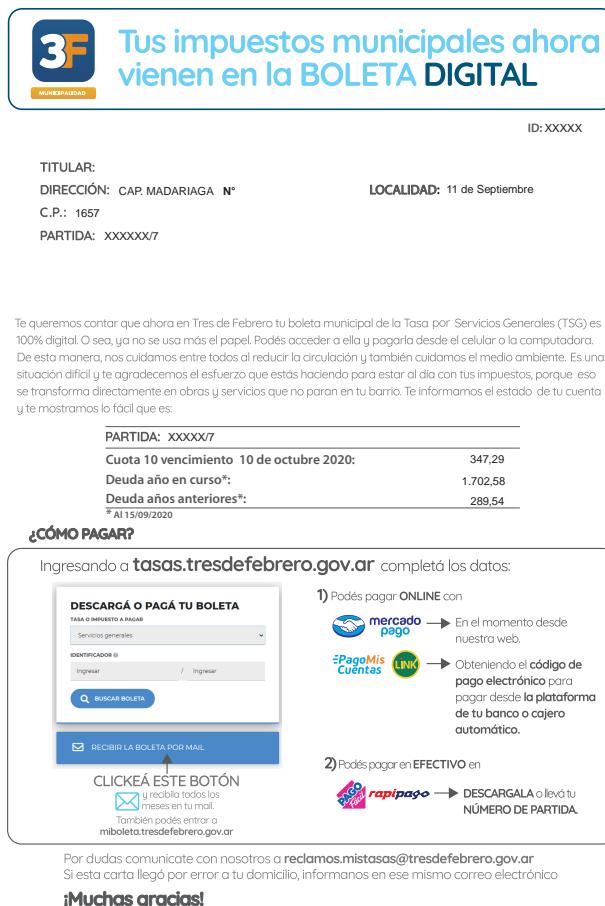
| Dependent variable:<br>Pr(subscribe to e-bill)            | Placebo:            |                       | Intervention:         |                       |
|---|---------------------|-----------------------|-----------------------|-----------------------|
|   | By 17 Sep           | Early                 | By due date           | By 31 Oct             |
|   | (1)                 | (2)                   | (3)                   | (4)                   |
| <b>A. Blocks with 20% treated (<math>T_g = 1</math>):</b> |                     |                       |                       |                       |
| Treated   | -0.0006<br>(0.0006) | 0.0015*<br>(0.0008)   | 0.0041***<br>(0.0015) | 0.0057***<br>(0.0019) |
| Untreated   | -0.0003<br>(0.0003) | 0.0005<br>(0.0004)    | -0.0006<br>(0.0006)   | -0.0009<br>(0.0009)   |
| <b>B. Blocks with 50% treated (<math>T_g = 2</math>):</b> |                     |                       |                       |                       |
| Treated   | 0.0001<br>(0.0004)  | 0.0018**<br>(0.0008)  | 0.0061***<br>(0.0015) | 0.0081***<br>(0.0018) |
| Untreated   | -0.0006<br>(0.0005) | 0.0010<br>(0.0006)    | 0.0020*<br>(0.0011)   | 0.0025*<br>(0.0013)   |
| <b>C. Blocks with 80% treated (<math>T_g = 3</math>):</b> |                     |                       |                       |                       |
| Treated   | -0.0002<br>(0.0004) | 0.0031***<br>(0.0006) | 0.0065***<br>(0.0009) | 0.0086***<br>(0.0012) |
| Untreated   | 0.0005<br>(0.0003)  | 0.0011<br>(0.0008)    | 0.0012<br>(0.0012)    | 0.0008<br>(0.0015)    |
| Mean of Pure Control at baseline ( $T_g = 0$ )            | 0.0425              | 0.0425                | 0.0425                | 0.0425                |
| Observations  | 137,612             | 137,612               | 137,612               | 137,612               |
| Number of clusters  | 3,981               | 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 day relative to September 27th, 2020 (baseline date). The estimates correspond exactly to the numbers shown in Figures (10) and (11). Column (1) shows the results for 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 by the due date of the October 2020 bill; Column (4) shows the results for subscriptions 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 ( $T_g = 0$ ) on September 27, 2020. Standard errors clustered by blocks are reported in parentheses. \* p<0.10, \*\* p<0.05, \*\*\* p<0.01

