TEXTO PARA DISCUSSÃO

No. 590

Audits or Distortions: The Optimal Scheme to Enforce Self-Employment Income Taxes

Eduardo Zilberman

DEPARTAMENTO DE ECONOMIA
www.econ.puc-rio.br

Audits or Distortions: The Optimal Scheme to Enforce Self-Employment Income Taxes*

Eduardo Zilberman
Department of Economics
PUC-Rio

November 2011

Abstract

I investigate the optimal auditing scheme for a revenue-maximizing tax-collection agency that observes not only reported profits, but also the level of employment at each firm. Each firm is owned by a single entrepreneur whose managerial ability is random. The optimal auditing scheme is discontinuous and non-monotone in ability. In intermediate audit costs, less-productive entrepreneurs face auditing probabilities that increase in ability, whereas the ablest ones are not audited. I argue that if the optimal auditing scheme were adopted in practice, net revenue collected from nonfarm sole proprietors would increase by at least 59 percent.

Keywords: optimal auditing, tax evasion, entrepreneurship, mechanism design.

JEL Classification: D21, H26, K42.

*E-mail: zilberman@econ.puc-rio.br. I thank Ennio Stacchetti for insightful guidance. I also thank Juliano Assunção, Tiago Berriel, Ryan Booth, Vinicius Carrasco, Bernardo da Silveira, Edoardo di Porto, Bruno Ferman, Raquel Fernández, Vivian Figer, Sean Flynn, Renato Gomes, Pedro Hemsley, Boyan Jovanovic, Alessandro Lizzeri, Thomas Mertens, Giri Parameswaran, Nicola Persico, Leonardo Rezende, Nikita Roketskiy, Tomasz Sadzik, Daniel Xu, Gianluca Violante, Joyce Wong, and participants at conferences and seminars for helpful comments and discussions. I am especially indebted to Saki Bigio, whose comments and discussions were invaluable. All errors are mine.

1 Introduction

The literature on optimal income tax enforcement has focused mainly on taxpayers whose income is exogenous or salaried workers.[1] However, in the U.S., taxes on wages and salaries are subject to employer withholding and, thus, almost perfectly enforced. Indeed, Slemrod [2007] reports that only one percent of wages and salaries are underreported to the tax-collection agency (henceforth the IRS). In contrast, 43 percent of individual business income is underreported.[2] This evidence suggests that, in order to enforce income taxes, the IRS should concentrate on developing a better strategy to monitor business proprietors.

This paper asks the following: Given that reported income and a **single factor of production** are observable, how should a revenue-maximizing IRS monitor heterogeneous entrepreneurs? The source of heterogeneity is a random managerial ability, which is private information. Once the monitoring strategy is conditioned on a single input, the IRS can indirectly distort production in order to provide incentives, which enriches its set of tools to enforce self-employment income taxes.

I interpret this single factor of production as labor input. Since wage taxes are almost perfectly enforced, the number of workers at each firm seems to be easily observable by the IRS. Using data on entrepreneurship and employment from the U.S., I argue that if distortions and audits were optimally combined, net revenue collected from nonfarm sole proprietors would increase by at least 59 percent.

In Section 2, I adapt the two-stage game developed in Bigio and Zilberman [2011] to study optimal self-employment income tax enforcement. A self-employed individual is a

[1] The seminal paper is Reinganum and Wilde [1985], inspired by the costly state verification model of Townsend [1979]. Notable contributions are Border and Sobel [1987], Melumad and Mookherjee [1989], Mookherjee and Png [1989], Cremer et al. [1990], Sánchez and Sobel [1993], Cremer and Gahvari [1996], Macho-Stadler and Pérez-Castrillo [1997], Chander and Wilde [1998], and Bassetto and Phelan [2008]. The first theoretical work on tax noncompliance is Allingham and Sandmo [1972], which builds on Becker [1968]'s work on the economics of crime. Recent surveys are Andreoni et al. [1998], Slemrod and Yitzhaki [2002], and Sandmo [2005].

[2] These figures account for ten billion and 109 billion dollars, respectively. Kleven et al. [2011], for instance, define tax-evasion rate as the share of reported income that is underreported and calculate it for Denmark. The tax-evasion rate is 14.9 percent for self-employment income, 1.1 percent for personal income, 41.6 percent for self-reported income, and 0.3 percent for third-party-reported income.

risk-neutral entrepreneur who owns and manages a single firm. An entrepreneur experiences a random managerial ability – her privately observed type – that enhances productivity in a plant exhibiting decreasing returns to scale. Production is carried out by a team of workers. Entrepreneurs can underreport income – that is, the profits generated by their own firms – in order to evade taxes. If a firm is audited, true income is uncovered, and the entrepreneur must pay a penalty.

In the first stage, given the distribution of managerial ability, the IRS commits to a costly monitoring strategy, dependent on both reported income and labor input, in order to maximize net revenue. In the second stage, entrepreneurs take into account this monitoring strategy and maximize expected profits by choosing labor input and reported income. Hence, labor is not only a factor of production, but also a signal of true income. At a production cost, labor can be strategically distorted to signal a lower income to the IRS.

To solve this model, I adopt a mechanism design approach, in which the choice of labor and reported income are delegated to the IRS, which is the principal. Due to the revelation principle, it is enough to focus within the class of direct mechanisms that respect incentive compatibility and individual rationality. That is, an agent reports her type (i.e., her managerial ability) and is then assigned a labor input to employ, an amount to report as profits, and a probability with which she will be audited. The mechanism is designed such that agents truthfully report their types and derive at least their reservation values.

The main finding is that the optimal auditing scheme may be **non-monotone** in ability. In intermediate audit costs, less-productive entrepreneurs face auditing probabilities that increase in managerial ability, whereas the ablest ones are not audited. At some threshold level of ability, the optimal auditing scheme discontinuously drops.

There are two driving forces behind this non-monotonicity result. The first is the reservation value, which is type-dependent. Since an entrepreneur can declare her true profits and pay the right amount of taxes, her reservation value is the post-tax truthfully declared profits, which is increasing in managerial ability. Hence, entrepreneurs face countervailing incentives. On the one hand, an entrepreneur is willing to understate her type in order to pay less in taxes. On the other hand, she is willing to overstate her type in order to be assigned a higher reservation value by the principal. Moreover, the incentives to understate (overstate) are relatively stronger for high-ability (low-ability) types. In this environment, a non-monotone monitoring strategy may be optimal as long as incentive compatibility is not violated.

The second driving force is the joint condition of the monitoring strategy on both reported income and labor input. If it depends only on reported income, this model collapses to a variant of Sánchez and Sobel [1993], in which the optimal monitoring strategy must be non-increasing in order to not violate incentive compatibility. Every entrepreneur below a certain threshold type is audited with constant probability, whereas those above it are not monitored at all. In contrast, if the monitoring strategy depends only on labor input, as in Bigio and Zilberman [2011], the optimal monitoring strategy is non-decreasing. Those types below a certain threshold are not monitored at all, whereas those above it are audited with constant probability. In Sections 3 and 4, I discuss the role of each driving force in detail.

The role of costly audits as a tool to maximize government revenue is twofold: First, it enforces taxes from those that are audited; second, it provides incentives by preventing misreport from other types, which allows the IRS to require higher income declarations from them. Similarly, labor distortions can be used to provide incentives, but only at a production cost that diminishes revenue collection. The optimal mechanism balances the use of these two tools in a way that preserves incentive compatibility and maximizes net revenue collection.

The optimal mechanism has the following properties: (1) As in a standard mechanism-design problem, the top-type is not distorted; (2) in the top range of the type distribution, audits are never used, and labor is distorted downwards to provide incentives; (3) below some threshold type, stronger incentives for entrepreneurs to overstate their types and, thus, to be assigned a higher reservation value limit the further use of distortions to provide incentives; (4) if the audit cost is too high, only labor distortions are used to provide incentives; (5) if the audit cost is not too high, audits and labor distortions are optimally combined to enforce taxes; in the bottom range of the type distribution, both the optimal monitoring strategy and labor schedule are increasing in ability; at the threshold type, they drop discontinuously; (6) every entrepreneur evades taxes; and (7) the effective tax rate is higher in the middle of the type distribution; thus, the overall regressive (or progressive) bias that arises from evasion is unknown.

In Section 4, I present an example that summarizes the intuition behind these results, which are analytically derived in Section 5. In Section 6, I discuss some generalizations to the model.

In Section 7, I quantitatively explore the implications of this model. In particular, I establish a lower bound to the net revenue increase if the mechanism developed in this paper were counterfactually adopted. Under the assumption that the U.S. is a "relatively distortion-free" competitive economy,[3] I use data on employment and entrepreneurship from the Survey of Consumer Finance to impose some discipline on the managerial-ability distribution.

Results suggest that, once adopted, the optimal mechanism can substantially increase revenue and reduce evasion in the U.S. Among nonfarm sole proprietors, for a conservative choice of parameters, revenue collection increases by at least 59 percent, and the fraction of reported income is at least 86 percent, as opposed to 43 percent documented in Slemrod [2007]. Section 8 concludes.

2 Model

I consider a two-stage game in which entrepreneurs remit taxes to the IRS. Taxes may potentially be evaded. In the first stage, the IRS commits to a monitoring strategy that depends both on labor input and reported income. In the second stage, entrepreneurs take into account the monitoring strategy and choose labor input and reported income.

[3] By relatively distortion-free, I mean an economy in which policies do not target the firm size. By competitive, I mean an economy in which entrepreneurs take prices as given and maximize expected profits.

There is a continuum of firms of measure one. Each firm is owned and managed by a single entrepreneur, who experiences a random managerial ability z, which is her privately observed type. I assume that z is independently and identically distributed according to G twice continuously differentiable, with density $g = G'$ uniformly bounded away from zero, and compact support $[\underline{z}, \overline{z}]$ with $\underline{z} \geq 0$.

There is a single good produced with a variable factor, labor n which is observable, and a fixed factor, managerial ability z. Hence, production displays decreasing returns to scale, which are important to generate positive profits in a competitive environment. Moreover, as Lucas [1978] points out, heterogeneity in managerial ability generates a firm-size distribution, which allows the IRS to screen over n.

The production technology is zn^{α}, with $\alpha \in (0,1)$ common to all firms. This functional form is chosen for tractability. I discuss the consequences of adopting a more general production function in Section 6.1. In particular, I show that the production technology can be generalized to $zn^{\alpha_0}\prod_{i=1}^{I} k_i^{\alpha_i}$, with $\sum_{i=0}^{I}\alpha_i \in (0,1)$ and $\alpha_i \geq 0$ for $i = 0, .., I$, as long as other inputs $\{k_i\}_{i=1}^{I}$ are not observable.[4] This generalization extends the scope of the model's application. In some contexts, labor might not be readily observable. However, n can be interpreted as any other observable factor of production.[5]

Let wages be the numeraire, and p be the price of the good; thus, pre-tax profits are $\pi(n,z) = pzn^{\alpha} - n$. Notice that the efficient level of employment is $n^*(z) = (\alpha p z)^{\frac{1}{1-\alpha}}$.

A profit tax rate τ is imposed exogenously by the government. After observing her own type z, the entrepreneur decides how much labor to hire and income to report to the IRS. Let reported profits be $x \geq 0$, so τx is the amount the entrepreneur pays out as taxes, and $\tau(\pi(n,z) - x)$ is the amount she evades. The IRS (the principal) costlessly observes labor n and reported income x. However, it is able to observe ability z and, hence, actual income,

[4]Even if other inputs are observable, it is enough to assume that the IRS does not condition its monitoring strategy on them.

[5]In some sectors, for instance, unskilled workers such as illegal immigrants are not readily observable. However, the IRS might still observe another input, such as electricity bills, skilled labor or an intermediate good bought from a formal firm.

only if it audits the firm at a constant cost $c > 0$. If an entrepreneur is audited, she is assessed by $\max\{\mu\tau(\pi(n,z) - x), 0\}$, where $\mu > 1$ is a linear penalty on the amount evaded.[6] Penalties are assumed to be linear for tractability; otherwise, I would not be able to rewrite the IRS problem in terms of informational rents, a trick that simplifies the solution.

Note that the IRS does not reward overreporting. Hence, without loss of generality, I restrict the set of reported income to be $[0, \pi(n,z)]$, and set $\max\{\mu\tau(\pi(n,z) - x), 0\} = \mu\tau(\pi(n,z) - x)$.

In this paper, the IRS is an agency responsible only for auditing and collecting taxes. Choosing tax rates and penalties is beyond its scope.[7] In particular, taxes would be fully enforced without cost if penalties were arbitrarily large – that is, $\mu \to \infty$. However, many authors argue that an abusive use of penalties is limited by other factors, such as a common ethical norm[8] or, more economically, the need to restrain the power of corruptible self-interested enforcers.[9]

The IRS knows the distribution of managerial ability, G. In the first stage, in order to maximize expected net revenue, the IRS commits to a monitoring strategy, which is an audit probability function, $\varphi(n, x)$, that depends on both employment and reported income.

As Andreoni et al. [1998] argue, assuming that the IRS objective is to maximize expected net revenue, instead of a welfare criterion, seems a reasonable positive description of how many tax agencies behave in practice. However, most tax agencies do not explicitly commit to a monitoring strategy that depends on available information. Thus, I justify this assumption on a normative ground. If net revenue collection is the main concern, as in periods of high

[6]Implicitly, I assume that all penalties are enforced even if $\mu\tau(\pi(n,z)-x) > \pi(n,z)-\tau x$; that is, penalties are higher than post-tax profits. If limited liability is a concern, I could assume that $\mu \in (1, \frac{1}{\tau}]$, which is enough to guarantee that penalties are payable only with post-tax profits.

[7]These variables are usually chosen by other government entities such as the Treasury or Congress. For example, in August 2007, the U.S. Government Accountability Office (U.S. GAO) published a report (http://www.gao.gov/new.items/d071062.pdf) suggesting that the Congress require the IRS to periodically adjust penalties for inflation.

[8]Rosen [2005], for example, argues that "existing penalty systems try to incorporate *just retribution*. Contrary to the assumptions of the utilitarian framework, society cares not only about the end result (getting rid of the cheaters) but also the processes by which the result is achieved."

[9]See Polinsky and Shavell [2000] for a survey.

budget deficits, the best the IRS can do is to commit to a monitoring strategy that depends on all costlessly observable variables.[10]

In the second stage, given φ, the entrepreneur's problem is to choose $n \geq 0$ and $x \in [0, \pi(n,z)]$ in order to maximize her expected profits:

$$\Pi(n, x, \varphi(n,x), z) \equiv \pi(n,z) - \tau x - \varphi(n,x)\mu\tau(\pi(n,z) - x).$$

Notice that labor is not only a factor of production, but also a signal of the true income. Hence, at a production cost, labor input can be strategically distorted to signal a lower income to the IRS.

Before proceeding with the analysis, I solve for the *full-information* case, in which z is observable. Let $\varphi^*(n,x,z)$ be the full-information monitoring strategy, which is also a function of z. In the second stage, an entrepreneur z weakly prefers to declare her true profits $\pi(n,z)$ rather than underreport $x < \pi(n,z)$ whenever $\varphi^*(n,x,z) \geq \frac{1}{\mu}$.[11] Similarly, the IRS weakly prefers that an entrepreneur z declares her true profits $\pi(n,z)$ rather than underreport $x < \pi(n,z)$ whenever it chooses $\varphi^*(n,x,z) \leq \frac{1}{\mu}$ in the first stage.[12]

Consequently, the best the IRS can do is to induce efficient production and truthful income declarations, without spending auditing resources. This is achieved by the following monitoring strategy:

$$\varphi^*(n,x,z) = \begin{cases} \frac{1}{\mu} & \text{if } x \neq \pi(n^*(z), z) \\ 0 & \text{otherwise} \end{cases}.$$

As long as the IRS commits to it, all taxes are enforced at no cost.[13] Information on employment at each firm is not necessary to achieve the optimum, which is implemented

[10]See Reinganum and Wilde [1986] and Erard and Feinstein [1994] for models in which the IRS does not commit to a monitoring strategy.

[11]Indeed, by comparing expected profits, $(1-\tau)\pi(n,z) \geq \pi(n,z) - \tau x - \varphi^*(n,x,z)\mu\tau(\pi(n,z) - x)$ if and only if $\varphi^*(n,x,z) \geq \frac{1}{\mu}$.

[12]Indeed, by comparing expected revenue, $\tau\pi(n,z) \geq \tau x + \varphi^*(n,x,z)\mu\tau(\pi(n,z) - x)$ if and only if $\varphi^*(n,x,z) \leq \frac{1}{\mu}$.

[13]Recall that $n^*(z)$ is the efficient labor, and it maximizes $(1-\tau)\pi(n,z)$. I use the superscript $*$ to denote the full-information solution in order to highlight that it induces efficient employment per firm.

only through off-equilibrium threats.

