OPTIMAL PENALTIES ON INFORMAL FIRMS[1]

APRIL 2020

ANDERS FREDRIKSSON

CORS — Center for Organization Studies, FEA-USP, Universidade de São Paulo, Av. Prof. Luciano Gualberto, 908, São Paulo CEP 05508-900, SP, Brazil

anders.fredriksson@usp.br

Abstract

What, if anything, should be done about the informal economy in developing countries? I study optimal penalties vis-à-vis informal firms in a simple capital accumulation model under three different government objectives: maximize formalization, maximize tax revenue from formalizing informal firms, and maximize welfare. A general conclusion, for all objectives, is that low productivity informal firms should be left alone. Higher productivity informal firms should instead face positive penalties. As the three objectives lead to differences in the range and severity of penalties towards such firms, however, the study also highlights the importance of discussing the appropriate policy objective vis-à-vis informality.

JEL classification: E26, H26, K42, O17

Keywords: Informality; penalties; enforcement; formalization; public policy; global

[1] I am much indebted to Paulo Furquim for discussions that helped shape this work. I also thank Sandro Cabral, Love Ekenberg, Shon Ferguson, Harry Flam, Johan Gars, Gustav Hansson, Ann Harrison, Bård Harstad, John Hassler, Per Krusell, Luiz Mesquita, Ted Miguel, Marcos Nakaguma, Annika Nilsson, José Mauricio Prado, Jesper Stage, Jakob Svensson, Kaj Thomsson, Ragnar Torvik and Fabrizio Zilibotti for helpful discussions, feedback and suggestions, as well as participants at the Berkeley development lunch, Stockholm University and IIES brown bag seminars, University of Gothenburg and Maastricht/UNU-MERIT seminars, and conference participants at: SUDSWEc, the Nordic Development Economics Conference (NCDE), the Research Workshop on Institutions and Organizations (RWIO-Brazil), the Association of Swedish Development Economists (ASWEDE) and the European Economic Association (EEA). Any errors are mine.

1. Introduction

What should governments do about informal production? This question is ever present in the development debate, at least since the recognition of an informal "sector", in the 1970's (ILO, 1972; Hart, 1973). According to several estimates, 35-40% of all economic activity in developing countries is in the informal sector, with a higher share, in some countries 70-80%, of employment (Bigsten et al, 2004; Chen, 2007; La Porta and Shleifer, 2014; Hassan and Schneider, 2016; Loayza, 2016B). Hence the question is highly relevant.

In this paper, I introduce a penalty policy vis-à-vis informal firms, in a simple model of capital accumulation, growth and potential formalization, and study how penalties should be optimally set, as a function of firm productivity. In the model, and similar to much of the literature on informality, formalization means paying a fixed entry cost to register with the government authorities. I analyze three different policy objectives: maximize formalization per se, maximize future tax contributions from informal firms that become formal, and maximize welfare, respectively. A general conclusion from the paper is that low productivity informal firms should not be penalized, which holds for all three objectives. Given informal sector productivity distributions in most developing countries, this result implies that the overwhelming majority of informal firms should be left alone. Higher productivity informal firms should instead face positive penalties. As the three policy objectives lead to differences in the range and severity of optimal penalties towards higher productivity informal firms, however, the study highlights the importance of also discussing the appropriate policy objective vis-à-vis informality.

Some developing country governments leave informal firms on their own. This can be because monitoring of such (typically) small firms is costly, the probability of an increased compliance is low, and, if achieved, would bring only minor benefits to the state. Alternatively, the government recognizes that most informal firms have very low productivity levels to start with, and penalization would make the firm/owners/workers even worse off. Other countries have a size/productivity dependent policy. Larger firms are monitored and penalized, if operating informally, with the goal of ultimately making these firms formal, whereas small informal firms are left alone.

