Spillover Effects of Tax Enforcement on Production Networks: Evidence from Paraguay
Pre-analysis Plan

Michael Carlos Best[*] Florian Grosset-Touba[†] Gastón Pierri[‡] Evan Sadler[§] Panos Toulis[¶]

June 5, 2025

This documents represents the intentions of the research team at the time it is filed. We may deviate from this plan if unexpected issues arise, but we will report a "populated PAP" if we materially deviate from it (Duflo et al., 2020).

1 Introduction

The creation of a tax system that raises substantial revenues efficiently and equitably is one of the central challenges in economic development. In large part, this relies on the creation of the capacity to enforce taxes effectively—to reduce tax evasion. Doing this requires a detailed understanding of the drivers of tax evasion and the optimal allocation of extremely scarce tax enforcement resources. Our project combines new theory and a series of three Randomized Controlled Trials (RCTs) in Paraguay to provide new insights into tax evasion by firms, the strength of enforcement spillovers through production networks, and how to optimally target enforcement activities.

This Pre-analysis Plan (PAP) presents the design and our plan for the analysis for the second experiment. The first experiment is registered at https://www.socialscienceregistry.org/trials/12208. Its pre-analysis plan is available at https://www.socialscienceregistry.org/versions/194767/docs/version/document. A midline report summarizing our main findings is available at https://michaelcbest.github.io/files/BGPST_Masivos_Midline.pdf. The third experiment is planned for June/July 2025 as described in the midline report.

Section 2 describes the context and experimental interventions. Section 3 describes our main outcomes and hypotheses. Section 4 describes our randomization design, while section 5 presents the empirical specifications and test statistics we will use to estimate treatment effects and perform inference.

[*] michael.best@columbia.edu. Columbia University, BREAD, CEPR, & NBER
[†] florian.grosset@ensae.fr. CREST, ENSAE, Institut Polytechnique de Paris
[‡] gpierri@iadb.org. Inter American Development Bank.
[§] es3668@columbia.edu. Columbia University
[¶] panos.toulis@chicagobooth.edu. University of Chicago Booth School of Business

2 Context and Experiment

2.1 Context

Paraguay is an upper-middle income country in South America. Informality and tax evasion are central barriers to the development of the economy: 65% of workers are informally employed, and tax revenues are a paltry 10.3% of GDP (World Bank, 2022). Moreover, the tax administration has extremely scarce resources with which to enforce the collection of taxes.

In this context, deciding how to deploy scarce audit resources is a first-order administrative and policy concern. Our project focuses on understanding how optimal targeting rules may change when enforcement activities have spillovers through production networks. The theoretical part of our project develops a tractable network model in which to study tax enforcement. The model-predicted optimal targeting rules will depend on a number of features of the production network and on a number of assumptions in the model. The experimental part of our project studies these questions empirically.

In partnership with Paraguay's tax authority, the *Dirección Nacional de Ingresos Tributarios* (DNIT), we are doing a series of three interlinked experiments. We developed a scalable tax enforcement intervention, building on the work of Carrillo, Pomeranz, and Singhal (2017) in Ecuador, and described in greater detail in section 2.3. The first experiment studies the direct effects of the intervention on the targeted taxpayers. The first experiment is registered at https://www.socialscienceregistry.org/trials/12208. Its pre-analysis plan is available at https://www.socialscienceregistry.org/versions/194767/docs/version/document. A midline report summarizing our main findings is availiable at https://michaelcbest.github.io/files/BGPST_Masivos_Midline.pdf.

The second experiment is described in this pre-analysis plan. The third experiment, planned for June/July 2025 will use the findings from the first two experiments and our theoretical model to derive and then experimentally test alternative tax enforcement targeting rules. Before turning to the details of the experiments, subsection 2.2 describes the administrative data we base our analysis on.

2.2 Data Sources

Through our partnership with the DNIT, we are able to use anonymized administrative tax return data on the universe of Value Added Tax (VAT) filers in the country. These represent our main data sources. We augment this with data on interactions between taxpayers and officials to measure the costs borne by the tax administration in administering the enforcement intervention. Specifically, the administrative data comes from four sources.

1. Every VAT-liable taxpayer files a tax return each month. From this tax return, we obtain the reported taxable sales (the total amount they sold in a month) and their reported taxable purchases (the total amount they purchased in a month).

2. If a taxpayer wishes to amend a previous return, they file a new form revising their return. We obtain these forms for all taxpayers for the period of our experiment.

3. Designated taxpayers also have to file a monthly "libro de compras" annex together with their return. In this dataset, the taxpayers list all of their purchases, who they bought from, and how much they paid. We obtain this data aggregated at the buyer-seller-month level.

4. Designated taxpayers also have to file a monthly "libro de ventas". In this dataset, the taxpayers list all of their sales, who they sold to, and how much was paid. We obtain this data aggregated at the buyer-seller-month level.

The DNIT determines which taxpayers have to submit the “libro de compras” and “libro de ventas” annexes. Most taxpayers who have to submit a “libro de compras” also have to submit a “libro de ventas”, and vice-versa. All large taxpayers, most medium-sized taxpayers, and selected important small taxpayers are information agents. Descriptive statistics for the sample of taxpayers included in each data source are provided in Table 1. We obtain the data from July 2018 onwards.

Table 1: Description of the tax payers

	VAT filers in 3rd-party reports	3rd-party reporters of purchases & sales	3rd-party reporters of only purchases or sales
# taxpayers	368,542	15,970	1,106
1[Small firm]	0.985	0.744	0.948
1[Medium firm]	0.012	0.212	0.047
1[Large firm]	0.002	0.043	0.005
# making 3rd-party sales reports	16,615	15,970	653
# making 3rd-party purchase reports	16,281	15,970	453
Taxable Sales in VAT return (PYG MM)	41.789	363.785	155.007
	(122.986)	(319.217)	(234.166)
Taxable Purchases in VAT return (PYG MM)	36.129	304.957	143.694
	(122.986)	(319.217)	(234.166)
Value added (PYG MM)	5.459	55.609	11.043
	(47.839)	(147.309)	(90.991)
VAT (PYG MM)	0.302	3.125	0.380
	(3.884)	(12.148)	(7.659)
# months / year file return	7.691	7.445	8.928
	(4.521)	(4.931)	(4.597)
third-party reported sales / self reported sales	0.627	0.747	0.484
	(0.472)	(0.683)	(0.522)
third-party reported purchases / self reported purchases	0.319	0.662	0.539
	(0.642)	(0.981)	(0.889)
# of taxpayers' suppliers making 3rd-party reports	21.347	195.918	81.128
	(58.405)	(203.758)	(144.356)
# of taxpayers' clients making 3rd-party reports	21.569	197.204	72.689
	(116.838)	(495.941)	(192.735)

Note: The data presented in this table covers the period from July 2018 to December 2021. Standard deviations in parenthesis. The final subsamples consist of entities that are also registered in the "Libro de Compras" and "Libro de Ventas". We excluded firms with zero reports in Compras, Ventas, or VAT Returns. To calculate the monthly averages of taxable sales in VAT returns, of taxable purchases in VAT returns, of value added in VAT returns, and VAT paid in VAT returns, we employed a technique called windsorizing. This involved replacing values below the 1st percentile and above the 99th percentile with the corresponding values at the 1st and 99th percentiles, respectively. This approach helps mitigate the influence of outliers and ensures more reliable and representative statistics. PYG stands for Paraguayan Guarani, the official currency of Paraguay. One dollar is equivalent to 7277.60 PYG approximately. MM stands for millions.