If the IRS does not observe z, an adverse selection problem arises. In order to increase her expected profits, an entrepreneur may distort her labor decision and report less income. To solve this problem, I adopt a mechanism-design approach. Since the IRS observes (n, x), which can be interpreted as the message sent by the agent, an entrepreneur z chooses:

$$(n(z), x(z)) \in \underset{n \geq 0, x \in [0, \pi(n,z)]}{\arg\max} \quad \pi(n, z) - \tau x - \varphi(n, x)\mu\tau(\pi(n, z) - x). \tag{1}$$

As opposed to other mechanism-design applications, such as monopoly screening, the principal does not control the agents' choice variables (n and x in this paper). Hence, (1) is equivalent to two sets of incentive compatibility constraints:

$$\begin{array}{ll} \text{(IC)} & \Pi(n(z), x(z), \phi(z), z) \geq \Pi(n(\tilde{z}), x(\tilde{z}), \phi(\tilde{z}), z), \forall (z, \tilde{z}) \in [\underline{z}, \overline{z}] \times [\underline{z}, \overline{z}] \\ \text{(IC-out)} & \Pi(n(z), x(z), \phi(z), z) \geq \Pi(n, x, \varphi(n, x), z), \forall (z, n, x) \text{ such that} \\ & z \in [\underline{z}, \overline{z}], n \geq 0, x \in [0, \pi(n, z)] \text{ and } (n, x) \neq (n(\tilde{z}), x(\tilde{z}))\ \forall \tilde{z} \in [\underline{z}, \overline{z}], \end{array}$$

where $\phi(z) = \varphi(n(z), x(z))$ is the direct monitoring strategy.

A variant of the revelation principle is derived, which states that it is enough to restrict attention to the class of direct mechanisms that not only induces truth-telling due to (IC), but also deters agents to choose out-of-equilibrium menus, which is captured by (IC-out). In other words, without lost of generality, an agent reports her type, say $\tilde{z}$, and is then assigned a menu $\{n(\tilde{z}),\ x(\tilde{z}),\ \phi(\tilde{z})\}$. The direct mechanism $\{n(z), x(z), \phi(z)\}_z$ is designed such that an entrepreneur reports her type truthfully (i.e., $\tilde{z} = z$), and gets a payoff higher than those associated with all possible off-equilibrium deviations.[14]

The set of out-of-equilibrium incentive compatibility constraints (IC-out) merits some digression. It is optimal for the IRS to relax (IC-out) as much as possible by punishing,

[14] Notice that reported income is an observed choice variable instead of an unobserved type or action. The revelation principle implies truthful revelation of types, so income can be misreported in equilibrium.

through sufficiently high monitoring probabilities $\varphi(n,x)$, off-scheduled deviations. However, one particular deviation does not depend on the auditing intensity; namely, to declare true profits, $x = \pi(n,z)$, and pay the right amount of taxes. If this is the case, post-tax profits are $(1-\tau)\pi(n,z)$, which are maximized at $n = n^*(z)$; thus, any mechanism must assign at least $(1-\tau)\pi(n^*(z),z)$ to the entrepreneur.

In principle, a type z entrepreneur could also deviate to other out-of-equilibrium allocations. However, given that any mechanism must assign at least $(1-\tau)\pi(n^*(z),z)$ to her, this problem is circumvented by setting $\varphi(n,x) = 1/\mu$ for all off-scheduled (n,x). Indeed, if $\varphi(n,x) = 1/\mu$, entrepreneurs prefer to declare their true profits instead of $x < \pi(n,z)$.[15] Consequently, (IC-out) can be replaced by

$$\text{(IR)} \quad \Pi(n(z), x(z), \phi(z), z) \geq (1-\tau)\pi(n^*(z), z), \forall z \in [\underline{z}, \overline{z}].$$

I call this set of constraints individual rationality (IR) because it states that any mechanism must assign at least a reservation value of $(1-\tau)\pi(n^*(z),z)$ to an entrepreneur z. It cannot be interpreted as a participation constraint, but rather as a constraint that preserves out-of-equilibrium incentive compatibility.

Figure 1 illustrates the role (IR) plays in the model. Suppose that the curve depicted on the left (n,x)-plan represents a truth-telling, direct mechanism, such that each point in this curve is associated with a single type z and, thus, with an audit probability $\phi(z)$. If any pair (n,x) outside this curve is audited with intensity $1/\mu$, (IR) ensures that entrepreneurs stick to the curve. In other words, through off-equilibrium threats, the IRS can implement any truth-telling, direct mechanism that respects (IR).

However, in practice, it is not sensible policy to intensively audit everyone who reports much more than expected. The right (n,x)-plan of Figure 1 shows an alternative way to implement a truth-telling, direct mechanism that respects (IR). Those that report a lower-than-expected income are audited intensively. Those that report above some threshold curve,

[15] Recall that $(1-\tau)\pi(n,z) \geq \pi(n,z) - \tau x - \varphi(n,x)\mu\tau(\pi(n,z) - x)$ if and only if $\varphi(n,x) \geq 1/\mu$.

the dashed line, are not audited. In the shaded region between the dashed and the solid lines, auditing intensities are set to deter off-scheduled deviations, but not necessarily equal to $1/\mu$.

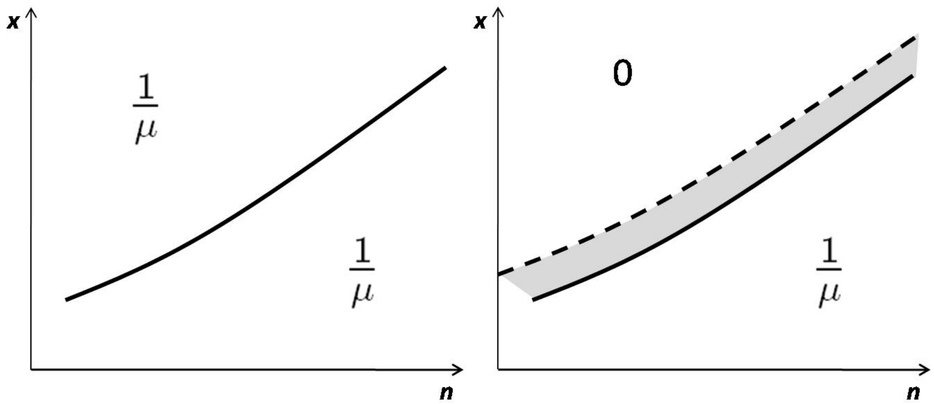

Figure 1: Implementation.

I am ready to state the IRS problem. Assume that n, x, and ϕ are piecewise continuously differentiable functions defined on $[\underline{z}, \overline{z}]$. Given G, the IRS problem is the following:

$$\max_{\{n(z),x(z),\phi(z)\}_z} \int_{\underline{z}}^{\overline{z}} \{\tau x(z) + \phi(z)[\mu\tau(\pi(n(z), z) - x(z)) - c]\}\, dG(z)$$

s.t.

(IC) $\Pi(n(z), x(z), \phi(z), z) \geq \Pi(n(\tilde{z}), x(\tilde{z}), \phi(\tilde{z}), z), \forall (z, \tilde{z}) \in [\underline{z}, \overline{z}] \times [\underline{z}, \overline{z}]$

(IR) $\Pi(n(z), x(z), \phi(z), z) \geq (1 - \tau)\pi(n^*(z), z), \forall z \in [\underline{z}, \overline{z}]$

(F) $\phi(z) \in [0, 1]\,, x(z) \in [0, \pi(n(z), z)], n(z) \geq 0, \forall z \in [\underline{z}, \overline{z}]$.

That is, the IRS solves for the direct mechanism $\{n(z), x(z), \phi(z)\}_z$ that maximizes expected net revenue subject to (IC), (IR) and feasibility (F), which requires that the set of offered menus corresponds to feasible probabilities, income declarations, and labor input.

Using terminology from Lewis and Sappington [1989]'s, this model displays countervailing incentives. On the one hand, an entrepreneur has incentives to underreport z and pay less in taxes. On the other hand, since the reservation value, $(1 - \tau)\pi(n^*(z), z)$, is increasing in

z, an entrepreneur is also tempted to overstate z and be assigned a higher value. That is, in order to satisfy (IR), the principal tends to design more favorable allocations towards higher types; thus, an entrepreneur has incentives to overstate z and get a better allocation.[16]

Sánchez and Sobel [1993] and Bigio and Zilberman [2011] study optimal enforcement policies in a similar environment. The former paper assumes that income is exogenous; thus, auditing probabilities cannot be conditioned on employment. The latter conditions the monitoring strategy only on labor input. In these papers, the IRS is assigned a fixed budget, which is exhausted in order to maximize revenue.[17] Moreover, Sánchez and Sobel [1993] consider a more general tax schedule, while Bigio and Zilberman [2011] consider more general production and audit cost functions. For simplicity, I specify functional forms for these primitives, although in Section 6, I discuss the extent to which they can be relaxed. See these papers for further discussion of most of the assumptions used here.

The next proposition, adapted from Sánchez and Sobel [1993], serves as a benchmark for the rest of the analysis. It assumes that labor input is not observable. The proof follows the steps in Bigio and Zilberman [2011] and is omitted. Let the superscript x denote the optimal solution when the monitoring strategy is conditioned only on reported income.

[16] Countervailing incentives imply that the solution is not necessarily characterized by the full informational rent extraction of the lowest type, so standard tricks in the literature are not readily applicable here.

[17] Operationally, both environments are similar since a budget constraint of the form $\int \phi(z)c dG(z) \leq C$, where C is the assigned budget, can be cast in the principal's problem using a Lagrange multiplier, say $\xi \geq 0$. Hence, a revenue maximizer IRS optimizes $\int \{\tau x(z) + \phi(z)[\mu\tau(\pi(n(z), z) - x(z)) - \xi c]\} dG(z) + \xi C$ subject to (F), (IR), and (IC).

Proposition 1. *(Sánchez and Sobel [1993], adapted) If the monitoring strategy does not depend on labor input, then there always exists a solution to the model of the following form:*

$$
\begin{aligned}
n^x(z) = &\ n^*(z) \quad \textit{if } \underline{z} \le z \le \overline{z} \ , \\
\phi^x(z) = &\begin{cases} 1/\mu & \textit{if } \underline{z} \le z < z_x \\ 0 & \textit{if } z_x \le z \le \overline{z} \end{cases} , \\
x^x(z) = &\begin{cases} \pi(n^*(z), z) & \textit{if } \underline{z} \le z < z_x \\ \pi(n^*(z_x), z_x) & \textit{if } z_x \le z \le \overline{z} \end{cases} ,
\end{aligned}
$$

where $z_x \in [\underline{z}, \overline{z}]$.

In words, there is a threshold type z_x, such that the IRS monitors every type below z_x in a way that generates a truthful income report. In contrast, every type greater than, or equal to, z_x is not audited and reports z_x's profits. As a consequence, the most productive entrepreneurs are the set of evaders. Note that policy cannot distort labor input, so production is carried out efficiently.[18]

3 Implementability

Let U denote the informational rent, which is the expected profits minus the reservation value, an agent gets. Hence,

$$U(z) = \max_{\tilde{z} \in [\underline{z}, \overline{z}]} \left\{ \pi(n(\tilde{z}), z) - \tau x(\tilde{z}) - \phi(\tilde{z}) \mu \tau (\pi(n(\tilde{z}), z) - x(\tilde{z})) - (1-\tau)\pi(n^*(z), z) \right\}. \quad (2)$$

The following lemma is standard and states necessary and sufficient conditions for the incentive-compatibility constraint be globally satisfied. The proof is omitted.

[18] If the monitoring strategy depends only on labor input, as in Bigio and Zilberman [2011], the IRS audits the most-productive entrepreneurs in a way that generates efficient production and truthful income report. In contrast, lower types are not audited and report zero profits. Moreover, some of the lower types have their labor input distorted away from its efficient level in order to prevent higher types from mimicking them, which generates a *missing middle* – that is, medium firms are scarce.

Lemma 1. *Incentive compatibility is verified if and only if*

$$\text{(LIC):} \quad \frac{dU}{dz}(z) = (1-\phi(z)\mu\tau)pn(z)^{\alpha} - (1-\tau)pn^{*}(z)^{\alpha} \ \text{a.e.},$$
$$\text{(M):} \quad (1-\phi(z)\mu\tau)n(z)^{\alpha} \ \text{is non-decreasing},$$

and that U is absolutely continuous.

These two conditions are crucial to understanding the results in this paper. The local incentive compatibility (LIC) follows from applying the envelope theorem[19] to (2) and evaluating the resulting equation at $\tilde{z} = z$. It specifies the required slope of the informational rent to induce truth-telling. Notice that (LIC) provides a clear interpretation of countervailing incentives. The first term captures the incentives to understate z and, thus, pay less in taxes, whereas the second term captures the temptation to overstate z and, thus, be assigned a higher reservation value.

(M) is a variant of the monotonicity condition present in the mechanism-design literature. If, for example, equilibrium audits are ruled out from the problem – that is, $\phi(z) = 0$ for all z – then (M) collapses to $n(z)$ being non-decreasing. Intuitively, by setting high enough off-equilibrium auditing intensities, the IRS can always shape reported income x to work as if it were compensatory transfers, as in a textbook mechanism-design problem (e.g., Fudenberg and Tirole [1991]). Hence, a non-decreasing labor schedule is sufficient for implementability.

In contrast, if the monitoring strategy does not depend on n, then labor distortions cannot be used to provide incentives. Therefore, (M) collapses to $\phi(z)$ being non-increasing. This is a standard property in the optimal-tax-enforcement literature.[20] It prevents higher types from mimicking a low type in order to pay less in taxes.

By combining the use of both labor distortions and auditing intensities to provide incentives, standard monotone conditions can be relaxed, while incentive compatibility is still

[19] See Milgrom and Segal [2002].

[20] When taxpayers are risk-averse, Mookherjee and Png [1989] show that the monotonicity of the monitoring strategy may not hold.

satisfied. Indeed, for a given z, it is possible to increase (or decrease) both $n(z)$ and $\phi(z)$ such that $(1-\phi(z)\mu\tau)n(z)^{\alpha}$ remains constant. In words, costly audits and labor distortions can be traded off without violating (LIC) and (M). How should audits and distortions be optimally combined in order to enforce taxes? I tackle this question in the next sections.

4 Example

This section shows an example that summarizes the intuition behind the results in this paper. I assume that z follows a uniform distribution with support $[2,3]$, $\alpha = 1/2$, $p = 1$, $\mu\tau = 1$, and $\tau = 0.25$. The next section provides an analytical solution.[21]

Figure 2 illustrates the solution for different values of c. Rows depict the behavior of employment n, reported income x, auditing intensity ϕ, and the informational rent U, respectively. The first column considers $c = 0$. Since the IRS faces no cost to audit entrepreneurs, the full-information revenue is recovered. In particular, audits suffice to fully enforce taxes, and labor is not distorted away from its efficient level. That is, $n(z) = n^*(z)$, $x(z) = \pi(n^*(z), z)$, and $\phi(z) = 1/\mu$ almost everywhere.

On the other hand, if c is high enough, as in the fourth column,[22] it is too costly to use auditing probabilities in equilibrium. Therefore, only labor distortions are used. As in a standard mechanism-design problem, the top type is not distorted away from the full-information case; that is, $n(\overline{z}) = n^*(\overline{z})$, while labor is distorted downwards to provide incentives. Below a threshold type, call it z_1, depicted by the dashed line, the individual rationality binds, which places a limit on the further use of distortions to provide incentives. In particular, employment has to be adjusted to keep $U(z) = 0$ for $z \leq z_1$. As opposed to problems without countervailing incentives, the individual rationality can bind at intermediate types.

The most interesting case is when c takes intermediate values, as in the second and third columns, for $c = 0.35$ and $c = 0.45$, respectively.

[21] In the next section, I state three additional assumptions on the primitives and one property that the solution must have to be optimal. Those are satisfied in this example.

[22] In this particular example, for $c \geq 0.54$.

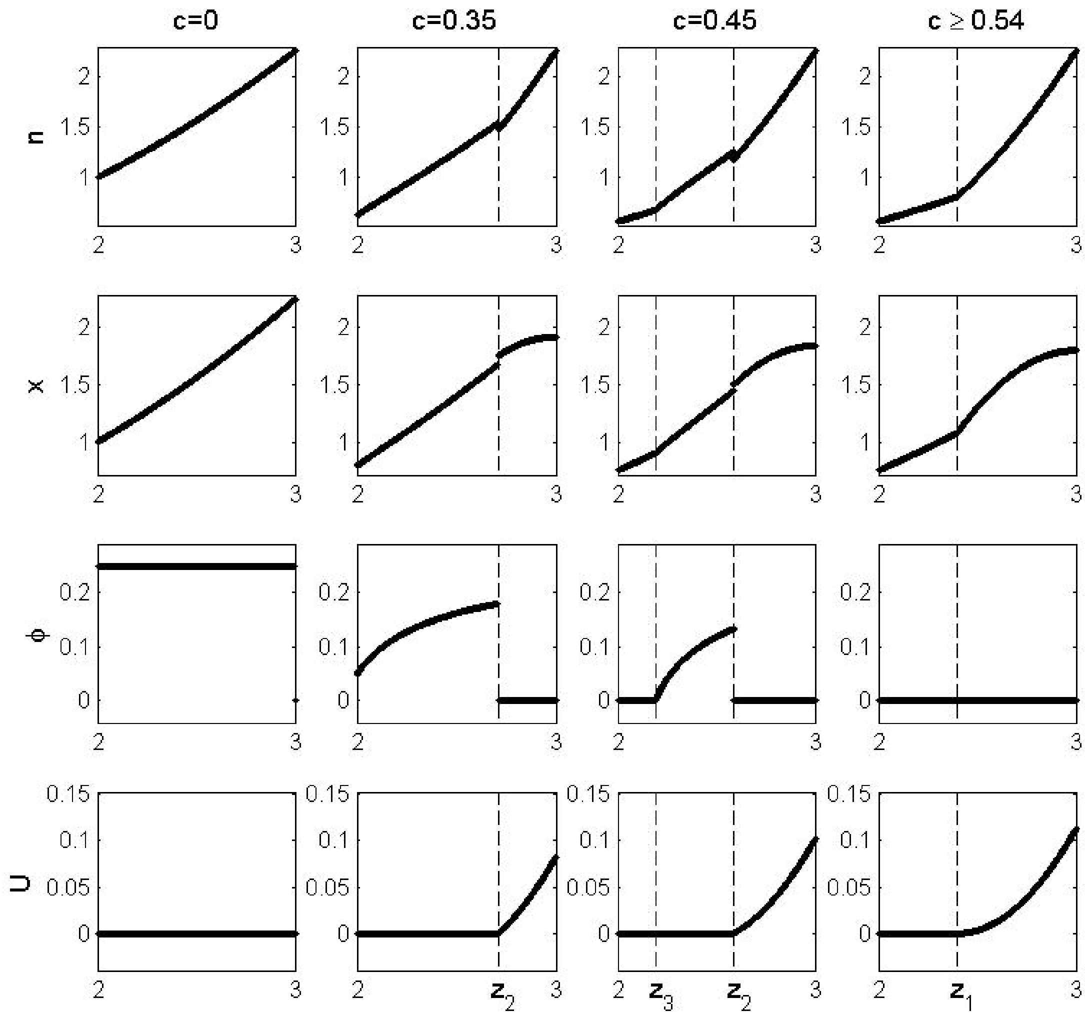

Figure 2: Optimal mechanism.

