Optimal Redistribution with a Shadow Economy

Paweł Doligalski — University of Bristol

Luis E. Rojas [*] — MOVE, UAB and Barcelona GSE

October 5, 2020

Abstract

We extend the theory of optimal redistributive taxation to economies with an informal sector. Crucially, we allow for *moonlighting* — a situation when a worker has a formal main job and an informal secondary job. The optimal tax formula contains two novel terms capturing informality responses on an intensive and an extensive margin. Both terms decrease the optimal tax rates. We estimate the model with Colombian data and find that informality strongly reduces tax rates at all income levels. In particular, the possibility to migrate to entirely informal employment restricts tax rates at low and medium income levels, while the possibility of moonlighting is relevant at higher levels of earnings. We also show that the informal sector is welfare improving when the preferences for redistribution are weak and welfare deteriorating when the preferences for redistribution are strong. To explain this result, we demonstrate that the informal sector increases labor efficiency at the expense of possible redistribution.

Keywords: informal sector, moonlighting, income taxation, redistribution.
JEL Codes: H21, H26, J46.

[*] Paweł Doligalski (corresponding author): Department of Economics, University of Bristol, The Priory Road Complex, Priory Road, BS8 1TU, Bristol, the United Kingdom, email: pawel.doligalski@bristol.ac.uk, tel: +44 117 954 6930. Luis E. Rojas: Building B - Campus UAB, 08193 Cerdanyola del Vallés, Barcelona, Spain, email: luis.rojas@MOVEbarcelona.eu, tel: (+34) 93 581 47 39 Ext. 4739. Luis Rojas acknowledges support from the ERC Advanced Grant (APMPAL) GA 324048. Paweł Doligalski thanks the Central Bank of Hungary for the possibility of working on this project during his stay there. We are grateful for useful comments of Árpád Ábrahám, Charles Brendon, Antoine Camous, Hal Cole, Mike Golosov, Piero Gottardi, Ramon Marimon, Wojciech Kopczuk, Claus Kreiner, Dirk Krueger, Etienne Lehmann, Humberto Moreira, Erwin Ooghe, Wojciech Paczos, Evi Pappa, Dominik Sachs, Julia Schmieder, Jon Temple, Yanos Zylberberg and anonymous referees. All errors remain our own.

1. Introduction

Informal activity, broadly defined as any economic endeavor which evades taxation, accounts for a large fraction of economic activity in both developing and developed economies. The share of informal production in GDP is consistently estimated to be on average above 10% in high income OECD countries and above 30% in developing and transition countries, in extreme cases reaching 70% (Schneider and Enste 2000; Schneider, Buehn, and Montenegro 2011). Globally, 2 billion workers are employed informally (ILO 2018). The shadow economy allows workers to earn additional income which is unobserved by the government. Intuitively, this additional margin of response to taxation makes income redistribution more difficult. Indeed, empirical studies document informality responses following tax reforms.[1] On the other hand, the informal jobs seem to be less productive and attract mostly the poor.[2] If the informal sector benefits those in need, perhaps it can be useful from the social welfare perspective.

Our aim is to study the informal sector within an optimal taxation framework. We derive the optimal non-linear income tax schedule in an economy with a shadow sector and characterize how informality determines its shape. We quantify the importance of our theoretical results by estimating the model with Colombian data. Informality turns out to be quantitatively important for both the optimal policy and social welfare. To further understand the social welfare impact of informality, we propose a novel decomposition of a welfare change into the efficiency and the redistributive components.

Building on the seminal work of Mirrlees (1971), we consider a framework with heterogeneous agents equipped with distinct formal and shadow productivities. Workers face an idiosyncratic fixed cost of working in the shadow economy, which may reflect either ethical or technological constraints. The government observes only formal incomes and introduces a non-linear tax to maximize its redistributive welfare criterion. Importantly, we allow workers to supply labor simultaneously to the formal sector and the shadow sectors. In this way we can study *moonlighting*, which we define as a situation when a worker with a formal primary job has an informal secondary job. Informal secondary employment is common and accounts for a substantial fraction of informal workers in many countries.[3] Furthermore, evidence suggests that starting a tax-advantaged secondary

[1] Gorodnichenko, Martinez-Vazquez, and Peter (2009) show that a reduction of tax rates decreased informality in Russia, see Monteiro and Assunção (2012) and Rocha, Ulyssea, and Rachter (2018) for evidence from Brazil. Kleven, Knudsen, Kreiner, Pedersen, and Saez (2011) document a positive impact of income tax rates on tax evasion in Denmark. Regarding other taxes, Berger, Fellner-Röhling, Sausgruber, and Traxler (2016) find a positive impact of TV license fees on their evasion in Austria.

[2] We find that in Colombia, focusing on the main jobs, the shadow economy accounts for 58% of jobs and 55% of hours but only 31.4% of earnings.

[3] Out of all workers engaged in informal work, the share of workers with a formal main job was more than 10% in Barbados, more than 20% in the Russian Federation and Lithuania (Hussmanns and Jeu 2002) and more than 50% in Poland (Statistics Poland 2019). 13.5% of married households in Romania had at least one moonlighting member (Kim 2005, based on Table 1) and 12-15% of workforce were moonlighting in Ukraine (Commander, Isachenkova, and Rodionova 2013). In Brazil, 37% of

job is an important margin of response to tax reforms (Tazhitdinova 2017).

To set the stage for our main results, we first examine the incentive-compatible income choices. Our setting is an example of a multidimensional taxation, or screening, model, since agents are heterogeneous with respect to the productivity and the fixed cost of shadow employment. Such models are notorious for intractability (Rochet and Choné 1998). Indeed, we find that the local incentive-compatibility constraints are not sufficient to ensure global incentive-compatibility. On the other hand, we manage to incorporate additional constraints which ensure global incentive-compatibility without adding much complexity to the analysis. First, we summarize the second dimension of heterogeneity — the idiosyncratic fixed cost of informal employment — by focusing on two classes of agents: the *high-cost types* with prohibitively high fixed cost who are always working exclusively formally, and the *low-cost types* with no fixed cost who can be either exclusively formal, exclusively informal or moonlighting. We find that the local incentive constraints are not sufficient to prevent deviations to formal income levels where income distributions of the two classes do not overlap. Income distributions may be non-overlapping at the bottom (where there may be only the low-cost workers), at the top or where the tax schedule is locally regressive (where there may be only the high-cost workers). We include additional incentive constraints to account for these deviations.

Our main theoretical result is a sufficient statistics formula for the optimal tax schedule in the economy with an informal sector. The formula contains two novel terms due to informality responses on the extensive margin (getting an informal job) and the intensive margin (shifting hours between a formal and an informal job). The extensive margin responses are typically modeled as binary: working or not working. In our setting it would correspond to agents being able to work only formally or only informally and, as a result, would rule out moonlighting. Instead, we allow workers to moonlight, which means that they can complement formal earnings with additional income from an informal job. Intuitively, these responses can be important for workers with well-paid formal jobs who face high marginal tax rates and for whom transitioning to entirely informal employment is too costly. The possibility of moonlighting also gives rise to informality responses on the intensive margin — shifting hours between the formal main job and the informal secondary job. We find that moonlighting workers respond on the intensive margin differently than formal workers. First, the formal earnings of moonlighting workers are more elastic. Second, moonlighting workers would never choose formal earnings where the tax schedule is locally regressive, i.e. where the marginal tax rates are decreasing. Consequently, if the tax schedule features regions of regressivity, the formal income choices of the moonlighting workers can become discontinuous to avoid these

all secondary jobs are micro-enterprises and can be classified as informal (Henley, Arabsheibani, and Carneiro 2009). Balán, Browning, and Jelin (1973) describe the case of a Mexican factory worker who had been moonlighting in his informal shoe store for three years before he was sure of the success of his venture and quited the factory job (pp. 216-217). Regarding Colombia, while we do not detect significant moonlighting given the current tax schedule, our model predicts that moonlighting becomes substantial once the tax schedule becomes more progressive.

regions. Then, following a change of the marginal tax rate, the moonlighting workers can respond on the intensive margin by jumping over the regressivity region to a discretely lower level of formal earnings. In contrast, the intensive margin responses of formal workers are always marginal. Even though informality responses may involve abrupt earnings changes, we summarize their impact on tax revenue with well-defined elasticities.

We analytically examine how informality affects the optimal tax rates in two ways. First, we fix the distribution of formal income and examine what happens if informality responses were ignored. We find that ignoring informality responses would result in higher tax rates. In other words, correctly accounting for work incentives in the presence of the informal sector leads to lower optimal tax rates. Second, we fix model primitives, such as the distribution of productivities in the formal sector, and compare the optimal top tax rate in the model with and without the shadow economy. This comparison is more challenging since the income distribution is allowed to freely adjust to tax policy. We analytically show that the optimal top tax rate in the model with a shadow economy is lower both due to the informality responses and due to the endogenous adjustment of the income distribution. In particular, once the top tax rate exceeds a certain tipping point, a large fraction of top earners start to moonlight and discretely reduce their formal earnings. Given a lower number of individuals with high formal earnings, it is optimal to set the top tax rate at a lower level.

We estimate the model with Colombian data. Colombia is an attractive case study for two reasons. First, it has a large informal sector: we find that 58% of main jobs are informal. Second, the level of informality in Colombia is very close to the average for the whole Latin America. We extract the information on formal and shadow incomes from the household survey and estimate the model by maximum likelihood. The model replicates well the empirical sorting of workers between the formal sector and the informal sector.

In the first quantitative exercise we compare the optimal tax schedule with the tax schedules chosen when various informality responses are ignored. Importantly, in this comparison we allow for the endogenous adjustment of the income distribution. We find that the possibility of workers to migrate to entirely informal employment restricts tax rates at low and medium income levels, while the possibility of moonlighting is relevant at higher levels of income. Specifically, if all informality responses are ignored, the marginal tax rates are overshot at all income levels and in particular at the bottom, where they are too high by 70 percentage points or more. As a result, the shadow economy doubles in size relative to the optimum, which has catastrophic welfare consequences. If instead it is acknowledged that workers can move to the shadow economy and only the moonlighting responses are ignored, the tax rates at the bottom are approximately optimal, but the rates above the median formal income are too high — by up to 20 percentage points — when preferences for redistribution are strong. That is because incentives for moonlighting are important higher in the income distribution compared to incentives for switching from entirely formal to entirely informal employment. When preferences for redistribution are strong, ignoring moonlighting responses substantially increases the incidence of moonlighting among the most productive workers. Thus, it leads to a large welfare loss, equivalent to 2.4% drop in consumption.

In the second quantitative exercise we compare the optimal tax schedule from our estimated model with the *Mirrleesian* schedule, defined as a tax schedule which is optimal in the otherwise identical economy where the informal sector does not exist. While at low income levels the Mirrleesian tax rates increase rapidly, the optimal rates are relatively constant. Consequently, the optimal tax rates are strictly lower than Mirrleesian rates over most of the income distribution. We find that the shadow economy increases the social welfare by 1% of consumption when the preferences for redistribution are weak and reduces the social welfare by 2.7% of consumption when the preferences for redistribution are strong. To understand these results, we propose a novel decomposition of the social welfare change into efficiency and redistributive impacts. We find that the informal sector in Colombia has a positive efficiency impact by providing less productive workers with relatively high shadow productivity and by reducing marginal tax rates in the formal sector. On the other hand, lower marginal tax rates reduce income redistribution, which generates a negative redistributive impact. The former effect is dominant when the social welfare function places a high weight on equality. Our results point out that even if the informal sector could be shut down at no cost, such policy would bring welfare gains only if the government had a sufficiently strong preference for redistribution.

Related literature. The most related paper is on income taxation with tax avoidance by Kopczuk (2001). First, he shows that the standard sufficient statistics formula for the optimal linear tax is still valid. In contrast, our results imply that the standard formula for the optimal *non-linear* tax is no longer valid in the presence of a shadow economy.[4] Second, he provides an example of welfare-improving tax avoidance. According to our welfare decomposition, tax avoidance in his example has a positive redistributive impact and a negative efficiency impact. In the Online Appendix we show that this is not the only possibility: the redistributive and efficiency impacts of the opportunity to avoid, or evade, taxes could each be positive or negative depending on the model primitives. A related literature study optimal income taxation with a possibility of shifting income between two tax bases (Piketty and Saez 2013; Piketty, Saez, and Stantcheva 2014). In particular, Selin and Simula (2020) derive the optimal non-linear tax schedules in such environment, but they effectively rule out partial shifting which would correspond to moonlighting in our framework. Beaudry, Blackorby, and Szalay (2009) study redis-

[4] Our settings is not identical to Kopczuk's, since we consider a fixed cost of shadow employment. In a previous working paper version (Doligalski and Rojas 2016), we show that the standard formula for the optimal non-linear tax is not valid even if we abstract from the fixed cost of shadow employment.

tribution with informal sector when both formal income and formal hours worked are observed. We instead maintain the Mirrleesian assumption of unobserved hours worked.

Another approach to study tax evasion, originating with Allingham and Sandmo (1972), uses a framework with probabilistic audits and penalties, taking a tax rate as given. Andreoni, Erard, and Feinstein (1998) and Slemrod and Yitzhaki (2002) review this strand of literature. We take a complementary approach and study the optimal non-linear tax schedule conditional on fixed tax evasion abilities of workers. Although we do not model tax audits and penalties explicitly, they are one of the possible justifications for different productivities in the formal and the shadow sector. Under this interpretation, our results on the welfare-improving informal sector can provide insights into the optimal design of tax audits. Some early results from merging both optimal taxation and optimal tax compliance policies were derived by Cremer and Gahvari (1996), Kopczuk (2001) and Slemrod and Kopczuk (2002). Leal Ordóñez (2014) and Di Nola, Kocharkov, Scholl, and Tkhir (2020) investigate tax and enforcement policies quantitatively in the dynamic incomplete markets models.

This paper is closely related to the literature on the optimal taxation with multiple sectors. Rothschild and Scheuer (2014) consider uniform taxation of multiple sectors when agents can work in many sectors simultaneously. Kleven, Kreiner, and Saez (2009), Scheuer (2014) and Gomes, Lozachmeur, and Pavan (2017) study differential taxation of broadly understood sectors (e.g. individual tax filers and couples, employees and entrepreneurs), when agents can belong to one sector only. Jacobs (2015) studies a complementary problem when all agents work in all sectors at the same time. Our analysis differs in that we consider a particular case of differential taxation — only one sector is taxed — when agents face an idiosyncratic fixed cost of participating in one of the sectors. This structure implies that some agents can effectively work in one sector only, while others are unconstrained in supplying labor to two sectors simultaneously.

Emran and Stiglitz (2005) and Boadway and Sato (2009) study commodity taxation in the presence of informality. Both papers assume that commodity taxes affects only the formal sector.[5] Hence, provided that formal and shadow goods are perfect substitutes, a consumption tax is equivalent to a proportional tax on formal income. Under these assumptions our focus on non-linear income tax is without loss of generality. Boadway, Marchand, and Pestieau (1994) and Huang and Rios (2016) study the optimal tax mix in the opposite case, when the consumption tax cannot be evaded. A related literature on the optimal commodity taxation with home production (Kleven, Richter, and Sørensen 2000; Olovsson 2015) studies the case of non-perfect substitutability between market and home produced goods.

[5] In principle, VAT taxation covers informal firms indirectly if they purchase intermediate goods from the formal firms. De Paula and Scheinkman (2010) show that exactly for this reason informal firms tend to make transactions with other informal firms. Bachas, Gadenne, and Jensen (2020) discuss more evidence that informal enterprises do not remit consumption taxes.

Structure of the paper. In the following section we introduce the framework and characterize the incentive-compatible allocations. In Section 3 we derive the optimal tax formula and show that the informal sector reduces the optimal tax rates. Section Section 4 is devoted to the quantitative exploration of our theoretical results. The last section provides conclusions. Most proofs are relegated to the appendix.

2. Framework

There is a continuum of agents with heterogeneous labor productivities. Each agent can work in the formal sector (formal economy), in the informal sector (shadow economy), or in both simultaneously. The fundamental difference between the two sectors is that formal earnings are observed by the tax authority and can be used to determine individual income tax payments, while informal earnings are hidden and cannot be used to determine taxes. In addition, individual labor productivity can differ between the sectors and participation in the informal sector is subject to a fixed cost, which we describe below. The possibility of simultaneous work in the two sectors allows us to capture *moonlighting*, which happens when a worker with a formal job has a secondary informal job.

Individuals are heterogeneous with respect to two privately observed characteristics: a productivity type θ and a cost type κ. The productivity type θ determines the labor productivity in the formal economy $w^f(\theta)$ and in the shadow economy $w^s(\theta)$. Earnings from each sector are a product of the sectoral productivity and the labor supplied to that sector. We assume that both productivity functions are non-negative and continuously differentiable with respect to θ and that the formal productivity is strictly increasing. θ is drawn from $[\underline{\theta}, \bar{\theta}], \bar{\theta} \leq \infty$, according to a cumulative distribution function $F(\theta)$ and a density $f(\theta)$.

The cost type κ is a fixed cost of engaging in informal employment. It can be interpreted either as a technological constraint on tax evasion or a utility cost of violating social norms.[6] The idiosyncratic fixed cost allows two agents of the same formal productivity to have different shadow employment opportunities, which is an important feature of the data.[7] Conditional on θ, the fixed cost is drawn from $[0, \infty)$ according to a cumulative

[6] In principle, we could introduce a fixed cost of formal employment as well. This would correspond to what Magnac (1991) calls a segmentation approach to informal labor markets, according to which shadow workers are restricted from formal employment by various labor regulations. An alternative, competitive approach is that individuals sort between the two sectors according to their individual advantage, which corresponds more closely to our framework. Magnac (1991) shows that the data on married women in Colombia favor the latter, competitive approach. It has been documented that informality is not driven by entry costs to the formal sector also in other setting, e.g. in Argentina (Pratap and Quintin 2006), Brazil (Rocha et al. 2018) and Sri Lanka (De Mel, McKenzie, and Woodruff 2013). Furthermore, Pratap and Quintin (2006) show that for the workers with low productivity the informal wage is larger than the formal wage, which could be explained by the entry cost in the informal sector.

[7] In Section 4 we show that observable individual characteristics alone are not sufficient to explain empirical informality patterns (see the second panel of Figure 5).

distribution function $G_\theta(\kappa)$ and a density $g_\theta(\kappa)$. For the model without the fixed cost of shadow employment, see the earlier working paper version (Doligalski and Rojas 2016).

The agents' utility over consumption c and labor n, net of the fixed cost of shadow employment, is $c - v(n)$, where v is increasing, strictly convex, twice differentiable and satisfies $v'(0) = 0$. This quasi-linear preference structure, which follows Diamond (1998), does not prevent us from studying income redistribution: we characterize the entire Pareto frontier which is invariant to any increasing transformation of the utility function. Hence, our results are applicable also with utility functions $\mathbb{G}(c - v(n))$, where $\mathbb{G}$ is a strictly increasing and concave function. Nevertheless, this approach rules out the income effect. The impact of the income effect on the optimal tax schedules is well understood since Saez (2001) and the analysis can be easily extended in this direction.

To solve the model with a continuum of types, we impose the *Spence-Mirrlees* single crossing condition. This property ensures that formal income is (weakly) increasing in productivity type θ even if agents are working informally.

Lemma 1. *Agents' preferences satisfy a strict Spence-Mirrlees single crossing condition if and only if $w^s(\theta)/w^f(\theta)$ is strictly decreasing with θ or $w^s(\theta) = 0$ for all θ.*

The single crossing requires that the comparative advantage in shadow labor is decreasing with formal productivity. This assumption is maintained throughout the paper.[8] In Section 4 we verify that it holds in the data for Colombia.