2.3 Tax Enforcement Intervention

Our intervention is based on the discrepancies between taxpayers' tax return declaration and their trading partners' informational declarations.

2.3.1 Discovering Reporting Discrepancies

For each focal taxpayer i in each month m, we compute the total amount that taxpayers who make informational declarations about their purchases report buying from the focal taxpayer. This is their third-party reported sales y^r_{im}. This is then compared with what the focal taxpayer declares in their tax return y^d_{im}. Since not all taxpayers are required to make informational declarations about their purchases, the focal taxpayer's declared sales should a-piori always be at least as large as third-party reported sales. That is, it should be the case that the reporting *discrepancy* $d_{im} = y^r_{im} - y^d_{im} \leq 0$. Note that when computed this way, it is natural for there to be discrepancies. Since only the most important taxpayers (mostly large- and medium-sized taxpayers) make third-party reports, only part of taxpayers' sales are covered by third-party reporting. As a result, we expect most discrepancies to be negative. Whenever this is not the case, there must be misreporting, either because the focal taxpayer has under-reported their sales, because one or more clients have over-reported their purchases, or because there has been a mistake in a declaration.[1]

Figure 1 shows the distribution of normalized discrepancies (since discrepancies vary tremendously in size, we normalize them by computing $\tilde{d}_{im} = d_{im}/\frac{1}{2}\left(y^r_{im} + y^d_{im}\right)$). Panel A shows the overall distribution of normalized discrepancies, removing cases where discrepancies are exactly zero. We see that the bulk of discrepancies are negative, but there is a meaningful mass of positive discrepancies that are potentially the result of tax misreporting. In Panel B we plot the distributions of strictly positive discrepancies separately for sales taxable at 10% (the standard rate) and sales taxable at 5% (the reduced rate applied to a subset of goods). Consistent with misreporting, we see that the distribution of 10% discrepancies features more medium-sized and large discrepancies while those taxable at 5% tend to be smaller and are less likely to feature completely unreported sales (a smaller mass at 2).

In Figure 2, we explore the incidence of positive discrepancies across taxpayers and over time. For each taxpayer, we compute the proportion of months in which they file a tax return that shows positive discrepancies compared to third-party reports of their sales. Panel A shows the distribution of this propensity across taxpayers, removing the 59% of taxpayers who never have positive discrepancies.[2] We see that there are many taxpayers whose returns always generate positive discrepancies, but also many taxpayers who only sometimes generate positive discrepancies. In Panel B we track the percentage of returns that have positive discrepancies over time. The green line shows the percentage of returns with any discrepancy, while the red line breaks out only discrepancies in sales taxable at 10% and the blue line shows discrepancies at 5%. The incidence of discrepancies is remarkably stable over time, with about 7% of returns featuring some discrepancy. However, there is no discernible seasonality, though there may be a small reduction starting at the onset of the pandemic.

Not all discrepancies are large enough to merit the tax authority's attention, so only taxpayer-months in which the positive discrepancy was at least 10M PYG (approximately USD 1,400) were considered eligible for our intervention.

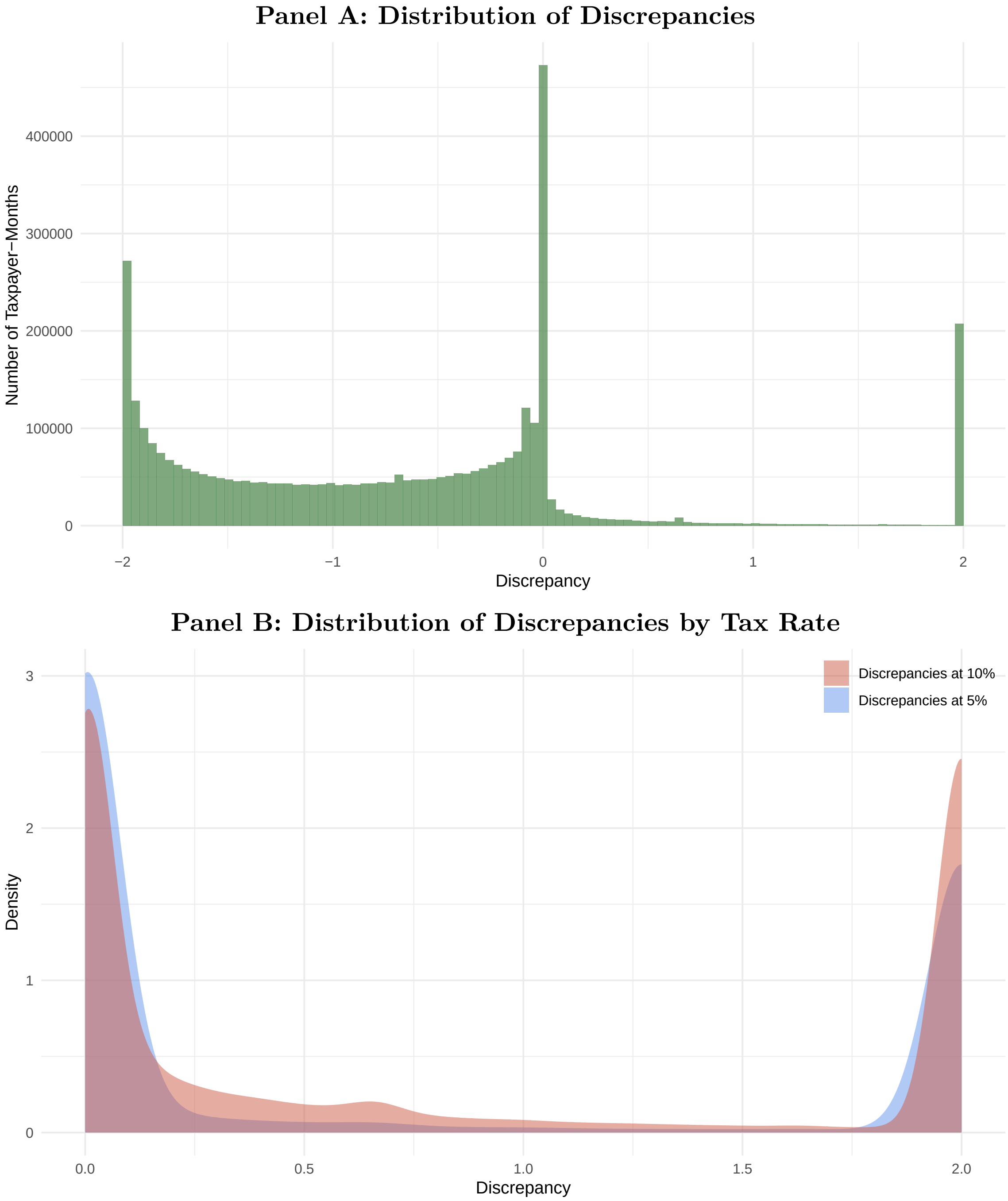
2.3.2 Intervention

Our intervention is a modified version of the intervention studied by Carrillo et al. (2017) in Ecuador. Among the eligible taxpayers, taxpayers were randomly selected (following the protocols described in section 4) to be sent a notice. The notice presents taxpayers with a summary by month of their reported sales, the purchases reported by their clients, and the

[1] Common mistakes discovered during the experiment include imports/exports being declared using the wrong exchange rate, invoices that are later canceled due to product returns, sales on credit reported as liquidated, and sales assigned to the incorrect month.