Similarly, the "non-distortion at the top" result also holds here, and labor is distorted downwards up to a threshold type, call it z_2, which is the (right) dashed line in the second (third) column. In this region, only distortions are used.

Below z_2, up to another threshold, call it z_3, which is the left dashed line in the second column,[23] the IRS combines both distortions and audits to enforce taxes. For $z \in [z_3, z_2)$, both $\phi(z)$ and $n(z)$ are increasing. At $z = z_2$, they drop discontinuously.[24] In this region,

[23]In this example, for all $c > 0.375$, $z_3 > \underline{z}$. If $c \leq 0.375$, $z_3 = \underline{z}$.

[24]These discontinuities might not be visible in Figure 2, but some of the plots are reproduced in Figure 3.

the individual rationality binds.

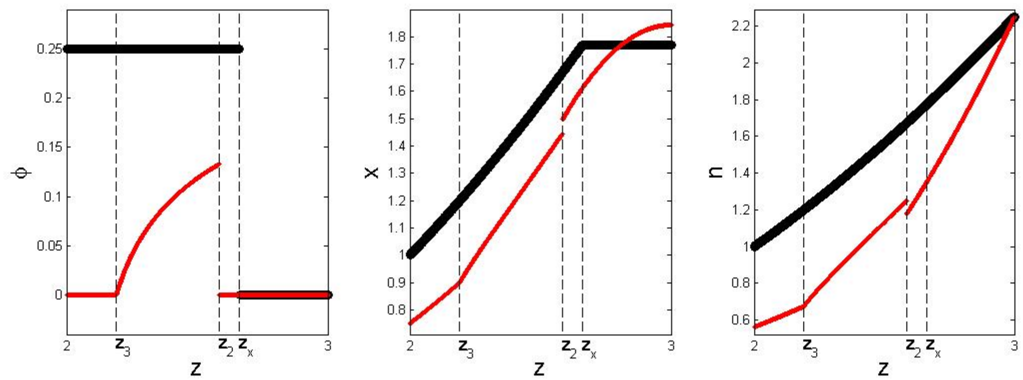

Figure 3: Optimal mechanism, c = 0.45.

In Figure 3 for $c = 0.45$, I compare the optimal mechanism that conditions the monitoring strategy on both reported income and labor input (thin lines), with the one that conditions it only on the former (thick lines), as in Proposition 1.[25] Once the IRS also screens over labor input, it is optimal to exchange costly audits for distortions in order to provide incentives and maximize net revenue.

To understand the intuition behind the jump at z_2, notice that, in order to increase net revenue, the use of audits are twofold: (1) it enforces taxes from those that are audited; and (2) it prevents deviations from other types, allowing the IRS to require higher income declarations from them. In the top range of the type distribution, entrepreneurs have strong incentives to understate their type and then pay less in taxes. Hence, monitoring those at the top is effective to enforce their taxes, but ineffective to prevent deviations from other types. Given that the audit cost is lumpy, the jump of the monitoring strategy at z_2 balances these two goals. It allows the IRS to audit and enforce taxes from intermediate productive types and, at the same time, to establish a lower bound on the income reported by the highest types.

If the monitoring strategy depended only on reported income, the same reasoning would

[25] Following the steps in Sánchez and Sobel [1993], z_x (from Proposition 1) is the unique root in $[2, 3]$ that solves $c = \mu\tau\frac{1-G(s)}{g(s)}pn^*(s) = \frac{(3-s)s}{2}$ in s, given $c \in [0, 1]$. That is, $z_x = \frac{3+\sqrt{9-8c}}{2}$. For $c > 1$, such a root does not exist; thus, $z_x = \underline{z} = 2$.

justify that it discontinuously drops at z_x. However, as depicted in Figure 3, the ablest entrepreneurs do not report, or bunch at, the threshold type income anymore. Intuitively, by distorting labor downwards to provide incentives, the IRS can separate the equilibrium and design an increasing reported income schedule. The jump in the labor schedule is necessary to keep U continuous and, thus, preserve incentive compatibility. At the same time, it also assigns fewer distortions for those below z_2, which increases production and net revenue collection.

To understand the intuition behind an increasing monitoring strategy at $[z_3, z_2)$, recall from Lemma 1 that the possibility to screen over n relaxes the monotonicity requirement that $\phi(z)$ is non-increasing. Hence, a non-monotone monitoring strategy is consistent with incentive compatibility. Moreover, incentives to overstate z and be assigned a higher reservation value are relatively stronger in the bottom range of the type distribution. Therefore, an increasing monitoring strategy at $[z_3, z_2)$ not only prevents type misreporting, but also allows the IRS to save on expenses by selecting a more productive group of entrepreneurs to audit. Finally, individual rationality, which is binding in this region, determines the amount of distortion used in equilibrium.[26]

In the third column of Figure 2, for $z \leq z_3$, feasibility imposes a limit on the use of auditing probabilities. Thus, labor distortions are adjusted to respect individual rationality, which is still binding in this region. In this case, the lowest types are not monitored in equilibrium.

Interestingly, if c is not too high, the model predicts a *missing middle* in the reported income space. In other words, some intermediate values of income are never reported to the IRS. Moreover, there is a region in which two different types employ the same amount of labor.

Finally, it is the possibility of setting off-equilibrium threats that makes audits a powerful tool to enforce taxes. As Figure 1 and the second row of Figure 2 illustrate, for any value of

[26] In the next section, I show that $U(z) = 0$ if and only if (LIC) is equalized to zero; that is, $(1 - \phi(z)\mu\tau)pn(z)^\alpha = (1-\tau)pn^*(z)^\alpha$. Given $\phi(z)$, this equation determines $n(z)$ wherever (IR) is binding.

c, the reported income schedule is positive, which translates into a positive lower bound on the net revenue collected.[27] In Section 7, I explore this insight to assess the revenue gains from adopting the optimal mechanism in practice.

In the next two subsections, I use this example to discuss further insights from the model.

4.1 Key driving forces

In this section, I argue that both countervailing incentives and observability of labor are the key driving forces behind the non-monotonicity of the optimal auditing scheme.

If $\tau = 1$,[28] the reservation value, $(1-\tau)\pi(n^*(z), z)$, ceases to be type-dependent and, thus, countervailing incentives are ruled out. I show below that, in this case, the optimal monitoring strategy is non-increasing. Every z below a threshold type is audited with intensity $1/\mu$, while every z above it is not monitored. Moreover, (LIC) and (M) become

$$\begin{aligned} (\mathrm{LIC}_{\tau=1}): \quad & \frac{dU}{dz}(z) = (1 - \phi(z)\mu) p n(z)^\alpha \text{ a.e.}, \\ (\mathrm{M}_{\tau=1}): \quad & (1 - \phi(z)\mu) n(z)^\alpha \text{ is non-decreasing,} \end{aligned}$$

respectively. Consequently, a non-monotone monitoring strategy could be implementable, although it is not optimal.

Therefore, a type-dependent reservation value is necessary to break the monotonicity result that $\phi(z)$ is non-increasing. But is it sufficient? No. If labor input is not observable, countervailing incentives are still present in the model, but Proposition 1 follows. Assume

[27] Similarly, if labor is not observable and c becomes arbitrarily large, taxpayers declare the lowest-type income (see Proposition 1).

[28] $\tau = 1$ describes a context in which the principal aims to fully appropriate the agent's profits. This action can be legitimate, as in the example of a holding company requiring reports on the profitability of its subsidiaries. But it can also be illegitimate, as in the example of a local mafia extorting business owners.

that $\tau < 1$ and labor is not observable, so (LIC) and (M) become

$$\begin{aligned}(\mathrm{LIC}_x): &\quad \frac{dU}{dz}(z) = (1-\phi(z)\mu\tau)pn^*(z)^\alpha - (1-\tau)pn^*(z)^\alpha \text{ a.e.},\\ (\mathrm{M}_x): &\quad \phi(z) \text{ is non-increasing,}\end{aligned}$$

respectively. The IRS cannot implement a non-monotone monitoring strategy.

4.2 Implications

In a risk-neutral environment, if the IRS commits to a monitoring strategy that depends only on reported income, the cut-off audit rule derived in Proposition 1 is a remarkably robust result. However, its policy implications are unsatisfactory for two reasons. First, the amount underreported as a fraction of income increases with income, introducing a regressive bias on effective taxes. Second, only those who declare income honestly – precisely the poorest taxpayers – will be audited. In this section, I show how these implications change once the monitoring strategy also depends on labor.

4.2.1 Underreported income

Let underreported income as a fraction of true income be $1 - x(z)/\pi(n(z), z)$. Figure 4 plots this variable for different values of c. The solid line represents the optimal mechanism when the monitoring strategy depends on both reported profits and labor input, while the dashed line conditions the monitoring strategy only on the former.

In contrast with the previous literature, every taxpayer evades in the model. Moreover, the relationship between income and the fraction of income that is underreported is non-monotone and discontinuous. In particular, those in the bottom and top underreport proportionally more than those in the middle range of the type distribution.

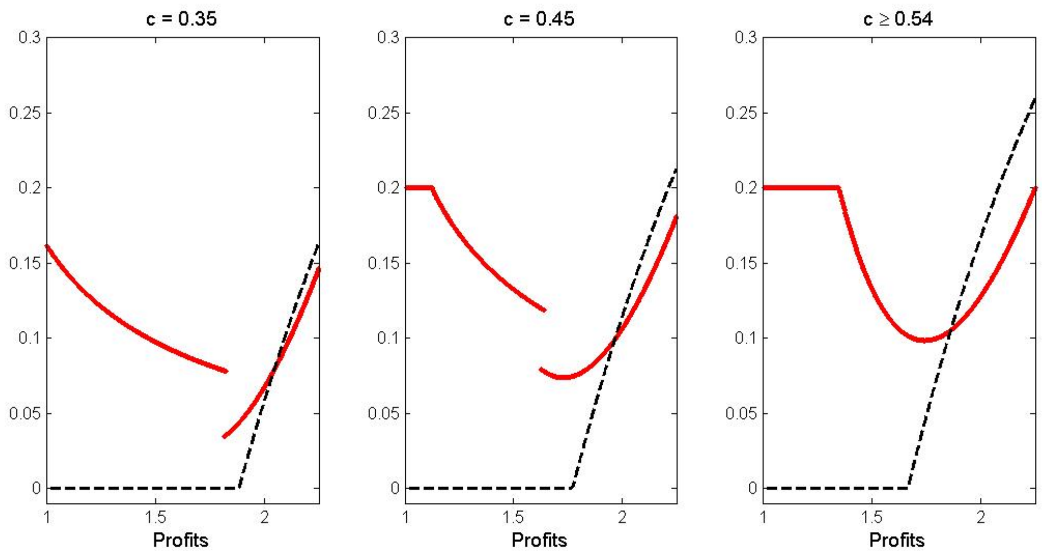

Figure 4: Underreported income as a fraction of true income.

4.2.2 Effective tax rate: regressive or progressive bias?

If the IRS screens only over reported income, effective taxes are regressive since the set of evaders consists of the most productive entrepreneurs. Once the monitoring strategy is conditioned on a signal of the true income, this regressive bias could be mitigated.[29]

Let the expected effective tax rate for a given type z be

$$\tau^e(z) \equiv \frac{\tau x(z) + \mu\tau\phi(z)(\pi(n(z), z) - x(z))}{\pi(n(z), z)}.$$

Figure 5 plots $\tau^e(z)$ against true profits for different values of c. The solid line represents the optimal mechanism when the monitoring strategy depends on both reported profits and labor input, while the dashed line conditions the monitoring strategy only on the former.

Effective taxes are unevenly distributed. On the one hand, the poorest entrepreneurs are paying proportionally less in taxes. On the other hand, effective taxes decrease in the top

[29]To my knowledge, Scotchmer [1987] was the first to formally point out this possibility. In her model, taxpayers are grouped into classes according to their income signal. As a result, effective taxes are progressive across classes, although regressive within. In her paper, both income and its signal are exogenous. See, also, Macho-Stadler and Pérez-Castrillo [2002]. In Bigio and Zilberman [2011], for instance, effective taxes are progressive since the lowest types are the set of evaders.

range of the distribution. Therefore, the overall bias from evasion on the progressiveness of taxes is unknown. Intuitively, by placing higher effective taxes in the middle range of the type distribution, the IRS not only prevents those in the top range from understating z, but also targets its audit expenses towards a more productive group of entrepreneurs.

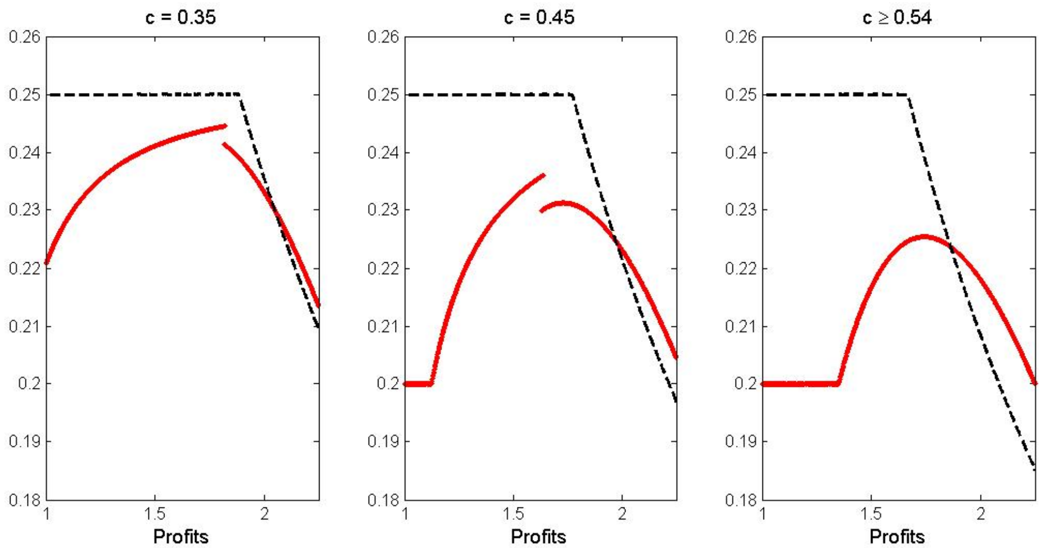

Figure 5: Expected effective tax rate.

4.2.3 Outcomes of audits

Once the IRS commits to the cut-off audit rule described in Proposition 1, all audited taxpayers are known to have reported honestly. This case is depicted in the first plot of Figure 6, where the thick line is the optimal monitoring strategy $\phi(z)$, and the thin line is the amount evaded $\tau(\pi(n(z), z) - x(z))$. Ex-post, audits do not generate revenue for the government. Hence, the IRS is tempted to deviate from its announced monitoring strategy and audit the ablest entrepreneurs.

Once the audit rule also depends on n, as in the second plot, audits generate some gross revenue, but not necessarily positive net revenue. Note that audits no longer target honest taxpayers. Hence, although there still exist ex-post profitable deviations, a stronger case for the IRS's ability to commit can be made.