# Supplementary Materials for: “A Framework for Partial Population Experiments”

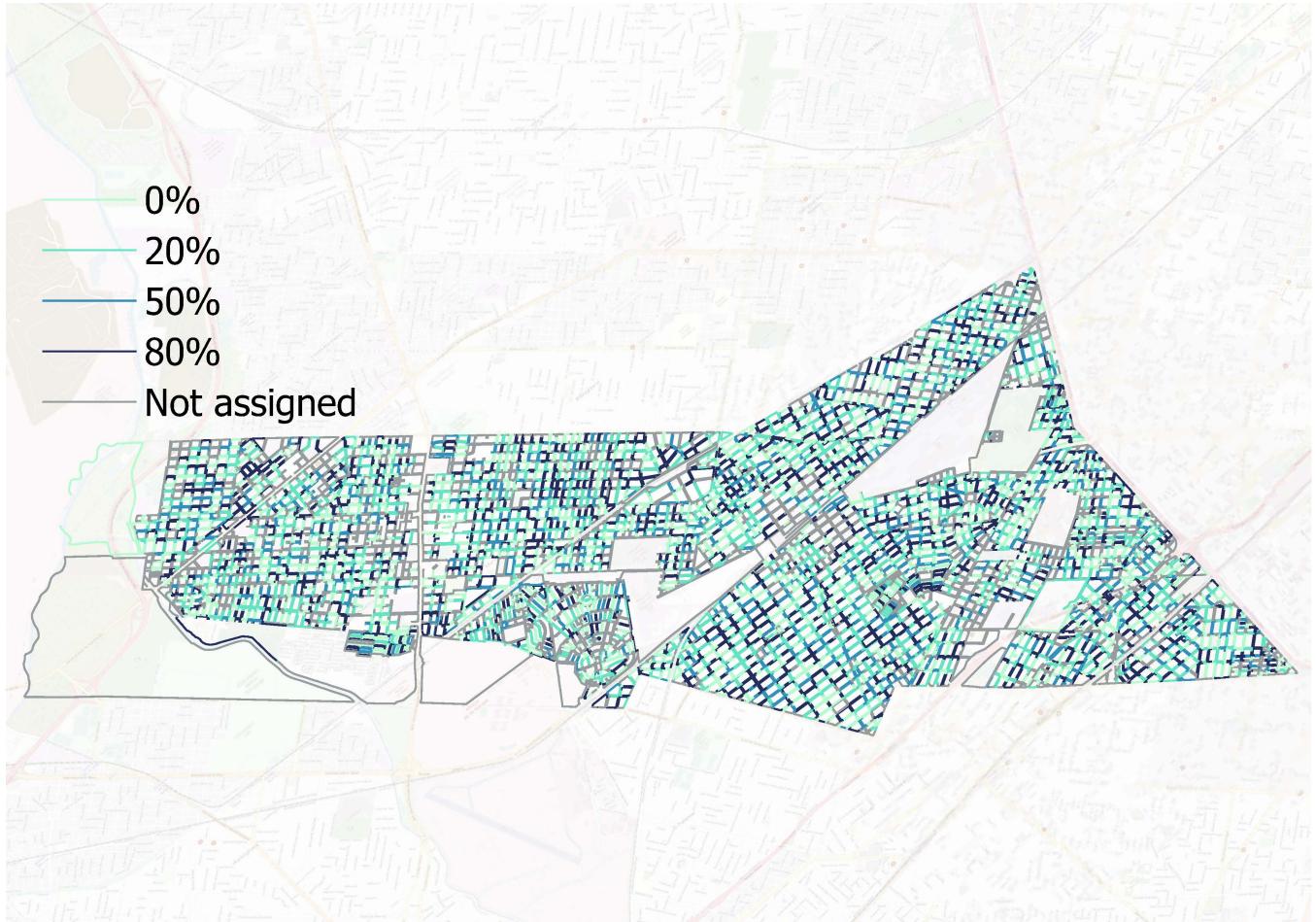
## A Additional figures and tables

Figure A.1: Example of the intervention letter



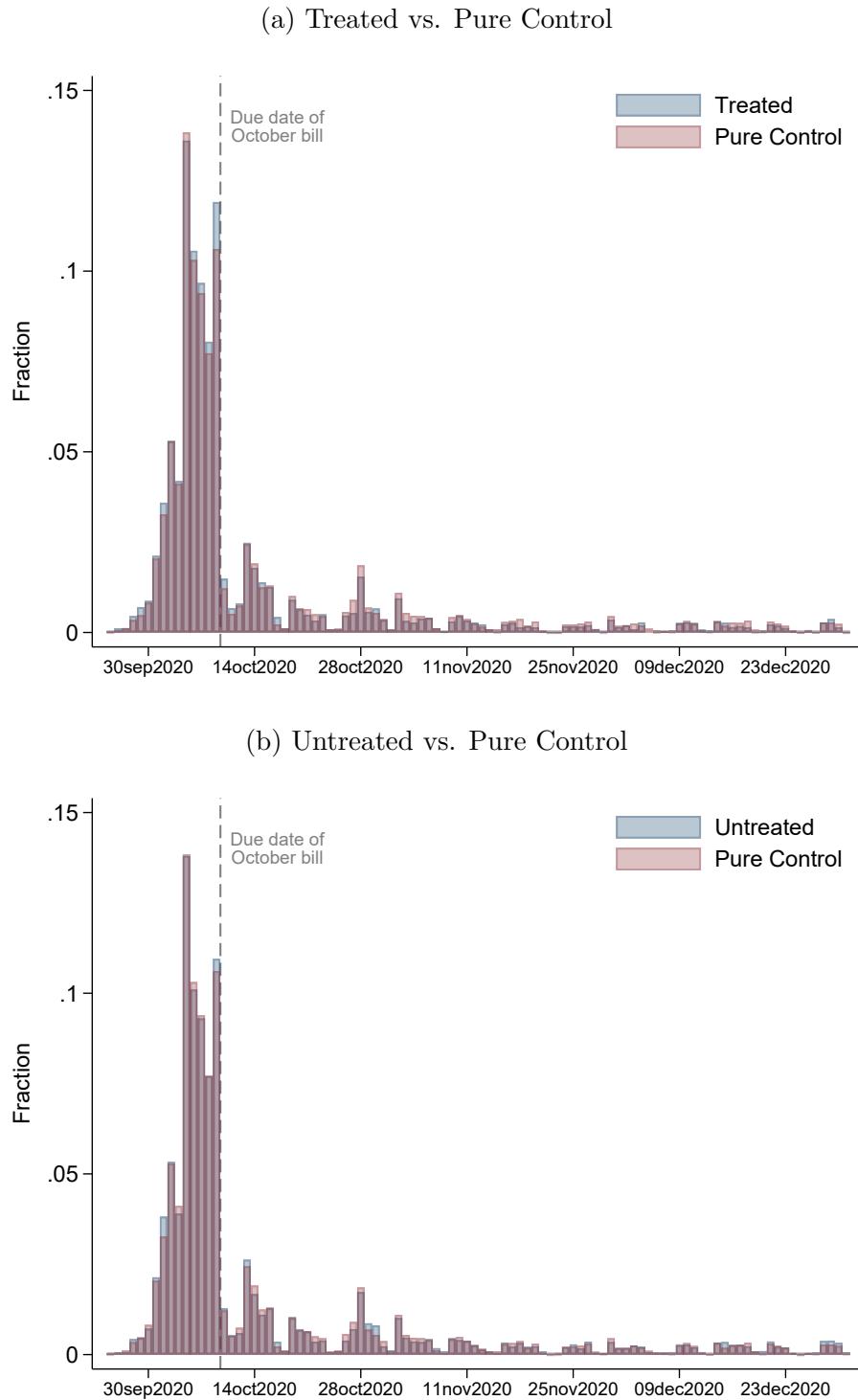
*Notes:* This figure shows an anonymized example of the letters sent during the intervention between September 28th and October 7th, 2020. The headline reads: “*Your municipal taxes are now available on the electronic bill.*” The information below the headline contains the name of the account holder,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