[2] This can happen either because none of their clients are third-party reporters or because they report sales totaling more than the available third-party reports.

Figure 1: Distribution of Discrepancies: 2019–2021



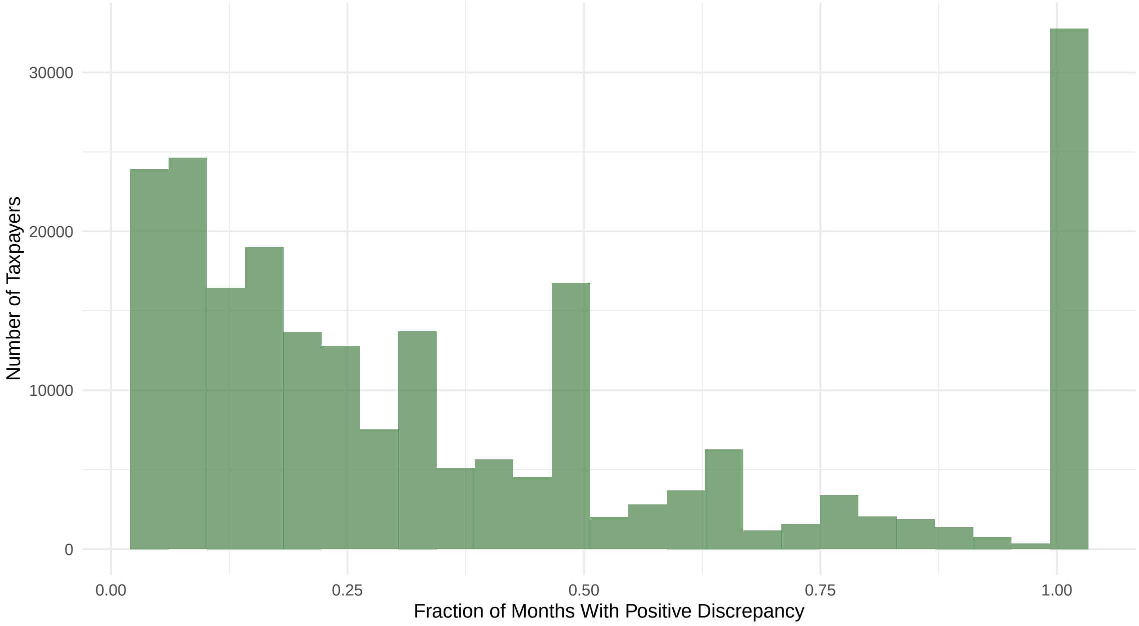
Panel A: Distribution of Discrepancies

Panel B: Distribution of Discrepancies by Tax Rate

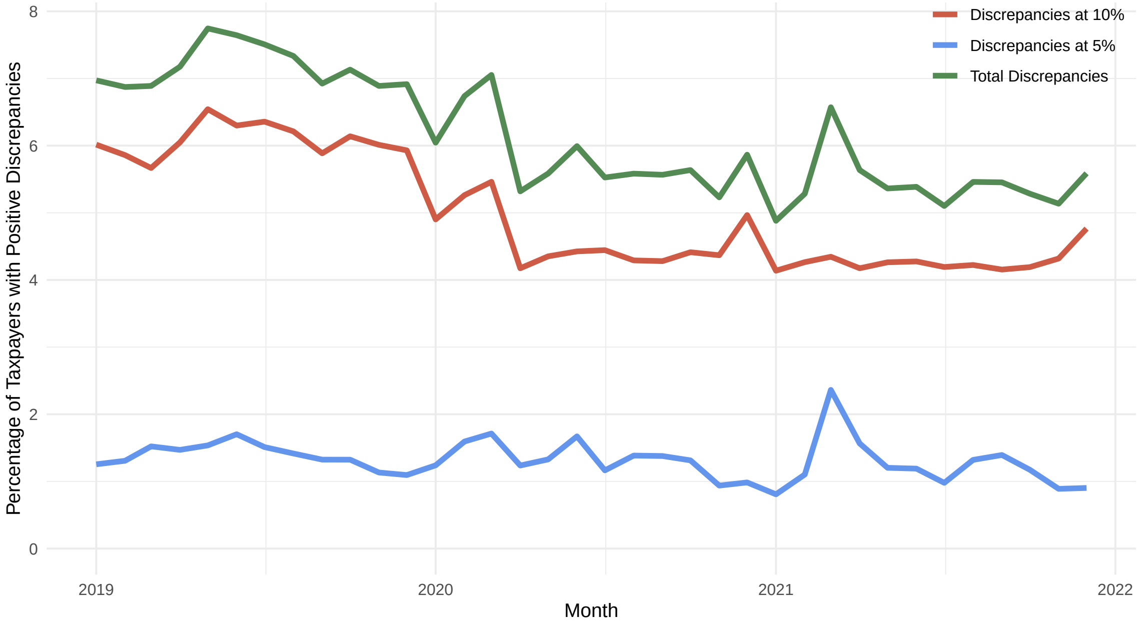
Notes: The figure shows the distribution of discrepancies between the sales declared by taxpayer i in month m in their VAT returns, y^d_{im}, and the total purchases from the taxpayer reported by their clients in their informational declarations y^r_{im}. We normalize these by the average of the two reports to account for the large differences in firm size: $\tilde{d}_{im} = y^r_{im} - y^d_{im}/\frac{1}{2}\left(y^r_{im} + y^d_{im}\right))$. Whenever $\tilde{d}_{im} > 0$, it is indicative of potential misreporting. Panel A shows the overall distribution of normalized discrepancies, removing cases where discrepancies are exactly zero. We see that the bulk of discrepancies are negative, but there is a meaningful mass of positive discrepancies. In Panel B we plot the distributions of strictly positive discrepancies separately for sales taxable at 10% (the standard rate) and sales taxable at 5% (the reduced rate applied to a subset of goods).