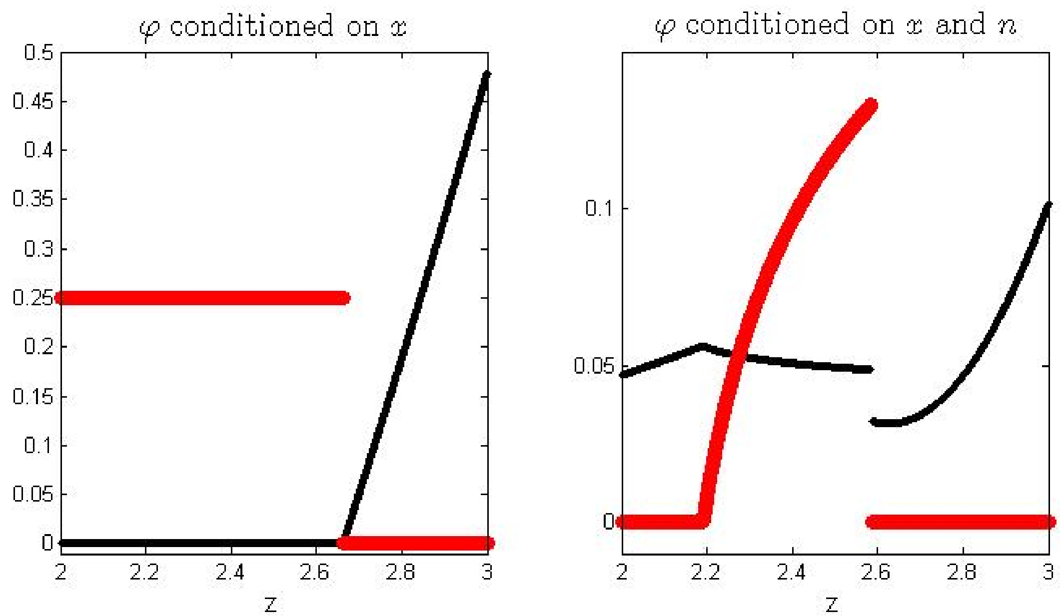

Figure 6: Outcomes of audits, $\phi(z)$ vs. $\tau(\pi(n(z), z) - x(z))$, c=0.45.

Finally, the richest taxpayers are never audited in both mechanisms. However, in contrast with the cut-off audit rule, the poorest taxpayers might also not be audited once the monitoring strategy is conditioned on n.

5 Solution

Given the characterization of (IC) in Lemma 1, the IRS problem is to solve

$$\max_{\{n(z),U(z),\phi(z)\}_z} \int_{\underline{z}}^{\overline{z}} [\pi(n(z), z) - U(z) - c\phi(z)]\, dG(z) - \Omega$$

s.t.

$$\text{(LIC)} \quad \frac{dU}{dz}(z) = (1 - \mu\tau\phi(z))pn(z)^{\alpha} - (1 - \tau)pn^*(z)^{\alpha}, \text{ a.e.}$$

$$\text{(IR)} \quad U(z) \geq 0, \forall z \in [\underline{z}, \overline{z}]$$

$$\text{(F)} \quad \phi(z) \geq 0, \forall z \in [\underline{z}, \overline{z}],$$

where $\Omega = \int_{\underline{z}}^{\overline{z}}(1 - \tau)\pi(n^*(z), z)dG(z)$.

To solve this problem, I ignore (M), $\phi(z) \leq 1$, $n(z) \geq 0$, $x(z) \geq 0$, and $x(z) \leq \pi(n(z), z)$, for all z, from the set of constraints. Notice that $x(z) \leq \pi(n(z), z)$ is implied by (IR)[30] and that $\phi(z) \leq 1$ never binds in equilibrium.[31] The remaining ignored constraints must be verified in equilibrium.

Let λ be the costate variable associated with the state variable U. The Hamiltonian is

$$H(U, n, \phi, \lambda, z) = [pzn^{\alpha} - n - U - c\phi]g(z) + \lambda[(1 - \mu\tau\phi)pn^{\alpha} - (1 - \tau)pn^{*}(z)^{\alpha}].$$

For a given type z, let $\omega(z)$ and $\theta(z)$ be the Lagrange multipliers associated with $\phi(z) \geq 0$ and $U(z) \geq 0$, respectively. The Lagrangian is

$$L(U, n, \phi, \lambda, z) = H(U, n, \phi, \lambda, z) + \theta U + \omega\phi.$$

Let the superscript o denote the optimum solution to the optimization problem stated above. Following Seierstad and Sydsæter [1987] (Theorem 2, page 361), given that H is concave in ϕ, the following set of conditions is sufficient for a global maximum.

1. $\left[z + (1 - \mu\tau\phi^{o}(z))\frac{\lambda(z)}{g(z)}\right] \alpha p n^{o}(z)^{\alpha-1} = 1$;
2. $cg(z) + \lambda(z)\mu\tau p n^{o}(z)^{\alpha} = \omega(z)$;
3. $\frac{d\lambda}{dz}(z) = g(z) - \theta(z)$;
4. $\frac{dU^{o}}{dz}(z) = (1 - \mu\tau\phi^{o}(z))pn^{o}(z)^{\alpha} - (1 - \tau)pn^{*}(z)^{\alpha}$;
5. $\omega(z) \geq 0$; $\phi^{o}(z) \geq 0$; $\omega(z)\phi^{o}(z) = 0$;
6. $\theta(z) \geq 0$; $U^{o}(z) \geq 0$; $\theta(z)U^{o}(z) = 0$;
7. $\lambda(\underline{z})U^{o}(\underline{z}) = 0; \lambda(\underline{z}) \leq 0; \lambda(\overline{z})U^{o}(\overline{z}) = 0; \lambda(\overline{z}) \geq 0$;
8. $\lambda(z^{-}) \geq \lambda(z^{+})$; $[\lambda(z^{-}) - \lambda(z^{+})]U^{o}(z) = 0$;

[30] Indeed, $x(z) \leq \pi(n(z), z)$ if and only if $U(z) \geq (1 - \tau)[\pi(n(z), z) - \pi(n^{*}(z), z)]$.

[31] Recall from Section 2 that $\phi(z) = 1/\mu$ suffices to generate truthful income report and to provide incentives. Since monitoring is costly, $\phi(z) \leq 1/\mu$ in equilibrium.

9. $-\lambda(z) \leq \frac{zg(z)}{(1-\mu\tau\phi^o(z))}$.

One and two, above, are the first order conditions with respect to n and ϕ, respectively. Three is the costate law of motion. Four is the local incentive-compatibility constraint. Five and six are the complementary slackness conditions that ensure feasibility and individual rationality, respectively. Seven is the set of transversality conditions. Eight needs to be satisfied if the costate λ is allowed to be discontinuous.[32] Finally, 9 guarantees that the Hamiltonian is concave in n. Notice that 9 can be ignored since it is implied by 1 and $\phi^o(z) \leq 1/\mu$.

The proof consists of a guess-and-verify method. The trick is to conjecture the subsets of the type space $[\underline{z}, \overline{z}]$, in which the inequalities $U^o \geq 0$ and $\phi^o \geq 0$ are binding. In other words, it is to set the appropriate values for the Lagrange multipliers θ and ω along the interval $[\underline{z}, \overline{z}]$.[33]

First, note that if the solution is continuously left differentiable at the top-type $\overline{z}$, then $n^o(\overline{z}) = n^*(\overline{z})$, $\phi^o(\overline{z}) = 0$, and $U^o(\overline{z}) > 0$. Moreover, $\lambda(\overline{z}) = 0$, $\omega(\overline{z}) > 0$ and $\theta(\overline{z}) = 0$.[34] This is a variant of the "non-distortion at the top" kind of result, in which the top type is not distorted away from the full-information case and gets positive informational rent.

By continuity, $U^o(z) > 0$ and $\omega(z) > 0$, which implies $\theta(z) = 0$ and $\phi^o(z) = 0$, for all $z < \overline{z}$ in a small neighborhood of $\overline{z}$. Moreover, by solving the differential equation in 3 with boundary $\lambda(\overline{z}) = 0$, one gets $\lambda(z) = G(z) - 1$. From 1, $n^o(z) = \left(\alpha p \left[z - \frac{1-G(z)}{g(z)}\right]\right)^{\frac{1}{1-\alpha}}$ in this neighborhood.

To proceed with the analysis, I assume one restriction on the distribution of types. Define

[32] Throughout the paper, I use the following notation: $h(z^-)$ is the left limit of h at z, $h(z^+)$ is the right limit of h at z, $\frac{d^-h}{dz}(z)$ is the left derivative of h at z, and $\frac{d^+h}{dz}(z)$ is the right derivative of h at z.

[33] This strategy is partially inspired by Maggi and Rodríguez-Clare [1995]'s analysis of a mechanism-design problem that features a type-dependent participation constraint. However, since this setup is different, the map from this paper to theirs is not perfect. In particular, here, the agent's objective is not quasi-linear and there are two decision variables. Moreover, for a certain range of values for c, the optimal mechanism is discontinuous. See Jullien [2000] for an alternative treatment of type-dependent participation constraints.

[34] Indeed, from 7, $\lambda(\overline{z}) \geq 0$. Hence, since $c > 0$, then $\omega(\overline{z}) > 0$ (from 2) and $\phi(\overline{z}) = 0$ (from 5). Moreover, $n^o(\overline{z}) \geq n^*(\overline{z})$ (from 1), which implies $\frac{d^-U^o}{dz}(\overline{z}) > 0$ (from 4). If, by contradiction, $U^o(\overline{z}) = 0$, then there exists $z < \overline{z}$ such that $U^o(z) < 0$, which violates 6. Hence, $U^o(\overline{z}) > 0$, which implies $\lambda(\overline{z}) = 0$ (from 7), $\theta(\overline{z}) = 0$ (from 6), and $n^o(z) = n^*(z)$ (from 1).

$h \equiv \frac{1-G}{g}$, which is one over the hazard rate, and $\gamma \equiv [1-(1-\tau)^{\frac{1-\alpha}{\alpha}}] \in (0,1)$.

Assumption 1. *$\gamma z - h(z)$ is non-decreasing.*

This assumption is a weaker version of the monotone hazard rate condition commonly assumed in the literature. Indeed, if $h(z)$ is non-increasing, then Assumption 1 holds. The role of this assumption is twofold: First, it ensures that (M) is satisfied without the additional expositional cost of dealing with bunching; second, it guarantees that the individual rationality constraints are not violated.

The task, now, is to find a region in the type space, such that either $\phi^o(z) \geq 0$ or $\theta(z) \geq 0$ ceases to hold with equality. Notice that for all z in a small neighborhood of $\overline{z}$, both

$$\frac{dU^o}{dz}(z) = p\left(\alpha p\left[z - h(z)\right]\right)^{\frac{\alpha}{1-\alpha}} - (1-\tau)p(\alpha p z)^{\frac{\alpha}{1-\alpha}} > 0 \text{ (from 4)} \tag{3}$$

and

$$cg(z) + \lambda(z)\mu\tau p n^o(z)^\alpha = g(z)\left[c - h(z)\mu\tau p\left(\alpha p\left[z - h(z)\right]\right)^{\frac{\alpha}{1-\alpha}}\right] > 0 \text{ (from 2)} \tag{4}$$

hold by continuity. I conjecture that $\phi^o(z) \geq 0$ or $\theta(z) \geq 0$ or both cease to hold with equality wherever one of the inequalities in (3) or (4) is strictly reversed. The solution displays the property that $U^o(z) = 0$ if and only if $\frac{d^- U^o}{dz}(z) = 0$.

Formally, define $A(z) \equiv \gamma z - h(z)$,[35] which is non-decreasing by Assumption 1, and let

$$z_1 = \sup_{s \in [\underline{z}, \overline{z}]} \{A(s) \leq 0\}. \tag{5}$$

The term in curly brackets in equation (5) is obtained from reverting the inequality in (3). By construction, if $\{s \in [\underline{z}, \overline{z}] : A(s) \leq 0\}$ is not empty, Assumption 1 implies that z_1 is the highest root that solves $A(s) = 0$, and that $\frac{dU^o}{dz}(z) > 0$ for all $z > z_1$. If it is empty, let $z_1 = \underline{z}$.

[35] Recall that $\gamma \equiv [1-(1-\tau)^{\frac{1-\alpha}{\alpha}}] \in (0,1)$.

Similarly, define $B(z) \equiv h(z)\mu\tau p\left(\alpha p\left[z-h(z)\right]\right)^{\frac{\alpha}{1-\alpha}}$, and let

$$z_2 = \sup_{s\in[\underline{z},\overline{z}]} \{B(s) > c\}. \tag{6}$$

The term in curly brackets in equation (6) is obtained from strictly reverting the inequality in (4). By construction, if $\{s \in [\underline{z}, \overline{z}] : B(s) > c\}$ is not empty, then $B(z_2) = c$. If it is empty, let $z_2 = \underline{z}$.

In words, analyzing from the right, if $z_1 \geq z_2$, then the differential equation in 4 would be equalized to zero before conditions 2 and 5 are violated. On the other hand, if $z_2 > z_1$, $\phi(z) = 0$ for some $z < z_2$ would violate conditions 2 and 5 before the differential equation in 4 is equalized to zero. In particular, I show that for $z \leq \max\{z_1, z_2\}$, $\theta(z) \geq 0$ or $\phi^o(z) \geq 0$ or both cease to hold with equality.

First, consider the case $z_1 \geq z_2$ or, equivalently, $c \geq \max_{s\in[z_1,\overline{z}]} B(s)$. The next proposition studies the case depicted in the fourth column in Figure 2, in which c is too high.

Proposition 2. *If $z_1 \geq z_2$ (that is, $c \geq \max_{s\in[z_1,\overline{z}]} B(s)$), then*

$$n^o(z) = \begin{cases} (1-\tau)^{\frac{1}{\alpha}}(\alpha p z)^{\frac{1}{1-\alpha}} & \text{if } \underline{z} \leq z < z_1 \\ \left(\alpha p\left[z-h(z)\right]\right)^{\frac{1}{1-\alpha}} & \text{if } z_1 \leq z \leq \overline{z} \end{cases},$$

$$\phi^o(z) = \; 0 \quad \text{if } \underline{z} \leq z \leq \overline{z} \;,$$

$$U^o(z) = \begin{cases} 0 & \text{if } \underline{z} \leq z < z_1 \\ p\int_{z_1}^{z}\left[\left(\alpha p\left[s-h(s)\right]\right)^{\frac{\alpha}{1-\alpha}} - (1-\tau)(\alpha p s)^{\frac{\alpha}{1-\alpha}}\right] ds & \text{if } z_1 \leq z \leq \overline{z} \end{cases},$$

$$\lambda(z) = \begin{cases} -\gamma z g(z) & \text{if } \underline{z} \leq z < z_1 \\ G(z) - 1 & \text{if } z_1 \leq z \leq \overline{z} \end{cases}.$$

Moreover, the solution is continuous.

Proof. For $z \geq z_1$, the proof is outlined in the text. Hence, if $z_1 = \underline{z}$, the result follows.

Assume that $z_1 > \underline{z}$. For $z < z_1$, set $\phi^o(z) = 0$, and solve 1 and 4 (equalized to zero) in

$n^o(z)$ and $\lambda(z)$. Hence, $\lambda(z) = -\gamma z g(z)$ and $n^o(z) = (1-\tau)^{\frac{1}{\alpha}}(\alpha p z)^{\frac{1}{1-\alpha}}$. By construction, $\lambda(z)$ is continuous, so 8 is satisfied.

Note that Assumption 1 implies that $-[\gamma g(z) + g'(z)h(z)] \leq g(z)$. From 3 and 6, $\theta(z) = g(z) - \frac{d\lambda}{dz}(z) \geq 0$ is equivalent to $-\gamma[g(z)+zg'(z)] \leq g(z)$, which is satisfied if $g(z)+zg'(z) \geq 0$. If $g(z)+zg'(z) \leq 0$, which implies that $g'(z) < 0$, it is enough to show that $-\gamma[g(z)+zg'(z)] \leq -[\gamma g(z) + g'(z)h(z)]$ or, equivalently, $\gamma z \leq h(z)$ for all $z \leq z_1$, which follows from z_1's definition and Assumption 1. Hence, $\lambda(\underline{z}) < 0$ and $U^o(z) = 0$ are consistent with condition 6, the differential equation in 4 equalized to zero, and condition 7.

It remains to show that conditions 2 and 5 hold; that is, $\omega(z) = cg(z) + \lambda(z)\mu\tau p(n^o(z))^\alpha \geq 0$. In fact, by plugging $\lambda(z)$ and $n^o(z)$ into this expression, one obtains $\gamma z \mu \tau p(1-\tau)(\alpha p z)^{\frac{\alpha}{1-\alpha}} \leq c$. Hence, it is enough to show that $\gamma z_1 \mu \tau p(1-\tau)(\alpha p z_1)^{\frac{\alpha}{1-\alpha}} \leq c$. Recalling the definition of z_1, this requirement collapses to $B(z_1) \leq c$, which follows from $z_1 \geq z_2$.

Finally, it is straightforward to verify that $\frac{dn^o}{dz}(z) \geq 0$, $\frac{d\phi^o}{dz}(z) = 0$, $\frac{dx^o}{dz}(z) \geq 0$, wherever these derivatives exist, and $n^o(\underline{z}) \geq 0$ and $x^o(\underline{z}) \geq 0$. Hence, the omitted constraints are satisfied. $\square$

Consider $z_2 > z_1$ instead. Hence, $\omega(z_2) = 0$ and $\frac{d^+U^o}{dz}(z_2) > 0$. One attempt to solve this case is to keep $\lambda(z) = G(z) - 1$ in a small neighborhood of z_2, and for $z < z_2$, let both $n^o(z)$ and $\phi^o(z)$ jointly solve conditions 1 and 2 (with $\omega(z) = 0$) in this neighborhood. However, the solution to this system implies that $\phi^o(z) < 0$ for some $z < z_2$ in any neighborhood of z_2.[36] Therefore, this approach does not work. To make further progress, $\lambda(z)$ needs to be changed for $z < z_2$, and from conditions 3, 7 and 8, it follows that $U^o(z_2) = 0$.