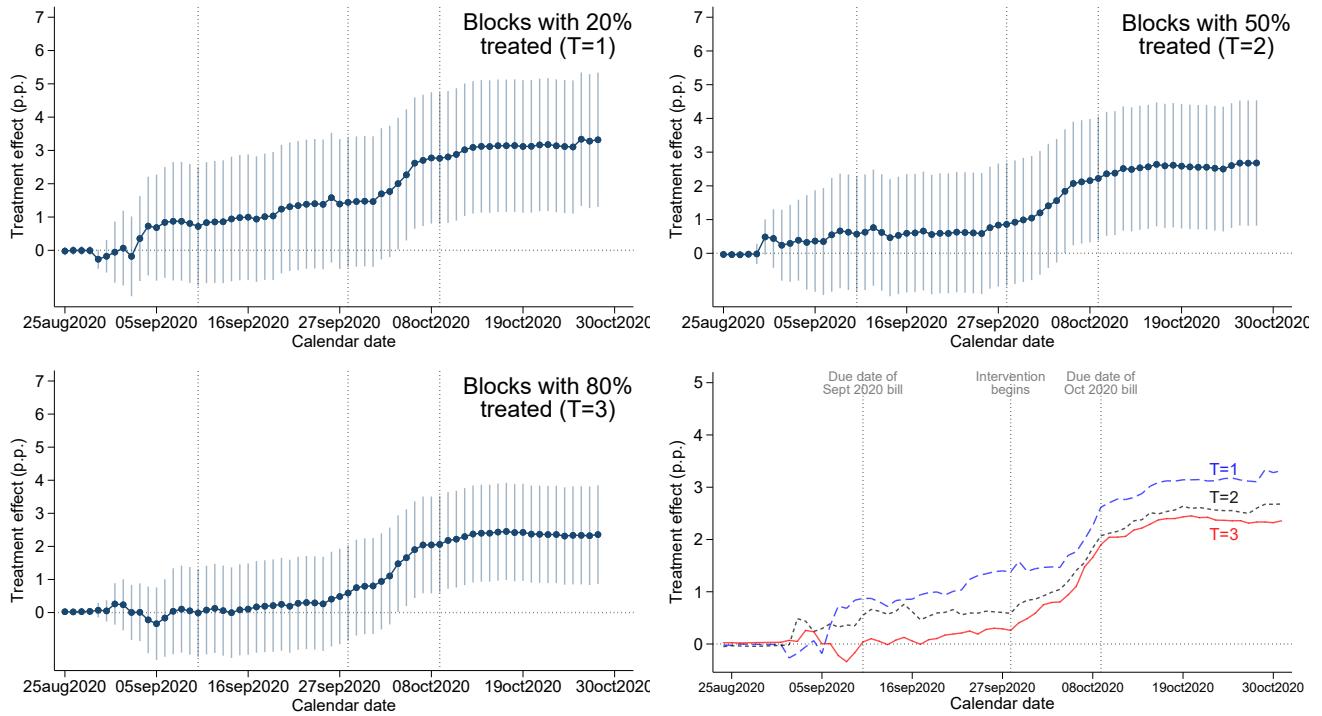
*Notes:* This figure shows a map of the municipality where the 2-level randomized communication campaign took place. We highlight the group-level assignment of blocks (*cuadras*) with different colors: pure control blocks with 0% treated (light green), blocks with 20% treated accounts (green), blocks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