2.1. Incentive-compatible allocations

Suppose that agents face a non-linear tax schedule T. The indirect utility an agent (θ, κ) derives from formal earnings y, given the optimal choice of shadow earnings, is

$$V(y, T, \theta, \kappa) \equiv \max_{y^s \geq 0} \left\{ y + y^s - T(y) - v\left(\frac{y}{w^f(\theta)} + \frac{y^s}{w^s(\theta)}\right) - \kappa \mathbb{1}_{y^s > 0} \right\} \tag{1}$$

where $\frac{y}{w^f(\theta)}$ and $\frac{y^s}{w^s(\theta)}$ correspond to the labor supplied to the formal and the informal sector, respectively. An *allocation* consists of an assignment of formal income to all types $y^f : [\underline{\theta}, \bar{\theta}] \times [0, \infty) \to \mathbb{R}_+$ and a tax schedule $T : \mathbb{R}_+ \to \mathbb{R}$.[9] An allocation (y^f, T) is *incentive-compatible* if given the tax schedule T the assignment of formal earnings y^f maximizes each agent's utility, i.e. if for all θ and κ

$$V(y^f(\theta, \kappa), T, \theta, \kappa) \geq V(y', T, \theta, \kappa) \text{ for all } y' \in \mathbb{R}_+. \tag{2}$$

[8] If the single-crossing condition did not hold, the local incentive-compatibility constraints would not be sufficient to prevent deviations within the cost classes, which would immensely complicate the analysis.

[9] Without loss of generality we focus on tax schedules with prohibitively high values at income levels which do not belong to the image of $y^f(\cdot, \cdot)$. It rules out deviations to formal income levels which are not earned by any agent.

We can characterize incentive-compatible allocations by focusing on two classes of agents: *low-cost workers*, defined as those with a zero fixed cost of shadow employment ($\kappa = 0$) and *high-cost workers*, defined by a prohibitively high fixed cost (denoted by $\kappa = \infty$). We will describe incentive-compatible formal income schedules of these agents shortly. For now, take as given the formal income schedule of the low-cost workers $y^f(\cdot, 0)$ and of the high-cost workers $y^f(\cdot, \infty)$ and suppose that they are incentive-compatible given a tax schedule T. Denote the informal earnings of the low-cost workers, implicit in the definition of their indirect utility, by $y^s(\cdot, 0)$.

Define a *formality threshold* $\tilde{\kappa}(\theta) = V(y^f(\theta, 0), T, \theta, 0) - V(y^f(\theta, \infty), T, \theta, \infty)$, where, for brevity, we suppress the dependence of the threshold on the allocation. This threshold is positive when the low-cost workers have some informal earnings and obtain a strictly higher utility than the high-cost workers of the same productivity type. Take a worker of an arbitrary type (θ, κ). Depending on whether the cost type κ is above (resp. below) the formality threshold $\tilde{\kappa}(\theta)$, this agent chooses earnings like a high-cost (resp. a low-cost) worker of the same productivity type:

$$\left(y^f(\theta, \kappa), y^s(\theta, \kappa)\right) = \begin{cases} \left(y^f(\theta, \infty), 0\right) & \text{if } \kappa \geq \tilde{\kappa}(\theta) \\ \left(y^f(\theta, 0), y^s(\theta, 0)\right) & \text{otherwise.} \end{cases} \tag{3}$$

The agents of type (θ, κ) where κ is above the formality threshold $\tilde{\kappa}(\theta)$ work only formally. The agents with a cost below the threshold supply some shadow labor: they can be either moonlighting or working exclusively informally.

We have described an incentive-compatible assignment of formal income to all agents conditional on the formal income schedules of the low-cost and the high-cost workers. Now we will characterize the income choices of these two classes of agents. Without loss of generality we focus on formal income schedules which are right-continuous — it means that agents indifferent between two formal earnings levels choose the higher one. In the typical optimal taxation or screening model it is enough to restrict attention to *local incentive-compatibility*, making sure that no agent has incentives to misreport his productivity type marginally (see e.g. Fudenberg and Tirole 1991).[10] Whereas the local incentive-compatibility constraints are not sufficient for the global incentive-compatibility in our setting, as we will demonstrate soon, they are still very relevant. To be precise, they are sufficient for almost all productivity types.[11]

We proceed to derive the local incentive-compatibility constraints. When the alloca-

[10] The local incentive-compatibility imposes two requirements. First, the indirect utility $V(y^f(\theta, \kappa), T, \theta, \kappa)$ is continuous with respect to the productivity type θ. Second, when the income schedule $y^f(\cdot, \kappa), \kappa \in \{0, \infty\}$, is differentiable, which happens almost everywhere, the allocation satisfies $\frac{d}{d\theta'} V\left(y^f(\theta', \kappa), T, \theta, \kappa\right)\Big|_{\theta'=\theta} = 0$. This last condition can be expressed as the first-order conditions in the main text.

[11] In Proposition 3 we show that local incentive constraints are not sufficient for types (θ, ∞), $(\theta, 0)$, and at any point of discontinuity of $y^f(\cdot, 0)$. Since $y^f(\cdot, 0)$ is increasing, there are at most countably many discontinuity points.

tion is locally differentiable, these constraints can be expressed as intuitive first-order conditions. The first-order condition of the high-cost worker with productivity type θ is

$$\left(1 - T'\left(y^f(\theta, \infty)\right)\right) w^f(\theta) = v'\left(\frac{y^f(\theta, \infty)}{w^f(\theta)}\right), \tag{4}$$

which means that the marginal return to formal labor — the product of the formal productivity and the net-of-tax rate — is equal to the marginal disutility from labor.

Let's focus on the low cost worker with productivity type θ and consider three cases. First, if a worker does not work in the shadow economy, he will choose the same earnings as his high-cost peer:

$$\left(1 - T'\left(y^f(\theta, 0)\right)\right) w^f(\theta) = v'\left(\frac{y^f(\theta, 0)}{w^f(\theta)}\right) > w^s(\theta). \tag{5}$$

Second, if the worker supplies labor only to the informal sector, his first-order condition equalizes the marginal return to shadow labor — the shadow productivity — and the marginal disutility from labor:

$$\left(1 - T'\left(y^f(\theta, 0)\right)\right) w^f(\theta) < v'\left(\frac{y^s(\theta, 0)}{w^s(\theta)}\right) = w^s(\theta), \tag{6}$$

Finally, suppose that the low-cost worker is moonlighting, i.e. supply labor to both sectors. His first-order condition equalizes the the marginal returns to formal and shadow labor with the marginal disutility from labor:

$$\left(1 - T'\left(y^f(\theta, 0)\right)\right) w^f(\theta) = v'\left(\frac{y^f(\theta, 0)}{w^f(\theta)} + \frac{y^s(\theta, 0)}{w^s(\theta)}\right) = w^s(\theta). \tag{7}$$

The above first-order condition has two important implications. First, the total labor supply of a moonlighting worker is determined by his shadow productivity and, hence, cannot be affected by taxes. What taxes affect is only the sectoral split of labor.

Second, moonlighting is closely related to tax progressivity. We can rearrange (7) as $T'\left(y^f(\theta, 0)\right) = 1 - \frac{w^s(\theta)}{w^f(\theta)}$, where the right-hand side is strictly increasing with θ by the single-crossing condition. Thus, the marginal tax rates faced by moonlighting workers are strictly increasing with their productivity type. The proposition below explores the implications of this result. We show that moonlighting happens only at the income levels where the tax is progressive, i.e. features increasing marginal tax rates, and a tax schedule which is regressive everywhere admits no moonlighting.

Proposition 1. *Suppose $T \in \mathcal{C}^2$.*

1. *If the tax schedule is strictly regressive locally at some $y > 0$, i.e. $T''(y) < 0$, then there is no moonlighting worker with formal earnings y.*

2. *If the tax schedule is regressive everywhere, i.e. $T''(y) \leq 0$ for all $y \in \mathbb{R}_+$, then no*

worker is moonlighting.

The intuition is simple. A worker will be moonlighting if the marginal benefit to supplying the formal labor relative to the informal labor is decreasing. Then the worker supplies formal labor at first, but as the marginal benefit decreases, he switches to the informal labor. That is exactly what happens when the tax schedule is progressive: low marginal tax rates at low income levels encourage formal labor at first, but high tax rates at higher level discourage it.

We can obtain stronger results under the following parametric assumptions: the productivity schedules are log-linear in θ, which we also assume in our quantitative model, and the tax function features a constant rate of progressivity p as in Feldstein (1969) and Benabou (2000). Then tax progressivity needs to be sufficiently strong to admit moonlighting.

Proposition 2. *Suppose that (i) productivity schedules follow $w^i(\theta) = w^i(0)e^{\rho^i\theta}$, $w^i(0) > 0$ for $i \in \{f, s\}$ where $\rho^f > \rho^s > 0$, (ii) the productivity type is unbounded: $\theta \in \mathbb{R}_+$, (iii) the disutility from labor features a constant Frisch elasticity ε: $v(n) = \frac{1+\varepsilon}{\varepsilon}n^{1+\frac{1}{\varepsilon}}$, (iv) the tax schedule satisfies $T(y) = y - \frac{1-\tau}{1-p}y^{1-p}$ where $1 \geq p > 0$.*

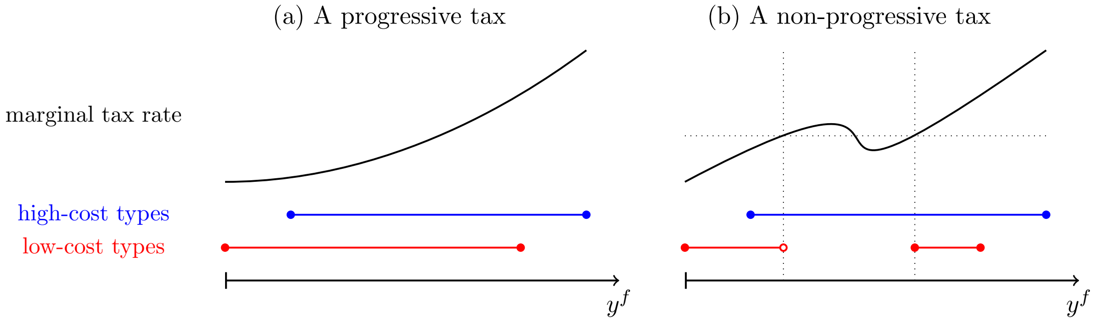
No worker is moonlighting if and only if

$$p \leq \frac{\rho^f - \rho^s}{\rho^f + \varepsilon \cdot \rho^s} \quad \text{and} \quad w^f(0)^{1-p} \geq \frac{w^s(0)^{1+p\varepsilon}}{1-\tau}. \tag{8}$$

According to Proposition 2, there is no moonlighting when two conditions in (8) are satisfied. The right inequality ensures that the bottom type has no incentives to moonlight. The more interesting left inequality ensures that incentives for moonlighting are not increasing with θ. This condition places an upper bound on the rate of progressivity which is consistent with no moonlighting. The upper bound is positive and increasing in the difference of growth rates of formal and shadow productivity $\rho^f - \rho^s$. When this difference is greater, high productivity types are comparatively worse at informal labor, which weakens their incentives for moonlighting. The upper bound is also decreasing with the Frisch elasticity of labor supply ε. Keeping the tax schedule fixed, higher elasticity means that very productive formal workers will choose higher earnings. Hence, they will be more exposed to increasing tax rates and have stronger incentives for moonlighting.

What happens with moonlighting when the tax schedule is neither progressive nor regressive everywhere, but features regions of local progressivity and regressivity? Empirical income tax and transfer schedules, which typically have increasing statutory income tax rates, often become locally regressive at the earnings level where transfers are phased-out. By Proposition 1, no moonlighting worker will be found in the regions of local regressivity. If such regions are surrounded by regions of local progressivity, then the formal income schedule of the moonlighting, low-cost workers can become discontinuous, as depicted in Figure 1.

Figure 1: Tax progressivity and continuity of $y^f(\cdot, 0)$


(a) A progressive tax

(b) A non-progressive tax

Note: The horizontal lines indicate whether there are workers of a given type at a given formal income level.

Since local tax regressivity may lead to a discontinuity in the formal income schedule of the low-cost workers, we need to include a local incentive-compatibility constraint for this case. Suppose that $y^f(\cdot, 0)$ increases discontinuously at θ_d. The local incentive-compatibility constraint ensures that the low-cost agent of type θ_d is indifferent between the two discontinuously different income levels. It is easy to show that then not only the marginal but also the *average* returns to formal and shadow labor coincide:

$$\left(1 - \frac{T(y^f(\theta_d, \kappa)) - T(y^f(\theta_d^-, \kappa))}{y^f(\theta_d, \kappa) - y^f(\theta_d^-, \kappa)}\right) w^f(\theta_d) = w^s(\theta_d). \tag{9}$$

Equations (4) to (9) constitute the local incentive-compatibility constraints. In the one-dimensional taxation or screening model the local incentive-compatibility constraints together with income monotonicity requirement are sufficient for the global incentive compatibility. This result has been extended to some environments with multidimensional heterogeneity.[12] Yet, it does not apply in our setting. Specifically, there exist formal income schedules of the low-cost and the high-cost workers that are increasing and satisfy the local incentive-compatibility constraints and yet violate some of the *global* incentive constraints.

For intuition, notice that the local incentive constraints prevent deviations to formal income levels earned by other agents from the same cost class (the class of the high-cost or the low-cost workers). The only deviations for which local incentive constraints can be insufficient are to the formal income levels earned by agents from the other cost class. Further note that if the formal income distributions of the high-cost and the low-

[12] Kleven et al. (2009), Scheuer (2014) and Gomes et al. (2017) recover sufficiency of local incentive constraints in two-dimensional settings with two sectors, under the assumption that agents can work in one sector at a time. Rothschild and Scheuer (2014) allow workers to supply labor in multiple sectors simultaneously, but the government observes and taxes the sum of all incomes, which implies that the local incentive constraints are sufficient.

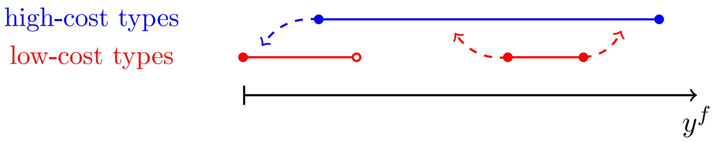
cost workers were perfectly overlapping, then the local constraints would be sufficient. Since these distributions are not necessarily perfectly overlapping, we need to include additional, non-local incentive constraints to cover formal income intervals earned by only one class of workers.

Proposition 3. *An allocation (y^f, T) is incentive-compatible if, and only if,*

1. *$y^f(\cdot, \infty)$ and $y^f(\cdot, 0)$ are increasing and satisfy local incentive-compatibility constraints.*
2. *$y^f(\theta, \cdot)$ is consistent with the formality threshold and satisfies (3) for all θ.*
3. *The type $(\underline{\theta}, \infty)$ cannot gain by deviating to any lower formal income.*
4. *The type $(\bar{\theta}, 0)$ cannot gain by deviating to any higher formal income.*
5. *If $y^f(\cdot, 0)$ is discontinuous at θ_d, then the type $(\theta_d, 0)$ cannot gain by deviating to any formal income from the interval $(y^f(\theta_d^-, 0), y^f(\theta_d, 0))$.*

In the proof of Proposition 3 we identify all cases in which formal income distributions of the high-cost and the low-cost workers can be non-overlapping, see Figure 2 for graphical depiction. First, notice that the formal earnings of the low-cost worker of type θ are weakly lower than that of the high-cost worker of type θ. Consequently, income distributions can be non-overlapping at the bottom, if there are only low-cost workers, or at the top, if there are only high-cost workers. Additional incentive constraints 3 and 4 address these cases. Second, income distributions can be non-overlapping if one of the income schedules is discontinuous where the other is not. We show that only discontinuities in the schedule of the low-cost types can lead to non-overlapping distributions. We capture these with the incentive constraint 5.

Figure 2: Deviations not prevented by the local incentive-compatibility constraints



Note: The horizontal lines indicate whether there are workers of a given type at a given formal income level. The arrows represent the deviations not prevented by the local incentive-compatibility constraints.

Whereas multidimensional screening problems where the local incentive-compatibility constraints are insufficient are notorious for intractability (Rochet and Choné 1998), that is not the case with our model. The non-local incentive-compatibility constraint 3 can be verified ex post and — if violated — it can be incorporated directly into the planner's problem, in the manner analogue to the monotonicity requirement in the standard one-dimensional problem. The non-local constraints 4 and 5 restrict the set of incentive-compatible allocations, but do not affect the fiscal or welfare impact of a small perturbation of the tax schedule, starting from the incentive-compatible allocation.[13] Consequently, they do not affect the optimal tax formulas. Nevertheless, these constraints are important to account for in the quantitative application of the model.

2.2. The planner's problem

The social planner maximizes the average of individual utilities weighted with Pareto weights $\lambda(\theta,\kappa)$. We normalize the weights such that $\mathbb{E}\{\lambda(\theta,\kappa)\}=1$ which implies that the Pareto weights and the marginal social welfare weights coincide.[14] The planner solves

$$\max_{\substack{y^f\,:\,[\underline{\theta},\bar{\theta}]\times[0,\infty)\to\mathbb{R}_+\\ T\,:\,\mathbb{R}_+\to\mathbb{R}}} \int_{\underline{\theta}}^{\bar{\theta}}\int_0^{\infty}\lambda(\theta,\kappa)V(y^f(\theta,\kappa),T,\theta,\kappa)dG_\theta(\kappa)dF(\theta) \tag{10}$$

subject to the incentive-compatibility constraints from Proposition 3 and the budget constraint

$$\int_{\underline{\theta}}^{\bar{\theta}}\int_0^{\infty}T(y^f(\theta,\kappa))dG_\theta(\kappa)dF(\theta)\geq E, \tag{11}$$

where E stands for exogenous government expenditures. By solving the planner's problem for arbitrary Pareto weights, we recover the entire Pareto frontier of the model without income effects.[15]

We proceed with the theoretical analysis under the standard assumption that the monotonicity constraints on formal income schedules are not binding, which means that there is no bunching along the productivity dimension alone. We rule out this bunching pattern because it is well understood from the one-dimensional models (Mussa and Rosen 1978; Ebert 1992) and it happens rarely.[16] Crucially, we allow for all other bunching

[13] In Section 3 we show that the fiscal cost of intensive margin responses of the moonlighting workers is independent of the magnitude of formal income adjustment. Namely, it does not matter for the tax revenue whether these workers adjust formal income marginally or jump to a discreetly lower formal income level. Given that we do not need to keep track of the exact formal income responses of the moonlighting workers, we can derive the optimal tax formulas without explicitly accounting for constraints 4 and 5.

[14] The marginal social welfare weights describe the welfare impact of marginally increasing consumption of a given type of agents, expressed in the units of tax revenue (see e.g. Piketty and Saez 2013). In our environment they are equal to $\lambda(\theta,\kappa)/\eta$, where η is the multiplier of the budget constraint. It is easy to show that at the optimum $\eta=\mathbb{E}\{\lambda(\theta,\kappa)\}$.