The optimal penalty policy obviously depends on the objective, and welfare maximization by a benevolent government/social planner is a natural starting point. However, welfare effects in models of informality are disputed. One line of argument holds that informal entrepreneurs act as unfair competition for formal firms, as the former do not comply with tax, labor and environmental regulation, which gives a cost advantage (Farrell, 2004). Relatedly, informal firms do not contribute to public goods, which indirectly affects formal firm productivity and other outcomes (e.g. Loayza, 1996; Johnson et al, 1997; Dessy and Pallage, 2003; Ihrig and Moe, 2004; Levy, 2008; Loayza and Rigolini, 2011). Other negative externalities may also be present. A different view stresses the entrepreneurial- and growth potential of informal firms, which are held back by government regulation (de Soto, 1989). Such regulation can imply costs to become formal ("entry costs") or to stay formal (e.g. taxes, labor costs etc.). Some models based on this argument do not consider any negative effects whatsoever from reducing entry costs, thus mechanically increasing welfare. This approach has been criticized from the perspective of a trade-off between ex-ante and ex-post transaction costs (Arruñada, 2007; Arruñada and Manzanares, 2015), hence rationalizing the regulation of firm entry into formality. Yet another perspective on informality, which dates further back, is that the informal sector mainly consists of individuals who cannot find other employment opportunities, either temporarily (because of labor market conditions related to the business cycle) or permanently (because of structural barriers to formal employment). The impact on formal firms is minimal.

Different variants of these opposing views—ultimately affecting how a model of informal-formal linkages and a welfare analysis should be constructed—are discussed in the literature. Bruhn, de Mel, McKenzie and Woodruff, with influential studies of informal firms in developing countries, use the label "Tokman vs. de Soto" (de Mel, McKenzie and Woodruff, 2010; Bruhn, 2013). Tokman (2007) sees informality as marginalized individuals conducting some economic activity in waiting for a formal sector job, of which there are too few. This differs sharply from the de Soto (1989) entrepreneurial perspective. Using Sri Lankan data on the personal characteristics of wage workers, firm owners and own-account workers, i.e. individuals conducting small-scale informal economic activity, in combination with a "species classification" approach, de Mel, McKenzie and Woodruff (2010) conclude that around 70% of own-account workers resemble wage workers, and 30% resemble entrepreneurs/firm owners. Similar exercises are conducted for Mexico (Bruhn, 2013) and Benin (Benhassine et al., 2016). This categorization

is kindred to the Global Entrepreneurship Monitor´s dichotomy of opportunity vs. necessity entrepreneurs, where the latter category is relatively more common in developing countries (Reynolds et al, 2001), especially in economic downturns (Loayza and Rigolini, 2011).

Using extensive data to take stock of the discussion on how to best describe and model developing country informality, La Porta and Shleifer (2008, 2014) argue for a "dual" model (inspired by Lewis, 1954), which contrasts both to de Soto´s (1989) "romantic" and Farrell´s (2004) "parasitic" view. According to La Porta and Shleifer, most informal firms are small, unproductive, sell low-quality products, paid in cash, to different customers than those targeted by formal firms, and can typically not compete in the formal sector. Informal firm managers are significantly less educated, and largely constitute a different pool of individuals, than those running formal sector businesses. The informal and formal sectors of the economy are largely separated. One consequence of this perspective is that most informal firms would not benefit from formalization. La Porta and Shleifer (2014, p. 125) recommend "extreme caution with policies that impose on them any kind of additional costs". Ulyssea (2018), in turn, estimates a model that includes informality on both the extensive and intensive margins (unregistered firms, and registered firms with unregistered workers, respectively), thus suggesting a unifying framework that contains the three classes of informal firms, i.e. subsistence/survival, non-compliance/parasitism and "de Soto" entrepreneurship. By undertaking a counterfactual analysis and comparing the resulting equilibria, the effects of different entry-, tax- and enforcement policies are studied.[2]

Differently from Ulyssea (2018), and the literature at large, this paper studies, in a capital accumulation model of informal firm investment, growth and potential formalization, how an enforcement policy vis-à-vis informal firms, here referred to as *penalties*, should be set, as a function of the informal firm´s productivity level. I take as given both the formal sector tax level and positive entry costs, and thus focus on the optimal penalty schedule. I first study the objective of speeding up formalization. Subsequently, I ask how penalties should be set to maximize a firm´s contribution to tax revenue, once formal. I then outline also how a welfare objective can be analyzed, within the same framework. The restriction to the penalty instrument

[2] Other studies incorporating the above empirical regularities are Ranis and Stewart (1999), considering traditional and modernizing components of the informal sector, and Loayza (2016A,B), with a rudimentary/modern division.

