Journal of Public Economics 95 (2011) 1021–1035

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Optimal self-employment income tax enforcement

Saki Bigio, Eduardo Zilberman *

Department of Economics, New York University, United States

ARTICLE INFO

Article history:
Received 18 June 2009
Received in revised form 28 February 2010
Accepted 10 June 2010
Available online 16 June 2010

JEL classifications:
D21
H26

Keywords:
Optimal auditing
Tax evasion
Informal sector
Missing middle
Entrepreneurship
Presumptive taxation

ABSTRACT

Most models of optimal income tax enforcement assume that income is either random or solely remunerates labor, neglecting that auditing strategies may depend on observable inputs. This paper outlines a model to optimally monitor self-employed entrepreneurs when, in addition to reported profits, the tax collection agency also observes the number of workers employed (or any other input variable) at each firm. We show that, by conditioning the monitoring strategy only on labor input, it is optimal for the IRS to audit firms in a way that generates some empirical regularities, like the *missing middle*. We also show that the optimal direct mechanism can be implemented by an indirect monitoring strategy that is consistent with actual IRS practices. In particular, the IRS calculates inputted income as function of labor. Whenever an entrepreneur reports profits that are lower than inputted income, she is randomly monitored. Finally, we formalize a model of optimal presumption taxation, in which inputted income is the tax base, to compare revenue collection across tax systems.

1. Introduction

The optimal income tax enforcement literature[1] mostly focuses on individual taxpayers whose income is either exogenous or solely remunerates labor/effort supply decisions.[2] However, Slemrod (2007) reports that in the U.S., only 1% of wages and salaries are underreported to the tax collection agency (henceforth the IRS), whereas 43% of self-employment business income is not.[3] This evidence suggests that the degree by which the IRS can enforce taxes varies across occupations. In particular, taxes on wages and salaries are more easily enforced than on self-employment business income.[4]

In practice, wage taxes are virtually always enforced. Thus, the number of workers at each firm seems to be costless information available to the IRS. This paper shifts the focus from individuals to firms, and asks the following: given that employment per firm is observable, what is the optimal scheme to monitor heterogeneous entrepreneurs?[5]

This paper makes three contributions. First, we show that, when the monitoring strategy is conditioned only on labor input, it is optimal for the IRS to audit firms in a way that generates some empirical regularities. In particular, the model generates a *missing middle*, that is the fact that medium-scale firms are scarce. Second, when auditing probabilities depends both on labor input and reported income, we show that optimal auditing theory is consistent with actual IRS practices. Finally, by allowing taxpayers to self-report their income subject to audits and penalties, we show that the IRS collects more revenue than by imposing a tax base presumed from labor input. On the technical side, we solve a mechanism design problem that goes beyond quasi-linearity and that features a type-dependent individual rationality constraint.

In Section 2, we introduce a general model, in which heterogeneous entrepreneurs (or self-employed business owners) can underreport their individual income, that is, the profits generated by their own firms. The source of heterogeneity is an unobserved random managerial ability that enhances productivity in a plant exhibiting

* Corresponding author.
E-mail addresses: sbigio@nyu.edu (S. Bigio), eduardo.zilberman@nyu.edu (E. Zilberman).

[1] The seminal paper is Reinganum and Wilde (1985), which is inspired by the costly state verification model of Townsend (1979). Notable contributions are Border and Sobel (1987), Mookherjee and Png (1989), Sánchez and Sobel (1993), Cremer and Gahvari (1996), Chander and Wilde (1998), and Bassetto and Phelan (2008). The first theoretical work on tax noncompliance is Allingham and Sandmo (1972). Recent surveys are Andreoni et al. (1998), Slemrod and Yitzhaki (2002), and Sandmo (2005).

[2] An exception is Parker (2010), who introduces entrepreneurship and dynamic incentives in a two-period model.

[3] These figures account for 10 and 109 billion dollars, respectively.

[4] Kleven et al. (2010) define tax evasion rate as the share of reported income that is underreported. Using data from Denmark, they estimate it as being 8.1% for self-employment income, 0.4% for earnings, 0.3% for third-party reported income, and 37% for self-reported income.

[5] In our model, firms and entrepreneurs make up a single unit. Henceforth, we use both terms interchangeably.

0047-2727/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2010.06.011

decreasing returns to scale. Entrepreneurs face an audit probability that depends on labor input and reported income, which are assumed to be costlessly observed by the IRS. If monitored, the firm must pay a linear penalty on the underreported amount.

The model is a two-stage game. In the first stage, given the managerial ability distribution, a budget-constrained IRS commits to a monitoring strategy that depends on employment and reported income. In the second stage, firms take into account this monitoring strategy and choose labor inputs and reported income. To solve this model, we invoke the revelation principle, and cast it into a direct mechanism design problem.

A novelty of this paper is that it allows the monitoring strategy to depend on an observable input choice, interpreted as labor. In the absence of strategic interactions, entrepreneurs with higher ability would hire more labor. However, employment also acts as a signal of the ability, thus entrepreneurs may distort their labor input strategically in order to signal a different ability.

It turns out that the most general version of the problem is hard to solve. More precisely, the problem's formulation implies that, on top of the multidimensional aspect, the agent's objective is not quasi-linear and the individual rationality constraint is type-dependent. Nonetheless, we characterize the solution for two subproblems, one in which the IRS conditions the monitoring strategy only on reported income, and another in which it only conditions on employment.