Figure 2: Frequency of Positive Discrepancies 2019–2021

Panel A: Incidence of Positive Discrepancies Across Taxpayers



Panel B: Percentage of Declarations with Positive Discrepancies



Notes: The figure shows the incidence of positive discrepancies across taxpayers and over time. For each taxpayer, we compute the fraction of months in which they file a tax return but it generates positive discrepancies when compared to third-party reports of their sales. Panel A shows the distribution of this propensity across taxpayers, removing the 59% of taxpayers who never have discrepancies. Panel B tracks the percentage of returns that have positive discrepancies over time. The green line shows the percentage of returns with any discrepancy, while the red line breaks out only discrepancies in sales taxable at 10% and the blue line shows discrepancies at 5%.

discrepancies between them. The invoices underlying the purchases reported by clients are attached as an annex to back up the discrepancies. The notice then requests that the taxpayer file amendments of the relevant tax returns to address the inconsistencies.

Carrillo et al. (2017) present somewhat disappointing results from the rollout of this type of intervention in Ecuador. Specifically, two issues arose. First, the overall response rate in Ecuador was fairly low: only 10–20% of notified taxpayers made an amendment. Second, those taxpayers that did respond did not become more compliant: each dollar of increases in reported sales tended to be accompanied by increases in reported purchases of up to 96 cents, leading to little or no change in reported value added and hence no detectable increase in tax liabilities or tax payments.

To address these issues we made two changes to the notices. First, we strengthened the language used. Specifically, the notices emphasized that the request was for the taxpayer to file an amendment to their tax return in which they increase their reported sales to account for the discrepancy, but not make any other amendments to their return. This was intended to avoid taxpayers simultaneously increasing their reported purchases.[3] Second, the notices were accompanied by greater follow-up and sanctions for non-response. In particular, auditors were assigned to follow up on each case. If the taxpayer responded but did not amend or amended only partially, the auditors reviewed the case and made a determination about whether to accept the partial response. If the taxpayer did not respond by the 10-day deadline, or responded unsatisfactorily, the auditor blocked the taxpayer's ID ("RUC"). Blocked IDs are not able to request more tax invoices, and are unable to perform a range of other processes with the tax authority (though they are still able to file and amend returns), potentially disrupting business activities for non-compliant taxpayers, especially larger ones that rely more on being able to issue tax invoices and have more complicated interactions with the tax authority.

The notifications for this experiment were sent to 5,000 taxpayers in December 2024. They were selected as described in section 4. Once sufficient time has passed, we will receive updated data in June 2025 with which to study the impacts as described in the following sections.

3 Outcomes and Hypotheses

3.1 Outcomes

In this project, we want to estimate the effects of tax audits on four main types of outcomes, for both the targeted taxpayers (direct effects) and their trading partners (spillover effects):

- Amendments: Do taxpayers amend their tax returns? If so, do they amend their reported sales? Reported purchases? Does this increase their tax liability and tax payments?
- Tax returns: In their tax returns filed after the intervention, do taxpayers report different amounts of sales? Purchases? Does this increase their tax liability and tax payments?
- Bilateral flows: Does the intervention change whom taxpayers report buying and selling from?
- Discrepancies: Is the discrepancy between the taxpayer's VAT total reported sales and the sum of its buyers' bilateral purchase reports reduced by the intervention? Is the discrepancy between each of the taxpayer's bilateral sales reports and the corresponding bilateral purchase report by its buyer reduced by the intervention? Are the related discrepancies on the side of the taxpayer's *purchases* changed by the intervention?

[3] Naturally, the tax authority cannot prevent taxpayers from filing amendments that amend many lines on their returns. Rather, the message was intended to be (highly) suggestive.

3.2 Hypotheses

In this project, we are generally interested in both the direct and spillover effects of the intervention. An earlier phase of the project focused on testing for the direct effects of the intervention. That phase is described in our previous pre-analysis plan for the experiment registered at https://www.socialscienceregistry.org/trials/12208 with pre-analysis plan available at https://www.socialscienceregistry.org/versions/194767/docs/version/document.

The main focus of this new phase of the project is to test for the existence and shape of spillover effects from tax audits. We will still test for direct effects, and adopt the same hypotheses as the earlier phase and PAP.

To express formally the null hypotheses of interest, let $Y_j(d_j, \mathbf{d}_{N_j})$ denote the potential outcome of taxpayer j. This definition depends on the taxpayer's own treatment status $d_j \in \{0, 1\}$ and the *neighborhood treatment*, $\mathbf{d}_{N_j}$. The neighborhood treatment is a type of "sufficient statistic" for the potential outcome, and is often referred to as *exposure mapping.* This particular specification assumes away any outcome dependence on second or higher-degree neighbors, but our framework could be expanded if necessary. To define neighborhood treatments in more detail, we will use $s_{ij} \in \{0, 1\}$ to indicate whether taxpayer i is a supplier to taxpayer j, $c_{ij} \in \{0, 1\}$ to indicate whether taxpayer i is a client of taxpayer j, ω^s_{ij} as the weight of taxpayer i among all suppliers to taxpayer j, and ω^c_{ij} as the weight of taxpayer i among all clients to taxpayer j.

Global effect. To test whether the treatment has any effect whatsoever on the outcomes, we will test the following null hypothesis:

$$H_0 : Y_j(\mathbf{d}) = Y_j(\mathbf{d}'), \text{ for all } j, \ \mathbf{d}, \mathbf{d}' \in \{0, 1\}^n. \tag{1}$$

Note that this hypothesis does not require that the neighborhood exposure mapping is correct since we are testing for a global treatment effect. To test H_0 we can use an ANOVA-like procedure where we fit two models: one model of outcomes Y on covariates X without including the treatments and a second model that includes treatments. The difference in the fit between the two models can be used as the test statistic. See Guo, Lee, and Toulis (2025) for details.

Direct effect. The direct effect of treatment aims to isolate the effect of the individual treatment assignment, while controlling for the treatment assignments of the individual unit's neighbors. Thus, we define:

$$H_0^{\text{dir}} : Y_j(1, \mathbf{d}_{N_j}) = Y_j(0, \mathbf{d}_{N_j}), \text{ for all } j, \ \mathbf{d} \in \{0, 1\}^n. \tag{2}$$

In this definition, the neighborhood treatment needs to be fixed at the same vector value to allow the imputation of potential outcomes under H_0^{dir}. In practice, this requires us to randomize treatment between taxpayers that are not connected to each other; i.e., they form an anticlique in the interfirm network. See Section 4.3 for details.

Heterogeneity of direct effect within subgroups. For any subgroup of taxpayers S (e.g., of particular size or economic sector), we will test the following hypothesis:

$$H_0^{\text{dir-het}} : Y_j(1, \mathbf{d}_{N_j}) - Y_j(0, \mathbf{d}_{N_j}) = \tau, \text{ for all } j \in S, \text{ and some constant } \tau \in \mathbb{R}. \tag{3}$$

The constant τ can be arbitrary. Rejecting $H_0^{\text{dir-het}}$ would imply that the treatment effect is heterogeneous within group S; e.g., it could be stronger for taxpayers of particular size and sector. In general, randomization tests of heterogeneity require a careful selection of test statistic that is sensitive to the variation of the treatment effect by taxpayer unit type. In Section 5, we propose outcome model specifications that could be used to select such test statistics.