[15] Suppose that the social welfare function is $\int_{\underline{\theta}}^{\bar{\theta}}\int_0^{\infty}\mathbb{G}\left(V(y^f(\theta,\kappa),T,\theta,\kappa)\right)dG_\theta(\kappa)dF(\theta)$, where $\mathbb{G}$ is an increasing and differentiable function. $\mathbb{G}$ is typically assumed to be strictly concave and it can represent either the decreasing marginal utility of consumption or the planner's taste for equality. In this case we find the optimal allocation iteratively. Start with an initial guess of the Pareto weights. In each step, find the optimum given the Pareto weights and set the new Pareto weights — to be used in the next step — according to $\lambda(\theta,\kappa)=\mathbb{G}'\left(V(y^f(\theta,\kappa),T,\theta,\kappa)\right)$, where the indirect utility function V is evaluated at the optimum found in this step.

[16] This type of bunching is more important in the setting without the fixed cost of shadow employment and we study it in detail in the earlier working paper version (Doligalski and Rojas 2016).

patterns. In particular, we allow for the bunching of agents with simultaneously different cost and productivity types, which happens when there are formal and moonlighting workers with the same formal earnings. We also assume that the non-local incentive-compatibility constraint 3 from Proposition 3 is not binding. We verify ex post that both assumptions are true in all our quantitative exercises.

3. Optimal tax formula

In this section we derive the optimal tax formula. We use the perturbation approach of Saez (2001), i.e. we consider a small variation of the marginal tax rate at some formal income level. In Online Appendix A we derive the tax formulas using the mechanism design approach and provide the exact correspondence between the two approaches.

From now on we will focus on the endogenous distribution of formal income. Denote the density of formal income by $h(\cdot)$. We can decompose it into the density of formal workers $h^f(\cdot)$ and the density of moonlighting workers $h^s(\cdot)$, such that at each income level $y > 0$ we have $h(y) = h^f(y) + h^s(y)$.[17] The distribution of formal income may involve a mass point at 0 due to workers engaged in the exclusively informal employment. Take some incentive-compatible allocation with twice-differentiable tax schedule T and perturb the marginal tax rate in the formal income interval $[y, y+dy]$ by a small $d\tau > 0$. This perturbation influences tax revenue via: *(i)* intensive margin responses of formal workers, *(ii)* intensive margin responses of moonlighting workers, *(iii)* extensive margin responses due to workers changing their informality status, *(iv)* mechanical and welfare effects. We describe these effects in turn below.

Intensive margin responses of formal workers. In response to the increase in the marginal tax rate, the formal workers with income y or slightly higher will reduce their formal earnings. The income reduction is standard and equal approximately to

$$h^f(y)\tilde{\varepsilon}^f(y)y\frac{d\tau dy}{1-T'(y)}, \quad \text{where } \tilde{\varepsilon}^f(y) = \left(\frac{1}{\varepsilon(y)} + \frac{T''(y)y}{1-T'(y)}\right)^{-1}. \tag{12}$$

$\tilde{\varepsilon}^f(y)$ is the elasticity of formal income of formal workers with respect to the marginal tax rate along the non-linear tax schedule. It depends both on $\varepsilon(y)$, the elasticity along the *linear* tax schedule, or the Frisch elasticity, and the local tax curvature. With a locally progressive tax ($T''(y) > 0$), an income increase in response to a tax rate cut is reduced, as a higher income leads to a higher tax rate. Hence, the local progressivity (resp. regressivity) of the tax schedule reduces (resp. increases) the elasticity of income.

[17] Formally, $h^f(\cdot)$ and $h^s(\cdot)$ are not densities as they do not integrate to 1, but rather to shares of formal and moonlighting workers in total employment, respectively. Keeping this slight abuse of terminology in mind, we will continue calling them densities.

Intensive margin responses of moonlighting workers. Suppose that there are some moonlighting workers with formal income y. The reduction of formal income of moonlighting workers is equal to

$$h^s(y)\tilde{\varepsilon}^s(y)y\frac{d\tau dy}{1-T'(y)}, \quad \text{where } \tilde{\varepsilon}^s(y) \equiv \frac{1-T'(y)}{T''(y)y} > \tilde{\varepsilon}^f(y). \tag{13}$$

The derivation of this elasticity is relegated to the proof of Theorem 1. The elasticity of the formal income of moonlighting workers is strictly greater than that of exclusively formal workers: $\tilde{\varepsilon}^s(y) > \tilde{\varepsilon}^f(y)$.[18] The intuition for this results is tightly related to the first-order conditions (4) and (7). An increase of the tax rate reduces the marginal benefit from supplying formal labor for formal and moonlighting workers in a symmetric manner. Both formal and moonlighting workers will reduce formal labor supply until the marginal benefit increases up to the level of the marginal cost. The difference between them is in the determination of the marginal cost of formal labor. For the formal worker the marginal cost is the marginal disutility of labor $v'(\cdot)$, which decreases as the total labor supply is reduced. For the moonlighting worker, however, the total labor supply is fixed and the tax perturbation affects only the sectoral split of labor. The marginal cost for these workers is the forgone informal income, which is equal to the shadow productivity $w^s(\theta)$. Given that the marginal cost of the moonlighting workers is constant in formal labor, rather than decreasing as in the case of the formal workers, they will adjust formal labor more than formal workers.

The formal income schedule of the moonlighting workers can become discontinuous when the tax schedule is not progressive everywhere. Consequently, we need to examine the case in which there are no moonlighting workers at formal earnings level y where we perturb the tax rate, but there are some with strictly higher and strictly lower formal earnings. Denote by $s(y)$ the formal income level at which moonlighting workers respond on the intensive margin to the perturbation of tax rate $T'(y)$.[19] Suppose that $s(y)$ is strictly greater than y. By the local incentive-compatibility constraint (9) the moonlighting worker with formal earnings $s(y)$ is indifferent between $s(y)$ and some lower formal earnings level, which we will denote by $s(y) - \Delta y$. Consider an increase in $T'(y)$. As the tax burden at $s(y)$ increases, the agent strictly prefers $s(y) - \Delta y$ to $s(y)$ and jumps to the lower level of formal earnings.

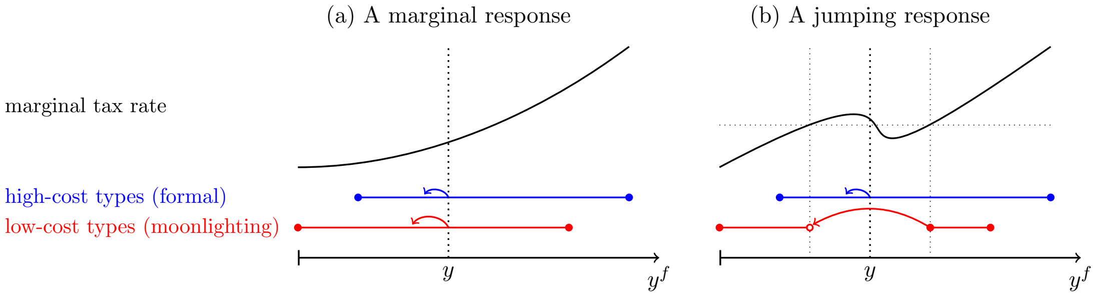
Figure 3 illustrates the two types of formal income responses of moonlighting workers. On the left panel, the tax schedule is locally strictly progressive and the formal income schedule of the moonlighting workers is locally continuous at y. Consequently, the moonlighting workers respond to an increase of $T'(y)$ by marginally reducing their formal income. On the right panel, the tax schedule is non-progressive and the formal income schedule of the moonlighting workers is discontinuous. In the response to an

[18] We do not need to consider a locally regressive tax ($T''(y) < 0$), since by Proposition 1 there will be no moonlighting workers with such formal earnings.

[19] One can show that $s(y) = \min_\theta\{y^f(\theta, 0) \text{ s.t. } y^f(\theta, 0) \geq y\}$.

increase in $T'(y)$ the moonlighting workers discretely jump to a lower formal income level.

Figure 3: Intensive margin responses of moonlighting workers


(a) A marginal response

(b) A jumping response

Note: The horizontal lines indicate whether there are workers of a given type at a given formal income level. The arrows represent the formal income responses to an increase of the marginal tax rate at formal earnings y.

Conveniently, the tax revenue impact of these jumping responses can still be described with the intensive margin elasticity $\tilde{\varepsilon}^s(\cdot)$. To see this, note that the perturbation increases the tax burden at $s(y)$ by $d\tau dy$ and makes some moonlighting workers discretely decrease their formal income from $s(y)$ to $s(y) - \Delta y$. The measure of workers that decides to jump is given by $h^s(s(y))ds(y)$. By differentiating (9) we obtain $ds(y) = [T''(s(y))\Delta y]^{-1} d\tau dy$.[20] Therefore, the overall income reduction is exactly as in the case when shadow workers adjust income marginally:

$$\Delta y h^s(s(y))ds(y) = h^s(s(y))\tilde{\varepsilon}^s(s(y))s(y)\frac{d\tau dy}{1 - T'(s(y))}. \tag{14}$$

The intuition is that, although each jumping individual makes a discrete income reduction Δy, the measure of jumping individuals is inversely proportional to the size of the income reduction. As a result, the overall income reduction is independent of Δy and such that the elasticity at $s(y)$ is finite and equal $\varepsilon^s(s(y))$.

Tax revenue impact of intensive margin responses. When the formal income adjustment is marginal, the tax revenue impact of the intensive margin responses is given by the product of the income adjustment and the marginal tax rate. However, as discussed above, some of the moonlighting workers may respond by jumping to discretely lower formal earnings. Conveniently, it follows from the first-order conditions (7) and (9) that the average tax rate between the two income levels is equal to their marginal tax rate. As a result, the tax revenue impact of the jumping responses is exactly the same as if these workers adjusted income marginally.

[20] Denote $s(y) - \Delta y$ as $\tilde{y}$. We can rewrite (9) as $(s(y) - \tilde{y})T'(s(y)) - (T(s(y)) - T(\tilde{y})) = 0$. Perturb the tax level at $s(y)$ by $d\tau dy$. By totally differentiating this equation we find that $\Delta y T''(s(y))ds(y) - d\tau dy = 0$.

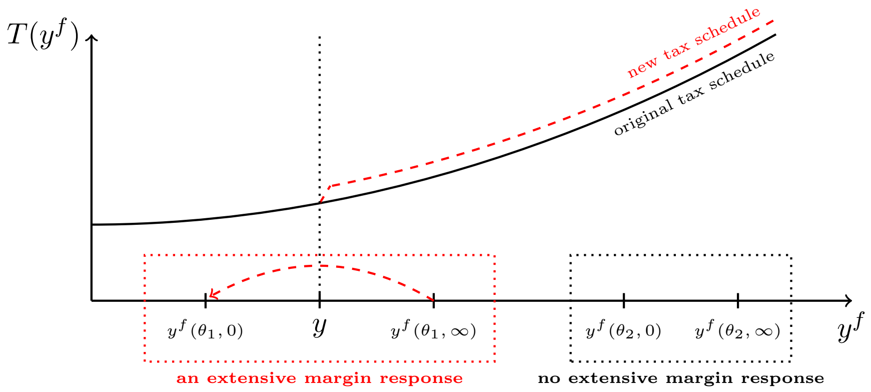
Therefore, when there are some low-cost workers with formal income equal to or higher than y, we can express the tax revenue impact from the intensive margin responses of formal and moonlighting workers, no matter whether they are responding marginally or jumping, as

$$-\left(\frac{T'(y)}{1-T'(y)}h^f(y)\tilde{\varepsilon}^f(y)y+\frac{T'(s(y))}{1-T'(s(y))}h^s(s(y))\tilde{\varepsilon}^s(s(y))s(y)\right)d\tau dy. \tag{15}$$

In the remaining case, when there are no low-cost workers at or above y, the second term in the bracket is set to zero.

Extensive margin responses. The extensive margin responses in our model consist of a switch from working exclusively formally to moonlighting or working exclusively informally. The possibility of moonlighting means that the extensive margin responses are not equivalent to responses on the formal participation margin. In particular, a worker who switches from exclusively formal employment to moonlighting continues to work in the formal sector and retains a fraction of his formal earnings. It has important implications for the incidence of the extensive margin responses following a tax reform, as depicted in Figure 4. The perturbation of $T'(y)$ increases the tax burden for workers with incomes above y. Consequently, it increases incentives for informality for agents who, conditional on working informally, would earn less than y in the formal sector. On the other hand, incentives for informality are unaffected for formal workers who, even if they moonlighted, would have formal earnings above y — they would need to pay a higher tax either way.

Figure 4: The incidence of extensive margin responses



Note: An increase of marginal tax rate at formal earnings y triggers an extensive margin response of a formal worker with productivity type θ_1, but does not trigger an extensive margin response of a formal worker with productivity type θ_2.

To capture the incidence of the extensive margin responses and their impact on the tax revenue, let's define the formal income gap between the high-cost and the low-cost workers in two ways. $\Delta_\infty(y')$ tells us by how much a formal worker with income y' would decrease his formal income if he had a lower realization of the fixed cost and worked informally. $\Delta_0(y')$ tells us by how much the moonlighting worker with formal income y' would increase his formal earnings if he had a higher realization of the fixed cost and did not work in the shadow economy.

Suppose that there are some low-cost workers with formal income above y. The perturbation of $T'(y)$ increases incentives for informality for formal workers in the income interval $(y, s(y+dy)+\Delta_0(s(y+dy)))$. Workers with income below y are unaffected, since their tax schedule is unchanged. Workers with income above $s(y+dy)+\Delta_0(s(y+dy))$ pay taxes higher by $d\tau dy$ no matter whether they stay formal or start to moonlight, so their incentives for informality are unchanged as well. In the following derivations we focus on a subinterval $[y+dy, s(y)+\Delta_0(s(y))]$, since the terms corresponding to the remaining parts of the original interval are of second order (i.e. proportional to dy^2) and vanish as we consider an arbitrarily small dy.

Define the *tax burden of staying formal* as $\Delta T(y') = T(y') - T(y' - \Delta_\infty(y'))$. It captures the tax revenue loss when a formal worker with income y' starts supplying informal labor. Furthermore, denote the impact of the tax perturbation on the density of formal workers at income y' by $dh^f(y')$. The tax revenue impact of the perturbation due to extensive margin responses is

$$\int_{y+dy}^{s(y)+\Delta_0(s(y))} dh^f(y')\Delta T(y')dy'd\tau dy = -\int_{y+dy}^{s(y)+\Delta_0(s(y))} \pi(y')h^f(y')dy'd\tau dy, \qquad (16)$$

where $\pi(y')$ is the elasticity of the density of formal workers at y' with respect to the tax burden of staying formal. Intuitively, the more elastic is the density of formal workers, the higher is the tax revenue loss due to increased participation in the shadow economy.

In the case when all the low-cost workers have formal incomes below y, the tax perturbation increases incentives for informality at all formal income levels above $y+dy$.

Mechanical and welfare impact. Consider the tax schedule at incomes above $y+dy$. The perturbation keeps the tax rate fixed, while increasing the tax level by $d\tau dy$. On the one hand, an increase in the tax level mechanically raises the tax revenue. On the other hand, it reduces utility of agents with higher incomes, resulting in a welfare loss. Denote the average Pareto weight at a given formal income level y' by $\bar{\lambda}(y')$. Ignoring the second-order terms, the combined mechanical and welfare impact of the perturbation is

$$\int_{y+dy}^{\infty} (1-\bar{\lambda}(y'))h(y')dy'd\tau dy. \qquad (17)$$

Optimal tax formulas. Optimality requires that no small tax perturbation can increase the welfare-adjusted tax revenue. Hence, the sum of all the impacts of the tax perturbation: (15) - (17), needs to be zero for any $d\tau$ and an arbitrary small dy.

Theorem 1. *Suppose that the bunching along the productivity dimension alone does not occur. When some low-cost workers have formal income greater than or equal to y, the optimal tax rate satisfies*

$$\frac{T'(y)}{1-T'(y)}h^f(y)\tilde{\varepsilon}^f(y)y + \frac{T'(s(y))}{1-T'(s(y))}h^s(s(y))\tilde{\varepsilon}^s(s(y))s(y)$$

$$= \int_y^{\infty}\left[1-\bar{\lambda}(y)\right]h(y)dy - \int_y^{s(y)+\Delta_0(s(y))}\pi(y')h^f(y')dy'. \quad (18)$$

When all low-cost workers have formal income below y, the optimal tax rate satisfies

$$\frac{T'(y)}{1-T'(y)}h^f(y)\tilde{\varepsilon}^f(y)y = \int_y^{\infty}\left[1-\bar{\lambda}(y')-\pi(y')\right]h(y')dy'. \quad (19)$$

The two formulas equate the costs and benefits from marginally increasing the tax rate $T'(y)$. Tax formula (18) applies at income levels such that there are some low-cost worker with higher formal earnings, otherwise the optimal tax rate is given by formula (19).

The left-hand side of formula (18) consists of the deadweight loss from distorting the formal workers and the moonlighting workers. The deadweight loss terms increase in *(i)* the marginal tax rate, as the reduction in formal income implies a higher tax loss if it is taxed at the higher rate, *(ii)* the density of formal income and *(iii)* the formal income reduction per worker in response to a higher tax rate, i.e. the product of formal income and the income elasticity. There are two important differences between the deadweight loss terms of formal and moonlighting workers. First, conditional on the local progressivity of the tax schedule, the formal income of the moonlighting workers is more elastic than the income of the formal workers. Second, unlike the distorted formal workers, the distorted moonlighting workers may have formal income that is strictly higher than y. The second tax formula (19) captures the case when no moonlighting workers are distorted by the tax rate perturbation, so only the deadweight loss of the formal workers is present.

The right-hand side of formula (18) captures the mechanical and the welfare impacts of the reform as well as the tax loss from increased participation in the shadow economy. Note that, according to formula (18), higher $T'(y)$ leads to greater incentives for informality only for workers with formal income between y and an upper bound $s(y) + \Delta_0(s(y)$. Formal workers with income higher than this upper bound do not face stronger incentives for informality, since even if they decided to work in the shadow economy their formal income would be high enough such that they would pay higher taxes anyway. In contrast, when formula (19) applies, the perturbation increases incentives for informality for all workers with formal income above y. That is because all of these workers, if they worked informally, would have formal income below the level at which the tax rate is perturbed.

The tax formulas are novel, as they incorporate moonlighting responses — complementing formal earnings with income from an informal job. As we discuss above, moonlighting modifies the incidence of extensive margin responses and can lead to intensive margin responses at higher income levels than the level at which the tax rate is perturbed. There are other settings that give rise to phenomena similar to moonlighting where such formulas could be applied, e.g. the model of home production, or the problem of a local tax authority which residents can work partly outside its jurisdiction as seasonal workers.

3.1. How does a shadow economy affect optimal tax rates?

We examine the impact of a shadow economy on the optimal tax rates in two ways, following the approach of Scheuer and Werning (2017). First, we fix the formal income distribution and other sufficient statistics and evaluate how the shadow economy alters the optimal tax schedule by switching off the components of the tax formulas which relate to informality responses. This analysis is most informative for choosing tax policy based on a given, observed formal income distribution. Second, we compare the optimal top tax rate with and without a shadow economy for given model primitives while allowing the formal income distribution to adjust. This comparison is useful for the counterfactual analysis. It informs us how the optimal top tax rate would change if we could costlessly shut down the informal sector.

3.1.1. Comparison for a fixed formal income distribution

Take as given the formal income distribution, the schedule of average Pareto weights at each formal income level, as well as all other sufficient statistics required to compute the optimal tax rates according to Theorem 1. We will compare the optimal tax formulas with two benchmark cases:

- I: the tax formula of the planner who ignores the possibility of moonlighting, but acknowledges the mobility between the formal and the informal sectors,
- II: the tax formula of the planner who ignores both the possibility of moonlighting and the mobility between the two sectors.