Consequently, a natural candidate for the optimal mechanism when $z < z_2$ and $z_2 > z_1$

[36]Indeed, fix $z < z_2$ in a small neighborhood of z_2 such that $B(z) > c$. By solving condition 2 (with $\omega(z) = 0$) at $n(z)$ and using $c < B(z)$, one obtains $n^o(z) < (\alpha p[z - h(z)])^{\frac{1}{1-\alpha}}$. An inspection of condition 1 shows that this inequality is true if and only if $\phi^o(z) < 0$.

is the solution to the following system of three equations in three unknowns (ϕ, n and λ).

$$\begin{aligned} \left[g(z)z + (1-\mu\tau\phi)\lambda\right]\alpha p n^{\alpha-1} - g(z) &= 0 \\ cg(z) + \lambda\mu\tau p n^{\alpha} &= 0 \\ (1-\mu\tau\phi)n^{\alpha} - (1-\tau)(\alpha p z)^{\frac{\alpha}{1-\alpha}} &= 0. \end{aligned} \tag{7}$$

These equations are condition 1, condition 2 (with $\omega = 0$), and the differential equation in condition 4 equalized to zero. The following assumption guarantees that if a solution to (7) exists, it is unique.

Assumption 2. $\tau \leq 1 - \left(\frac{2\alpha}{1+\alpha}\right)^{\frac{\alpha}{1-\alpha}}$.

If the tax rate is 25 percent, any $\alpha \geq 0.22$ satisfies this assumption. Similarly, if $\alpha = 2/3$, then it is satisfied for any $\tau \leq 0.36$. Therefore, Assumption 2 holds for empirically plausible values of τ and α. However, if $\tau > 0.4$, this assumption is violated for any value of α.

Note that the derivative of the reservation value with respect to z, $(1-\tau)pn^*(z)^{\alpha}$, is decreasing in τ. Therefore, this assumption ensures that countervailing incentives are strong enough.

The following lemma states sufficient conditions for existence and uniqueness of a solution to the system in (7).

Lemma 2. *For $\phi \in [0, 1/\mu]$ and $n \geq 0$:*

If $c > \frac{\mu\tau}{\alpha}[(1-\tau) - (1-\tau)^{\frac{1}{\alpha}}](\alpha p z)^{\frac{1}{1-\alpha}}$, then the system of equations in (7) does not have a solution.

If $c \leq \frac{\mu\tau}{\alpha}[(1-\tau) - (1-\tau)^{\frac{1}{\alpha}}](\alpha p z)^{\frac{1}{1-\alpha}}$, it has a unique solution.

The proof is in Appendix A. This system might not have a solution for some small values of z. Define $z_3 \in [0, \infty)$ as being the unique root that solves the following equation in s.

$$\frac{\mu\tau}{\alpha}[(1-\tau) - (1-\tau)^{\frac{1}{\alpha}}](\alpha p s)^{\frac{1}{1-\alpha}} = c.$$

If $z_3 < \underline{z}$, redefine $z_3 = \underline{z}$. Therefore, for all $z \in [z_3, z_2)$, the system in (7) has a unique solution. Let it be denoted by $\{\hat{n}(z), \hat{\phi}(z), \hat{\lambda}(z)\}_{z\in[z_3,z_2)}$.

The following lemma states some of the properties of this solution.

Lemma 3. $\{\hat{n}(z), \hat{\phi}(z), \hat{\lambda}(z)\}_{z\in[z_3,z_2)}$ *has the following properties:*

1. $\hat{n}(z) \in [(1-\tau)n^*(z), n^*(z)]$; $\hat{\phi}(z) \in [0, 1/\mu]$; $\hat{\lambda}(z) < 0$;

2. *if* $z_3 \in (\underline{z}, z_2)$, *then* $\hat{n}(z_3) = (1-\tau)^{\frac{1}{\alpha}}(\alpha p z_3)^{\frac{1}{1-\alpha}}$; $\hat{\phi}(z_3) = 0$; $\hat{\lambda}(z_3) = [(1-\tau)^{\frac{1-\alpha}{\alpha}} - 1] z_3 g(z_3)$;

3. $\frac{d\hat{n}}{dz}(z) \geq 0$ *and* $\frac{d\hat{\phi}}{dz}(z) \geq 0$;

4. $\hat{n}(z)$ *and* $\hat{\phi}(z)$ *do not depend on the distribution of types;*

5. $(1 - \mu\tau\hat{\phi}(z))\hat{n}(z)^{\alpha} = (1-\tau)(\alpha p z)^{\frac{\alpha}{1-\alpha}}$.

This lemma is proved in Appendix A. The first and second properties state that the candidate for the optimal mechanism is feasible and continuous at z_3 if $z_3 \in (\underline{z}, z_2)$, respectively. The third and fourth properties say that the labor input and auditing probabilities that solve (7) are increasing in z and do not depend on G, although z_2 depends. Finally, the fifth is the last equation of the system in (7), which guarantees that (M) is satisfied.

At this degree of generality, it is not possible to characterize closed-form solutions to $\{\hat{n}(z), \hat{\phi}(z), \hat{\lambda}(z)\}_{z\in[z_3,z_2)}$ for all possible values of α.[37] Hence, to proceed with the analysis, I state a property that this solution might or might not have.

Property 1. *For all* $z \in [z_3, z_2)$, $\frac{d\hat{\lambda}}{dz}(z) \leq g(z)$.

If Property 1 holds, it is possible to characterize the optimal mechanism. This property is needed to ensure that $\theta(z) \geq 0$ for all $z \in [z_3, z_2)$, which guarantees that the individual rationality is satisfied (see conditions 3 and 6).

The following lemma gives a sufficient condition to ensure that Property 1 holds. Define $\rho \equiv \frac{1+\alpha}{\alpha}(1-\tau)^{\frac{1-\alpha}{\alpha}} - 2$, and note that $\rho \geq 0$ from Assumption 2.

[37] If $\alpha = 1/2$, for example, the system of equations in (7) has a closed-form solution. However, the formulas are too convoluted, since a solution to a third degree polynomial equation is required. Hence, I solve this system numerically in Section 4.

Lemma 4. *If* $\gamma \leq \rho + z\frac{g'(z)}{g(z)}\gamma\rho$ *for* $z \in [z_3, z_2)$*, then Property 1 holds.*

The proof is in Appendix A. This sufficient condition imposes a joint restriction on α, τ, and G. Although not readily interpreted, it can be useful to check if Property 1 is satisfied. If z follows a uniform distribution, for example, Property 1 holds if $\tau \leq 1 - \left(\frac{3\alpha}{2\alpha+1}\right)^{\frac{\alpha}{1-\alpha}}$.[38]

Finally, one last assumption is needed to characterize the optimal mechanism.

Assumption 3. $z_3 \leq z_1$.

Assumption 3 is sufficient to characterize the optimal allocation for $z \leq z_3$. In particular, for $z \leq z_3$, the optimal allocation has the same closed form as the solution in Proposition 2 for $z \leq z_1$. This assumption is far from being restrictive. A sufficient condition, for example, is that $B(z)$ single crosses c.[39]

The cases depicted in the second and third columns of Figure 2 illustrate the following proposition.

[38] If $\alpha = 1/2$, as in Section 4, then any $\tau \leq 1/4$ ensures that Property 1 is satisfied.

[39] A weaker sufficient condition is if $B(z_1) < c$, then $z_1 \geq z_2$. Indeed, assume, by contradiction, that $z_3 > z_1$. Hence, using the definitions of z_1 and z_3, one obtains $B(z_1) < c$, which implies that $z_1 \geq z_2$.

Proposition 3. *If $z_2 > z_1$ (that is, $c < \max_{s\in[z_1,\overline{z}]} B(s)$) and Property 1 is satisfied, then*

$$n^o(z) = \begin{cases} (1-\tau)^{\frac{1}{\alpha}}(\alpha p z)^{\frac{1}{1-\alpha}} & \text{if } \underline{z} \le z < z_3 \\ \hat{n}(z) & \text{if } z_3 \le z < z_2 \\ (\alpha p\,[z - h(z)])^{\frac{1}{1-\alpha}} & \text{if } z_2 \le z \le \overline{z} \end{cases},$$

$$\phi^o(z) = \begin{cases} 0 & \text{if } \underline{z} \le z < z_3 \\ \hat{\phi}(z) & \text{if } z_3 \le z < z_2 \\ 0 & \text{if } z_2 \le z \le \overline{z} \end{cases},$$

$$U^o(z) = \begin{cases} 0 & \text{if } \underline{z} \le z < z_2 \\ p\int_{z_2}^{z}\left[(\alpha p\,[s-h(s)])^{\frac{\alpha}{1-\alpha}} - (1-\tau)(\alpha p s)^{\frac{\alpha}{1-\alpha}}\right] ds & \text{if } z_2 \le z \le \overline{z} \end{cases},$$

$$\lambda(z) = \begin{cases} -\gamma z g(z) & \text{if } \underline{z} \le z < z_3 \\ \hat{\lambda}(z) & \text{if } z_3 \le z < z_2 \\ G(z) - 1 & \text{if } z_2 \le z \le \overline{z} \end{cases}.$$

The solution is discontinuous at $z = z_2$. In particular, $n^o(z_2^-) > n^o(z_2^+)$, and $\phi^o(z_2^-) > \phi^o(z_2^+) = 0$.

The proof is in Appendix A. The next corollary assumes that $c \to 0$, which is depicted in the first column of Figure 2.

Corollary 1. *If $c \to 0$, then*

$$\{n^o(z), \phi^o(z), U^o(z)\}_z \to \{n^*(z), 1/\mu, 0\}_{z<\overline{z}} \cup \{n^*(z), 0, 0\}_{z=\overline{z}}.$$

Proof. At $c = 0$, $\hat{\phi}(z) = 1/\mu$, $\hat{n}(z) = (\alpha p z)^{\frac{1}{1-\alpha}}$, $\hat{\lambda}(z) = 0$, for all z, $z_2 = \overline{z}$, and $z_3 = \underline{z}$. □

Finally, I also consider $\tau = 1$, which describes a context in which the principal aims to fully appropriate the agent's profits. If $\tau = 1$, Assumption 2 is violated; thus, Lemmas 2, 3, and 4, and Proposition 3 are not valid. In contrast, Proposition 1, which assumes $z_1 \ge z_2$, is still valid. If $z_2 > z_1$, I rely only on Assumption 1 and Property 1 to prove the following proposition.

Proposition 4. *If $\tau = 1$ and $z_2 > z_1$ (that is, $c < \max_{s\in[z_1,\overline{z}]} B(s)$), then*

$$n^o(z) = \begin{cases} n^*(z) & \text{if } \underline{z} \le z < z_2 \\ (\alpha p\,[z - h(z)])^{\frac{1}{1-\alpha}} & \text{if } z_2 \le z \le \overline{z} \end{cases},$$

$$\phi^o(z) = \begin{cases} 1/\mu & \text{if } \underline{z} \le z < z_2 \\ 0 & \text{if } z_2 \le z \le \overline{z} \end{cases},$$

$$U^o(z) = \begin{cases} 0 & \text{if } \underline{z} \le z < z_2 \\ p\int_{z_2}^{z}(\alpha p\,[s - h(s)])^{\frac{\alpha}{1-\alpha}} ds & \text{if } z_2 \le z \le \overline{z} \end{cases}.$$

The solution is discontinuous at $z = z_2$. In particular, $n^o(z_2^-) > n^o(z_2^+)$ and $\phi^o(z_2^-) > \phi^o(z_2^+) = 0$.

Proof. For $z \ge z_2$, the proof is outlined in the text. If $z_2 = \underline{z}$, the result follows. Assume that $z_2 > \underline{z}$. For $z < z_2$, the unique solution of the system of equations in (7) is $\hat{\phi}(z) = 1/\mu$, $\hat{n}(z) = (\alpha p z)^{\frac{1}{1-\alpha}} = n^*(z)$, and $\hat{\lambda}(z) = -cg(z)/\mu p n^*(z)^{\alpha}$. Since $n^o(z_2^-) > n^o(z_2^+)$, $\lambda(z_2^-) > \lambda(z_2^+)$ and 8 is satisfied. The remaining conditions follow from Property 1, which implies $\theta(z) \ge 0$ and $U^o(z) = 0$.[40] □

This is the case discussed in Section 4.1.

6 Extensions

To solve the problem, I specify functional forms for five objects: (1) the production technology, $F(z,n,K) = zn^{\alpha}$, where K is a vector of other inputs; (2) the utility function,

[40] In this case, Property 1 implies the following assumption on the distribution of types: $g'(z) \ge \left[\frac{1}{z}\frac{\alpha}{1-\alpha} - \frac{\mu p(\alpha p z)^{\frac{\alpha}{1-\alpha}}}{c}\right] g(z)$, for all $z \le z_2$. In the example, in Section 4, this inequality verifies if and only if $c \le 8$.

$u(y) = y$, where y is income or, equivalently, consumption in a static environment; (3) the penalty function, $M(e) = \mu e$, where e is the amount evaded; (4) the tax schedule, $T(x) = \tau x$; and (5) the audit cost function, $C(z, n) = c$. In this section, I argue whether or not these assumptions can be modified without substantially changing the results.

6.1 Production technology

In this section, I discuss two possible generalizations for the production technology. First, I consider a general function of the form $zf(n)$. Second, I show how to extend the model to accommodate multiple inputs, given that only one is costlessly observable.

6.1.1 General functional form

Recall that Assumptions 1, 2 and 3, which are crucial to prove the results in this paper, depend on the technological parameter α. Consequently, in order to validate Propositions 2 and 3 under a more general production technology, it is necessary to adapt these assumptions.

Assume, for instance, that the production technology takes the form $zf(n)$, where $f' > 0$, $f'' < 0$, and $f(0) = 0$. By applying the solution method developed in Section 5, one can derive the same qualitative results but at an expositional cost, as ad-hoc, and somewhat convoluted, restrictions on f and its derivatives would be imposed. It is challenging to characterize the optimal mechanism when these restrictions are violated.[41] Instead, I opt to explore the Cobb-Douglas production function.

However, even for the Cobb-Douglas case, the optimal mechanism is not fully characterized. Indeed, although realistic, Assumption 2 restricts the set of values α can take for a given τ. It is important to understand how the optimal mechanism would behave if this assumption were relaxed. Proposition 4 is a small step in this direction. I leave the characterization of the mechanism when Assumption 2 is violated as an open question for further

[41] For the specific case of Proposition 2, in which c is high enough such that $\phi(z) = 0$ for all z, the problem is isomorphic to a standard mechanism-design problem with a type-dependent reservation value. Jullien [2000] provides a comprehensive characterization of the optimal mechanism for this case.

research.

6.1.2 Multiple inputs

In this section, I show how the production technology can be generalized to multiple inputs, as long as only one is costlessly observable by the IRS. Let $F(z, n, k_1, ..., k_I) = zn^{\alpha_0} \prod_{i=1}^{I} k_i^{\alpha_i}$, where $\{k_i\}_{i=1}^{I}$ are the inputs that are not observable by the IRS.[42] Assume that F displays decreasing returns to scale, so $\sum_{i=0}^{I} \alpha_i < 1$, and $\alpha_i \geq 0$ for $i = 0, .., I$. Finally, let r_i be the price of input k_i, which is bought in a competitive market.

In Appendix A, given that $\{k_i\}_{i=1}^{I}$ are chosen in the second stage of the game, I show that pre-tax profits can be written as

$$\pi(n, z) = \zeta(z; p, \{r_i\}, \{\alpha_i\})n^{\alpha} - n, \tag{8}$$

where $\alpha = \alpha_0 / \left(1 - \sum_{i=1}^{I} \alpha_i\right) \in (0, 1)$, and ζ is a function of $\{\alpha_i\}_{i=1}^{I}$, $\{r_i\}_{i=1}^{I}$, p, and z. Consequently, the distribution of z, G, induces a distribution of ζ, say $\hat{G}$, and the mechanism developed above can be applied directly to ζ. However, Assumptions 1 and 3 and Property 1 need to be restated in terms of ζ and $\hat{G}$. Finally, although pre-tax profits are $\zeta n^{\alpha} - n$, the output produced is not ζn^{α}.

In Section 7, where I pursue an empirical evaluation of the mechanism developed in this paper, this extension will be useful for two reasons. First, it accounts for a more realistic production technology. Second, if z is either log-normally or Pareto distributed, which is commonly assumed in the literature, then ζ also is.

6.2 Audit cost

Without compelling empirical evidence, it is hard to inspect the shape of the audit cost $C(z, n)$.

[42] Similarly, even if some of these inputs are observable, it is enough to assume that the IRS does not condition its monitoring strategy on them.

One might argue, for instance, that large firms take longer to monitor than smaller firms, which justifies $\frac{\partial C}{\partial n} \geq 0$. In contrast, there may be a visibility effect that reduces the informational cost associated with monitoring larger firms, so it is also possible that $\frac{\partial C}{\partial n} \leq 0$.

Similarly, a high-ability entrepreneur could find it easier to circumvent the law and hide her income, which justifies $\frac{\partial C}{\partial z} \geq 0$. However, if high ability translates into more-complex business operations, the need to use accounting books could make $\frac{\partial C}{\partial z} \leq 0$. Along these lines, as Kleven et al. [2009] argue, if these books are known to many employees, because of whistleblowing rewards, the entrepreneur is less likely to hide them successfully from the IRS, so $\frac{\partial C}{\partial n \partial z} \leq 0$.