Heterogeneity of direct effect with paired subgroups. Consider two subgroups of taxpayers, S_1 and S_2; e.g., S_1 could be the set of small taxpayers and S_2 the set of medium-sized taxpayers. We would like to test whether the treatment effects are constant within each subgroup and that these constant effects are equal between the two groups. That is,

$$\begin{aligned} H_0^1 :&Y_j(1, \mathbf{d}_{N_j}) - Y_j(0, \mathbf{d}_{N_j}) = \tau_1, \text{ for all } j \in S_1. \\ H_0^2 :&Y_i(1, \mathbf{d}_{N_i}) - Y_i(0, \mathbf{d}_{N_i}) = \tau_2, \text{ for all } i \in S_2. \\ H_0^{\text{pair}} :&H_0^1 \cap H_0^2 \cap \{\tau_1 = \tau_2\}. \end{aligned} \tag{4}$$

In this formulation, we would like to create a test that is valid for the composite hypothesis H_0^{pair} but powerful only against the alternative $H_a : \tau_1 \neq \tau_2$. To that end, we could consider using the difference in treatment effects between the two subgroups as the test statistic.

Note that H_0^{pair} could be rejected if the effects are heterogeneous within each group (even if they are on average equal between the two groups considered). As such, testing first for the heterogeneity of direct effects within subgroups is necessary: if the null hypothesis of homogeneity is rejected, it is necessary to find smaller groups (that are homogeneous) before testing for the heterogeneity of effects within paired subgroups.

Spillover effect. In contrast to the direct effect, the spillover effect isolates the effect of neighbors' treatment assignment, while controlling for the individual treatment assignment. Thus, we want to test:

$$H_0^{\text{spill}} : Y_j(0, \mathbf{d}_{N_j}) = Y_j(0, \mathbf{d}'_{N_j}), \text{ for all } j, \ \mathbf{d}, \mathbf{d}' \in \{0,1\}^n. \tag{5}$$

Here, the individual treatment needs to be fixed (either control or treatment) in order to isolate the effect from the individual unit's neighbors on the unit's outcome. Following the idea described in the global null, in order to test H_0^{spill} we could fit two models: one that regresses outcomes on covariates excluding neighborhood treatments and another model including them. The difference in fit between the two models measures the strength of the spillover effect.

Heterogeneity of spillover effect within subgroups. For any subgroup of taxpayers S (e.g., of particular size or economic sector), we would like to test the following hypothesis:

$$H_0^{\text{spill-het}} : Y_j(0, \mathbf{d}_{N_j}) - Y_j(0, \mathbf{0}) = \tau, \text{ for all } j \in S, \ \mathbf{d} \in \{0,1\}^n, \text{ and some constant } \tau \in \mathbb{R}. \tag{6}$$

The constant τ can be arbitrary. Rejecting $H_0^{\text{spill-het}}$ would imply that the spillover effect is heterogeneous within group S, say, depending on the size or sector of the focal taxpayer. As in the case of heterogeneous direct effects, our tests for spillover heterogeneity will generally rely on test statistics that are sensitive to the variation of spillover effect by focal unit type.

To uncover heterogeneity in the spillover effect due to the focal units-neighbors interaction, a variant of this hypothesis could also include the type of neighbor taxpayers as well as its interaction with the type of their focal taxpayers. For instance, we could test whether the spillover effects on a focal taxpayer j are stronger if they originate from j's supplier taxpayers or j's client taxpayers.

Specifically, let $C_j = \{i : c_{ij} > 0\}$ and $S_j = \{i : s_{ij} > 0\}$ be the set of clients and suppliers of j, respectively. We want to ask:

A. *For a taxpayer j, does auditing its suppliers affect its own outcomes?*

$$H_0^A : Y_j(0, \mathbf{d}_{N_j}) = Y_j(0, \mathbf{d}'_{N_j}), \text{ for all } j, \mathbf{d}, \mathbf{d}' \in \{0,1\}^n, \text{ such that } \mathbf{d}_{C_j} = \mathbf{d}'_{C_j}. \tag{7}$$

A variation of this hypothesis is to consider a weighted linear exposure mapping, and then test the following null hypothesis:

$$H_0^{A'} : Y_j(0, \mathbf{d}_{N_j}) = Y_j(0, \mathbf{d}'_{N_j}), \text{ for all } j, \mathbf{d}, \mathbf{d}' \in \{0,1\}^n, \text{ such that } \sum_{i \neq j} c_{ij} \omega_{ij}^c d_i = \sum_{i \neq j} c_{ij} \omega_{ij}^c d'_i.$$

B. *For a taxpayer j, if some of its suppliers are treated, does the type of its treated suppliers (e.g., its most or least important suppliers in value) affect its outcomes?*

$$H_0^B : Y_j(0, \mathbf{d}_{N_j}) = Y_j(0, \mathbf{d}'_{N_j}), \text{ for all } j,\ \mathbf{d}, \mathbf{d}', \text{ such that } \mathbf{d}_{C_j} = \mathbf{d}'_{C_j}, \text{ and } \sum_{i \neq j} s_{ij}\mathbf{d}_i = \sum_{i \neq j} s_{ij}\mathbf{d}'_i.$$

Rejecting H_0^B implies that the spillover effect on focal taxpayer j from j's suppliers does not depend only on the number of treated suppliers, but potentially also on the types of j's treated suppliers.

C. A natural variation of H_0^A and $H_0^{A'}$ is to test these hypotheses with the roles of suppliers and clients reversed. Similarly, a natural variation of H_0^B is to test whether the type of client taxpayers affects the spillover effects, controlling for the total number of treated neighbor clients. This is testing H_0^B with the roles of suppliers and clients reversed.

3.3 Dimensions of Heterogeneity

As discussed above, we seek to test hypotheses related to the heterogeneity of the direct and spillover effects of the notifications. In particular, the dimensions of heterogeneity that we will consider for each taxpayer are:

- legal form
- region
- economic activity
- firm age
- firm size
- history of amendments
- history of tax declarations (sales, purchases, tax liability, credits etc)
- network position summarized using network centrality statistics (weighted by trading volumes)

We will also use these same observables for the treated taxpayers' trading partners (their clients and suppliers). And we will allow for rich interactions between all of the above, following the methods described below in section 5. In the event that we receive additional data with which to search for heterogeneity we will use it and update this Pre-Analysis Plan.

4 Randomization Design

The randomization design implemented in this second phase of the larger project is illustrated in its entirety in Figure 3. The figure shows that our design is comprised of three key components: the stratified randomized design implemented to detect spillover effects, the stratified randomized design to detect direct effects, and the algorithm (marked with "ALGORITHM" in the figure) utilized to split the eligible and non-eligible taxpayers in order to implement these two randomized designs. Before we present details about all three key components in the subsections that follow, we first give some background information on pre-processing the phase-2 data.