In case I, when the planner ignores only the moonlighting responses, the tax formula is given by

$$\frac{T_I'(y)}{1-T_I'(y)}h(y)\bar{\varepsilon}(y)y = \int_y^\infty \left[1-\bar{\lambda}(y')\right] h(y')dy' - \int_y^\infty \pi(y')\cdot \mathbb{1}_{\Delta_\infty(y')=y'} h^f(y')dy', \quad (20)$$

where $\bar{\varepsilon}(y)$ is the average formal earnings elasticity at formal income y and $\pi(y') \cdot \mathbb{1}_{\Delta_\infty(y')=y'}$ is the elasticity of formal labor market participation at formal income y' with respect to the tax level $T_I(y')$. The indicator function $\mathbb{1}_{\Delta_\infty(y')=y'}$ makes sure that only the extensive margin responses which reduce formal earnings to zero are accounted for.[21] This formula coincides with the tax formula derived by Jacquet, Lehmann, and Van der Linden (2013) in the model with intensive margin responses and endogenous participation in the labor market.

In case *II*, when the planner ignores all informality responses, the tax formula is given by

$$\frac{T'_{II}(y)}{1 - T'_{II}(y)} h(y)\bar{\varepsilon}(y)y = \int_y^\infty \left[1 - \bar{\lambda}(y')\right] h(y')dy'. \tag{21}$$

Here, the planner effectively believes in an extreme version of the segmented market hypothesis, where the allocation of workers to the formal and the informal sectors is given and policy invariant. In this view, the tax schedule affects only the labor supply of formal workers on the intensive margin. Hence, this his tax formula coincides with the formula of Diamond (1998) and Saez (2001), derived in the model with intensive margin of labor supply alone.

The following proposition compares the optimal tax rates, $T'(y)$, with the tax rates in to the two benchmark cases.

Proposition 4. *Fix the schedule of average Pareto weights $\bar{\lambda}(\cdot)$, the distribution of formal income $h^f(\cdot)$ and $h^s(\cdot)$, intensive margin elasticities $\tilde{\varepsilon}^f(\cdot)$ and $\tilde{\varepsilon}^s(\cdot)$, extensive margin elasticities $\pi(\cdot)$, formal income gaps $\Delta_\infty(\cdot)$ and $\Delta_0(\cdot)$ and the mapping $y \mapsto s(y)$. Suppose that bunching along the productivity dimension alone does not occur, $\pi(y) \geq 0$ for all y and the optimal tax rate $T'(y) \geq 0$ for all y. Then $T'(y) \leq T'_I(y) \leq T'_{II}(y)$.*

We obtain a clear ordering of marginal tax rates at each income level. The optimal tax formula prescribes the lowest rates, followed by the rates set when moonlighting is ignored, and the highest rates are chosen when all informality responses are ignored.

The intuition is simple. The optimal tax formula correctly incorporates the entire fiscal cost of raising tax rates at the given income level. In contrast, the formula which ignores moonlighting (case I) is missing the fiscal cost of some extensive margin responses — the ones when a formal worker starts to moonlight — as well as of the intensive margin responses of moonlighting workers which happen at a higher level of formal income. Hence, it prescribes tax rates which are (weakly) too high. The formula which ignores all informality responses (case II) is, in addition, missing the fiscal cost of all extensive margin responses. As a result, it prescribes tax rates which are (weakly) higher than these implied by the other two formulas.[22]

[21] It is easy to show that $y' > s(y) + \Delta_\infty(s(y))$ implies that $\Delta_\infty(y') < y'$. Therefore, the effective range over which the extensive margin responses are integrated is not larger than in the formula (18).

[22] The prescriptions of the three formulas may coincide at some income levels. It happens when all workers of a given productivity type are formal, since then the tax perturbation triggers no intensive margin responses of moonlighting workers nor any extensive margin responses.

3.1.2. Comparison for fixed primitives

Take as given model primitives: the distribution of productivity and cost types, the productivity schedules and the schedule of Pareto weights. In the following proposition we compare the optimal top tax rate from the model with a shadow economy, denoted by $T'(\infty)$, with the the optimal top rate in the model where the informal sector does not exist, denoted by $T'_M(\infty)$. Since our model without the informal sector is just the standard Mirrlees model, we call $T'_M(\infty)$ a Mirrleesian top tax rate. In contrast to the previous comparison, here we allow the for the income distribution and all the other sufficient statistics to endogenously adjust to the top tax rate. Let's first determine how the top tax rate influences the shape of the upper tail of the formal income distribution with a shadow economy.

Lemma 2. *Suppose that (i) the formal productivity distribution has a Pareto tail:* $\lim_{\theta\to\bar{\theta}} \frac{f(\theta)w^f(\theta)}{1-F(\theta)}\left(\frac{dw^f(\theta)}{d\theta}\right)^{-1} = \alpha$, *(ii) the fixed cost of shadow employment has a Pareto tail:* $\forall_\theta \lim_{\kappa\to\infty} \frac{\kappa g_\theta(\kappa)}{1-G_\theta(\kappa)} = \gamma$, *(iii) the Frisch elasticity of labor supply is* ε. *Then the tail parameter of the formal income distribution* $\alpha_y = \lim_{y\to\infty} \frac{h(y)y}{1-H(y)}$ *satisfies*

$$\alpha_y = \begin{cases} \frac{\alpha}{1+\varepsilon} & \text{if } 1-T'(\infty) \geq \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}, \\ \frac{\alpha}{1+\varepsilon}+\gamma & \text{otherwise.} \end{cases} \tag{22}$$

The tail parameter α_y describes the *thinness* of the upper tail of the formal income distribution. When the top tax rate is sufficiently low, none of the most productive types work informally and the thinness of the formal income tail is exactly the same as in the standard Mirrlees model. However, as soon as the top tax rate crosses a tipping point $1 - w^s(\bar{\theta})/w^f(\bar{\theta})$, a positive fraction of top earners starts informal employment. As a result, the thinness of the upper tail increases discretely by γ, the tail parameter of the fixed cost distribution. Intuitively, if γ is high, there are many workers with a low fixed cost of shadow employment who reduce their formal income and join the shadow economy. If instead γ is low, there are few workers with a low fixed cost of shadow employment and the formal income distribution is less affected.

The following proposition shows that the shadow economy leads to a (weakly) lower optimal top tax rate, conditional on other primitives of the economy.

Proposition 5. *Suppose that the assumptions of Lemma 2 hold and additionally the Pareto weight* $\lambda(\theta,\kappa)$ *converges to* $\lambda \in [0,1)$ *as* $\theta \to \bar{\theta}$ *for all* κ. *Then*

$$T'(\infty) \leq T'_M(\infty) = 1 - \frac{\alpha\varepsilon}{(1-\lambda)(1+\varepsilon)+\alpha\varepsilon}. \tag{23}$$

In particular, there exists a threshold $\tilde{\gamma} > 0$ *such that*

$$1 - T'(\infty) = \begin{cases} 1 - T'_M(\infty) & \text{if } 1 - T'_M(\infty) \geq \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}, \\ \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})} & \text{if } 1 - T'_M(\infty) < \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})} \text{ and } \gamma \geq \tilde{\gamma}, \\ \frac{\alpha\varepsilon + \gamma\varepsilon(1+\varepsilon)}{(1-\lambda)(1+\varepsilon) + \alpha\varepsilon + \gamma\varepsilon(1+\varepsilon) - \gamma\delta} & \text{if } 1 - T'_M(\infty) < \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})} \text{ and } \gamma \leq \tilde{\gamma}, \end{cases} \tag{24}$$

where $\delta = (1+\varepsilon)^2 \frac{T'(\infty)}{1-T'(\infty)} \left(\left(\frac{w^s(\bar{\theta})}{w^f(\bar{\theta})} \frac{1}{1-T'(\infty)} \right)^{1+\varepsilon} - 1 \right)^{-1} > 0.$

Suppose that the Mirrleesian top tax rate is above the tipping point $1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}$ and would push some top productivity types to the shadow economy. From Lemma 2 we know that even a marginal increase of the top tax rate above the tipping point entails a discrete fiscal cost, as a thinness of the upper tail of formal income distribution is increased by γ. Hence, when γ is sufficiently large, the top tax rate is optimally set exactly at the tipping point $1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}$, i.e. at the highest level which does not give incentives for informality at the top. In contrast, when γ is relatively low, the benefits of higher tax rate dominate the cost of the increased thinness of the formal income tail and some top workers will optimally work in the shadow economy. The optimal rate still falls short of the Mirrleesian rate for two reasons. First, since the upper tail of formal income distribution is thinner, the gains from increasing the top tax rate are reduced — see the terms $\gamma\varepsilon(1+\varepsilon)$ in the third case of (24). Second, increasing the top tax rate is more costly due to the top productivity types who respond on the extensive margin and join the shadow economy, which is captured by the term $-\gamma\delta$.

Finally, when the the Mirrleesian top tax rate is below the tipping point $1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}$, then the optimal top tax rate coincides with the Mirrleesian rate. Intuitively, in this case the incentives for informality are not binding at the top.

4. Quantitative analysis

In this section we explore the quantitative importance of our theoretical results. We estimate the model using the household survey from Colombia. Using the estimated model, we first analyze the impact of the informality responses on the the optimal tax schedule. Second, we examine the welfare impact of the existence of the informal sector. Additional quantitative results are available in Online Appendix B, where we show that although the actual tax schedule in Colombia is Pareto efficient, the pattern of the implied Pareto weights suggests that informality responses were not properly accounted for.

4.1. Estimation

Although we have expressed the optimal tax rates in terms of sufficient statistics, some of these statistics are very local in nature. In particular, the density of formal income of moonlighting workers is very responsive to the shape of the tax schedule. Even if this density was reliably estimated, it would be of limited use unless we knew exactly how it changes with the tax schedule. To overcome this obstacle, we follow the suggestion of Chetty (2009) and estimate the structural model to extrapolate the values of sufficient statistics out of sample.

We estimate the model using survey data from Colombia. Whereas our estimation strategy can be applied to a wide set of countries, we focus on a region with high informality which is sufficiently developed to use a non-linear income tax and transfer schedule: Latin America. Among the Latin American countries Colombia is a very attractive candidate since its informality rate is very close to the mean and the median for the whole region.[23]

Below we explain how we identify informality in the data and introduce our estimation strategy. The detailed description of the data and of the estimation procedure is provided in Appendix C.

Identifying informality. We identify the main job of a given worker as informal if the worker reports not contributing to the mandatory social insurance programs. Since the social insurance contributions are paid jointly with payroll taxes and the withheld part of the personal income tax, a worker who pays contributions is automatically subject to income taxation. Thus, this approach is particularly well suited for our exercise.[24] We find that 58% of all workers in Colombia in 2013 were employed informally at the main job, a result consistent with other indicators of informality in Colombia.[25] The average wage in the informal sector is about half of the average wage in the formal sector and the distributions of wages in the two sectors overlap significantly (see Figure 12 in Appendix C).

Out of workers with a formal main job about 6% report to have a secondary job. Some of this workers could be moonlighting in the shadow economy. However, the available

[23] Based on ILO (2018), the national share of informal employment in total employment in Latin America has a mean of 58.3% and a median of 59%, while it is equal to 60.6% in Colombia. This result differs slightly from our estimate of the size of the informal sector due to a different time period considered.

[24] Detecting informality via social security contributions is broadly consistent with the methodology of the International Labour Organization (ILO 2013) and is used by the Ministry of Labor of Colombia (ILO 2014), as well as by Goldberg and Pavcnik (2003), Guataquí, García, and Rodríguez (2010) and Mora and Muro (2017) in the studies of Colombia.

[25] The official statistical agency of Colombia (DANE) follows an alternative measure of informality based on size of the establishment, status in employment and educational level of workers. They find that 57.3% and 56.7% of workers were informal in the first two quarters of 2013 (ILO 2014), which is very close to 58% we find for the entire 2013.

data does not allow us to identify the sector of work in the second job. Hence, we treat the informality status of the second job as a latent variable.

Sample selection. We restrict attention to individuals aged 24-50 years without children (34,000 individuals). We choose this sample because these workers face a tax and transfer schedule which is not means-tested and does not depend on choices absent from our modeling framework, such as the number of children or college attainment.

Estimation strategy. The main challenge in estimating the model is identifying the joint distribution of formal and shadow productivities. For each worker we observe the hourly wage at the main job, which we interpret as productivity, and the sector of the main job, which can be either formal or informal. Crucially, we do not observe the counterfactual productivity in the sector in which the worker is not employed at the main job. Heckman and Honore (1990) and French and Taber (2011) show that the data on wages and the sector in which workers' participate is in general not sufficient to identify the sectoral productivity profiles, since workers self-select to a sector in which they are more productive. Heckman and Honore (1990) prove that the model can be identified with additional regressors which influence wages. We follow this approach. Denote the vector of regressors, which includes workers' and jobs' characteristics, by X.[26] We assume that X is informative about the worker's productivity type: $\theta \sim N(X\beta, \sigma_\theta^2)$, where vector β and scalar σ_θ are parameters to be estimated. This assumption allows us to match similar individuals who, due to different realizations of the fixed cost of informal employment, ended up having the main job in different sectors. Given that, we can infer a counterfactual shadow productivity of each formal worker from the observed shadow productivity of the matched informal workers, and *vice versa*.

Additionally, we assume that *(i)* the sectoral log-productivity schedules $\log w^f(\cdot)$ and $\log w^s(\cdot)$ are affine with respect to the productivity type,[27] *(ii)* the fixed cost of shadow employment κ is drawn from a generalized Pareto distribution, the parameters of which are allowed to vary with the productivity type θ, *(iii)* disutility from labor is given by $v(n) = \Gamma\frac{n^{1+1/\varepsilon}}{1+1/\varepsilon}$, implying a constant intensive margin labor elasticity ε which we fix to 0.33 following Chetty (2012). The support of the productivity type $[\underline{\theta}, \overline{\theta}]$ is normalized to $[0, 1]$. We obtain the density of the productivity type $F(\theta)$ with kernel density estimation and we fit a Pareto tail to the distribution of top wages. Given these assumptions, we formulate the likelihood function and estimate the model using maximum likelihood. The likelihood function and the parameter estimates are available in Appendix C.

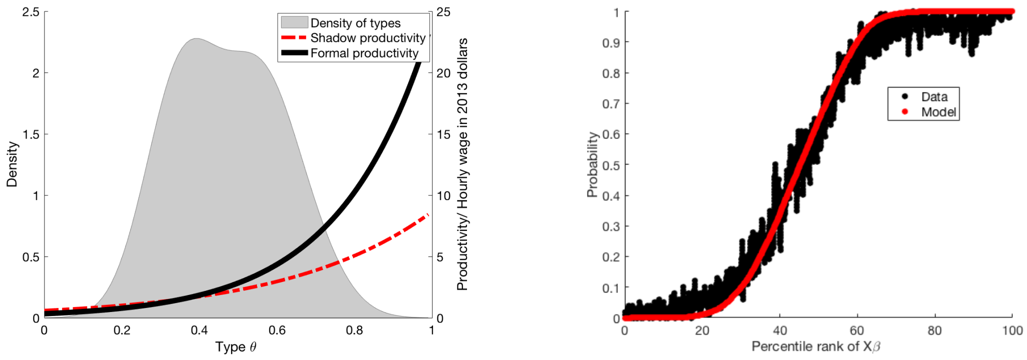
[26] In our estimation the vector X contains typical regressors from Mincerian wage equations such as age, gender, education level and experience. Following Pratap and Quintin (2006), who emphasize the importance of the establishment size to explain the differences of average wages across the formal and the informal sectors, we also include job and firm characteristics such as the task performed by the worker and the size of the firm.

[27] This implies a log-linear specification of wages and observables, which is widely used in empirical earnings equations. See Heckman, Lochner, and Todd (2006) for a discussion.

Moonlighting cannot be recovered from the survey directly. We do not impose, however, that workers with a formal main job are exclusively formal. Instead, we treat the moonlighting margin as an unobservable in the estimation of the model. The estimated model will then imply a moonlighting behavior which is consistent with the data.

Estimation results. The left panel of Figure 5 presents the estimated productivity profiles and the density of productivity types. The bottom 25% of workers are more productive in the shadow sector while the median worker is 6% more productive formally. We find that the comparative advantage in the shadow economy decreases with the productivity type.[28] Thus, as assumed in the theoretical analysis, the single crossing condition holds. The density of productivity types in the main part of the distribution is approximately normal, which means that sectoral wages are distributed approximately log-normally with a Pareto tail.

Figure 5: Estimation results



(a) Productivity profiles and type distribution

(b) Probability of having a formal main job

(a) Kernel density estimate of the productivity type distribution obtained from the observed $X\beta$ in our sample (left axis). (b) The data counterpart is the fraction of individuals with a formal main job in a rolling window of 200 workers centered around each $X\beta$ in our sample.

The right panel of Figure 5 shows the estimated probability of having a main job in the formal sector for each percentile of $X\beta$. The probability of having a formal main job is increasing and covers the whole range from 0% to 100%. To illustrate the fit of the model we also plot the share of shadow workers in a rolling window of 200 workers centered around each observed $X\beta$ in the sample. The model tracks the data well, showing that our parametric specification is compatible with the observed sorting of workers across sectors.

The estimated model also predicts that the actual tax schedule is not progressive enough to provide incentives for moonlighting. The low-cost workers ($\kappa = 0$) below the 65th

[28] Under our parametric assumptions, the comparative advantage in the shadow economy follows $w^s(\theta)/w^f(\theta) = w^s(0)/w^f(0)\exp\left\{\left(\rho^s - \rho^f\right)\theta\right\}$ and is decreasing when $\rho^s - \rho^f < 0$. The point estimate of $\rho^s - \rho^f$ is -1.74 with a standard error of 0.08.

percentile of the $X\beta$ distribution prefer to be exclusively informal, while those above decide to be exclusively formal. Hence, the bottom 65% of agents work either only formally or informally and sort according to their cost type κ, while the top 35% of agents work only formally.

4.2. Optimal tax schedule and the role of the informal sector

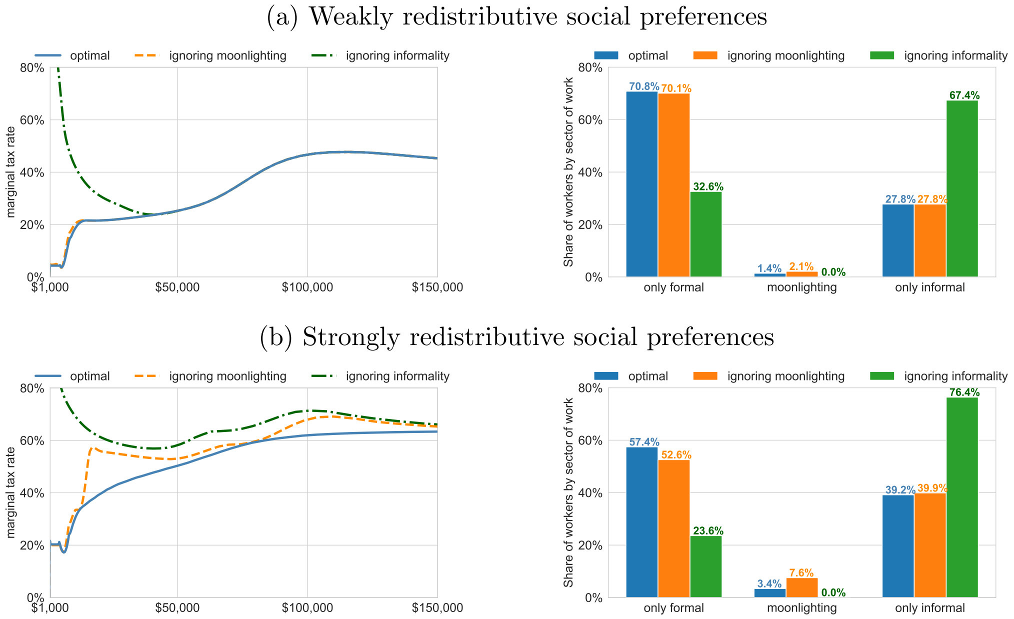
In this subsection we derive the optimal tax schedules for Colombia and compare them to tax schedules obtained when various informality responses are ignored. We consider two benchmark cases, similarly as in Proposition 4. In the first case the planner ignores moonlighting but acknowledges the mobility of workers between sectors, with the tax formula given by (20). In the second case the planner ignores all informality responses: both the moonlighting and the mobility between the two sectors, in which case the tax formula is given by (21). The latter case can be interpreted as a belief in an extreme version of the segmented market hypothesis, where the allocation of workers between the sectors of work is immutable. Importantly, we allow the income distribution to endogenously adjust to the chosen tax schedule.