Consequently, I adopt an agnostic view about the audit cost. In particular, I look for restrictions on its partial derivatives that are sufficient to support the qualitative results from the previous section.

For simplicity, I assume that the concavity of the Hamiltonian with respect to n is preserved.[43] Hence, from the set of sufficient conditions for an optimum, only items 1 and 2 change.

$$1'. \quad \left[z + (1 - \mu\tau\phi^o(z))\frac{\lambda(z)}{g(z)}\right]\alpha p n^o(z)^{\alpha-1} = 1 + \frac{\partial C}{\partial n}(z, n^o(z))\phi^o(z)$$

$$2'. \quad C(z, n^o(z))g(z) + \lambda(z)\mu\tau p n^o(z)^{\alpha} = \omega(z).$$

It is easy to verify that Proposition 2, which assumes that $z_1 \geq z_2$,[44] would still be valid whenever the LHS of $2'$ is greater than, or equal to, zero for all $z \leq z_1$.[45]

If $z_2 \geq z_1$, for $z \in [z_3, z_2]$, the optimal mechanism solves the system of equations in (7). To account for a general audit cost, the first and second equations in (7) need to be substituted for $1'$ and $2'$ (with $\omega(z) = 0$). A close inspection of Appendix A, especially the proofs of Lemmas 2 and 3, reveals that the extra term $\frac{\partial C}{\partial n}(z, n^o(z))\phi(z)$ in $1'$ complicates

[43] A sufficient condition is $\frac{\partial^2 C}{\partial n^2} \geq 0$.

[44] Note that the definition of z_1 does not change, but z_2 needs to be redefined. In particular, $z_2 = \sup_{s \in [\underline{z}, \overline{z}]}\{B(s) < C(s, (\alpha p[s - h(s)])^{\frac{1}{1-\alpha}})\}$.

[45] A sufficient condition is $z\frac{\partial C}{\partial n}(z, n^o(z))\frac{dn^o}{dz}(z) + z\frac{\partial C}{\partial z}(z, n^o(z)) \leq \frac{1}{1-\alpha}\frac{\mu\tau}{\alpha}\gamma(1-\tau)(\alpha p z)^{\frac{1}{1-\alpha}}$.

the characterization of the mechanism. Thus, I set $\frac{\partial C}{\partial n} = 0$ and study the case in which the audit cost C depends only on z.[46]

Under the additional assumption that $zC'(z) \leq \frac{1}{1-\alpha}C(z)$, one can follow the steps in Section 5 and verify that the characterization of the mechanism is qualitatively the same. Note that this assumption can accommodate a non-monotone audit cost.

If $zC'(z) > \frac{1}{1-\alpha}C(z)$ at some range, the characterization of the mechanism would be more complicated. Intuitively, if the cost to audit increases at a high rate, the IRS might prefer to save audit expenses by following a non-monotone monitoring strategy in $[z_3, z_2)$.

6.3 Tax schedule

By imposing a tax schedule of the form $T(x) = \tau x$, I rule out possible interactions between non-linearities on the tax system and the optimal monitoring strategy.[47]

Consider a general tax schedule T, twice continuous differentiable. For simplicity, I assume that the concavity of the Hamiltonian with respect to n is preserved. From the set of sufficient conditions for an optimum, the first-order condition with respect to n becomes

$$\left[z + (1 - \mu T_o'(z)\phi^o(z))\frac{\lambda(z)}{g(z)}\right] \alpha p n^o(z)^{\alpha-1} = 1 + \mu T_o''(z)\phi^o(z)\frac{\lambda(z)}{g(z)}\left[\alpha p z n^o(z)^{\alpha-1} - 1\right] p n^o(z)^\alpha,$$

where $T_i(z) = T(\pi(n^i(z), z))$, $T_i'(z) = T'(\pi(n^i(z), z))$, and $T_i''(z) = T''(\pi(n^i(z), z))$, for $i = o, *$. Unfortunately, n^o enters this expression in a convoluted way, jeopardizing any attempt to extend the analytical results in Proposition 3 to a general tax schedule.

If c is high enough, such that $\phi^o(z) = 0$ for all z, an analogous proposition to Proposition 2, in which $T_*'(z)$ plays the role of τ, can be derived.[48] The crucial step is to show that $\frac{d\lambda}{dz}(z) \leq g(z)$, which is true under the assumptions that $T' \in [0, 1)$, $T'' > 0$, and $h(z)$ is

[46] z_3 also needs to be redefined. In particular, $z_3 = \sup_{s\in[\underline{z},\overline{z}]}\{\frac{\mu\tau}{\alpha}\gamma(1-\tau)(\alpha p s)^{\frac{1}{1-\alpha}} < C(s)\}$.

[47] If $T(x) = \tau_0 + \tau x$, the characterization of the optimal mechanism does not change. However, net revenue collection is increasing in τ_0.

[48] In this case, z_1 can be defined in a similar way, and for $z \leq z_1$, $\lambda(z) = -\gamma(z) z g(z)$, where $\gamma(z) = [1 - (1 - T_*'(z))^{\frac{\alpha}{1-\alpha}}]$.

non-increasing.

6.4 Penalty and utility functions

Linear penalties and risk neutrality are the hardest assumptions to relax. Consider penalties, for instance. Under the assumption that the penalty function M is differentiable, the local incentive-compatibility constraint can be rewritten as

$$\frac{dU}{dz}(z) = (1 - \phi(z)M'(e(z))\tau)pn(z)^{\alpha} - (1 - \tau)pn^{*}(z)^{\alpha}, \text{ a.e.,}$$

where $e(z) = \tau(\pi(n(z), z) - x(z))$. If $M'(e)$ depended on e, the IRS problem could not be rewritten in terms of informational rent U since x would still pop out in the Hamiltonian. However, this trick substantially facilitates the use of optimal control techniques in order to solve the problem. A similar argument can be developed for risk neutrality.[49]

7 A quantitative exploration

The message of this paper is that the IRS can increase net revenue if it is willing to impose distortions almost everywhere. Hence, in order to inform policy, it is desirable to quantify this trade-off.

Ideally, one would like to embed the IRS's actual practiced monitoring strategy into a general model and use data on reported profits, actual profits, inputs, and audits to pin down the distribution of managerial ability and the parameters of the model, such as the audit cost c. Thus, it would be possible to counterfactually assess the implications of the mechanism developed in this paper.

[49] In a standard mechanism-design problem, the principal's objective is usually written in terms of informational rents, instead of compensatory transfers. Consequently, transfers can not pop out in the set of local incentive-compatibility constraints. This is obtained under the commonly used assumption that utility is quasi-linear in transfers. In this paper, both linear penalties and risk neutrality allow reported income to play a role similar to that of transfers.

However, this approach poses two challenges. First, in most countries, including the U.S., the audit scheme is strictly guarded or too obscure.[50] Hence, it cannot be used as a benchmark. Second, even in countries in which the tax-collection agency commits to a publicly known monitoring strategy, such as in Italy,[51] the lack of public data limits this approach.

Therefore, in order to provide some assessment of the trade-off between revenue collection and efficiency, I follow an alternative approach. In particular, I focus on the potential revenue gains from adopting the optimal mechanism. Let R^o be the revenue collection generated by the optimal mechanism, and R^a be the actual revenue collected by the IRS from some set of self-employed entrepreneurs. I establish a lower bound to R^o/R^a.

7.1 Back-of-the-envelope calculation

Let Y^i and X^i, $i = a, o, *$, be aggregate pre-tax profits and reported income, respectively. Recall that a stands for the actual figures, o for the optimal mechanism outcome, and $*$ for the full-information outcome, in which production is carried out efficiently. Notice that $Y^* \geq \max\{Y^a, Y^o\}$. Moreover, let $rep^i \equiv X^i/Y^i$ be the fraction of aggregate income that is reported to the IRS. Finally, let $\overline{\phi}$ be the maximum probability that a firm is actually audited by the IRS. Consequently,

$$\begin{aligned} \frac{R^o}{R^a} &\equiv \frac{\tau X^o + \int \phi^o(z)[\mu\tau(\pi(n^o(z), z) - x^o(z)) - c]dG(z)}{\tau X^a + \int \phi^a(z)[\mu\tau(\pi(n^a(z), z) - x^a(z)) - c]dG(z)} \geq \\ &\geq \frac{X^o}{X^a + \mu\overline{\phi}(Y^a - X^a)} = \frac{rep^o \times \frac{Y^o}{Y^a}}{rep^a + \mu\overline{\phi}(1 - rep^a)} \geq \frac{rep^o \times \frac{Y^o}{Y^*}}{rep^a + \mu\overline{\phi}(1 - rep^a)}. \end{aligned} \tag{9}$$

[50] Andreoni et al. [1998] describe how audit policy is conducted in the U.S. for individual income tax returns. In a first stage, intensive audits are conducted on a stratified random sample. Then, these results are used to assess the likelihood that a report contains evasion. Slightly over one-half of all audit selections are based at least partly on this method. However, the rules used to assign each report a likelihood that it contains irregularities are strictly guarded.

[51] In Italy, for a given class of firms under the *Studi di settore*, presumed sales proceeds are statistically inferred from easily observable variables, such as the surface area of offices and warehouses, the number of employees, the type of customers, and so on. In particular, a small- or medium-sized firm can be audited if it reports sales proceeds that are lower than a presumed level. See Arachi and Santoro [2007] and Santoro [2008] for more details.

The first inequality imposes that c is high enough, such that $\phi^o(z) = 0$ for all z, which underestimates the net revenue gain from adopting the optimal mechanism. Consequently, the RHS of (9) is a lower bound to R^o/R^a. Since information on μ, $\overline{\phi}$, and rep^a can be gathered, it remains to calculate rep^o and Y^o/Y^*, which I do next.

7.2 Calculating Y^*, Y^o, and rep^o

Recall from Section 6.1 that the model accommodates a production technology of the form $zn^{\alpha_0} \prod_{i=1}^{I} k_i^{\alpha_i}$, as long as only labor n is observable. Consequently, pre-tax profits can be rewritten as

$$\zeta(z; p, \{r_i\}_{i=1}^{I}, \{\alpha_i\}_{i=1}^{I})n^{\alpha} - n,$$

where $\alpha = \alpha_0 / \left(1 - \sum_{i=1}^{I} \alpha_i\right) \in (0, 1)$, and ζ is a function of the prices, the technological parameters, and the managerial ability.[52] If z follows a truncated Pareto distribution, which is assumed from now on, ζ also does. Let β be the shape parameter of ζ's underlying distribution, $\hat{G}$, which has support $[\underline{\zeta}, \overline{\zeta}]$. The idea is to apply the optimal mechanism directly to $\hat{G}$.

To calculate Y^*, Y^o, and rep^o, I follow an approach similar to that of Guner et al. [2008]'s analysis of policies that depend on firm size. In particular, I assume that the U.S. is a "relatively distortion-free" competitive economy and use employment data to impose some discipline on the distribution of managerial ability. By relatively distortion-free, I mean an economy in which policies do not target the firm size. By competitive, I mean an economy in which entrepreneurs take prices as given and maximize expected profits. Hence,

$$n^*(\zeta) = (\alpha\zeta)^{\frac{1}{1-\alpha}}. \tag{10}$$

Note that distortions that do not target firm size, such as an input linear tax, are innocuous.

[52] Recall that production $zn^{\alpha_0} \prod_{i=1}^{I} k_i^{\alpha_i}$ is not equal to $\zeta(z; p, \{r_i\}_{i=1}^{I}, \{\alpha_i\}_{i=1}^{I})n^{\alpha}$. Therefore, without specifying values to prices, it is not possible to identify the efficiency loss induced by the optimal mechanism.

Indeed, in addition to prices, technological parameters, and managerial ability, ζ would also absorb these distortions.

Note that I am implicitly assuming that either the actual monitoring strategy does not depend on labor input (or any other proxy for firm size), or entrepreneurs do not internalize it whenever they make employment decisions. Although it is likely that the IRS uses information on business size to select audits, many authors have argued that taxpayers have little knowledge of the audit function. In particular, they tend to overestimate the probability of an IRS audit. See, for example, Andreoni et al. [1998].

By combining data on employment at each firm with equation (10), I set the parameters $\underline{\zeta}$, $\overline{\zeta}$, and β to match some properties of the data. Hence, the assumption that the U.S. is a relatively distortion-free competitive economy is needed only to impose some discipline on the parameters underlying the distribution $\hat{G}$, which are subject to sensitivity analysis in Appendix B.

Define $\eta \equiv 1 - \sum_{i=0}^{I} \alpha_i$, which is the share of output that goes to the entrepreneur as pre-tax profits, so that $\alpha = \alpha_0/(\alpha_0 + \eta)$. By calibrating η, α_0, $\underline{\zeta}$, $\overline{\zeta}$, and β, I can calculate the full-information aggregate profits, $Y^* \equiv \int [\zeta n^*(\zeta)^\alpha - n^*(\zeta)] d\hat{G}(\zeta)$.

To calculate Y^o and rep^o and, thus, generate the counterfactual, I make two additional assumptions. First, I rule out general equilibrium effects through prices; that is, I assume that p and $\{r_i\}_{i=1}^{I}$ are fixed. Second, I assume that the distribution of occupations is fixed, such that entrepreneurs are not allowed to become workers and vice-versa. These assumptions make the distribution of ζ, $\hat{G}$, invariant to policy. Thus, the optimal mechanism can be applied directly to $\hat{G}$, which was backed out from employment data in a previous step.

If the demand for inputs in the corporate sector is relatively large, once compared with the set of firms considered in this exercise, general equilibrium effects through prices is a minor issue. Moreover, given that prices are fixed and, due to withholding or third-party report, wage taxes are fully enforced without cost, the possibility of moving across occupations would further increase net revenue. Indeed, once adopted, the optimal mechanism reduces profits in the entrepreneurial sector. Therefore, a standard occupational-choice model would predict more workers fully complying with taxes and fewer entrepreneurs managing firms, which reinforces the aim of this exercise in providing a lower bound to R^o/R^a.[53]

Finally, in order to establish the RHS of (9), I assume that c is high enough, such that only labor distortions are used in equilibrium, which underestimates the net revenue gain from adopting the optimal mechanism. Consequently, to calculate Y^o and X^o, I use the optimal allocation derived in Proposition 2. Notice that I do not need to specify values for the audit cost c and the linear penalty μ to calculate these figures.

7.3 Calibration

Following Guner et al. [2008], I set $\eta = 0.2$, and $\alpha_0 = (1 - \eta) \times 0.6$.

To calibrate $\underline{\zeta}$, $\overline{\zeta}$, and β, I use data on employment and entrepreneurship from the 2001 Survey of Consumer Finance (SCF). In particular, I restrict the sample to families in which one member actively manages and owns only one business, as in the theory presented above, and employs at least one worker. I consider only nonfarm sole proprietors, a group that underreported 57 percent of its income in 2001, as documented in Slemrod [2007]. This procedure leads to 825 observations. Table 1 shows some descriptive statistics for the number of workers at each firm.

	Mean	St. dev.	Min	Max	Obs
# workers	6.92	26.68	1	299	825

Table 1: Descriptive statistics (weighted). Source: SCF 2001.

# workers	[1,5)	[5,10)	[10,20)	[20,100)	[100,299]
Data	80.0%	13.1%	2.9%	2.5%	1.4%
Model	77.1%	10.9%	5.9%	5.2%	1.0%

Table 2: Firm size distribution in the data (weighted) and in the model that assumes a relatively distortion-free competitive economy. Source: SCF 2001 and author's own calculations.

[53] This argument assumes that both self-employers and employees face the same income tax rate τ.

I set $\underline{\zeta} = 1.42$ and $\overline{\zeta} = 7.58$, which implies that $n^*(\underline{\zeta}) = 1$ and $n^*(\overline{\zeta}) = 299$, as in the sample. Similarly, $\beta = 3.07$ generates the average employment per firm in the data. Table 2 shows that this simple strategy matches the observed firm size distribution reasonably well.

Efficient aggregate profits, Y^*, are 14.4, which is reasonable once wages are the numeraire. Under the assumptions stated above, to calculate Y^o and rep^o, I need to specify a value for τ and then use the formulas from Proposition 2.[54]

Piketty and Saez [2007] document that the average federal individual income tax rate was 11.5 percent in 2004. If payroll taxes are added, the tax rate increases to 20.8 percent. Since this figure ignores both local and state taxes, I generate results for τ ranging from 0.15 to 0.35.

I turn to the choice of μ and $\bar{\phi}$. According to the U.S. code, title 26,6663, "if any part of any underpayment of tax required to be shown on a return is due to fraud, there shall be added to the tax an amount equal to 75 percent of the portion of the underpayment which is attributable to fraud." However, typically, penalties are assessed at a rate of 20 percent of the amount underpaid (Andreoni et al. [1998]).

Similarly, the IRS does not seem to rely on an intensive use of audits to enforce taxes. According to the 2001 IRS Data Book, fewer than two percent of the schedule C returns were audited.[55] Hence, setting $\mu\bar{\phi} = 0.15$ seems a conservative choice, although I also generate results for $\mu\bar{\phi}$ ranging from 0.05 to 0.25.[56]

It is worth mention that if non-pecuniary penalties, such as the possibility of imprisonment, were properly accounted for in the model, R^o/R^a would increase even more. The same argument is valid for other types of cost, such as the financial cost of hiring professional assistance or the moral cost of being an outlaw.