Initially, we create two sets for the taxpayers that may be eligible to be treated in phase 2 (marked as "E", $N = 62,344$) and those that are not eligible for treatment (marked as "NE", $N = 332,649$). Upon inspection, we removed 6,819 taxpayers from the E set that were already treated

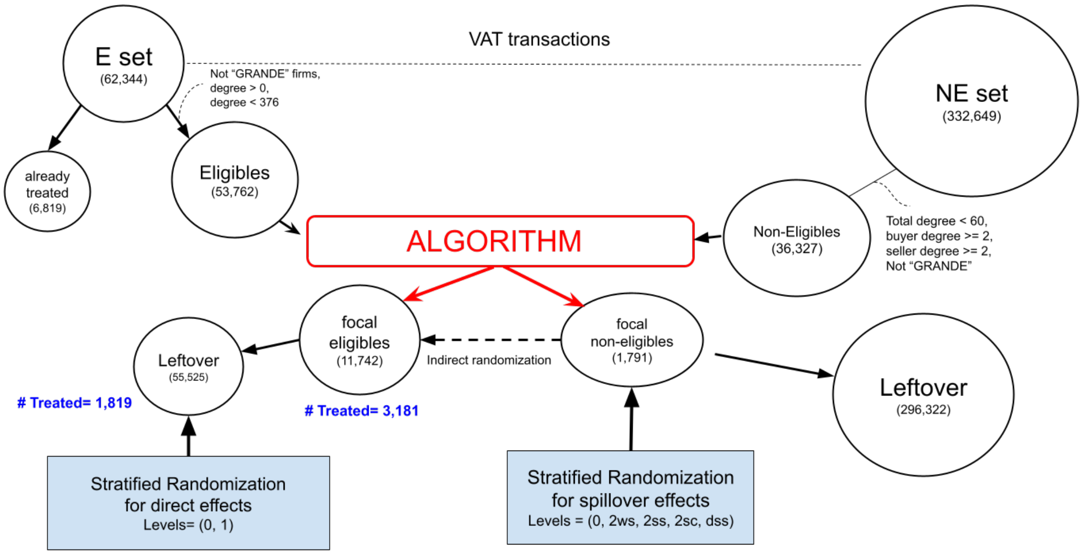


Figure 3: Schematic Representation of the Randomization Design in Phase 2.

in previous waves. Moreover, to reduce the chance of saturating the taxpayers with spillovers, we removed taxpayers of type "GRANDE" and degree larger than 376, which corresponds to the 97.5th percentile of the empirical degree distribution. This left us with a set of 53,762 taxpayers that were eligible for treatment in phase 2 (marked as "Eligibles").

In the NE set, we followed a similar approach, and removed taxpayers (i) of type "GRANDE", (ii) degree > 60, (iii) buyer/seller degree smaller than 2. The buyer (or seller) degree of a taxpayer is equal to the number of transactions the taxpayer made as a buyer (or seller). As explained in later sections, thresholding the buyer/seller degree allows us to experiment with directional treatment effects, and, in particular, to distinguish between seller-induced or buyer-induced spillover effects. This filtering of the NE set left us with a set of 36,327 taxpayers ("Non-Eligibles"), that were not eligible for treatment, a roughly 90% reduction compared to the original set.

4.1 Algorithm for Focal Set Creation

Given the sets of Eligibles and Non-Eligibles described above, the next step was to create the set of focal taxpayers for the experimental randomization. These focal taxpayers will also be used for the randomization-based analysis of the experimental results. An illustration of this algorithm is shown in Figure 4.

The algorithm begins with an empty *focal non-eligible set*, say $\text{FNE} = \emptyset$. Let NonNE denote all the 36,327 non-eligible taxpayers initially in Wave 5. The *focal eligible set* is initially set equal to the entire set of eligibles, i.e., $\text{FE} = \text{Eligibles}$. The algorithm then proceeds in the following steps.

Algorithm 1 Choose Focal Non-eligibles

```
1. Choose a taxpayer randomly from NonNE and calculate all it's first-order neighbors, say $\mathcal{N}(i)$, in the Eligibles set.
2. Add $i$ in the focal non-eligible set: $\text{FNE} \leftarrow \text{FNE} \cup \{i\}$.
3. Add the neighborhood of $i$ to the focal eligible set: $\text{FE} = \text{FE} \cup \mathcal{N}(i)$.
4. Remove from NonNE all first-order connections to $\mathcal{N}_i$: $\text{NonNE} \leftarrow \text{NonNE} \setminus \mathcal{N}(\mathcal{N}(i))$.
5. Repeat Step 1 unless NonNE is empty.
```

The output of this algorithm consists of the sets FNE and FE. The set of focal non-eligibles,

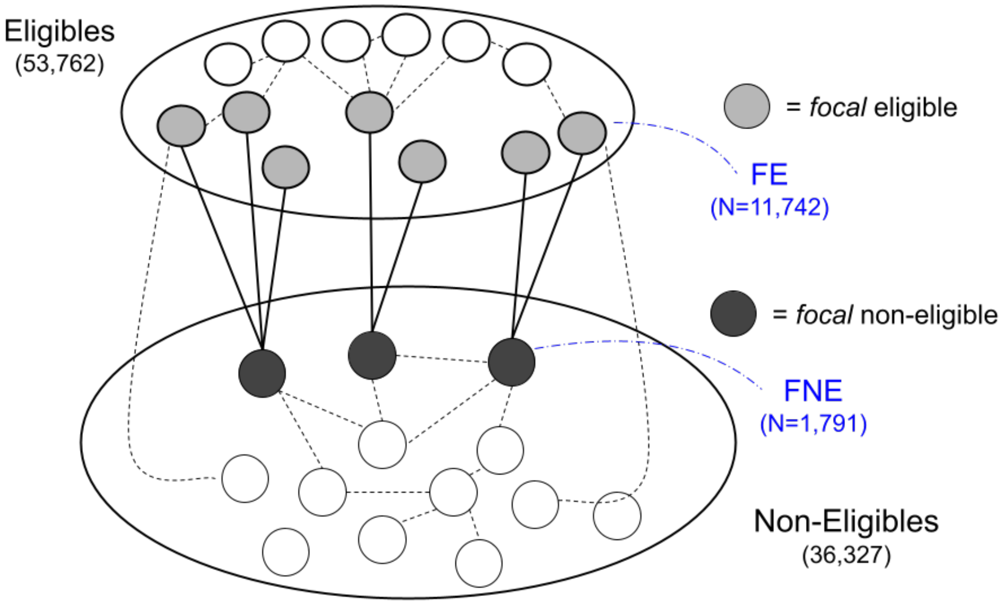


Figure 4: Algorithm producing the focal units for neighborhood treatment randomization. Solid lines correspond to connections between taxpayers, and dashed lines to potential connections. By construction, there is only one solid line from a focal eligible (FE) taxpayer to non-eligible taxpayers (FNE).

FNE, is the focal set of taxpayers that will be used to test for spillovers effects. The set of focal eligibles, FE, is the focal set of taxpayers that will be used to randomize treatment. The key idea behind the algorithm is that each taxpayer in FE is connected *only to a* single taxpayer in FNE. As a result, we can randomize "neighborhood treatments" *independently* on the non-focal eligible taxpayers in FNE.