The tax schedules we present for the two benchmark cases follow the notion of the Self-confirming Policy Equilibrium (SCPE), developed by Rothschild and Scheuer (2016). Since the distribution of income is endogenous to tax policy, we find the tax schedules implied by each formula iteratively: a tax schedule implies an income distribution which, together with a tax formula, results in a new tax schedule. A SCPE is a fixed point of this mapping. In such equilibrium, the income distribution and the tax schedule are consistent with the beliefs of the planner. The planner has no incentives to adjust the policy and does not discover its misperceptions, which in our case correspond to unawareness of various informality responses. In principle, each tax formula can admit multiple SCPE. We report the equilibrium which yields the highest welfare. Each tax schedule is required to generate the same revenue as the actual Colombian income tax.

We assume that Pareto weights follow $\lambda(\theta) = r(1-F(\theta))^{r-1}$ as in Rothschild and Scheuer (2013). The parameter $r \geq 1$ captures the strength of the redistributive preferences and is equal to the Pareto weight placed on the least productive agents. The average weight is always equal to 1 and the weight of the most productive agents converges to 0 when $r > 1$. We consider two cases of social preferences: $r = 1.1$ and $r = 1.7$, which we call weakly and strongly redistributive, respectively. The Pareto weights placed on the 90th percentile of θ are approximately 0.9 for the weakly redistributive and 0.3 for strongly redistributive social preferences.

Figure 6 depicts the optimal tax schedules and the tax schedules chosen when either moonlighting responses or all informality responses are ignored (left column), as well as the implied distribution of workers between the sectors of work (right column). The rows correspond to different social welfare functions. Additional statistics, including welfare

Figure 6: Equilibrium tax schedules and the distribution of workers across sectors



(a) Weakly redistributive social preferences

(b) Strongly redistributive social preferences

Note: The label 'ignoring moonlighting' corresponds to ignoring only the moonlighting responses, while 'ignoring informality' corresponds to ignoring all informality responses. In the optimum with weakly (strongly) redistributive social preferences the 50th, 95th and 99th percentiles of formal income are approx. $10,500 ($9,400), $45,000 ($40,000) *and* $87,000 ($78,000), *respectively.*

Table 1: Welfare comparison and aggregate statistics

		weakly redistributive			**strongly redistributive**		
		optimal	ign. moonlighting	ign. informality	optimal	ign. moonlighting	ign. informality
welfare loss		0. %	0. %	13.5%	0. %	2.4%	24.8%
	only formal	91.8%	91.7%	66.8%	84.9%	80.8%	47.7%
total income share	moonlighting	0.6%	0.9%	0. %	3.4%	9.8%	0.1%
	only shadow	7.6%	7.4%	33.2%	11.8%	9.4%	52.2%
	only formal	0.%	0.%	0.%	0.%	0.%	0.%
informal income share	moonlighting	4.4%	6. %	0. %	8.7%	22.2%	0. %
	only shadow	95.6%	94. %	100. %	91.3%	77.8%	100. %
	only formal	63.4%	63.8%	79.9%	68.7%	66.9%	75.7%
median θ percentile	moonlighting	20.8%	28.8%	0. %	65.7%	78.4%	99.7%
	only shadow	13. %	13. %	32.9%	18.5%	18.8%	38. %

Note: The label 'ign. moonlighting' corresponds to ignoring only the moonlighting responses, while 'ign. informality' corresponds to ignoring all informality responses.

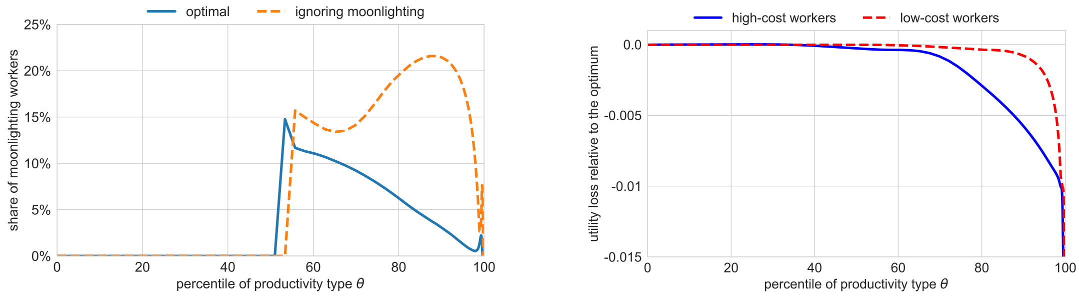
comparisons, are reported in Table 1. Remarkably, the order of tax rates predicted by Proposition 4 continues to hold, even though the assumption of identical income distributions is clearly not satisfied. We find that ignoring all informality responses leads to higher tax rates at each income level than ignoring only moonlighting responses, while the optimal tax rates are the lowest.

The optimal tax schedules are close to fully progressive: the marginal tax rates almost always increase with income. At low income levels tax rates are low and roughly constant, they start to rise close to the median income (approx. $10,000) and stabilize at the top. A stronger taste for redistribution shifts the schedule up while roughly preserving this shape. Thus, the optimal fraction of workers with exclusively formal employment decreases with the strength of redistributive preferences from 71% to 57%. The bulk of the remaining workers are employed exclusively informally. The share of moonlighting individuals is small and increases from 1.4% to 3.4% as redistributive preferences become stronger. The shadow workers' income share is much lower than their population share, as mostly the least productive agents are working in the informal sector.

When all informality responses are ignored, the tax schedules feature very high marginal tax rates at low income levels, approaching 100% at the bottom. The tax rates are decreasing through the most of the income distribution and increase again as they approach the top income tail, generating a U-shape familiar from the works of Diamond (1998) and Saez (2001). High tax rates push most of the low and medium productivity workers to the shadow sector. Nevertheless, from the planner's perspective the tax schedule seems optimal. That's because the implied density of formal income at low and medium income levels — and, hence, the perceived deadweight loss from taxation — is, in fact, low. We find that ignoring all informality responses when setting the tax policy effectively doubles the share of shadow workers relative to the optimum. Note that although taxes rates are on average higher than in the optimum, tax progressivity is actually lower. That is because the marginal tax rates increase the most at low earnings. Furthermore the share of moonlighting workers decreases to zero: all shadow workers are exclusively informal. This drop in the share of moonlighting workers is consistent with our theoretical findings linking moonlighting and tax progressivity. The welfare loss from ignoring informality responses is catastrophic and ranges from 13.5% to 24.8% of consumption depending on the social welfare function, as reported in Table 1. In other words, accounting for all informality responses brings a huge welfare gain.

The importance of accounting for moonlighting can be inferred by comparing the optimal tax schedule with the tax schedule when only the moonlighting responses are ignored. The impact of moonlighting depends crucially on the preferences for redistribution. When preferences for redistribution are strong, the moonlighting responses reduce the marginal tax rates above the median formal income by up to 20 percentage points. The moonlighting responses — agents starting to complement formal income with additional informal earnings — are important higher in the income distribution,

Figure 7: Consequences of ignoring moonlighting



(a) Incidence of moonlighting

(b) Utility loss from ignoring moonlighting

compared to the responses of switching from entirely formal to entirely informal employment. Intuitively, a secondary informal job is tempting for workers with well-paid formal jobs who face high marginal tax rates and for whom transitioning to entirely informal employment is too costly. On the other hand, when preferences for redistribution are weak, the moonlighting responses have little effect on the optimal tax schedule. In this case the tax rates for high productivity workers are not high enough to create incentives for informality.

When the preferences for redistribution are strong, ignoring moonlighting results in a share of moonlighting workers that more than doubles the optimal value, with a large welfare loss equivalent to 2.4% drop in consumption. Since the the tax schedule is excessively progressive, with the tax rates too high above median formal income but approximately optimal below, we should expect moonlighting to become more prevalent. However, why is the increased moonlighting so damaging for social welfare? We find that the sorting of workers across sectors is substantially different in comparison to the optimal allocation (see Figure 7, panel a). Relative to the optimum, not accounting for moonlighting responses induces moonlighting among workers with higher productivity, mostly from the top quartile of the productivity distribution. The median percentile of productivity type of moonlighting workers increases from 66% at the optimum to 78%. As the most productive workers who face high marginal tax rates replace a part of their formal earnings with shadow earnings, the tax revenue is substantially eroded. In fact, although the overall level of taxes is substantially higher at high income levels (e.g. the average tax rate at the 95th percentile of the formal income distribution is higher by 8 percentage points), the overall tax revenue is actually slightly lower. It means that the least productive workers receive a lower transfer. Furthermore, since the tax schedule chosen when ignoring the moonlighting responses generates a lower tax revenue while imposing higher distortions, it is Pareto inefficient. Indeed, all agents in the economy loose relative to the optimum, although losses are concentrated among the most productive workers (see Figure 7, panel b).

To conclude this investigation, we find that the shadow economy in Colombia has important implications for the optimal design of the tax schedule. In particular, the possibility of workers to migrate to entirely informal employment restricts tax rates at low and medium income levels, while the possibility of moonlighting is relevant at higher levels of income.

4.3. Welfare impact of the shadow economy

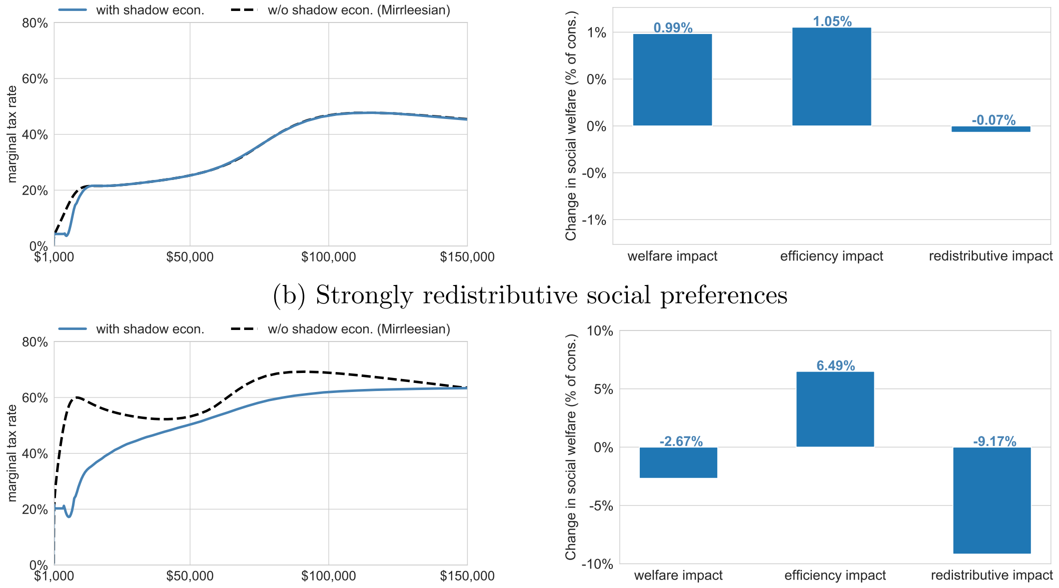
In this subsection we study the implications of the *existence* of the informal sector: we compare the optimal allocation in the calibrated economy with the optimal allocation in the counterfactual economy where the informal sector does not exist — the *Mirrleesian* economy.

In the left column of Figure 8 we compare the optimal tax schedules in the two economies. The Mirrleesian schedule features steeply increasing marginal tax rates at low income levels. In contrast, when the informal sector is present, optimal tax rates are relatively constant at the low income levels and start rising only close to the median income (approx. $10,000). For weakly redistributive social preferences, the two tax schedules coincide above the median income. When preferences for redistribution are strong, the existence of the informal sector depresses tax rates at virtually all income levels, apart from the very top. Given the functional form of the productivity schedules, Proposition 5 implies that the shadow economy does not affect the top tax rate.

The right column of Figure 8 demonstrates the difference in social welfare between the calibrated and the Mirrleesian economy. When the preferences for redistribution are weak, the shadow economy improves welfare by an amount equivalent to a 1% increase in consumption. In contrast, when the preferences for redistribution are strong, the shadow economy reduces welfare by an amount equivalent to more the 2.7% drop in consumption.

Welfare decomposition. To understand these welfare results, we propose a novel decomposition of the change of social welfare into the *efficiency impact* and the *redistributive impact*. It is standard to compare welfare of the two economies using the value of the social welfare function or the shape of the entire Pareto frontier. However, both of these methods blend together efficiency and redistributive considerations. Namely, we do not know if social welfare increases because individual choices are less distorted or because the tax schedule redistributes more resources to needy individuals. Our decomposition answers this question. It allows us to quantify which part of the social welfare difference can be attributed to more efficient labor choices, and which part to a more justly distributed tax burden. While we apply the decomposition to analyze the welfare impact of the shadow economy, it is straightforward to apply it in the context of other structural changes of the economic environment.

Figure 8: Optimal tax schedules in the model with and without the shadow economy


(a) Weakly redistributive social preferences

(b) Strongly redistributive social preferences

Proposition 6. *First, consider an allocation from the model with an informal sector* (y^f, T) *with an associated schedule of shadow earnings* y^s*. Denote total income by* $y(\theta,\kappa) = y^f(\theta,\kappa) + y^s(\theta,\kappa)$ *and total labor supply by* $n(\theta,\kappa) = \frac{y^f(\theta,\kappa)}{w^f(\theta)} + \frac{y^s(\theta,\kappa)}{w^s(\theta)}$*. Second, consider the Mirrleesian allocation* (y_M, T_M) *and denote the Mirrleesian labor supply by* $n_M(\theta) = \frac{y^M(\theta)}{w^f(\theta)}$*. Define the welfare impact of the shadow economy as*

$$\mathcal{WI} = \int_{\underline{\theta}}^{\overline{\theta}} \int_0^\infty \lambda(\theta)[U(y(\theta,\kappa) - T(y^f(\theta,\kappa)), n(\theta,\kappa)) - \kappa \mathbb{1}_{y^s(\theta,\kappa)>0} - U\left(y_M(\theta) - T_M(y_M(\theta)), n_M(\theta)\right)] dG_\theta(\kappa) dF(\theta). \tag{25}$$

The welfare impact can be represented as $\mathcal{WI} = \mathcal{EI} + \mathcal{RI}$*, where the efficiency impact* $\mathcal{EI}$ *and the redistributive impact* $\mathcal{RI}$ *are given by*

$$\mathcal{EI} = \int_{\underline{\theta}}^{\overline{\theta}} \int_0^\infty \lambda(\theta) \left[U(y(\theta,\kappa), n(\theta,\kappa)) - \kappa \mathbb{1}_{y^s(\theta,\kappa)>0} - U\left(y_M(\theta), n_M(\theta)\right)\right] dG_\theta(\kappa) dF(\theta), \tag{26}$$

$$\mathcal{RI} = \int_{\underline{\theta}}^{\overline{\theta}} \int_0^\infty \lambda(\theta) \left[T_M(y_M(\theta)) - T(y^f(\theta,\kappa))\right] dG_\theta(\kappa) dF(\theta). \tag{27}$$

Proof. It follows from linearity of preferences with respect to consumption: $U(y - T, n) = U(y,n) - T$. The decomposition can be easily generalized to utility functions which are strictly concave in consumption and non-separable between consumption and labor supply.[29] □

[29] Suppose that preferences over consumption and labor supply are given by $U(c, n)$, where $U_{cc} \leq 0$ and

The efficiency impact $\mathcal{EI}$ is the difference in the social welfare between the optimal and the Mirrleesian allocations if each agent where to consume her total income. Intuitively, it is a measure of social welfare before redistribution. It captures the influence of the informal sector on the allocation of labor supply, including the fixed cost of shadow employment. Suppose that some workers are more productive in the informal sector than in the formal sector and that for the others the marginal tax rates are lower than in the Mirrleesian economy. Then, as long as the aggregate fixed cost of shadow employment of non-formal workers is not too large, the efficiency impact will be positive. In that case the informal sector enhances the efficiency of labor supply. Alternatively, if the shadow productivity is relatively low and the fixed cost of shadow employment is high, the efficiency impact can be negative.

The redistributive impact $\mathcal{RI}$ is the difference in welfare-weighted taxes and transfers. It captures the influence of the informal sector on the optimal allocation of tax burden among workers. If in the absence of the informal sector the planner is able to reduce taxes for individuals with high Pareto weights (e.g. low productivity individuals) and increase taxes for individuals with low Pareto weights (e.g. high productivity individuals), then the redistributive impact will be negative. In that case the informal sector restricts redistribution. Alternatively, if the shadow economy allows the planner to raise transfers at low income levels, the redistributive impact can be positive.

Kopczuk (2001) proposed an example where tax avoidance increases welfare by improving redistribution at the cost of efficiency. It may suggest that the possibility of avoiding or evading taxes can improve social welfare by allowing for more even division of a smaller aggregate output. We show that such scenario is only one of many possibilities. In Online Appendix C we construct a simple model with two types and show analytically that the signs of $\mathcal{RI}$ and $\mathcal{EI}$ are, in general, ambiguous: depending on the schedules of formal and shadow productivity, the informal sector can reduce or enhance welfare along the dimensions of efficiency or redistribution independently.

Our calibrated economy provides an good example of nontrivial welfare implications of informality. We find that the Colombian informal sector has a positive efficiency impact and a negative redistributive impact (see the right column of Figure 8). There are two channels driving these results. First, the least productive agents are more productive informally than in the formal economy, which boosts the efficiency impact. Second, the shadow economy leads to lower tax rates. On the one hand, it implies lower labor distortions in the formal sector, which contributes to higher efficiency impact.

U_{cn} can be non-zero. Define the redistributive impact as below, the efficiency impact is then given by the difference between the welfare impact and the redistributive impact.

$$\mathcal{RI} = \int_{\underline{\theta}}^{\overline{\theta}} \int_0^\infty \lambda(\theta) \Big\{ U\left(y(\theta,\kappa) - T(y^f(\theta,\kappa)), n(\theta,\kappa)\right) - U\left(y(\theta,\kappa), n(\theta,\kappa)\right) \\ - \left[U\left(y_M(\theta) - T(y_M(\theta)), n_M(\theta)\right) - U\left(y_M(\theta), n_M(\theta)\right)\right] \Big\} dG_\theta(\kappa) dF(\theta). \tag{28}$$

On the other hand, it substantially reduces redistribution. Naturally, a reduction of redistribution hurts more with stronger preferences for redistribution. Hence, when the preferences for redistribution are strong, the negative redistribution impact dominates the positive efficiency impact and the informal sector reduces the overall welfare.