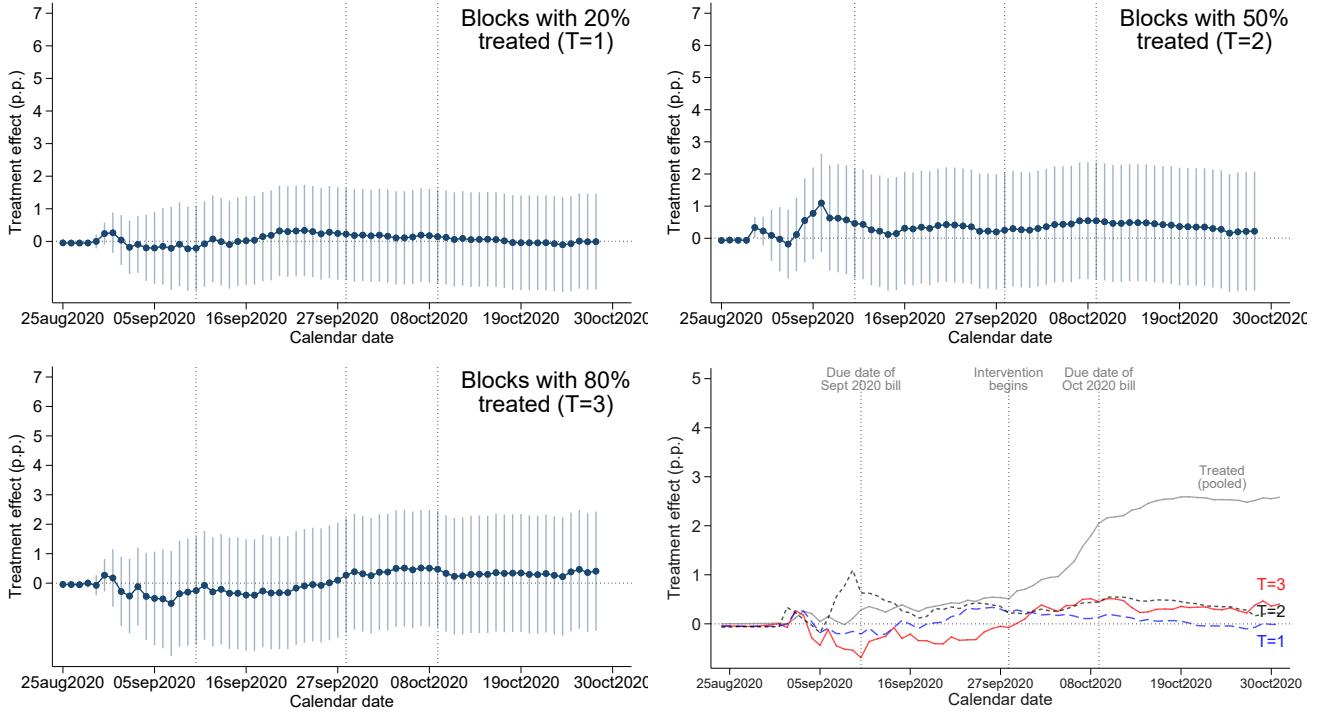
*Notes:* These figures show the fraction of individuals paying the October 2020 bill before and after the due date (October 9th, 2020). Panel (a) shows the distribution of payments for treated units (in blue) relative to pure control units (in red). We pool together treated units from  $T_g = 1, 2, 3$ . 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: Difference between treated and pure control groups (Sep'20 bill)



*Notes:* These figures show the coefficients and 95% confidence intervals from a saturated regression that computes, at each calendar date, the payment rate difference between each treated group and the pure control group. The top left panel corresponds to treated units in group  $T_g = 1$  (blocks with 20% treated). The top right panel corresponds to treated units in group  $T_g = 2$  (blocks with 50% treated). The bottom left panel corresponds to treated units in group  $T_g = 3$  (blocks with 80% treated). The bottom right panel superimposes the point estimates from the previous three panels. Standard errors are clustered by block. The first vertical bar denotes the due date of the September 2020 bill. The second vertical bar denotes the start of the intervention. The due date for the October 2020 bill was October 9th and is indicated with a third vertical bar. The letters were delivered between September 28th and October 7th.

Figure A.5: Difference between untreated and pure control groups (Sep'20 bill)



*Notes:* These figures show the coefficients and 95% confidence intervals from a saturated regression that computes, at each calendar date, the payment rate difference between each untreated group and the pure control group. The top left panel corresponds to untreated units in group  $T_g = 1$  (blocks with 20% treated). The top right panel corresponds to untreated units in group  $T_g = 2$  (blocks with 50% treated). The bottom left panel corresponds to untreated units in group  $T_g = 3$  (blocks with 80% treated). The bottom right panel superimposes the point estimates from the previous three panels. Standard errors are clustered by block. The first vertical bar denotes the due date of the September 2020 bill. The second vertical bar denotes the start of the intervention. The due date for the October 2020 bill was October 9th and is indicated with a third vertical bar. The letters were delivered between September 28th and October 7th.