Figure 4 visualizes these points and how the algorithm works. The set of focal eligible taxpayers, FE, corresponds to the light gray taxpayers in the set of Eligibles. The set of focal non-eligible taxpayers, FNE, corresponds to the dark gray taxpayers in the set of non-Eligibles. In our data, the algorithm calculates a total of $11,742$ FNE taxpayers and $1,791$ FE taxpayers. It is important to note that taxpayers in FNE can be connected *only to one* taxpayer in FE, but they can be connected to each other and also to non-FNE taxpayers in the set of non-Eligibles. This is illustrated in the figure through the dashed lines. Taxpayers in the FNE set are also allowed to be connected to each other although these connections tend to be rare.

4.2 Randomization of Neighborhood Treatments

After the algorithm has been applied, and sets FE and FNE have been generated, the next step is the randomization of neighborhood treatments. We use 5 different levels for these neighborhood treatments:

(i) **0**: No neighbor is treated.

(ii) **2ws**: Treat 2 weakest suppliers (seller neighbors to ego taxpayer).

(iii) **2ss**: Treat 2 strongest suppliers .

(iv) **2sc**: Treat 2 strongest clients (buyer neighbors from ego taxpayer).

(v) **dss**: Treat d strongest suppliers.

In these definitions, the "strength" of a connection is calculated based on weighing the transactions between taxpayers. In particular, for each taxpayer we calculate two weights corresponding to the taxpayer operating as a buyer or a seller. To calculate the strength of supplier j to taxpayer i, we divide the transaction value between buyer i and seller j by the total number of transactions in which taxpayer i acts as a buyer. The definition for the strength i's clients is completely symmetrical.

We selected the particular 5 neighborhood treatment levels listed above based on which type of heterogeneity in the spillover effects we aimed to be able to study through the experiment. In particular, comparing treatment **0** with other levels identifies, marginally, the existence of spillover effects. Comparing **2ws** with **2wc** could potentially identify whether treating buyer or seller neighbors matters for the spillover effect. Comparing **2ws** with **2ss** could potentially identify whether the strength of interactions between taxpayers matters for the spillover effect. Comparing **dss** with **2ss** could potentially identify whether the number of treated neighbors, while controlling for the strength of interactions, matters for the spillover effect. Finally, we note that in these definitions we did not include “weak clients” (treatment arm **2wc**) because preliminary analysis and theory suggest that the spillover effect should mainly originate from supplier neighbors.

Stratified randomization of neighborhood treatments. The final stage in our design is the actual randomization. We follow a stratified design, where we initially stratify taxpayers based on whether they have “MEDIANO” neighbors, their baseline amendment rate and whether they are mostly buyers or sellers. More concretely, we define these strata as follows. First, for each taxpayer in FNE we calculate whether there is a neighbor of the taxpayer in the FE set that has size “MEDIANO”. Second, the baseline amendment rate takes three levels according to whether the average *pre-treatment* amendment value of a taxpayer is positive, negative, or zero. Third, for each taxpayer in FNE we calculate whether in the majority of recorded transactions the taxpayer acts as a buyer or a seller. As a result, based on the possible values for “mediano neighbors” × baseline amendment × “majority buyer”, there are $3 \times 3 \times 2 = 12$ possible strata of FNE taxpayers.

Under this stratification, we randomize the 5 neighborhood treatments listed above in a completely randomized manner directly on the focal non-eligible taxpayers (FNE). We emphasize that no actual treatment is applied on the non-eligible taxpayers. Instead, the neighborhood treatments on the focal non-eligible taxpayers determine what actual randomization needs to occur on the focal eligible taxpayers (FE). For example, when a taxpayer i in FNE is assigned to **2wc**, then all taxpayers that bought from i are ordered according to their buyer strength with respect to i, as described above. Then, the two taxpayers with the lowest strength are treated while the rest are set in control. This scenario, along with other example neighborhoods treatments, is illustrated in Figure 5. In the resulting randomized design a total of 3,181 eligible taxpayers were treated.

4.3 Randomization for Direct Effects

The randomization design described above is used to identify spillover effects. For the direct effects, we utilize a similar design as in the first phase of the project (as described in the pre-analysis plan available at https://www.socialscienceregistry.org/versions/194767/docs/version/document), leveraging anticlique sets between eligible taxpayers. That is, we start from a subset of taxpayers in Eligibles that (i) are not connected to any other eligible taxpayer and (ii) are not connected to any taxpayer in FNE. The size of this subset of taxpayers is equal to 14,556, and since they form an anticlique (no connections between any two taxpayers in the set), they can be used to test for direct effects.

In particular, we follow a stratified design as before, where we initially stratify the remaining taxpayers in the Eligibles set based on degree, baseline amendment rate and size. More concretely,

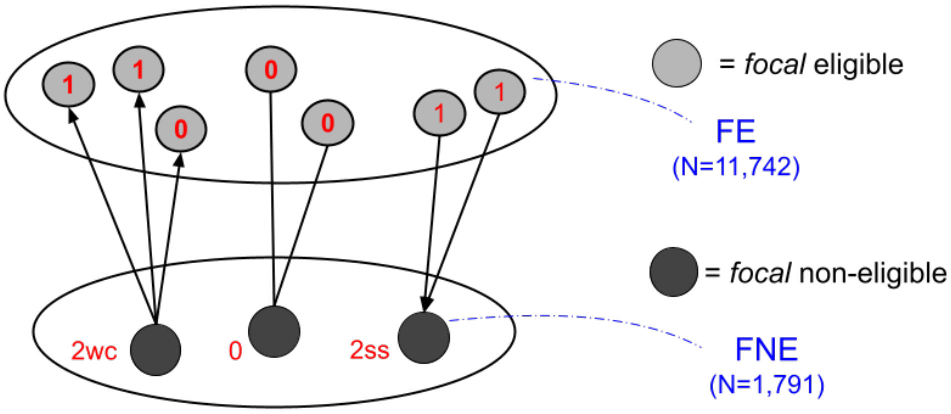


Figure 5: Stratified randomization for spillover effects. The design randomizes the 5 neighborhood treatments symbolically on focal non-eligible taxpayers in the FNE set. These neighborhoods treatments then can be translated into actual randomization on the focal eligible set (FE).

we define these strata as follows. First, we split the set of non-Eligibles in 10 bins according to their degree distribution, and we use the bin classification as the first factor in our stratification. Second, as in the spillovers case, the baseline amendment rate takes three levels according to whether the average *pre-treatment* amendment value of a taxpayer is positive, negative, or zero. Third, the size of the taxpayer is classified as whether it is "PEQUENO" or not. As a result, based on the possible values for degree × baseline amendment × size, there are $10 \times 3 \times 2 = 60$ possible strata. Several of these strata are empty, and are thus removed from the design.