5. Conclusions

This paper studies the optimal income taxation when agents can earn incomes in a shadow economy which are unobserved by the government. We show theoretically and quantitatively that the optimal tax schedule which accounts for informality responses features lower tax rates throughout the income distribution. Furthermore, in the model calibrated with the Colombian data we find that the shadow economy strengthens efficiency of labor supply at the expense of possible redistribution. When preferences for redistribution are weak, the former channel dominates and the existence of the shadow economy is welfare improving. These results highlight the non-trivial welfare implications of informality. To reduce informality is a common policy objective, included for instance among the Sustainable Development Goals.[30] We instead caution against unconditional implementation of policies aimed at reducing informality.

Our analysis could be extended in several directions. First, suppose that the government can use audits and penalties to differentially affect tax evasion opportunities of different agents.[31] The optimal design of tax audits could, rather than minimizing overall tax evasion, tailor individual evasion opportunities to maximize the welfare improving potential of the shadow economy. Second, the theoretical tools we developed could be used in other settings. Our tax formula applies when agents can simultaneously work in two, broadly understood, sectors and the tax schedule can be optimized over the income from only one of the sectors. Examples of such settings are the model of home production or the problem of a local tax authority which residents can work partly outside its jurisdiction. Furthermore, our welfare decomposition can be used to uncover the efficiency and the redistributive impacts of changes in the economic environment which are unrelated to informality, such as, for instance, changes in the productivity distribution due to structural change or education policies.

[30] Sustainable Development Goal 8 (Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all), target indicator 8.3.1 (Proportion of informal employment in non-agriculture employment, by sex), see UN General Assembly (2017). For another example, consider the Programme for the Promotion of Formalization in Latin America and the Caribbean (FORLAC) run by the International Labour Organization.

[31] For instance, conducting tax audits at medium levels of formal income restricts tax evasion of highly productive agents, but not of low productivity workers who would never choose such income level. See Cremer and Gahvari (1996) for the analysis of tax audits in the optimal taxation model with two types.

A. Proofs from Section 2

Proof of Lemma 1. The strict *Spence-Mirrlees* single crossing condition holds if, keeping the formal income level fixed, the marginal rate of substitution $v'\left(\frac{y^f}{w^f(\theta)}+\frac{y^s}{w^s(\theta)}\right)\Big/w^f(\theta)$ is strictly decreasing with θ. For formal workers it follows from the strict convexity of v. For workers that supply labor to informal sector we have $v'(n)=w^s(\theta)$ and the single-crossing follows from $w^s(\theta)/w^f(\theta)$ being strictly decreasing. □

Proof of Proposition 1. Regarding 1, the second-order condition of the moonlighting θ worker is $-T''(y^f(\theta,0))\leq 0$. It cannot be satisfied if $T''(y^f(\theta,0))<0$.

Regarding 2, suppose that worker of type $(\theta,0)$ is moonlighting and supplies informal labor n^s. Denote by $\bar{y}=y^f(\theta,0)+w^f(\theta)n^s$ the level of formal earnings this agent would obtain if he shifted the informal labor to the formal economy. Then

$$\bar{y}-T(\bar{y})-\left(y^f(\theta,0)-T(y^f(\theta,0))\right)=\int_{y^f(\theta,0)}^{\bar{y}}1-T'(y)dy\geq\int_{y^f(\theta,0)}^{\bar{y}}1-T'(y^f(\theta,0))dy$$
$$=(1-T'(y^f(\theta,0)))\left(\bar{y}-y^f(\theta,0)\right)=w^s(\theta)\frac{\bar{y}-y^f(\theta,0)}{w^f(\theta)},\tag{29}$$

where the inequality follows from tax regressivity and the rightmost equality is implied by (7). Then $\bar{y}-T(\bar{y})\geq y^f(\theta,0)-T(y^f(\theta,0))+w^s(\theta)n^s$, where the left-hand side is consumption when the informal labor is shifted to the formal economy, while the right-hand side is consumption in the original arrangement. Hence, the worker is at least weakly better off shifting the entire informal labor to the formal sector. Following our convention that agents indifferent between two formal earnings levels choose a higher one, there will be no moonlighting. □

Proof of Proposition 2. First, notice that when $p>0$, the marginal tax rate at the bottom is infinite: $\lim_{y\downarrow 0}T'(y)=\infty$. Consequently, all agents have incentives to supply at least a little bit of formal labor and workers engage in informality only via moonlighting.

By (4), the formal income of the high-cost θ worker satisfies

$$(1-\tau)y^f(\theta,\infty)^{-p}=\left(\frac{y^f(\theta,\infty)}{w^f(\theta,\infty)}\right)^{\frac{1}{\varepsilon}}\implies y^f(\theta,\infty)=\left((1-\tau)w^f(\theta)^{1+\frac{1}{\varepsilon}}\right)^{\frac{1}{p+\frac{1}{\varepsilon}}}.\tag{30}$$

The low-cost θ worker has no incentives to moonlight if $(1-T'(y^f(\theta,\infty)))w^f(\theta)\geq w^s(\theta)$. Plugging in the expression for $y^f(\theta,\infty)$, as well as the tax and the productivity schedules, we can express it as

$$(1-\tau)^{\frac{1}{1+p\varepsilon}}\cdot\frac{w^f(0)^{\frac{1-p}{1+p\varepsilon}}}{w^s(0)}\geq e^{\left(\rho^s-\frac{1-p}{1+p\varepsilon}\rho^f\right)\theta}.\tag{31}$$

Evaluate it at $\theta = 0$ to obtain the second inequality in (8). To ensure that this condition holds also for higher types, we need to restrict the exponent on the right-hand side to be non-positive: $\rho^s - \frac{1-p}{1+p\varepsilon}\rho^f \leq 0$. This inequality is always true if $\rho^s \leq 0$. If $\rho^s > 0$, it is equivalent to the first inequality in (8). □

Proof of Proposition 3. Given the single crossing condition, the necessity of constraint 1 (i.e. increasing formal income schedule and local incentive-compatibility constraints) follows from Theorem 7.2 in Fudenberg and Tirole (1991). By Theorem 7.3 in Fudenberg and Tirole (1991), constraint 1 is sufficient to prevent deviations within the cost type, i.e. deviations of some high-cost (low-cost) worker to formal income level earned by another high-cost (low-cost) worker. Additionally, constraint 2 is clearly necessary and sufficient to prevent deviations between different cost types for a fixed productivity type. Below we first show that constraints 3-5 are sufficient to prevent simultaneous deviations between the cost and the productivity types. Finally, we provide a graphical example of the insufficiency of local incentive-compatibility constraints.

Denote the image of formal income schedule of types with fixed cost $\kappa \in \{0, \infty\}$ by $Y(\kappa) \equiv \{y \in \mathbb{R}_+ : \exists_{\theta \in [\underline{\theta}, \bar{\theta}]} y^f(\theta, \kappa) = y\}$. Deviations between the cost classes may arise if the formal income schedules of the two classes do not have identical images: $Y(0) \neq Y(\infty)$. The difference in images may occur when suprema or infima of the two sets do not coincide: either $y^f(\underline{\theta}, 0) < y^f(\underline{\theta}, \infty)$ or $y^f(\bar{\theta}, 0) < y^f(\bar{\theta}, \infty)$. Constraints 3 and 4 take care of these deviations. Alternatively, one of the income schedules can exhibit a discontinuous jump where the other schedule remains continuous. Condition 5 prevents potential deviations when $y^f(\cdot, 0)$ is discontinuous and $y^f(\cdot, \infty)$ remains continuous.[32] Below we show that the reverse situation never happens: when there is $y \in (y^f(\underline{\theta}, \infty), y^f(\bar{\theta}, 0))$ such that $y \in Y(0)$, then always $y \in Y(\infty)$.

We will show that for any θ such that $y^f(\theta, 0) \in (y^f(\underline{\theta}, \infty), y^f(\bar{\theta}, 0))$ we can find $\tilde{\theta}$ such that $y^f(\tilde{\theta}, \infty) = y^f(\theta, 0)$. Take some incentive-compatible allocation (y^f, T) and choose any θ such that $y^f(\theta, 0) > y^f(\underline{\theta}, \infty)$ and $y^s(\theta, 0) > 0$. Consider a productivity type $\tilde{\theta}$ such that

$$\frac{v'(y^f(\theta, 0)/w^f(\tilde{\theta}))}{w^f(\tilde{\theta})} = \frac{w^s(\theta)}{w^f(\theta)}. \tag{32}$$

We will show that $y^f(\tilde{\theta}, \infty) = y^f(\theta, 0)$. It means that at any formal income level above $y^f(0, \infty)$ which is chosen by some low-cost worker there is also some high-cost worker.[33]

Consider indifference curves of agents $(\theta, 0)$ and $(\tilde{\theta}, \infty)$ depicted in Figure 9. The indifference curve of the low-cost θ-worker and the high-cost $\tilde{\theta}$-worker are tangential at formal

[32] Note that it is sufficient to impose additional constraints only on particular types: $(\underline{\theta}, 1)$, $(\bar{\theta}, 0)$, or a type at the discontinuity. If these constraints hold, no other type is tempted by a deviations due to a single-crossing condition.

[33] If $w^f(\underline{\theta}) > 0$, we need to make sure that $\tilde{\theta}$ always exists. Suppose on the contrary that $v'(y^f(\theta, 0)/w^f(\underline{\theta}))/w^f(\underline{\theta}) < w^s(\theta)/w^f(\theta)$, so that there is no $\tilde{\theta} \geq \underline{\theta}$ which satisfies (32). One can then show that it implies that if agent $(\theta, 0)$ prefers $y^f(\theta, 0)$ to $y^f(\underline{\theta}, \infty)$, so does agent $(\underline{\theta}, \infty)$. It is a contradiction, since $y^f(\theta, 0) > y^f(\underline{\theta}, \infty)$ and the allocation is incentive-compatible.

Figure 9: Indifference curves.

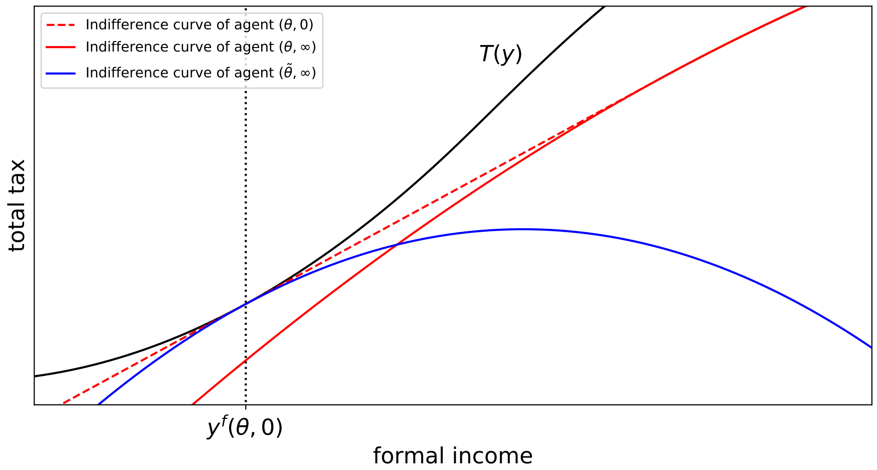
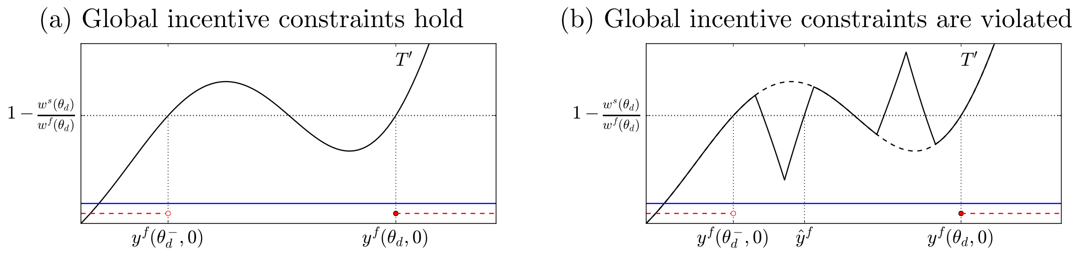


Figure 10: Insufficiency of the local incentive-compatibility constraints


(a) Global incentive constraints hold

(b) Global incentive constraints are violated

Note: The horizontal lines indicate whether at a given formal income level there are high-cost workers (solid, blue) or low-cost workers (dashed, red). In both panels agent $(\theta_d, 0)$ is indifferent between $y^f(\theta_d^-, 0)$ and $y^f(\theta_d, 0)$. Hence, the local incentive constraint (9) holds. However, in the right panel the worker strictly prefers formal income level $\hat{y}^f$, since the average tax rate between $y^f(\theta_d^-, 0)$ and $\hat{y}^f$ is below $1 - w^s(\theta_d)/w^f(\theta_d)$.

income $y^f(\theta, 0)$. Furthermore, the indifference curve of the low-cost worker is a straight line whenever this agent supplies shadow labor, while the indifference curve of the high-cost worker is strictly concave. Finally, the indifference curves of agents $(\theta, 0)$ and $(\tilde{\theta}, \infty)$ never cross. Otherwise, the indifference curves of agents $(\tilde{\theta}, \infty)$ and (θ, ∞) would cross more than once and the single crossing condition would be violated. Altogether, it means that $y^f(\theta, 0)$ is the incentive-compatible formal income choice of the high-cost $\tilde{\theta}$-worker. Suppose on the contrary that agent $(\tilde{\theta}, \infty)$ prefers some $\tilde{y}^f \neq y^f(\theta, 0)$. This is a profitable deviation for agent $(\theta, 0)$ as well, since his indifference curve is weakly higher. It contradicts the original assumption of incentive-compatibility of $y^f(\cdot, 0)$.

We demonstrate the insufficiency of local incentive-compatibility constraints in Figure 10. Consider an interval of moonlighting low-cost workers. As the marginal tax rates are not monotone increasing in formal income, the formal income schedule of these workers must be discontinuous. The local incentive constraint of the agent at the discontinuity $(\theta_d, 0)$ (given by equation (9)) implies that the agent is indifferent between the two formal income levels across the discontinuity: $y^f(\theta_d^-, 0)$ and $y^f(\theta_d, 0)$. In the right panel we modify the marginal tax rates in a way that total tax levels at $y^f(\theta_d^-, 0)$ and $y^f(\theta_d, 0)$ do not change. Thus, the local incentive constraint of the agent $(\theta_d, 0)$ still holds. However, this agent has a profitable deviation to $\hat{y}^f$. The average tax rate between $y^f(\theta_d^-, 0)$ and $\hat{y}^f$ is below $1 - w^s(\theta)/w^f(\theta)$, which implies that the utility from deviation to $\hat{y}^f$ is higher than the utility at $y^f(\theta_d, 0)$. Therefore, the new allocations is not globally incentive-compatible: it violates constraint 5 from Proposition 3. It is easy to construct similar examples of locally incentive-compatible allocations violating constraints 3 and 4. □

B. Proofs from Section 3

Proof of Theorem 1. The proof follows the main text in Section 3. Here we formally define and derive the sufficient statistics used.

Let's derive the elasticities of formal income. First, consider the first-order condition of the high-cost worker with θ productivity type. For brevity, denote the formal income of this worker by y and the labor supply by n. Then

$$y = w^f(\theta) \cdot v'^{-1}\left((1 - T'(y))w^f(\theta)\right). \tag{33}$$

Perturb the net-of-tax rate $1 - T'(y)$ by a small $d(1 - T'(y))$. If the tax schedule is locally linear then the corresponding income adjustment and the elasticity will be

$$dy = \frac{w^f(\theta)^2}{v''(n)} d(1 - T'(y)), \tag{34}$$

$$\varepsilon^f(y) = \frac{dy}{d(1 - T'(y))} \frac{1 - T'(y)}{y} = \frac{v'(n)}{nv''(n)}, \tag{35}$$

where the last equality applies the first-order condition again. Suppose now that the tax schedule is not linear. Then the perturbation of the net-of-tax rate will lead to an additional term, capturing the impact of local tax progressivity:

$$dy = \frac{w^f(\theta)^2}{v''(n)} d(1 - \tau) - T''(y) \frac{w^f(\theta)^2}{v''(n)} dy = \frac{w^f(\theta)^2}{v''(n)} \frac{d(1 - T'(y))}{1 + w^f(\theta)^2 \frac{T''(y)}{v''(n)}}, \tag{36}$$

$$\tilde{\varepsilon}^f(y) = \frac{dy}{d(1 - T'(y))} \frac{1 - T'(y)}{y} = \frac{\varepsilon^f(y)}{1 + w^f(\theta)^2 \frac{T''(y)}{v'(n)}} = \left[\frac{1}{\varepsilon^f(y)} + \frac{T''(y)}{1 - T'(y)} y\right]^{-1}. \tag{37}$$

Second, consider a moonlighting worker with productivity type θ. Again, for brevity denote the formal income of this worker by y. The relevant first-order condition is

$$\left(1 - T'(y)\right) w^f(\theta) = w^s(\theta). \tag{38}$$

Furthermore, the second-order condition is $T''(y) \geq 0$, which means that we can focus attention on the tax schedules which are weakly locally progressive. First, suppose that the tax schedule is linear. Then, the elasticity of formal income is infinite. To see this, note that if $(1 - T'(y))\, w^f(\theta) > w^s(\theta)$, the worker will shift the entire labor supply into the formal sector, while if $(1 - T'(y))\, w^f(\theta) < w^s(\theta)$, the worker will shift the entire labor supply to the shadow economy. Second, suppose that the tax schedule is non-linear and progressive. Then the formal income and the elasticity are

$$dy = \frac{1}{T''(y)} d(1 - T'(y)), \tag{39}$$

$$\tilde{\varepsilon}^s(y) = \frac{1 - T'(y)}{T''(y)y}. \tag{40}$$

We derive the elasticities of formal income with respect to the formal productivity in an analogous way:

$$\varepsilon^f_{w^f}(y) = 1 + \varepsilon^f(y), \quad \tilde{\varepsilon}^f_{w^f}(y) = \frac{\tilde{\varepsilon}^f(y)}{\varepsilon^f(y)} \varepsilon^f_{w^f}(y), \quad \tilde{\varepsilon}^s_{w^f}(y) = \left(1 - \frac{\rho^s(\theta)}{\rho^f(\theta)}\right) \tilde{\varepsilon}^s(y), \tag{41}$$

where $\rho^x(\theta) = \frac{dw^x(\theta)}{d\theta} \frac{1}{w^x(\theta)}$. Denote the derivative of formal income with respect to the productivity type along the non-linear tax schedule as

$$\tilde{y}^f_\theta(\theta, \kappa) \equiv \begin{cases} \tilde{\varepsilon}^f_{w^f}(y^f(\theta, \kappa)) \rho^f(\theta) y^f(\theta, \kappa) & \text{if } \kappa \geq \tilde{\kappa}(\theta), \\ \tilde{\varepsilon}^s_{w^f}(y^f(\theta, \kappa)) \rho^f(\theta) y^f(\theta, \kappa) & \text{otherwise.} \end{cases} \tag{42}$$

The density of formal workers at formal income $y^f(\theta, \infty)$, scaled by the share of formal workers, is defined as $h^f(y^f(\theta, \infty)) = (1 - G_\theta(\tilde{\kappa}(\theta))) f(\theta) / \tilde{y}^f_\theta(\theta, \infty)$ and $h^f(y^f) = 0$ for $y^f \notin y^f([\underline{\theta}, \bar{\theta}], \infty)$. The density of moonlighting workers at formal income $y^f(\theta, 0) > 0$, scaled by the share of moonlighting workers, is $h^s(y^f(\theta, 0)) = G_\theta(\tilde{\kappa}(\theta)) f(\theta) / \tilde{y}^f_\theta(\theta, 0)$ and $h^s(y^f) = 0$ for $y^f \notin y^f([\underline{\theta}, \bar{\theta}], 0)$. The density of formal income is then $h(y) = h^f(y) + h^s(y)$. The mean elasticity at income level y is $\bar{\varepsilon}(y) \equiv h^f(y)\tilde{\varepsilon}^f(y) + h^s(y)\tilde{\varepsilon}^s(y)$. The mass of workers working exclusively informally is $H_0 = \int_{\underline{\theta}}^{\bar{\theta}} G_\theta(\tilde{\kappa}(\theta)) \cdot \mathbb{1}_{y^f(\theta,0)=0} dF(\theta)$ and the cdf of formal income is $H(y) = H_0 + \int_0^y h(y) dy$.