[54] Given the values chosen for the parameters, the truncated Pareto distribution does not satisfy Assumption 1. However, for the purpose of this empirical exercise, an inspection of the proof of Proposition 2 reveals that this assumption can be relaxed as follows: 1. $z - h(z)$ is non-decreasing for $z \geq z_1$; 2. $\gamma z - h(z) \leq 0$ for $z \leq z_1$; and 3. $-\gamma[g(z) + zg'(z)] \leq g(z)$ for $z \leq z_1$. These properties are satisfied in this section.

[55] Schedule C returns are those filed by nonfarm self-employed taxpayers.

[56] Even if some taxpayers face high probabilities of being audited, by plugging a smaller value for $\bar{\phi}$, it is still likely that the inequality in (9) is satisfied, especially if these taxpayers represent a small fraction of the population.

Finally, following Slemrod [2007], I set $rep^a = 0.43$, which is the fraction of nonfarm sole-proprietor income reported to the IRS in 2001. This figure is provided by the IRS, which combines information from a program of random intensive audits, ongoing enforcement activities, and other special studies about particular sources of income (such as cash earnings) that is unlikely to be uncovered even in an intensive audit. Alternatively, I also generate results for $rep^a = 0.65$, which Feldman and Slemrod [2007] estimate by using unaudited tax returns data for 1999 in the U.S. Under the assumption that charity donations are unrelated to the source of income, these authors adapt the econometric approach in Pissarides and Weber [1989] to estimate self-employment income underreporting.[57]

7.4 Results

Table 3 displays the results for $rep^a = 0.43$. It reports the lower bound to R^o/R^a, which is the RHS of (9).

			lower bound to R^o/R^a				
τ	rep^o	$\frac{Y^o}{Y^*}$	$\bar{\phi}\mu = 0.05$	$\bar{\phi}\mu = 0.10$	$\bar{\phi}\mu = 0.15$	$\bar{\phi}\mu = 0.20$	$\bar{\phi}\mu = 0.25$
0.15	0.91	0.98	1.95	1.83	1.73	1.64	1.56
0.20	0.88	0.97	1.87	1.76	1.66	1.57	1.50
0.25	0.86	0.96	1.79	1.68	1.59	1.51	1.43
0.30	0.83	0.94	1.71	1.61	1.52	1.44	1.37
0.35	0.81	0.92	1.62	1.53	1.45	1.37	1.30

Table 3: Results for $rep^a = 0.43$.

A conservative choice of the parameters indicates substantial gains, in terms of net revenue, from adopting the optimal mechanism. If $\tau = 0.25$ and $\bar{\phi}\mu = 0.15$, for instance, net revenue collection from this set of entrepreneurs would increase at least by 59 percent. For nonfarm proprietor business income, $rep^a = 0.43$ is associated with a tax gap of \$68 billion, as reported in Slemrod [2007].[58] Consequently, the extra revenue collected could potentially

[57] In their original approach, Pissarides and Weber [1989] assume that food consumption, instead of charity donations, is unrelated to the source of income. These authors estimate that income underreporting for British self-employed individuals was approximately 35 percent in 1982.

[58] A tax gap of \$68 billion accounts for only individual federal income taxes, ignoring payroll, local, and state taxes.

be \$30 billion, which was 0.3 percent of the GDP in 2001 or, equivalently, three percent of the total federal individual income taxes collected in 2001.[59] Interestingly, even in the worst scenario – that is, $\tau = 0.35$ and $rep^o = 0.81$ – the degree of evasion is substantially lower than the one observed in the data.

Table 4 displays the results for $rep^a = 0.65$. As expected, the gains from adopting the mechanism are smaller, but still potentially large. For example, if $\tau = 0.2$ and $\bar{\phi}\mu = 0.10$, net revenue collection from this set of entrepreneurs would increase by 25 percent.

			lower bound to R^o/R^a				
τ	rep^o	$\frac{Y^o}{Y^*}$	$\bar{\phi}\mu = 0.05$	$\bar{\phi}\mu = 0.10$	$\bar{\phi}\mu = 0.15$	$\bar{\phi}\mu = 0.20$	$\bar{\phi}\mu = 0.25$
0.15	0.91	0.98	1.34	1.30	1.27	1.24	1.21
0.20	0.88	0.97	1.28	1.25	1.22	1.19	1.16
0.25	0.86	0.96	1.23	1.20	1.17	1.14	1.11
0.30	0.83	0.94	1.17	1.14	1.11	1.09	1.06
0.35	0.81	0.92	1.12	1.09	1.06	1.03	1.01

Table 4: Results for $rep^a = 0.65$.

In Appendix B, I do some sensitivity analysis by varying both β and α. Results are remarkably robust to variations of β, but net revenue collected would increase even more for higher values of α.

Leaving aside explanations based on moral, psychological, or social factors, the discrepancy between the model and the data highlights the inability of the IRS to deter evasion through audits. Given that in the U.S., the strategy to select audits is strictly guarded, these results suggest that the IRS can substantially fight evasion and, thus, increase revenue collection by committing to an optimal monitoring strategy that depends on proxies for business size. Recall that this exercise assumes that c is high enough, such that in equilibrium, $\phi^o(z) = 0$ for all z; thus, these results do not rely on an intensive use of audits, which is consistent with actual practices.

[59] A tax gap of \$68 billion is an aggregation across heterogeneous classes of entrepreneurs, including those that usually do not hire workers, such as independent contractors. Consequently, the potential \$30 billion of extra revenue should be viewed as an imperfect aggregator across classes, given that optimal monitoring strategies were designed for each class based on a single observable input.

The adoption of such mechanism raises other issues, such as the distortions it imposes or its implications for horizontal equity. From an ethical point of view, an IRS behavior that not only tolerates, but also explicitly, induces evasion would be questionable.

Finally, it is important to emphasize that this exercise is somewhat limited, as a non-linear tax schedule has implications for the optimal mechanism, the empirical strategy, and the calibration procedure. It is not clear what kind of bias would arise if non-linearities in the tax schedule were accounted for. I leave it as an open question for further research.

8 Final comments

The relevance of the theory presented in this paper hinges on the plausibility of two hypotheses. First, employment at each firm is easily observable by the IRS. Second, audits are relevant to explain self-employment income tax evasion.

At least in developed countries, to a first approximation, there is empirical evidence suggesting that labor is readily available information to the IRS. Only a tiny percent of wages and salaries are not reported to the IRS in the U.S. (Slemrod [2007]) and in Denmark (Kleven et al. [2011]).[60] If income taxes are subject to third-party report, as in the case of employers withholding taxes on wages and salaries, it is unlikely that the parties involved would collude to evade taxes.[61] Moreover, as long as wages are partially declared, the IRS still has information about the employee and the firm for which she works. Even if unskilled workers, such as illegal immigrants, can be hidden, the extension in Section 6.1 shows that the optimal mechanism can be applied as long as a single input is easily observable. Examples of such input are skilled labor or an intermediate good bought from a formal firm.

What if reasons other than audits are the main determinants of self-employed income tax evasion? Entrepreneurs, for example, might be caught in a web of business-to-business

[60]In contrast, di Porto [2011] documents that in Italy, artisan firms hide their workers in order to pay less in social security taxes.

[61]In the U.S., for instance, firms issue W-2 forms, one for each worker, detailing her identity and the amount of wages paid. Each form is sent to the IRS and the relevant employee. The latter uses the W-2 form to file her income tax return. See Logue and Slemrod [2010] for more details.

transactions that facilitate enforcement for tax reasons.[62] Similarly, access to the financial sector generates information that the IRS can use to enforce taxes (Gordon and Li [2009]). Finally, when the use of accounting books, necessary to run complex business operations, are known to many employees, the entrepreneur is less likely to hide them successfully from the IRS (Kleven et al. [2009]).

Nonetheless, recent evidence from field experiments, in which auditing probabilities are exogenously controlled, shows that audits have a strong impact on self-reported income. See Slemrod et al. [2001] for an experiment in the U.S., and Kleven et al. [2011] for another in Denmark. Consequently, the design of monitoring strategies can play an important role in increasing revenue collection, as the exercise in Section 7 highlights.

[62]For example, whenever a downstream firm buys from an upstream firm, value-added taxes along the production chain generate tax credits to be used against future tax liabilities. Thus, this transaction is observable by the IRS, and compliant firms have an incentive to deal among themselves (Kopczuk and Slemrod [2006] and de Paula and Scheinkman [2010]).

Appendix

A Proofs

The following proofs involve a change of variables that reduces the number of equations in (7). Define $\Phi \equiv 1 - \mu\tau\phi$ and $N \equiv n^{\alpha}$. Using the first and the second equations to eliminate λ from the problem, one reaches the following system of two equations in two unknowns (Φ and N).

$$\begin{aligned} \Phi &= \frac{\mu\tau}{\alpha c}\left[\alpha pzN - N^{\frac{1}{\alpha}}\right] \\ \Phi &= \frac{(1-\tau)[\alpha pz]^{\frac{\alpha}{1-\alpha}}}{N} \end{aligned} \tag{11}$$

Moreover, $\phi \in [0, 1/\mu]$ implies that $\Phi \in [1-\tau, 1]$.

Let $\Phi_1(N) = \frac{\mu\tau}{\alpha c}[\alpha pzN - N^{\frac{1}{\alpha}}]$, $\Phi_2(N) = \frac{(1-\tau)(\alpha pz)^{\frac{\alpha}{1-\alpha}}}{N}$, and define $N^* \equiv (\alpha pz)^{\frac{\alpha}{1-\alpha}}$. Since Φ_2 is strictly decreasing, then any solution to this system implies that $N \in [(1-\tau)N^*, N^*]$.

A.1 Proof of Lemma 2

Notice that Φ_1 is strictly concave, and Φ_2 is strictly convex. Therefore, $\Phi_1 = \Phi_2$ at most at two values of N. Moreover, $\Phi_1(N^*) = 0 < (1-\tau) = \Phi_2(N^*)$; thus, this system has one solution if $\Phi_1((1-\tau)N^*) > \Phi_2((1-\tau)N^*) = 1$. A little algebra shows that this requirement holds if and only if $c < \frac{\mu\tau}{\alpha}[(1-\tau) - (1-\tau)^{\frac{1}{\alpha}}](\alpha pz)^{\frac{1}{1-\alpha}}$.

This system does not have a solution if and only if $\Phi_1(N) < \Phi_2(N)$ for all $N \in [(1-\tau)N^*, N^*]$. Note that $\Phi_1(N) < \Phi_2(N)$ if and only if

$$\frac{\mu\tau[\alpha pzN^2 - N^{\frac{1+\alpha}{\alpha}}]}{\alpha(1-\tau)(\alpha pz)^{\frac{\alpha}{1-\alpha}}} < c.$$

Taking the first-order condition of the LHS with respect to N, one gets $\hat{N} = \left(2\frac{\alpha}{1+\alpha}\alpha pz\right)^{\frac{\alpha}{1-\alpha}}$.

But Assumption 2 implies that $\hat{N} \leq (1-\tau)N^*$. Hence, there are two possible candidates for a maximum: $(1-\tau)N^*$ and N^*. Plugging them into the LHS and choosing the one that yields the highest value, one concludes that this system does not have a solution if and only if $\frac{\mu\tau}{\alpha}[(1-\tau)-(1-\tau)^{\frac{1}{\alpha}}](\alpha pz)^{\frac{1}{1-\alpha}} < c$.

Finally, when $c = \frac{\mu\tau}{\alpha}[(1-\tau)-(1-\tau)^{\frac{1}{\alpha}}](\alpha pz)^{\frac{1}{1-\alpha}}$, $N = (1-\tau)N^*$ and $\Phi = 1$ is the only solution to the system of equations above.

A.2 Proof of Lemma 3

Items 1, 4 and 5 are immediate. Item 2 follows from the fact that $c = \frac{\mu\tau}{\alpha}[(1-\tau)-(1-\tau)^{\frac{1}{\alpha}}](\alpha pz_3)^{\frac{1}{1-\alpha}}$; thus, $N(z_3) = (1-\tau)N^*(z_3)$ and $\Phi(z_3) = 1$. It remains to show 3 – that is, $\frac{dN}{dz} \geq 0$ and $\frac{d\Phi}{dz} \leq 0$.

Equalizing both equations in (11), one obtains

$$\alpha pzN^2 - N^{\frac{1+\alpha}{\alpha}} = \frac{\alpha c}{\mu\tau}(1-\tau)(\alpha pz)^{\frac{\alpha}{1-\alpha}}.$$

Differentiating it with respect to N and z leads to

$$\frac{dN}{dz} = \frac{(1-\tau)\frac{\alpha c}{\mu\tau}\frac{\alpha}{1-\alpha}\frac{(\alpha pz)^{\frac{\alpha}{1-\alpha}}}{z} - \alpha pN^2}{\alpha pz2N - \frac{1+\alpha}{\alpha}N^{\frac{1}{\alpha}}}.$$

For $\frac{dN}{dz} \geq 0$, it is enough to show that both the denominator and the numerator are negative. Indeed, the numerator is smaller than zero if and only if

$$(1-\tau)\frac{\alpha c}{\mu\tau}\frac{\alpha}{1-\alpha}\frac{(\alpha pz)^{\frac{\alpha}{1-\alpha}}}{z} - \alpha pN^2 < 0.$$

Therefore, by plugging in the smallest possible value for N, $(1-\tau)N^*$, and after some manipulations,

$$c < \frac{1-\alpha}{\alpha}(1-\tau)\frac{\mu\tau}{\alpha}(\alpha pz)^{\frac{1}{1-\alpha}}.$$

Since $c \leq \frac{\mu\tau}{\alpha}[(1-\tau)-(1-\tau)^{\frac{1}{\alpha}}](\alpha pz)^{\frac{1}{1-\alpha}}$, it is enough to show that

$$1-(1-\tau)^{\frac{1-\alpha}{\alpha}} < \frac{1-\alpha}{\alpha},$$

which follows from Assumption 2.

The denominator is smaller than zero if and only if

$$\alpha pz 2N - \frac{1+\alpha}{\alpha}N^{\frac{1}{\alpha}} \leq 0 \iff \alpha pz 2\frac{\alpha}{1+\alpha} \leq N^{\frac{1-\alpha}{\alpha}}.$$

Hence, it is enough to plug the smallest possible value for N, $(1-\tau)N^{*}$, into this inequality and check that it holds. Indeed,

$$2\frac{\alpha}{1+\alpha} \leq (1-\tau)^{\frac{1-\alpha}{\alpha}}$$

is true by Assumption 2.

Similarly, from (11),

$$\Phi = \frac{\mu\tau}{\alpha c}\left[(1-\tau)\frac{(\alpha pz)^{\frac{1}{1-\alpha}}}{\Phi} - (1-\tau)^{\frac{1}{\alpha}}\frac{(\alpha pz)^{\frac{1}{1-\alpha}}}{\Phi^{\frac{1}{\alpha}}}\right].$$

Differentiating it with respect to Φ and z, and after some manipulation, leads to

$$\frac{d\Phi}{dz} = \frac{(1-\tau)\frac{\mu\tau}{\alpha c}\frac{1}{1-\alpha}\frac{(\alpha pz)^{\frac{1}{1-\alpha}}}{z}\left[\Phi^{\frac{1-\alpha}{\alpha}} - (1-\tau)^{\frac{1-\alpha}{\alpha}}\right]}{\frac{1+\alpha}{\alpha}\Phi^{\frac{1}{\alpha}} - \frac{\mu\tau}{\alpha c}(1-\tau)(\alpha pz)^{\frac{1}{1-\alpha}}\frac{1-\alpha}{\alpha}\Phi^{\frac{1}{\alpha}-2}}.$$

Therefore, given $\Phi \geq (1-\tau)$, for $\frac{d\Phi}{dz} \leq 0$, it is enough to show that the denominator of the previous equation is negative. That is,

$$\frac{1+\alpha}{\alpha}\Phi^{\frac{1}{\alpha}} - \frac{\mu\tau}{\alpha c}(1-\tau)(\alpha pz)^{\frac{1}{1-\alpha}}\frac{1-\alpha}{\alpha}\Phi^{\frac{1}{\alpha}-2} \leq 0 \iff c \leq \frac{\mu\tau}{\alpha}(1-\tau)(\alpha pz)^{\frac{1}{1-\alpha}}\frac{1-\alpha}{1+\alpha}\Phi^{-2}.$$

Since $c \leq \frac{\mu\tau}{\alpha}[(1-\tau)-(1-\tau)^{\frac{1}{\alpha}}](\alpha pz)^{\frac{1}{1-\alpha}}$, it is enough to show that

$$1-(1-\tau)^{\frac{1-\alpha}{\alpha}} \leq \frac{1-\alpha}{1+\alpha},$$

which follows from Assumption 2.