allows for a detailed analysis of the differential dynamic effects penalties have on firms of different productivity, and the incentive to formalize. The capital accumulation model incorporates the above La Porta and Shleifer (2014) concern about low productivity levels, as, for such firms, penalties act as a disincentive to invest. Indeed, Loayza (2016A) and Ulyssea (2018) raise similar concerns, recognizing the detrimental effects increased penalties can have on (a range of) informal firms´ profits, welfare and other variables, even if such penalties succeed in reducing informality. To elaborate more on this point, to which I also return in the discussion section, consider the typical penalty specifications used in the literature. These entail either linear penalties (e.g. Loayza, 1996; Johnson et al, 1997; Ihrig and Moe, 2004; Prado, 2011; Nguimkeu, 2016) or convex penalties (e.g. Ulyssea, 2018; and papers referenced therein), where the latter approach can be rationalized by a probability of detection that increases in informal firm size. Yet, it is very likely that the least productive and smallest informal firms will respond to any penalty level by downsizing their operation (even further), perhaps even shut down the business altogether. In the literature this has been called hiding, lurking in the shadows, staying under the radar, operating at a suboptimally small scale, etc. There may also be general equilibrium effects (as in Ulyssea, 2018). Even with a policy objective that aims at increasing formalization per se, it would seem reasonable, for the firms that would not formalize under any circumstance, to leave them alone, at least if there are no sizeable negative externalities. The standard specifications of linear or convex penalties cannot capture such eventualities. As for more productive firms, the modeling approach with productivity-dependent penalties is also in line with the argument of de Mel, McKenzie and Woodruff (2010), that the growth potential of the informal firms that do resemble formal firm owners should not be neglected.

The recent empirical literature, summarized by Bruhn and McKenzie (2014), is largely disappointing with respect to de Soto´s vision, yet provides further modeling insights. Most reforms aiming at incentivizing firm formalization produce meager results, with perhaps a tenth of firms responding to different incentives. Bruhn (2013) finds that Mexican own account workers classified as resembling formal entrepreneurs are more likely to respond to formalization incentives, than are those classified as wage workers. Benhassine et al. (2016) find similar results for Benin. These results support the above view of (at least) two different types of informal entrepreneurship. An atypical study, with quite large estimated effects, is de Andrade et al. (2014). Higher enforcement levels were randomly assigned, in a sample of informal firms in

Belo Horizonte, Brazil (the firms had average monthly profits of USD 1000). The formalization rate increased 21-27 percentage points as a result of the intervention, suggesting that, for such (quite large informal) firms, increased enforcement can be a policy that works. Bruhn and McKenzie (2014) argue that "the key question for policymakers is then what, if anything, they should attempt to do about this vast quantity of small-scale informal firms" (p. 187). Their discussion, and the above references and empirical regularities, do provide tentative answers. First, formalization "per se" is questionable as a policy objective, although one rationale is that a large informal sector may undermine rule of law in general. Second, increasing tax revenue is a much more legitimate objective, and formalization efforts should then probably target the "relatively well-off" informal firms. Third, increased enforcement may be a good idea, but the authors are skeptical about attempting to formalize subsistence enterprises (pp. 198-199). With the above discussion, it is not surprising that a derivation of optimal policies under these objectives delivers a "zero first, then increasing" penalty, as a function of informal firm productivity. In addition, a welfare objective also leaves the least productive firms alone.

The paper proceeds as follows. In section 2 I introduce a simple capital accumulation model where an informal firm can achieve a productivity gain, but only after paying a fixed formalization fee. I study if and when the firm formalizes, and analyze this decision as a function of the firm´s productivity level and the informal-formal productivity differential. Section 3 then discusses two kindred problems. First, I solve for the informal firm productivity level for which the time until formalization is as low as possible, then for the productivity level that maximizes (present value) tax contributions from the firm, once having formalized. Next, section 4 solves for the productivity-dependent penalties needed to achieve the optimal productivity levels, for these two objectives. Section 4 also discusses the role and impact of enforcement/monitoring costs vis-à-vis informal firms. Section 5 proceeds to generalize the model of sections 2-4, further discusses some of the assumptions made and the validity of the results, and also examines the implications of the results in relation to actual firm productivity distributions. Section 6 analyses welfare. Section 7 discusses the results, with some of the derivations, and two model extensions, in the appendix.