The elasticity of the density of formal workers with respect to the tax burden of staying formal $\Delta T(y)$ is defined as

$$\pi(y^f(\theta, \infty)) = g_\theta(\tilde{\kappa}(\theta)) \frac{\Delta T(y^f(\theta, \infty))}{1 - G_\theta(\tilde{\kappa}(\theta))}. \tag{43}$$

The average welfare weights of formal and shadow workers at a given formal income are defined as

$$\bar{\lambda}^f(y^f(\theta, \infty)) = \int_{\tilde{\kappa}(\theta)}^{\infty} \lambda(\theta, \kappa) \frac{dG_\theta(\kappa)}{1 - G_\theta(\tilde{\kappa}(\theta))}, \quad \bar{\lambda}^s(y^f(\theta, 0)) = \int_0^{\tilde{\kappa}(\theta)} \lambda(\theta, \kappa) \frac{dG_\theta(\kappa)}{G_\theta(\tilde{\kappa}(\theta))}. \tag{44}$$

Then the average welfare weight at formal income y is $\bar{\lambda}(y) = \left(h^f(y)\bar{\lambda}^f(y) + h^s(y)\bar{\lambda}^s(y)\right)/h(y)$. Finally, the mapping $\theta \mapsto s(\theta)$ is defined as $s(y) = \min_\theta\{y^f(\theta,0) \text{ s.t. } y^f(\theta,0) \geq y\}$. □

Proof of Proposition 4. By assumptions made, the fiscal impacts of the extensive margin responses and of the intensive margin responses to increasing a marginal tax rate are non-negative. We can distinguish four cases and determine the order of tax rates in each of them.

1. There are some low-cost workers above y and ...
   - a) $\Delta_\infty(y) = 0:$ $\quad T'(y) = T'_I(y) = T'_{II}(y),$
   - b) $\Delta_\infty(y) > 0$ and $s(y) = y$: $\quad T'(y) \leq T'_I(y) \leq T'_{II}(y),$
   - c) $\Delta_\infty(y) > 0$ and $s(y) > y$: $\quad T'(y) < T'_I(y) \leq T'_{II}(y).$
2. There are no low-cost workers above y: $\quad T'(y) \leq T'_I(y) \leq T'_{II}(y).$

Consider these cases successively. 1a) $\Delta_\infty(y) = 0 \Leftrightarrow \Delta_0(y) = 0 \implies s(y) = y \wedge h^s(y) = 0$, which means that all workers at y are formal, there are no intensive margin responses of shadow workers and there are no extensive margin responses. Both (18) and the formula I collapse into the formula II. 1b) Since $s(y) = y$, the average intensive margin elasticity at y is sufficient to capture the intensive margin responses of formal and shadow workers. However, the formula II captures none of the extensive margin responses, while the formula I captures only a fraction of them. 1c) Neither the formula I nor II capture the intensive margin responses of the shadow workers, since they happen at income level higher than y. Analogous to the previous case with respect to the extensive margin responses. In the case 2, the formulas I and II correctly capture the intensive margin responses, but, analogously to the two previous cases, they miss a part of the extensive margin responses. □

Proof of Lemma 2. If all top workers are formal, i.e. $1 - T'(\infty) \geq \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}$, the distribution of formal income satisfies

$$\lim_{y\to\infty} \frac{1-H(y)}{h(y)y} = \lim_{\theta\to\bar{\theta}} \frac{1-F(\theta)}{f(\theta)w^f(\theta)} \frac{dw^f(\theta)}{d\theta} \frac{dy^f(\theta,\infty)}{dw^f(\theta)} \frac{w^f(\theta)}{y^f(\theta,\infty)} = \frac{1+\varepsilon}{\alpha}. \tag{45}$$

When there are some informal workers among the top productivity types, we have

$$\lim_{y\to\infty} \frac{1-H(y)}{h(y)y} = \lim_{\theta\to\bar{\theta}} \frac{1-\int_0^\theta (1-G_{\theta'}(\tilde{\kappa}(\theta')))dF(\theta')}{(1-G_\theta(\tilde{\kappa}(\theta)))\, f(\theta)w^f(\theta)} \frac{dw^f(\theta)}{d\theta} \frac{dy^f(\theta,\infty)}{dw^f(\theta)} \frac{w^f(\theta)}{y^f(\theta,\infty)}. \tag{46}$$

One can show that the formality threshold $\tilde{\kappa}(\theta)$ is asymptotically proportional to $w^f(\theta)^{1+\varepsilon}$:

$$\lim_{\theta\to\bar{\theta}} \frac{\tilde{\kappa}(\theta)}{w^f(\theta)^{1+\varepsilon}} = \frac{1}{1+\varepsilon}\left(\left(\frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}\right)^{1+\varepsilon} - \left(1-T'(\infty)\right)^{1+\varepsilon}\right). \tag{47}$$

Consequently, $1 - G_\theta(\tilde{\kappa}(\theta))$ is asymptotically proportional to $w^f(\theta)^{-\gamma(1+\varepsilon)}$ and

$$\lim_{\theta\to\bar{\theta}} \frac{1 - \int_0^\theta (1 - G_{\theta'}(\tilde{\kappa}(\theta')))dF(\theta')}{(1 - G_\theta(\tilde{\kappa}(\theta)))\, f(\theta) w^f(\theta)} \frac{dw^f(\theta)}{d\theta} = \lim_{w^f\to\infty} \frac{\int_{w^f}^\infty 1/(w)^{1+\alpha+\gamma(1+\varepsilon)} dw}{1/(w^f)^{1+\alpha+\gamma(1+\varepsilon)} w^f} \tag{48}$$

which is equal to $(\alpha + \gamma(1+\varepsilon))^{-1}$. Plugging that into (46), we get $\lim_{y\to\infty} \frac{1-H(y)}{h(y)y} = \frac{1+\varepsilon}{\alpha+\gamma(1+\varepsilon)}$. □

Proof of Proposition 5. If all the top productivity workers, including all the cost types, are formal then the optimal tax formula (18) in the limit as $\theta \to \bar{\theta}$ implies

$$\frac{T'(\infty)}{1 - T'(\infty)} \frac{\alpha\varepsilon}{1+\varepsilon} = 1 - \lambda \implies 1 - T'(\infty) = \frac{\alpha\varepsilon/(1+\varepsilon)}{1 - \lambda + \alpha\varepsilon/(1+\varepsilon)}. \tag{49}$$

This happens either if the shadow economy does not exist or if the shadow economy exists, but the top workers have not incentives to work informally: $(1 - T'(\infty))w^f(\bar{\theta}) \geq w^s(\bar{\theta})$.

Suppose on the contrary that $(1 - T'(\infty))w^f(\bar{\theta}) < w^s(\bar{\theta})$, which means that some top productivity workers work informally. In Lemma 2 we determined that the tail parameter of the productivity of formal workers is $(\alpha + \gamma(1+\varepsilon))^{-1}$. Furthermore, define the following function

$$\lim_{\theta\to\bar{\theta}} \frac{\Delta T(y^f(\theta,\infty))}{\tilde{\kappa}(\theta)} = (1+\varepsilon)\frac{T'(\infty)}{1 - T'(\infty)} \left(\left(\frac{w^s(\bar{\theta})/w^f(\bar{\theta})}{1 - T'(\infty)} \right)^{1+\varepsilon} - 1 \right)^{-1} \equiv \tilde{\delta}(T'(\infty)), \tag{50}$$

where $\tilde{\delta}(\tau) > 0$ for $\tau > 1 - w^s(\bar{\theta})/w^f(\bar{\theta})$ and $\tilde{\delta}(\tau)$ diverges to $+\infty$ as τ converges to $1 - w^s(\bar{\theta})/w^f(\bar{\theta})$ from the right. Notice that $\tilde{\delta} = (1+\varepsilon)\delta$, where δ is defined in the text of the proposition. Then the elasticity of the density of formal workers at the top converges to

$$\lim_{\theta\to\bar{\theta}} \frac{g_\theta(\tilde{\kappa}(\theta))\tilde{\kappa}(\theta)}{1 - G_\theta(\tilde{\kappa}(\theta))} \frac{\tilde{\delta}T(y^f(\theta,\infty))}{\tilde{\kappa}(\theta))} = \gamma\tilde{\delta}(T'(\infty)). \tag{51}$$

As $\theta \to \bar{\theta}$, the optimal tax formula (19) implies

$$\frac{T'(\infty)}{1 - T'(\infty)} \left(\frac{\alpha\varepsilon}{1+\varepsilon} + \gamma\varepsilon \right) = 1 - \lambda - \gamma\tilde{\delta}(T'(\infty))$$

$$\implies 1 - T'(\infty) = \frac{\alpha\varepsilon/(1+\varepsilon) + \gamma\varepsilon}{1 - \lambda + \alpha\varepsilon/(1+\varepsilon) + \gamma\varepsilon - \gamma\tilde{\delta}(T'(\infty))}. \tag{52}$$

To characterize the top tax rate, define the following auxiliary functions

$$\Phi_M(\tau) = 1 - \lambda - \frac{\tau}{1-\tau}\frac{\alpha\varepsilon}{1+\varepsilon}, \tag{53}$$

$$\Phi(\tau,\gamma) = \begin{cases} \Phi_M(\tau) & \text{if } \tau \leq 1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}, \\ 1 - \lambda - \gamma\tilde{\delta}(\tau) - \frac{\tau}{1-\tau}\left(\frac{\alpha\varepsilon}{1+\varepsilon} + \gamma\varepsilon\right) & \text{if } \tau > 1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}. \end{cases} \tag{54}$$

$\Phi_M(\tau)$ is the marginal social benefit of increasing the top tax rate from the level τ in the standard Mirrlees model. $\Phi_M(\cdot)$ is strictly decreasing, strictly concave and naturally $\Phi_M(T'_M(\infty)) = 0$. $\Phi(\tau,\gamma)$ is the marginal social benefit of increasing the top tax rate from the level τ in the model with a shadow economy when the fixed cost distribution has a tail parameter γ. $\Phi(\cdot,\gamma)$ is discontinuous at $1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}$, where for any positive γ it falls down to $-\infty$. On the interval $\left(1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}, 1\right)$ the function $\Phi(\cdot,\gamma)$ is, for any positive γ, first increasing and then decreasing, strictly concave and bounded from above by $\Phi_M(\cdot)$.

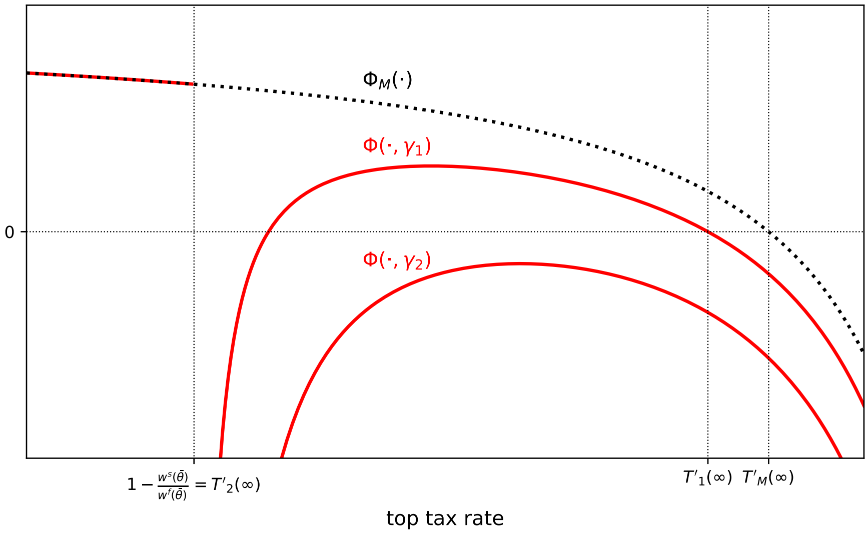
Suppose that $T'_M(\infty) > 1 - w^s(\bar{\theta})/w^f(\bar{\theta})$. The optimal top tax rate $T'(\infty)$ satisfies

$$T'(\infty) = \arg \max_{\tau^* \geq 1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}} \int_{1-\frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}}^{\tau^*} \Phi(\tau,\gamma) d\tau \tag{55}$$

$$= \arg \max_{\tau^* \geq 1 - \frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}} \int_{1-\frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}}^{\tau^*} \Phi_M(\tau) d\tau - \gamma \int_{1-\frac{w^s(\bar{\theta})}{w^f(\bar{\theta})}}^{\tau^*} \left(\tilde{\delta}(\tau) + \frac{\tau\varepsilon}{1-\tau}\right) d\tau. \tag{56}$$

There are two possible candidates for the optimal top tax rate: *(i)* $1 - w^s(\bar{\theta})/w^f(\bar{\theta})$ and *(ii)* $\tilde{\tau}$ which satisfies $\Phi(\tilde{\tau},\gamma) = 0$ and $\partial\Phi(\tau,\gamma)/\partial\tau \mid_{\tau=\tilde{\tau}} < 0$ (see Figure 11). Suppose that at some γ the solution is equal to $1 - w^s(\bar{\theta})/w^f(\bar{\theta})$. Since $\tilde{\delta}(\tau) + \frac{\tau\varepsilon}{1-\tau} > 0$ for all $\tau > 1 - w^s(\bar{\theta})/w^f(\bar{\theta})$, the solution is unchanged for any higher values of γ. It proves the existence of threshold $\tilde{\gamma}$. □

Figure 11: Determining the optimal top tax rate



Note: $\Phi_M(\tau)$ is the marginal social benefit of increasing the top tax rate from the level τ in the standard Mirrlees model, while $\Phi(\tau,\gamma)$ is marginal social benefit of increasing the top tax rate in the model with a shadow economy when the distribution of the fixed cost has a tail parameter γ. We consider two values of the tail parameter of the cost distribution: γ_1 and γ_2, $\gamma_2 > \tilde{\gamma} > \gamma_1$, where $\tilde{\gamma}$ is a threshold from Proposition 5. $T'_k(\infty)$ is the optimal top tax rate with γ_k, $k \in \{1,2\}$.

C. Estimation details

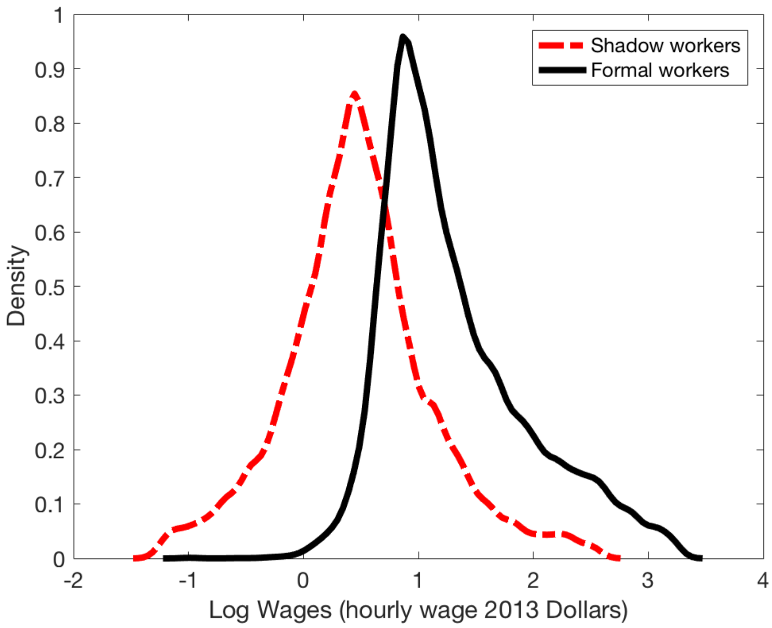
First we describe the data and explain how we recover wages and sectoral participation. Second, we list the identifying assumptions and formulate the likelihood function. Last we present the parameter estimates.

Data. We use the 2013 wave of the household survey by the official statistical agency of Colombia (DANE). We restrict attention to individuals aged 24-50 years without children (34,000 individuals). We choose this sample, since these workers face a tax and transfer schedule which is not means-tested and does not depend on choices absent from our modeling framework, such as a number of children or college attainment.

The information we use in the estimation is given by a sample $\left\{\omega_i, \iota_i^f, x_i, s_i\right\}_{i=1}^N$ of the random variables $\left\{\mathcal{W}, I^f, X\right\}$, where $\mathcal{W}$ is the hourly wage of worker before taxes; I^f an indicator variable for having a main job in the formal sector; X a vector of worker characteristics; and s_i the sampling weight of observation i and N the total number of observations in our sample. The indicator variable I^f is set equal to one if the worker reports to be affiliated to all three components of social security: pension system, health insurance and labor accidents insurance. A fraction (about 3%) of workers also have a second job. If the first job is formal we cannot identify if the worker's second job is shadow or formal. Therefore I^f indicates formality of the main job and does not imply that the worker is exclusively formal.

We use two questions of the survey to construct our measure of the hourly wage $\mathcal{W}$. First, the worker is asked what was her income at the main job last month. Second, what is the number of hours she 'normally' works at that job. We use the ratio of the reported income and hours in those questions to compute our measure of the hourly wage. Since the 'normal' number of hours need not to correspond to last month's number of hours we use our measure as a noisy measure of productivity in the model.[34] If the worker is identified to be formal at the main job we include the statutory payroll taxes that are paid by the employer in the computation of the pre-tax income at the main job. In Figure 12 the distribution of log-wages is presented for each sector. Variables included in vector X are listed in Table 2.

Figure 12: Density of observed log-wages in the formal and the informal sectors



Kernel density estimation of the wage distribution obtained from the observed wages in our sample.

Modeling assumptions. We assume that productivity in the participating sector is equal to the measured hourly wage $\mathcal{W}$ plus a normally distributed measurement error $u \sim N(0, \sigma_u)$. Also, productivity in each sector $j \in \{s, f\}$ features a constant, sector specific growth rate ρ^j with respect to the productivity type θ:

$$\log\left(w^j(\theta)\right) = \log\left(w^j(0)\right) + \rho^j\theta, \quad j \in \{s, f\}. \tag{57}$$

[34] We further assume that survey respondents correctly reveal their gross income from the main job, regardless of whether the main job is formal or informal. Other papers making this assumption include Meghir, Narita, and Robin (2015) for Brazil and López García (2015) for Chile.