A.3 Proof of Lemma 4

Notice that $\lambda(z) = -\frac{cg(z)}{\mu\tau p N(z)}$. Taking derivatives and reorganizing the terms, one obtains

$$\frac{d\lambda}{dz}(z) = \frac{1}{N(z)}\frac{c}{\mu\tau p}\left[\frac{N'(z)}{N(z)}g(z) - g'(z)\right].$$

If $\frac{N'(z)}{N(z)}g(z) - g'(z) \leq 0$, then the assertion is trivial. On the other hand, if $\frac{N'(z)}{N(z)}g(z) - g'(z) > 0$,

$$\begin{aligned} \frac{d\lambda}{dz}(z) &\leq \frac{z}{N(z)}[(1-\tau)-(1-\tau)^{\frac{1}{\alpha}}](\alpha pz)^{\frac{\alpha}{1-\alpha}}\left[\frac{N'(z)}{N(z)}g(z) - g'(z)\right] \leq \\ &\leq \gamma\left[\frac{N'(z)}{N(z)}zg(z) - zg'(z)\right], \end{aligned} \tag{12}$$

where the first inequality follows from $c \leq \frac{\mu\tau}{\alpha}[(1-\tau)-(1-\tau)^{\frac{1}{\alpha}}](\alpha pz)^{\frac{1}{1-\alpha}}$ for all $z \in [z_3, z_2)$, while the second follows from $N(z) \in [(1-\tau)N^*, N^*]$ for all $z \in [z_3, z_2)$.[63]

Note that

$$\begin{aligned} z\frac{N'(z)}{N(z)} &= \frac{\alpha pzN(z) - (1-\tau)\frac{\alpha c}{\mu\tau}\frac{\alpha}{1-\alpha}(\alpha pz)^{\frac{\alpha}{1-\alpha}}\frac{1}{N(z)}}{\frac{1+\alpha}{\alpha}N(z)^{\frac{1}{\alpha}} - 2\alpha pzN(z)} < \\ &< \frac{\alpha pz}{\frac{1+\alpha}{\alpha}N(z)^{\frac{1-\alpha}{\alpha}} - 2\alpha pz} \leq \frac{1}{\frac{1+\alpha}{\alpha}(1-\tau)^{\frac{1-\alpha}{\alpha}} - 2} = \frac{1}{\rho}. \end{aligned} \tag{13}$$

Consequently, by plugging (13) into (12), it is straightforward to verify that

$$\gamma \leq \rho + z\frac{g'(z)}{g(z)}\gamma\rho \Rightarrow \frac{d\lambda}{dz}(z) \leq g(z).$$

[63] Recall that $\gamma \equiv 1-(1-\tau)^{\frac{1-\alpha}{\alpha}}$.

A.4 Proof of Proposition 3

For $z \geq z_2$, the proof is outlined in the text. Hence, if $z_2 = \underline{z}$, the result follows. Assume that $z_2 > \underline{z}$. For $z \in [z_3, z_2)$, set $\phi^o(z) = \hat{\phi}(z)$, $n^o(z) = \hat{n}(z)$, and $\lambda^o(z) = \hat{\lambda}(z)$, which satisfy conditions 1, 2, and 5. Moreover, Property 1 and condition 3 imply $\theta(z) \geq 0$; thus, from 6, I set $U(z) = 0$, which agrees with the differential equation in 4 equalized to zero, given that $U(z_3) = 0$ is the boundary condition. If $z_3 = \underline{z}$, since $\hat{\lambda}(z_3) < 0$, condition 7 is satisfied.

It remains to show that 8 is satisfied. Note that at $z = z_2$, both conditions 1 and 2 (with $\omega = 0$) hold with equality, but $0 = \frac{d^- U^o}{dz}(z_2) < \frac{d^+ U^o}{dz}(z_2)$. Hence, at $z = z_2$, $n^o(z)$ and $\phi^o(z)$ are discontinuous.

If the domain of N and Φ are extended to $[0, N^*]$ and $[1-\tau, \infty)$, respectively, by following similar steps to the ones in the proof of Lemma 2, one can verify that this system always has two solutions if $0 < c < \frac{\mu\tau}{\alpha}[(1-\tau) - (1-\tau)^{\frac{1-\alpha}{\alpha}}](\alpha p z)^{\frac{1}{1-\alpha}}$. Index these solutions by l and h. At $z = z_2$, it is easy to verify that, without lost of generality, $\hat{n}_l(z_2) < n^o(z_2^+) < \hat{n}_h(z_2)$, $\hat{\phi}_l(z_2) < \phi^o(z_2^+) < \hat{\phi}_h(z_2)$, and $\hat{\lambda}_l(z_2) < \lambda(z_2^+) < \hat{\lambda}_h(z_2)$. By restricting the domain of N and Φ to be between $[(1-\tau)N^*, N^*]$, and $[(1-\tau), 1]$, respectively, the solution indexed by l is lost. Consequently, 8 is satisfied.

Finally, for $z < z_3$, given that $z_3 > \underline{z}$, the proof is similar to the one in Proposition 2, provided that Assumption 3 holds.

A.5 Working out equation (8)

If $F(z, n, k_1, ..., k_I) = z n^{\alpha_0} \prod_{i=1}^{I} k_i^{\alpha_i}$, expected profits are

$$pzn^{\alpha_0} \prod_{i=1}^{I} k_i^{\alpha_i} - n - \sum_{i=1}^{I} r_i k_i - \tau x - \mu\tau\varphi(n, x) \left(pzn^{\alpha_0} \prod_{i=1}^{I} k_i^{\alpha_i} - n - \sum_{i=1}^{I} r_i k_i - x \right). \quad (14)$$

By taking the first-order condition of (14) with respect to k_I, which is chosen in the last stage of the game, one obtains

$$k_I = \left(\frac{\alpha_I}{r_I} p z n^{\alpha_0} \prod_{i=1}^{I-1} k_i^{\alpha_i} \right)^{\frac{1}{1-\alpha_I}}.$$

By plugging it back into (14), expected profits can be rewritten as

$$\zeta(z; p, r_I, \alpha_I) n^{\tilde{\alpha}_o} \prod_{i=1}^{I-1} k_i^{\tilde{\alpha}_i} - n - \sum_{i=1}^{I-1} r_i k_i - \tau x - \mu \tau \varphi(n, x) \left(\zeta(z; p, r_I, \alpha_I) n^{\tilde{\alpha}_o} \prod_{i=1}^{I-1} k_i^{\tilde{\alpha}_i} - n - \sum_{i=1}^{I-1} r_i k_i - x \right),$$

where $\tilde{\alpha}_i = \frac{\alpha_i}{1-\alpha_I}$, and $\zeta(z; p, r_I, \alpha_I) = \left[\left(\frac{\alpha_I}{r_I} \right)^{\frac{\alpha_I}{1-\alpha_I}} - r_I \left(\frac{1}{r_I} \right)^{\frac{\alpha_I}{1-\alpha_I}} \right] (pz)^{\frac{\alpha_I}{1-\alpha_I}}$.

Proceeding iteratively, one reaches (8). Moreover, if z is either log-normally or Pareto distributed, then ζ also is.

B Sensitivity analysis

In this section, I check the robustness of the empirical results obtained in Section 7 by varying β, η, and α_o. In particular, I generate results for β equal to 4.13 and 2.49, which are consistent with an average employment of four and ten, respectively. I also set $\eta = 0.12$ and $\alpha_o = (1 - \eta) \times 0.64$, which is in line with the calibration by Kitao [2008] in another context.

Tables 5 and 6 report the results for $rep^a = 0.43$ and $rep^a = 0.65$, respectively. Notice that the results are remarkably robust to variations of β, but net revenue collected would increase even more for higher values of α.

		lower bound to R^o/R^a							
		$\eta = 0.2$ and $\alpha_0 = (1-\eta) \times 0.6$				$\eta = 0.12$ and $\alpha_0 = (1-\eta) \times 0.64$			
	τ	rep^o	$\bar{\phi}\mu = 0.05$	$\bar{\phi}\mu = 0.10$	$\bar{\phi}\mu = 0.15$	rep^o	$\bar{\phi}\mu = 0.05$	$\bar{\phi}\mu = 0.10$	$\bar{\phi}\mu = 0.15$
$\beta =$ 4.13	0.15	0.91	1.94	1.83	1.73	0.92	1.98	1.86	1.76
	0.20	0.88	1.86	1.75	1.66	0.90	1.90	1.79	1.69
	0.25	0.86	1.78	1.68	1.58	0.87	1.83	1.72	1.63
	0.30	0.83	1.70	1.60	1.51	0.85	1.75	1.65	1.56
$\beta =$ 3.07	0.15	0.91	1.95	1.83	1.73	0.92	1.98	1.86	1.76
	0.20	0.88	1.87	1.76	1.66	0.90	1.91	1.80	1.70
	0.25	0.86	1.79	1.68	1.59	0.88	1.84	1.73	1.63
	0.30	0.83	1.71	1.61	1.52	0.86	1.76	1.66	1.57
$\beta =$ 2.49	0.15	0.91	1.95	1.83	1.73	0.92	1.98	1.86	1.76
	0.20	0.88	1.87	1.76	1.66	0.90	1.91	1.80	1.70
	0.25	0.86	1.79	1.68	1.59	0.88	1.84	1.73	1.64
	0.30	0.83	1.71	1.61	1.52	0.86	1.77	1.67	1.57

Table 5: Results for $rep^a = 0.43$.

		lower bound to R^o/R^a							
		$\eta = 0.2$ and $\alpha_0 = (1-\eta) \times 0.6$				$\eta = 0.12$ and $\alpha_0 = (1-\eta) \times 0.64$			
	τ	rep^o	$\bar{\phi}\mu = 0.05$	$\bar{\phi}\mu = 0.10$	$\bar{\phi}\mu = 0.15$	rep^o	$\bar{\phi}\mu = 0.05$	$\bar{\phi}\mu = 0.10$	$\bar{\phi}\mu = 0.15$
$\beta =$ 4.13	0.15	0.91	1.34	1.30	1.27	0.92	1.36	1.32	1.29
	0.20	0.88	1.28	1.25	1.22	0.90	1.31	1.27	1.24
	0.25	0.86	1.22	1.19	1.16	0.87	1.27	1.22	1.19
	0.30	0.83	1.16	1.13	1.11	0.85	1.20	1.17	1.14
$\beta =$ 3.07	0.15	0.91	1.34	1.30	1.27	0.92	1.36	1.32	1.29
	0.20	0.88	1.28	1.25	1.22	0.90	1.31	1.28	1.24
	0.25	0.86	1.23	1.20	1.17	0.88	1.26	1.23	1.20
	0.30	0.83	1.17	1.14	1.11	0.86	1.21	1.18	1.15
$\beta =$ 2.49	0.15	0.91	1.34	1.30	1.27	0.92	1.36	1.32	1.29
	0.20	0.88	1.28	1.25	1.22	0.90	1.31	1.28	1.25
	0.25	0.86	1.23	1.20	1.17	0.88	1.26	1.23	1.20
	0.30	0.83	1.18	1.15	1.12	0.86	1.22	1.18	1.15

Table 6: Results for $rep^a = 0.65$.

References

M. G. Allingham and A. Sandmo. Income tax evasion: A theoretical analysis. *Journal of Public Economics*, 1(3-4):323–338, November 1972.

J. Andreoni, B. Erard, and J. Feinstein. Tax compliance. *Journal of Economic Literature*, 36(2):818–860, June 1998.

G. Arachi and A. Santoro. Tax enforcement for smes: Lessons from the italian experience? *eJournal of Tax Research*, 5(2):225–243, 2007.

M. Bassetto and C. Phelan. Tax riots. *Review of Economic Studies*, 75(3):649–669, July 2008.

G. S. Becker. Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2):169–217, March/April 1968.

S. Bigio and E. Zilberman. Optimal self-employment income tax enforcement. *Journal of Public Economics*, 95(9-10):1021–1035, October 2011.

K. C. Border and J. Sobel. Samurai accountant: A theory of auditing and plunder. *Review of Economic Studies*, 54(4):525–540, October 1987.

P. Chander and L. L. Wilde. General characterization of optimal income tax enforcement. *Review of Economic Studies*, 65(1):165–183, January 1998.

H. Cremer and F. Gahvari. Tax evasion and the general income tax. *Journal of Public Economics*, 60(2):235–249, May 1996.

H. Cremer, M. Marchand, and P. Pestieau. Evading, auditing and taxing: The equity-compliance tradeoff. *Journal of Public Economics*, 43(1):67–92, October 1990.

A. de Paula and J. A. Scheinkman. Value added taxes, chain effects and informality. *American Economic Journal: Macroeconomics*, 2(4):195–221, 2010.

E. di Porto. Undeclared work, employer tax compliance, and audits. *Public Finance Review*, 39(1):75–102, January 2011.

B. Erard and J. Feinstein. Honesty and evasion in the tax compliance game. *RAND Journal of Economics*, 25(1):1–19, Spring 1994.

N. E. Feldman and J. Slemrod. Estimating tax noncompliance with evidence from unaudited tax returns. *Economic Journal*, 117(518):327–352, March 2007.

D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1991.

R. Gordon and W. Li. Tax structures in developing countries: Many puzzles and a possible explanation. *Journal of Public Economics*, 93(7-8):855–866, August 2009.

N. Guner, G. Ventura, and Y. Xu. Macroeconomic implications of size-dependent policies. *Review of Economic Dynamics*, 11(4):721–744, October 2008.

B. Jullien. Participation constraints in adverse selection models. *Journal of Economic Theory*, 93(1):1–47, July 2000.

S. Kitao. Entrepreneurship, taxation and capital investment. *Review of Economic Dynamics*, 11(1):44–69, January 2008.

H. J. Kleven, C. T. Kreiner, and E. Saez. Why can modern governments tax so much? an agency model of firms as fiscal intermediaries. NBER Working Paper 15218, 2009.

H. J. Kleven, M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez. Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica*, 79(3):651–692, May 2011.

W. Kopczuk and J. Slemrod. Putting firms into optimal tax theory. *American Economic Review Papers and Proceedings*, 96(2):130–134, May 2006.

T. R. Lewis and D. E. M. Sappington. Countervailing incentives in agency problems. *Journal of Economic Theory*, 49(2):294–313, December 1989.

K. D. Logue and J. Slemrod. Of coase, calabresi, and optimal tax liability. *Tax Law Journal*, Summer 2010.

R. E. Lucas. On the size distribution of business firms. *Bell Journal of Economics*, 9(2): 508–523, Autumn 1978.

I. Macho-Stadler and J. D. Pérez-Castrillo. Optimal auditing with heteregeneous income sources. *International Economic Review*, 38(4):951–968, November 1997.

I. Macho-Stadler and J. D. Pérez-Castrillo. Audting with signals. *Economica*, 69(273):1–20, February 2002.

G. Maggi and A. Rodríguez-Clare. On countervailing incentives. *Journal of Economic Theory*, 66(1):238–263, June 1995.

N. D. Melumad and D. Mookherjee. Delegation as commitment: The case of income tax audits. *RAND Journal of Economics*, 20(2):139–163, Summer 1989.

P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, March 2002.

D. Mookherjee and I. Png. Optimal auditing, insurance, and redistribution. *Quarterly Journal of Economics*, 104(2):399–415, May 1989.

T. Piketty and E. Saez. How progressive is the u.s. federal tax system? a historical and international perspective. *Journal of Economic Perspectives*, 21(1):3–24, Winter 2007.

C. A. Pissarides and G. Weber. An expenditure-based estimate of britain's black economy. *Journal of Public Economics*, 39(1):17–32, June 1989.

A. M. Polinsky and S. Shavell. The economic theory of public enforcement of law. *Journal of Economic Literature*, 38(1):45–76, March 2000.

J. F. Reinganum and L. L. Wilde. Income tax compliance in a principal-agent framework. *Journal of Public Economics*, 26(1):1–18, February 1985.

J. F. Reinganum and L. L. Wilde. Equilibrium verification and reporting policies in a model of tax compliance. *International Economic Review*, 27(3):739–760, October 1986.

H. S. Rosen. *Public Finance*. McGraw-Hill, 2005.

I. Sánchez and J. Sobel. Hierarchical design and enforcement income tax policies. *Journal of Public Economics*, 50(3):345–369, March 1993.

A. Sandmo. The theory of tax evasion: A retrospective view. *National Tax Journal*, 58(4): 643–663, December 2005.

A. Santoro. Taxpayer's choices under studi di settore: What do we know and how we can interpret it? *Giornale degli Economisti e Annali di Economia*, 67(2):161–184, July 2008.

S. Scotchmer. Audit classes and tax enforcement policy. *American Economic Review Papers and Proceedings*, 77(2):229–233, May 1987.

A. Seierstad and K. Sydsæter. *Optimal Control Theory with Economic Applications*. North-Holland, 1987.

J. Slemrod. Cheating ourselves: The economics of tax evasion. *Journal of Economic Perspectives*, 21(1):25–48, Winter 2007.

J. Slemrod and S. Yitzhaki. Tax avoidance, evasion and administration. In A. J. Auerbach and M. Feldstein, editors, *Handbook of Public Economics*, volume 3, chapter 22, pages 1423–1470. Elsevier B.V., 2002.

J. Slemrod, M. Blumenthal, and C. Christian. Taxpayer response to an increased probability of audit: Evidence from a controlled experiment in minnesota. *Journal of Public Economics*, 79(3):455–483, March 2001.

R. M. Townsend. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory*, 21(2):265–293, October 1979.

Departamento de Economia PUC-Rio
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marques de Sâo Vicente 225 - Rio de Janeiro 22453-900, RJ
Tel.(21) 35271078 Fax (21) 35271084
www.econ.puc-rio.br
flavia@econ.puc-rio.br