## B 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  $\varepsilon_{ig}$  to be correlated within blocks and use a cluster-robust variance estimator. In this regression,  $\theta_t$  captures the average difference of  $X_{ig}$  of untreated units in groups with  $T_g = t$  relative

to the pure control group and  $\tau_t$  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.

Table A1: Balance test saturated regressions

|   | Property<br>Value | Front<br>Metres     | House<br>type     | Tenant<br>Male    | Tenant<br>Age     | Bill<br>amount    | N Bills<br>paid 2019 | Digital<br>payment |
|---|-------------------|---------------------|-------------------|-------------------|-------------------|-------------------|----------------------|--------------------|
|   | (1)               | (2)                 | (3)               | (4)               | (5)               | (6)               | (7)                  | (8)                |
| <b>A. Blocks with 20% treated (<math>T_g = 1</math>):</b> |                   |                     |                   |                   |                   |                   |                      |                    |
| Treated   | 0.022<br>(0.019)  | 32.569*<br>(16.790) | -0.009<br>(0.006) | 0.008<br>(0.011)  | 0.100<br>(0.541)  | 5.940<br>(9.550)  | 0.073<br>(0.119)     | -0.011<br>(0.012)  |
| Untreated   | 0.024<br>(0.017)  | 19.137<br>(14.046)  | -0.007<br>(0.005) | -0.006<br>(0.007) | 0.124<br>(0.403)  | 1.320<br>(7.765)  | 0.005<br>(0.092)     | 0.001<br>(0.008)   |
| <b>B. Blocks with 50% treated (<math>T_g = 2</math>):</b> |                   |                     |                   |                   |                   |                   |                      |                    |
| Treated   | 0.013<br>(0.019)  | 12.645<br>(20.384)  | -0.005<br>(0.006) | -0.004<br>(0.010) | -0.467<br>(0.498) | 1.164<br>(9.213)  | 0.032<br>(0.114)     | 0.001<br>(0.011)   |
| Untreated   | 0.011<br>(0.020)  | 25.302<br>(20.664)  | -0.003<br>(0.006) | -0.000<br>(0.010) | -0.419<br>(0.484) | 1.880<br>(9.665)  | 0.015<br>(0.113)     | 0.012<br>(0.011)   |
| <b>C. Blocks with 80% treated (<math>T_g = 3</math>):</b> |                   |                     |                   |                   |                   |                   |                      |                    |
| Treated   | 0.008<br>(0.017)  | -8.269<br>(17.775)  | -0.001<br>(0.004) | -0.005<br>(0.008) | -0.140<br>(0.405) | 2.813<br>(7.813)  | 0.052<br>(0.094)     | -0.002<br>(0.008)  |
| Untreated   | 0.001<br>(0.019)  | -1.760<br>(20.704)  | 0.001<br>(0.006)  | 0.001<br>(0.011)  | -0.530<br>(0.530) | 6.272<br>(12.949) | -0.062<br>(0.122)    | -0.002<br>(0.012)  |
| Mean Pure Control   | 13.643            | 841.496             | 0.913             | 0.616             | 19.145            | 368.659           | 6.714                | 0.352              |
| Observations  | 64,932            | 68,808              | 68,808            | 46,419            | 52,714            | 68,808            | 68,808               | 38,112             |
| N 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 2015 (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 ( $T_g = 0$ ). Standard errors clustered by blocks are reported in parentheses. \* p<0.10, \*\* p<0.05, \*\*\* p<0.01

## C Numerical Illustration

Table A2 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 \(forthcoming\)](#).

Table A2: Sample sizes in existing literature

|  | Sample size | No. of groups | Ave. group size | Sd. group size |
|--|-------------|---------------|-----------------|----------------|
| <a href="#">Giné and Mansuri (2018)</a>              | 2,736       | 67            | 39.4            | 16.7           |
| <a href="#">Haushofer and Shapiro (2016)</a>         | 1,440       | 123           | 23.4            | 14.8           |
| <a href="#">Ichino and Schündeln (2012)</a>          | 868         | 39            | 22.3            | 9.6            |
| <a href="#">Imai, Jiang and Malani (forthcoming)</a> | 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 A3. When the intraclass correlation is low ( $\rho = 0.1$ ), accounting for group size heterogeneity increases standard errors and MDEs between 6.8% and 14.5%. The problem worsens for larger intraclass correlations. When  $\rho = 0.5$ , adjusted standard errors and MDEs are between 8.3% and 19.6% larger, and between 8.5% and 20.2% larger when  $\rho = 0.8$ .