Table 2: Variables included in X

Variable	Description	Values
Individual characteristics		
Gender	Dummy variable equal to 1 for women	0-1
Age	Age of the worker	16-90
Age^2	Age squared	
Educ	Number of education years	0-26
Degree	Highest degree achieved (No degree to Doctorate)	1-5
Work	Number of months worked in the last year	1-12
Exper	Number of months worked in the last job	0-720
1stJob	Dummy for the first job (1 if it is the first job)	0-1
Job characteristics		
S-Man	Dummy for the manufacturing sector	0-1
S-Fin	Dummy for financial intermediation	0-1
S-Ret	Dummy for the sales and retailers sector	0-1
B-city	Dummy for a firm in one of the two largest cities	0-1
Size	Categories for the number of workers	1-9
Lib	Dummy for a liberal occupation	0-1
Admin	Dummy for an administrative task	0-1
Seller	Dummy for sellers and related	0-1
Services	Dummy for a service task	0-1
Worker-firm relationship		
Union	Dummy for labor union affiliation (1 if yes)	0-1
Agency	Dummy for agency hiring (1 if yes)	0-1
Senior	Number of months of the worker in the firm	0-720

The above assumption is not restrictive for the unconditional distribution of formal wages, as long as we are free to choose any distribution of the productivity types $F(\theta)$. This assumption, however, restricts the joint distribution of formal and shadow wages. The comparative advantage in the shadow economy becomes

$$\frac{w^s(\theta)}{w^f(\theta)} = \frac{w^s(0)}{w^f(0)} \exp\left\{\left(\rho^s - \rho^f\right)\theta\right\}. \tag{58}$$

The fixed cost of shadow employment κ follows a generalized Pareto distribution with density

$$g_\theta(\kappa) = \frac{1}{\sigma_\kappa\left(w^f(\theta) - w_\kappa\right)^{\alpha_\kappa}}\left(1 + \frac{\kappa}{\sigma_\kappa\left(w^f(\theta) - w_\kappa\right)^{\alpha_\kappa}}\right)^{-2}, \tag{59}$$

where parameters σ_κ, α_κ and w_κ determine how the distribution of the fixed cost is affected by the productivity type θ.

Proposition 7. *The model given by (57), (59) and an unrestricted distribution of types $F(\theta)$ is not identified with data on workers wages and sectoral choice.*

Proof. The model is not identified as any distribution of wages could have been generated by a version of the model where participation costs are irrelevant and all workers are sorted only according to their relative productivities. We assume the empirical marginal tax rates are non-negative and bounded away from 100%.

Consider the following parametrization of the model: $w^s(0) = \bar{w}$, $w^f(0) = \underline{w}^2/\bar{w}$, $\rho^s = -\rho^f = 2\ln(\underline{w}) - 2\ln(\bar{w})$, where $\bar{w}$ is an upper bound on the support of wages and $\underline{w} \in (0,1)$ is a lower bound. The support of θ is $[0,1]$ and the distribution of the fixed cost is collapsed to zero. Under this parametrization formal productivity is increasing in type θ, shadow productivity decreasing, and they cross at productivity equal to $\underline{w}$ for type $\theta = 0.5$.

Let $F_{W,s}$ be the cumulative density of wages of the participants in the shadow sector, $F_{W,f}$ that of the participants in the formal sector and μ_s the mass of individuals in the shadow sector. Any joint distribution of (w, I^f) can be replicated by setting the cumulative distribution of types as follows:

$$F(\theta) = \begin{cases} \mu_s F_{W,s}\left(\bar{w}\exp\{\rho^s\theta\}\right) & \text{if } \theta \in [0, 0.5] \\ \mu_s + (1-\mu_s)F_{W,f}\left(\frac{\underline{w}^2}{\bar{w}}\exp\{\rho^f\theta\}\right) & \text{if } \theta \in (0.5, 1] \end{cases}$$

Finally, to guarantee that workers with $\theta \in (0.5, 1]$ self-select to be formal workers, set the lower bound $\underline{w}$ to be the product of the lowest observed formal wage and the lowest possible net-of-tax rate: $\underline{w} = \min(w^f) \times \min_{y\geq 0}\{1 - T'(y)\}$. It guarantees that the after-tax formal wage is never below the shadow wage. □

Proposition 7 is a particular instance of the results of Heckman and Honore (1990) and French and Taber (2011), where it is shown that the data on wages and the sectoral participation is in general not sufficient to identify the productivity profiles. Heckman and Honore (1990) also prove that the model can be identified with additional regressors that affect the location parameters of the skill distribution. Motivated by this approach we include a vector of regressors X that can potentially convey information about the workers productivity and assume the following relationship:

$$\theta \sim N(X\beta, \sigma_\theta^2), \tag{60}$$

where β is a vector of parameters. We obtain $F(\theta)$ using (60) and a kernel density estimation of the $X\beta$ distribution. To capture the right tail of the wage distribution, we fit a Pareto distribution with parameter α_w to the top 1% of formal wages. Finally, we assume that agents' preferences over labor supply follow

$$v(n) = \Gamma \frac{n^{1+1/\varepsilon}}{1+1/\varepsilon}, \tag{61}$$

where ε is the common elasticity of labor supply which we fixed at 0.33 following Chetty (2012). Together, assumptions (57), (59), (60) and (61) identify the model. We estimate the model by Maximum Likelihood.

Likelihood function. We can decompose the mixed joint density of a given realization $\{\omega, \iota^f, x\}$ of the random variables $\{\mathcal{W}, I^f, X\}$ into three elements:

$$f_{\mathcal{W},I^f,X}(\omega, \iota^f, x; B) = P(X = x) \times P_{I^f|X}(I^f = \iota^f \mid X = x; B) \times f_{\mathcal{W}|I^f,X}(\omega \mid \iota^f, x; B)$$

where B is the vector of parameters

$$B = \left(\beta, \varepsilon, \Gamma, \gamma_0^s, \gamma_1^s, \gamma_0^f, \sigma_\theta, \sigma_u, \sigma_\kappa, w_\kappa\right)$$

and the elements correspond to:

- $P(X = x_i)$ is the sampling weight s_i.
- $P_{I^f|X}(I^f = \iota^f \mid X = x; B)$ is the probability that someone with characteristics x takes the participation decision ι^f. The decision to participate in the formal sector ι^f depends on the productivity type θ and the participation cost κ. Let $\mathbf{i}(\theta, \kappa)$ denote the optimal participation decision. Then this probability can be rewritten as

$$\begin{aligned} P_{I^f|X}(I^f = \iota^f \mid X = x; B) &= \int_0^1 P_{I^f|\theta}\left(I^f = \iota^f \mid \theta; B\right) f_{\theta|X}(\theta \mid x; B) d\theta \\ &= \int_0^1 \left(\int_0^\infty I_{(\mathbf{i}(\theta,\kappa)=\iota^f)} g_\theta(\kappa) d\kappa\right) f_{\theta|X}(\theta \mid x; B) d\theta \\ &= \int_0^1 \left(\iota^f + (-1)^{\iota^f} G_\theta(\tilde{\kappa})\right) f_{\theta|X}(\theta \mid x; B) d\theta \end{aligned}$$

where $I_{(a)}$ is an indicator function that takes the value of 1 if the condition a is satisfied; $\tilde{\kappa}$ is the threshold value of the participation cost; $f_{\theta|X}$ is given by a normal distribution $N(X\beta, \sigma_\theta)$; and $g_\theta(\tilde{\kappa})$ is given by (59).

- $f_{\mathcal{W}|I^f,X}(\omega \mid \iota^f, x; B)$ is the likelihood that a worker with characteristics x and observed participation ι^f has a measured wage of ω (at the sector indicated by ι^f). This probability can be written as

$$f_{\mathcal{W}|I^f,X}(\omega \mid \iota^f, x; B) = \int_0^1 f_{\mathcal{W}|I^f,\theta}\left(\omega \mid \iota^f, \theta; B\right) f_{\theta|I^f,X}(\theta \mid \iota^f, x; B) d\theta$$

where

$$f_{\mathcal{W}|I^f,\theta}\left(\omega \mid \iota^f, \theta; B\right) = \begin{cases} N\left(\log(\omega) - \log(w^f(0) - \rho^f\theta, \sigma_u\right) & \text{if } \iota^f = 1 \\ N\left(\log(\omega) - \log(w^s(0) - \rho^s\theta, \sigma_u\right) & \text{else} \end{cases}$$

and

$$f_{\theta|I^f,X}(\theta \mid \iota^f, x; B) = \frac{P_{I^f|\theta}\left(I^f = \iota^f \mid \theta; B\right) f_{\theta|X}(\theta \mid x; B)}{P_{I^f|X}(I^f = \iota^f \mid X = x; B)}$$

Parameter estimates. The parameter estimates are reported in Table 3. The estimated density of types as well as the fit of the model along the shadow economy participation margin are shown in Figure 5 in the main text.

Table 3: Parameter estimates

preferences		productivity schedules					distributions of θ and κ				
ε	Γ	$w^s(0)$	ρ^s	$w^f(0)$	ρ^f	α_w	σ_θ	σ_κ	α_κ	w_κ	σ_u
0.33	0.032	0.006	2.90	0.003	4.64	2.25	0.09	1.38	0.88	0.018	0.53
(-)	(8e-4)	(1e-4)	(.06)	(1e-4)	(.06)	(.03)	(2e-3)	(0.03)	(.01)	(2e-4)	(3e-3)
β individual characteristics								β worker-firm			
Gender	Age	Age^2	Educ	Degree	Work	Exper	1stJob	Union	Agency		
-0.08	0.04	-5e-4	0.02	0.05	0.02	6e-5	2e-4	0.17	-0.015		
(2e-3)	(1e-4)	(6e-6)	(5e-4)	(1e-3)	(7e-4)	(6e-5)	(2e-5)	(4e-3)	(3e-4)		
β job characteristics											
S-Man	S-Fin	S-Ret	B-city	Size	Lib	Admin	Seller	Services	Senior		
-0.04	0.14	-0.012	0.10	0.11	0.25	-5e-3	4e-3	-0.02	7e-4		
(1e-3)	(3e-3)	(3e-4)	(3e-3)	(2e-3)	(6e-3)	(1e-4)	(1e-4)	(6e-4)	(1e-5)		

Standard errors are reported in brackets. Standard errors are obtained by Case Resampling Bootstrap using 150 draws.

References

Allingham, M. G. and A. Sandmo (1972): "Income tax evasion: A theoretical analysis," *Journal of Public Economics*, 1, 323–338.

Andreoni, J., B. Erard, and J. Feinstein (1998): "Tax compliance," *Journal of Economic Literature*, 36, 818–860.

Bachas, P. J., L. Gadenne, and A. Jensen (2020): "Informality, Consumption Taxes and Redistribution," .

Balán, J., H. L. Browning, and E. Jelin (1973): "Men in a developing society," .

Beaudry, P., C. Blackorby, and D. Szalay (2009): "Taxes and employment subsidies in optimal redistribution programs," *American Economic Review*, 99, 216–42.

Benabou, R. (2000): "Unequal societies: Income distribution and the social contract," *American Economic Review*, 90, 96–129.

Berger, M., G. Fellner-Röhling, R. Sausgruber, and C. Traxler (2016): "Higher taxes, more evasion? Evidence from border differentials in TV license fees," *Journal of Public Economics*, 135, 74–86.

Boadway, R., M. Marchand, and P. Pestieau (1994): "Towards a theory of the direct-indirect tax mix," *Journal of Public Economics*, 55, 71–88.

Boadway, R. and M. Sato (2009): "Optimal tax design and enforcement with an informal sector," *American Economic Journal: Economic Policy*, 1, 1–27.

Chetty, R. (2009): "Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods," *Annual Review of Economics*, 1, 451–488.

——— (2012): "Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply," *Econometrica*, 80, 969–1018.

Commander, S., N. Isachenkova, and Y. Rodionova (2013): "Informal employment dynamics in Ukraine: An analytical model of informality in transition economies," *International Labour Review*, 152, 445–467.

Cremer, H. and F. Gahvari (1996): "Tax evasion and the optimum general income tax," *Journal of Public Economics*, 60, 235–249.

De Mel, S., D. McKenzie, and C. Woodruff (2013): "The demand for, and consequences of, formalization among informal firms in Sri Lanka," *American Economic Journal: Applied Economics*, 5, 122–50.

De Paula, A. and J. A. Scheinkman (2010): "Value-added taxes, chain effects, and informalit," *American Economic Journal: Macroeconomics*, 2, 195–221.

Di Nola, A., G. Kocharkov, A. Scholl, and A.-M. Tkhir (2020): "The Aggregate Consequences of Tax Evasion," .

Diamond, P. A. (1998): "Optimal income taxation: an example with a U-shaped pattern of optimal marginal tax rates," *American Economic Review*, 83–95.

Doligalski, P. and L. Rojas (2016): "Optimal Redistribution with a Shadow Economy," *EUI Working Papers.*

Ebert, U. (1992): "A reexamination of the optimal nonlinear income tax," *Journal of Public Economics*, 49, 47–73.

Emran, M. S. and J. E. Stiglitz (2005): "On selective indirect tax reform in developing countries," *Journal of Public Economics*, 89, 599–623.

Feldstein, M. S. (1969): "The effects of taxation on risk taking," *Journal of Political Economy*, 77, 755–764.

French, E. and C. Taber (2011): "Identification of models of the labor market," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 4, 537–617.

Fudenberg, D. and J. Tirole (1991): "Game theory," *Cambridge, MA: MIT Press.*

Goldberg, P. K. and N. Pavcnik (2003): "The response of the informal sector to trade liberalization," *Journal of Development Economics*, 72, 463–496.

Gomes, R., J.-M. Lozachmeur, and A. Pavan (2017): "Differential taxation and occupational choice," *The Review of Economic Studies*, rdx022.

Gorodnichenko, Y., J. Martinez-Vazquez, and K. S. Peter (2009): "Myth and Reality of Flat Tax Reform: Micro Estimates of Tax Evasion Response and Welfare Effects in Russia," *Journal of Political Economy*, 117, 504–554.

Guataquí, J. C., A. F. García, and M. Rodríguez (2010): "El Perfil de la Informalidad Laboral en Colombia," *Perfil de Coyuntura Económica.*

Heckman, J. J. and B. E. Honore (1990): "The empirical content of the Roy model," *Econometrica*, 1121–1149.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006): "Earnings functions, rates of return and treatment effects: The Mincer equation and beyond," *Handbook of the Economics of Education*, 1, 307–458.

Henley, A., G. R. Arabsheibani, and F. G. Carneiro (2009): "On defining and measuring the informal sector: Evidence from Brazil," *World development*, 37, 992–1003.

Huang, J. and J. Rios (2016): "Optimal tax mix with income tax non-compliance," *Journal of Public Economics*, 144, 52–63.

Hussmanns, R. and B. d. Jeu (2002): "ILO Compendium of official statistics on employment in the informal sector," Ilo working papers, International Labour Organization.

ILO (2013): "Measuring informality: A statistical manual on the informal sector and informal employment," Tech. rep., International Labour Organization.

——— (2014): "Trends in informal employment in Colombia: 2009 - 2013," Tech. rep., International Labour Organization.

——— (2018): "Women and men in the informal economy: a statistical picture (third edition)," Tech. rep., International Labour Organization.

Jacobs, B. (2015): "Optimal Inefficient Production," *mimeo, Erasmus University Rotterdam.*

Jacquet, L., E. Lehmann, and B. Van der Linden (2013): "Optimal redistributive taxation with both extensive and intensive responses," *Journal of Economic Theory*, 148, 1770–1805.

Kim, B.-Y. (2005): "Poverty and informal economy participation: Evidence from Romania," *Economics of Transition*, 13, 163–185.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011): "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark," *Econometrica*, 79, 651–692.

Kleven, H. J., C. T. Kreiner, and E. Saez (2009): "The optimal income taxation of couples," *Econometrica*, 77, 537–560.

Kleven, H. J., W. F. Richter, and P. B. Sørensen (2000): "Optimal taxation with household production," *Oxford Economic Papers*, 52, 584–594.

Kopczuk, W. (2001): "Redistribution when avoidance behavior is heterogeneous," *Journal of Public Economics*, 81, 51–71.

Leal Ordóñez, J. (2014): "Tax collection, the informal sector, and productivity," *Review of Economic Dynamics*, 17, 262–286.

López García, I. (2015): "Human capital and labor informality in Chile: a life-cycle approach," Tech. rep., RAND working paper series.

Magnac, T. (1991): "Segmented or competitive labor markets," *Econometrica*, 165–187.

Meghir, C., R. Narita, and J.-M. Robin (2015): "Wages and informality in developing countries," *American Economic Review*, 105, 1509–1546.

Mirrlees, J. A. (1971): "An exploration in the theory of optimum income taxation," *The Review of Economic Studies*, 175–208.

Monteiro, J. C. and J. J. Assunção (2012): "Coming out of the shadows? Estimating the impact of bureaucracy simplification and tax cut on formality in Brazilian microenterprises," *Journal of Development Economics*, 99, 105–115.

Mora, J. and J. Muro (2017): "Dynamic Effects of the Minimum Wage on Informality in Colombia," *LABOUR*.

Mussa, M. and S. Rosen (1978): "Monopoly and product quality," *Journal of Economic Theory*, 18, 301–317.

Olovsson, C. (2015): "Optimal taxation with home production," *Journal of Monetary Economics*, 70, 39–50.

Piketty, T. and E. Saez (2013): "Optimal Labor Income Taxation," in *Handbook of Public Economics*, ed. by A. J. Auerbach, R. Chetty, M. Feldstein, and E. Saez, Newnes, vol. 5, 391.

Piketty, T., E. Saez, and S. Stantcheva (2014): "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities," *American Economic Journal: Economic Policy*, 6, 230–271.

Pratap, S. and E. Quintin (2006): "Are labor markets segmented in developing countries? A semiparametric approach," *European Economic Review*, 1817–1841.

Rocha, R., G. Ulyssea, and L. Rachter (2018): "Do lower taxes reduce informality? Evidence from Brazil," *Journal of Development Economics*, 134, 28–49.

Rochet, J.-C. and P. Choné (1998): "Ironing, sweeping, and multidimensional screening," *Econometrica*, 783–826.

Rothschild, C. and F. Scheuer (2013): "Redistributive taxation in the Roy model," *The Quarterly Journal of Economics*, 128, 623–668.

——— (2014): "A Theory of Income Taxation under Multidimensional Skill Heterogeneity," Tech. rep., National Bureau of Economic Research.

——— (2016): "Optimal taxation with rent-seeking," *The Review of Economic Studies*, 83, 1225–1262.

Saez, E. (2001): "Using elasticities to derive optimal income tax rates," *The Review of Economic Studies*, 68, 205–229.

Scheuer, F. (2014): "Entrepreneurial taxation with endogenous entry," *American Economic Journal: Economic Policy*, 6, 126–163.

Scheuer, F. and I. Werning (2017): "The taxation of superstars," *The Quarterly Journal of Economics*, 132, 211–270.

Schneider, F., A. Buehn, and C. E. Montenegro (2011): "Shadow Economies all over the World: New Estimates for 162 Countries from 1999 to 2007," in *Handbook on the shadow economy*, ed. by F. Schneider, Edward Elgar Cheltenham, 9–77.

Schneider, F. and D. H. Enste (2000): "Shadow economies: Size, causes, and consequences," *Journal of Economic Literature*, 38, 77.

Selin, H. and L. Simula (2020): “Income shifting as income creation?” *Journal of Public Economics*, 182, 104081.

Slemrod, J. and W. Kopczuk (2002): “The optimal elasticity of taxable income,” *Journal of Public Economics*, 84, 91–112.

Slemrod, J. and S. Yitzhaki (2002): “Tax avoidance, evasion, and administration,” *Handbook of Public Economics*, 3, 1423–1470.

Statistics Poland (2019): “Unregistered employment in Poland in 2017,” Tech. rep., Statistics Poland.

Tazhitdinova, A. (2017): “Increasing Hours Worked: Moonlighting Responses to a Large Tax Reform,” *Available at SSRN 3047332.*

UN General Assembly (2017): “Work of the Statistical Commission Pertaining to the 2030 Agenda for Sustainable Development (a/res/71/313),” *UN General Assembly: New York, NY, USA.*