In the remaining strata, we follow a completely randomized design, treating roughly a fixed proportion % of taxpayers within each stratum. To generate a spillover effect from this randomization step as well, we decided to treat the stratum with taxpayers having no connections to non-Eligibles with a slightly lower rate than those taxpayers that are connected to some non-Eligibles (4% against 17%, respectively. In the resulting stratified randomized design, a total of 1,819 taxpayers were treated. Combining these with the 3,181 taxpayers selected using the procedure described in section 4.2, we arrived at a total of 5,000 treated taxpayers.

5 Design Specifications

Our main objective, described in Section 3, is to test for the direct and spillover effects of tax audits—both their existence and shape. Below we propose some example specifications that could be useful to estimate effects and conduct the randomization tests of Section 3.2.

5.1 Linear models

To estimate direct effects or spillover effects, given our experimental design, we will use classic tools from applied microeconomics to compare relevant groups of taxpayers—including ANCOVA and DiD specifications. For the direct effects, we can focus on eligible taxpayers and compare treated and control units.

For the spillover effects, we can compare non-eligible taxpayers with different types of exposure to treated taxpayers.[4] The idea is to start with a linear model of spillovers,

$$Y_j(0, d_{N_j}) = Y_j(0,0) + \beta_0 + \beta_s(\sum_i w_{ij}^s d_i) + \beta_c(\sum_i w_{ij}^c d_i) + \beta_n \sum_{i \in N_j} d_i,$$

parametrized by $\boldsymbol{\beta} = (\beta_0, \beta_c, \beta_s, \beta_n)$. In this specification, β_0 is a constant spillover effect, β_s is the supplier spillover effect, β_c is the client spillover effect, and β_n is the general neighbor spillover effect. We can adjust this specification to incorporate a time dimension. In particular, if multiple waves of audits are implemented within this phase of the project, or when pooling the two phases of the project in the analysis, we will adapt the specification described above to the stacked difference-in-differences methodology used in Cengiz, Dube, Lindner, and Zipperer (2019). This method, also described and compared to other recent approaches in the two-way fixed effects difference-in-differences literature (Baker, Larcker, & Wang, 2022), allows us to flexibly account for the staggered implementation of the treatment across multiple waves, with varying set of eligible taxpayers and treatment assignment procedures across waves.

Our specifications will account for the fact that taxpayers' expected treatment exposure is not constant across taxpayers, due to both our experimental design and the network structure driving the potential spillover effects. Our experiment has been specifically designed to yield finite-sample valid inference for our null hypotheses of interest under randomization inference and reasonable assumptions. Such a randomization inference approach, also advocated by Ding, Feller, and Miratrix (2016) and Guo et al. (2025), is also valuable since outcomes are likely to be correlated in non-trivial ways across trading partners.

To be specific, note that we can use a Fisherian Randomization Test (FRT) to test any value $\boldsymbol{\beta} = \mathbf{b}$. Variations of this specification are possible; e.g., by diving the exposures defined above by the network degree of j. Let pval($\mathbf{b}$) denote the p-value from such a test. This p-value is finite-sample valid for any $\mathbf{b}$. We can invert the test and define:

$$\mathcal{B}_\alpha = \{\mathbf{b} \in \mathbb{R}^4 : \text{pval}(\mathbf{b}) > \alpha\}.$$

The set $\mathcal{B}_\alpha$ is the $100(1-\alpha)\%$ randomization-based confidence set for the spillover parameters defined above.

A key focus of this experiment, which will influence the optimal targeting, lies in the heterogeneity of both the direct and spillover treatment effects. To do this, we will first test for our null hypotheses, defined above, for each relevant subgroup. We plan to correct for multiple testing issues by following procedures similar to those advocated by Benjamini, Krieger, and Yekutieli (2006); List, Shaikh, and Xu (2019).

We will also apply *causal trees* to identify heterogeneity via recursive partitioning (Athey & Imbens, 2016; "Machine learning who to nudge: Causal vs predictive targeting in a field experiment on student financial aid renewal", 2025) and other methods from the causal ML literature (Chernozhukov, Demirer, Duflo, & Fernández-Val, 2018). However, these methods have not been designed for use in settings with spillovers. We will thus consider the use of Graphical Neural Networks as a flexible model of spillover effects, extending the linear model defined above—but recognize that these methods are mostly new, so their applicability in our setting in not guaranteed.

5.2 Recurrent event analysis

The main outcome data are the tax amendments made by taxpayers over time. Each amendment may be viewed as a time-stamped event, while multiple amendment events can be made

[4] We will estimate these spillover effects by focusing on the taxpayers whose neighborhood treatment has been randomized, as discussed above. We will also leverage the random assignment across all eligible taxpayers to estimate the same effects across all non-eligible taxpayers connected to an eligible taxpayer, to try to increase statistical power.

by any given taxpayer. A reasonable modeling approach for this type of data is therefore to employ recurrent event analysis, which is usually done by adapting Cox proportional hazards models. This framework could reveal how taxpayer covariates and treatment influence the rate of amendment activity while addressing dependencies between repeated observations.

One simple adaptation of the Cox model is the *Andersen-Gill* model which defines

$$\lambda_{ik}(t) = \lambda_0(t) \exp(X_{ik}'\beta),$$

as the hazard function for taxpayer i at the kth event. The remaining specification of the amendments simply follows the classical Cox proportional hazards formulation, where the likelihood over β can be expressed by the partial likelihood:

$$L(\beta) = \prod_{k \in \mathcal{K}} \frac{\exp(X_{i(k)k}'\beta)}{\sum_j \exp(X_{jk}'\beta)}$$

where $\mathcal{K}$ is the set of observed amendment events, and $i(k)$ denotes the taxpayer that amended in event k. Extensions of this model include tied events (i.e., when taxpayers amend on the same day) and competing risks between various kinds of event types (e.g. amending reported purchases or reported sales). Both types of extensions can be handled in a straightforward way by statistical packages in `R` (such as `coxph`) and `Stata`. In addition, we could search for heterogeneous effects by including interactions in the above model.

References

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Baker, A. C., Larcker, D. F., & Wang, C. C. Y. (2022, May). How much should we trust staggered difference-in-differences estimates? *J. Financ. Econ.*, *144*(2), 370–395.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*, 491–507.

Carrillo, P., Pomeranz, D., & Singhal, M. (2017). Dodging the taxman: Firm misreporting and the limits to tax enforcement. *American Economic Journal: Applied Economics*, *9*, 144–164.

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019, August). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, *134*(3), 1405–1454.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernández-Val, I. (2018). *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india.* (NBER Working Paper 24658)

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(3), 655–671.

Duflo, E., Banerjee, A., Finkelstein, A., Katz, L. F., Olken, B. A., & Sautmann, A. (2020). *In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics.* (NBER Working Paper No. 26993)

Guo, W., Lee, J., & Toulis, P. (2025). Ml-assisted randomization tests for detecting treatment effects in a/b experiments. *arXiv preprint arXiv:2501.07722.*

List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*, 773–793.

Machine learning who to nudge: Causal vs predictive targeting in a field experiment on student financial aid renewal. (2025). *Journal of Econometrics*, *249*, 105945. doi: https://doi.org/10.1016/j.jeconom.2024.105945