Figure 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 A2 for the group size distribution. The ratio of standard errors increases rapidly for values of  $\rho$ , 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.

Table A3: 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 |

## D 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. Table A4 shows the expected (approximate) sample sizes.

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

Table A4: Approximate sample sizes

|           | Blocks | Control Obs | Treated Obs |
|-----------|--------|-------------|-------------|
| $T_g = 0$ | 1088   | 18808       | 0           |
| $T_g = 1$ | 1124   | 15547       | 3886        |
| $T_g = 2$ | 644    | 5565        | 5565        |
| $T_g = 3$ | 1124   | 3886        | 15547       |
| Total     | 3980   | 43806       | 24998       |

For power calculations, Figure D.6 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.
- $\text{ICC} = 0.1$  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.

## E Simulations

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:

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

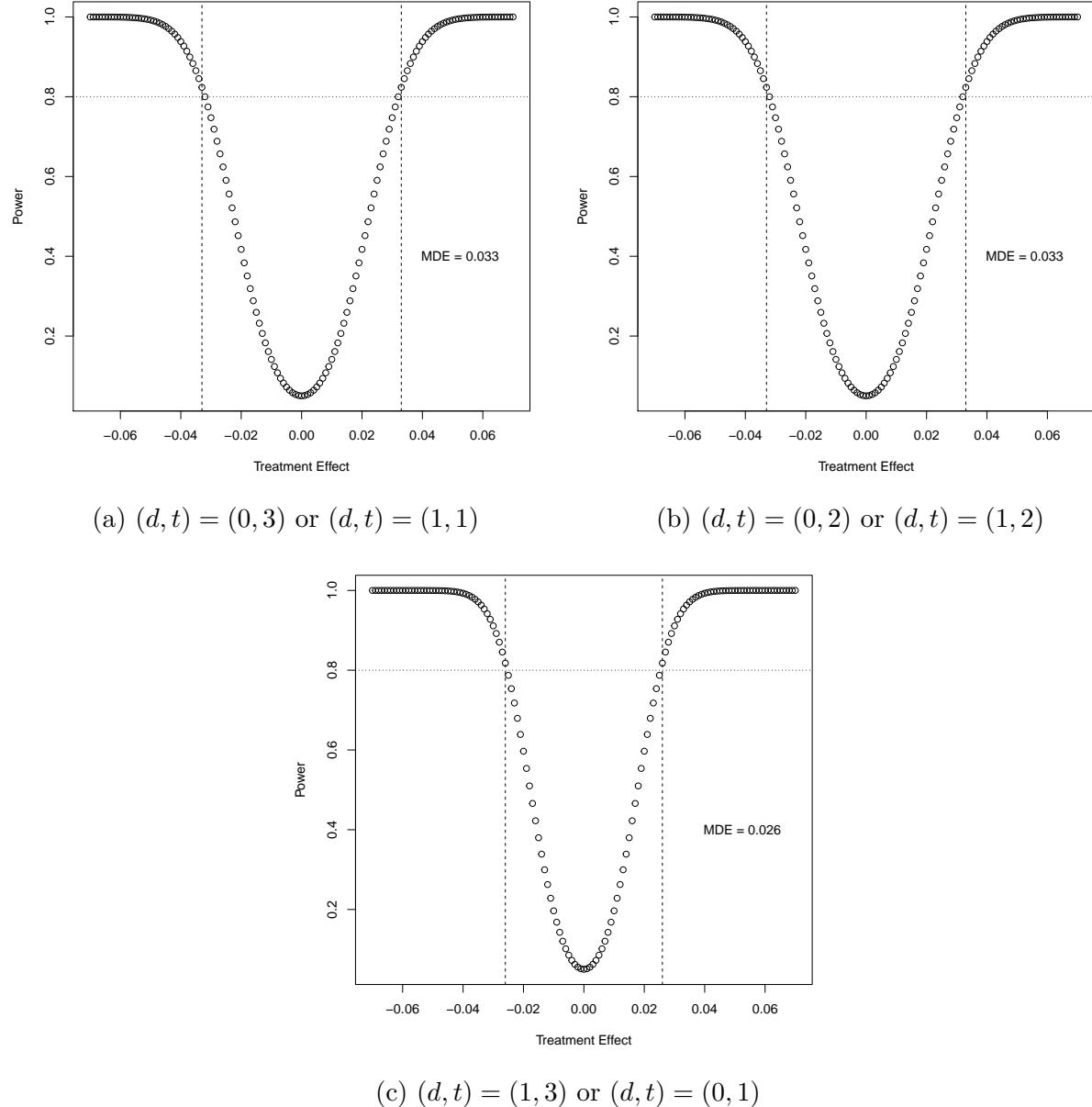
with  $U_g \sim \text{Uniform}(0, 1)$ . The individual treatment indicator is assigned according to the rule:

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

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

Figure D.6: Power functions



baseline June 2019 outcome  $Y_{ig}^{base}$ , the potential outcomes are constructed in the following way:

$$Y_{ig}(0, 0) = Y_{ig}^{base}$$

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

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

$$\mathbb{E}[Y_{ig}(0, 0)] = \mu_0$$

$$\mathbb{E}[Y_{ig}(d, t)] = c_{dt} + \mu_0 k$$

$$\text{Cov}(Y_{ig}(0, 0), Y_{ig}(d, t)) = k\mu_0(1 - \mu_0)$$

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  $\rho$  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 A5. The last parameter is set to zero to simulate the probability of type I error.

Table A5: Simulation results

|            | True value | Prob(reject) |
|------------|------------|--------------|
| $\theta_1$ | 0.021      | 0.812        |
| $\theta_2$ | 0.026      | 0.798        |
| $\theta_3$ | 0.027      | 0.791        |
| $\tau_1$   | 0.028      | 0.801        |
| $\tau_2$   | 0.026      | 0.800        |
| $\tau_3$   | 0.000      | 0.045        |

The simulation results are in line with the ones from 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.

## F Supplemental Econometric Appendix

### F.1 Within-Group Assignment Mechanisms

#### F.1.1 Fixed Margins

The within-group treatment is often assigned by choosing a fixed number of treated units within each group depending on  $p_t$ . However, depending on group size  $n_g$ , assigning exactly  $n_g p_t$  units to treatment is not possible when  $n_g p_t$  is not an integer. One possible way to deal with this issue is the following. Let  $N_g^1$  be the number of treated units in group  $g$ . Define a binary random variable  $\xi_g$  and let:

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

so that  $\xi_g$  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:

$$\mathbb{P}_g[\xi_g = 1 | T_g = t] = n_g p_t - \lfloor n_g p_t \rfloor.$$

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

### F.1.2 Bernoulli Trials

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

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

which do not vary over  $g$ . It follows that:

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

Then the variances are approximated by:

$$\mathbb{V}[\hat{\beta}_{0t}] \approx \frac{\sigma^2(0t)}{n q_t (1 - p_t)} \left\{ 1 + \rho_{0t} (1 - p_t) \left( \frac{\sum_g n_g^2}{n} - 1 \right) \right\} + \frac{\sigma^2(00)}{n q_0} \left\{ 1 + \rho_{00} \left( \frac{\sum_g n_g^2}{n} - 1 \right) \right\}$$

and

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

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

$$\begin{aligned} \Omega_n &= \frac{1}{n} \sum_g \mathbb{E}[\mathbb{1}'_g \mathbf{u}_g \mathbf{u}'_g \mathbb{1}_g] \\ W_n &= \frac{1}{n} \sum_g \mathbb{E}[\mathbb{1}'_g \mathbb{1}_g] \\ \widetilde{V}_n &= W_n^{-1} \Omega_n W_n^{-1}. \end{aligned}$$

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  $\lambda$  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 intracluster correlation satisfies  $|\rho| < 1$ . More generally, this condition restricts the amount of intracluster 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.

### F.3 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 F.2 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,

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

Under Assumptions 1, 2 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  $R_n$ ,

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

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:

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

where:

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

and

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

which completes the proof.  $\square$

## F.4 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:

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

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.  $\square$