The first problem has been studied previously (e.g., Sánchez and Sobel (1993)). In this context, the optimal monitoring strategy is to randomly audit firms which report below some threshold level of profits. In equilibrium, only low-ability entrepreneurs report honestly. This result is quickly reviewed in Section 3.1. In this mechanism, larger firms are the set of evaders, and every firm produces efficiently.

In Section 3.2, we study the opposite case in which the auditing probability depends solely on labor input. In contrast, the optimal strategy is to randomly monitor firms that employ labor above some threshold, while the rest are not monitored at all. Intuitively, labor input is a signal about the true managerial ability. Since profits increase with ability, by monitoring larger firms, the IRS collects taxes from the most profitable ones. As a result, smaller firms are the set of evaders. Moreover, in order to avoid detection, some of the firms distort labor input.

Hence, in contrast with the optimal tax enforcement literature, this mechanism is consistent with three empirical regularities: (1) a negative relationship between firm size, measured by employment, and amount evaded (Dabla-Norris et al. (2008)); (2) some of the evaders, if interpreted as "informals," producing inefficiently (la Porta and Shleifer (2008)); and (3) the existence of a missing middle (Tybout (2000)).

Not surprisingly, some papers in the development literature explain these facts in an adapted version of the Lucas (1978) model of entrepreneurial choice that allows an informal sector.[6] Our model is closely related to this literature, except that auditing policies are optimal here which, in principle, could revert these theoretical results. Consequently, it is optimal for the IRS to audit firms in a way that corroborates these facts.

In Section 4, we provide a partial characterization of the general solution that explains auditing practices in many countries. In particular, we restrict the monitoring strategy to be bang-bang, as the solutions to the previous two particular cases suggest. We show that any solution to the direct mechanism can be implemented by an indirect monitoring strategy as follows: The IRS calculates inputted income as a function of labor. Whenever an entrepreneur reports profits that are higher than inputted income, she is not monitored. If reported income is lower than inputted income, the IRS randomly monitors her. This auditing scheme is consistent with actual IRS practices, as we illustrate in Section 4.1, with anecdotal evidence.

The use of presumptions of income to tax or audit (as in this result) is the central principle of presumptive taxation.[7] Although it has been widely used in many countries, there is little theory formalizing optimal presumptive methods. This paper is a step towards this direction.[8]

A natural question arises in this framework. What if inputted income, as a function of labor, is designed to tax instead of audit entrepreneurs? In Section 5, we formalize and solve a model of presumptive taxation, in which the inputted income is the tax base. The optimal solution is characterized for tax systems in which presumptives of income are, and are not, rebuttable.[9] In particular, whenever presumptives of income are irrebuttable, the problem fits a standard mechanism design problem. Hence, labor is not distorted only at the top-type, and the informational rent is fully appropriated at the bottom-type. In contrast, in the rebuttable tax system, the reservation value is type-dependent, and standard tricks in the literature are not readily applicable. As before, employment is distorted everywhere except at the top-type. However, the IRS potentially appropriates the full informational rent not only at the bottom-type, but also among a continuum of low-types.

The mathematical structure of the problem allows us to compare revenue collection across tax systems, even under the unsolved general case, in which the monitoring strategy depends both on labor and reported income. In Section 6, we show that the revenue collected in the rebuttable presumptive tax system is a lower bound on the revenue collected in the general case. To derive this result, we assume that there is no cost to claim and prove that someone's income was presumed wrongly. If this assumption is violated, presumptive tax methods might generate more revenue. This comparison suggests some explanations on why modern tax systems are the rule, but presumptive taxation is still observed in many countries (see Bird and Wallace (2004) for a list).

2. General model

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

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. There is a single good produced with a single input, labor n. The production technology is $zf(n)$ common to all firms. Wages are the numeraire, and p is the price of the good. Therefore, pre-tax profits are given by $\pi(n,z)=pzf(n)-n$.

Assumption 1. The function $f:N\to\mathbb{R}_+$ is strictly increasing, strictly concave, and twice continuously differentiable, with $f(0)=0$ and $N\subseteq\mathbb{R}_+$.

[6] Rauch (1991) is the seminal contribution. See also Fortin et al. (1997), Choi and Tum (2005), Amaral and Quintin (2006), Antunes and Cavalcanti (2007), Dabla-Norris et al. (2008), and de Paula and Scheinkman (2009).

[7] There is a large practitioner literature on presumptive taxation (e.g., Tanzi and de Jantscher (1989) and Thuronyi (2004)).

[8] Scotchmer (1987) and Macho-Stadler and Pérez-Castrillo (2002) also consider a framework in which the IRS observes a signal of taxpayers' true income. In contrast, income is exogenous in both papers. In Scotchmer (1987), taxpayers are grouped into classes, and auditing probabilities depend on reported income and the class to which the individual belongs. Macho-Stadler and Pérez-Castrillo (2002) assume taxpayers do not know the realization of the signal.

[9] Under a rebuttable presumptive tax system, a taxpayer can claim and prove that her presumptive income is higher than her actual income. In contrast, under an irrebuttable presumptive tax system, taxpayers cannot claim any revision. See Tanzi and de Jantscher (1989) for more details.