2. A model of informal firm formalization

This section sets up a dynamic model of firm investment, growth and possible formalization. Starting out as informal, the question in this section is whether, when and at what size the firm will become formal, and how this decision depends on the informal firm's productivity level, A^i. The derivation arrives at expressions 9-12, which are then analyzed in the sections that follow.[3] The production function is simple: output is linear in the capital stock, k_t. As informal, the firm produces $A^i k_t$, if it has formalized, production is $A^f k_t$, with $A^f > A^i$. Thus, formality is desirable. The firm can grow by investing (i_t) in its capital stock, with a convex investment cost, $\frac{z}{2} i_t^2$. This gives a profit flow, $\pi_t = A^i k_t - \frac{z}{2} i_t^2$. The capital stock depreciates at rate δ, and therefore accumulates as $k_t' = i_t - \delta k_t$. To get access to the higher productivity, the firm must pay a formalization fee F, at some time T. After formalization, flow profits equal $A^f k_t - \frac{z}{2} i_t^2$. The firm discounts future profits at rate ρ. Starting with capital k_0, the informal firm chooses an investment path, whether it should become formal and the time of formalization (T).

Before proceeding, I briefly discuss three aspects of the model. First, real-life formal sector productivity is not constant, as will be the case below. The specification in sections 2-4 serves to get a tractable dynamic model, and to discuss the main dynamic effects affecting firm formalization. In section 5, I discuss a more general specification in which an underlying entrepreneurial ability parameter affects a firm´s productivity, if operating informally as well as formally, such that A^i and A^f both increase in ability. Section 5 discusses the conditions under which the results of sections 2-4 carry over to such a model. It also further motivates the informal-formal productivity differential. Second, production is a function of capital only. As the focus is on dynamic aspects, I exclude labor from the production function. I show in the last section of the appendix, however, that the results remain unchanged if there is also a labor decision. Third, I think of each firm as owning its capital stock. One interpretation of such a model is that of households receiving dividends from firms. Maximizing such dividends is the same problem as maximizing firm profits (Barro and Sala-i-Martin, 1999, pp. 119-121).

[3] Some parts of the modeling are inspired by the framework in Harstad and Svensson (2011). McKenzie and Sakho (2010), in an otherwise empirical paper, refer to a dynamic firm optimization problem similar to the one specified here.

Returning to the model specification, the profit maximization problem can be written as:

Choose i_t, T to

Max $\left[\int_0^T (A^i k_t - \frac{z}{2} i_t^2)\, e^{-\rho t} dt + \int_T^\infty (A^f k_t - \frac{z}{2} i_t^2)\, e^{-\rho t} dt - F e^{-\rho T}\right]$ s.t. $k_t' = i_t - \delta k_t$ and $k(0) = k_0$.

The problem is solved in two steps. First, the principle of optimality is used to solve backwards for the formal and then for the informal investment path (assuming T exists). We also derive the investment path if T does not exist. Under the assumption that formalization does take place, we then solve for the formalization time T. If no such T exists, the firm is informal forever.

2.1 Optimal investments

First assume T exists. Solving backwards, the "formal problem" takes the capital stock at time T, defined as $\widetilde{k_T}$, as an initial condition, and is solved for the investment path from T to ∞. We get an investment function i^{formal} and a continuation value $V^{formal}(T, \widetilde{k_t})$, which is the optimal profit from T and onwards. The profit maximization problem is:

Choose i_t to Max $\int_T^\infty (A^f k_t - \frac{z}{2} i_t^2)\, e^{-\rho t} dt$ s.t. $k_t' = i_t - \delta k_t$ and $k(T) = \widetilde{k_T}$. (1)

By defining the present-value Hamiltonian $H(t, i, k, \lambda) = \left(A^f k_t - \frac{z}{2} i_t^2\right) e^{-\rho t} + \lambda_t (i_t - \delta k_t)$, where λ_t is the present value Lagrange multiplier on the capital accumulation constraint, and applying the first-order conditions $\frac{\partial H(..)}{\partial i} = 0, \frac{\partial H(..)}{\partial k} = -\frac{\partial \lambda}{\partial t}$ and the transversality condition $Lim_{t\to\infty}(\lambda_t k_t) = 0$, we get:

$$i^{formal} = \frac{A^f}{z(\delta+\rho)},\ k_t^{formal} = \widetilde{k_T} e^{-\delta(t-T)} + \frac{A^f}{z\delta(\delta+\rho)}\left(1 - e^{-\delta(t-T)}\right),$$

$$V^{formal} = e^{-\rho T}\left(\frac{A^f \widetilde{k_T}}{\delta+\rho} + \frac{(A^f)^2}{2z\rho(\delta+\rho)^2}\right). \quad (2)$$

The firm invests a constant amount each "period". The capital stock converges to its steady state value, $k_\infty^{formal} = \frac{A^f}{z\delta(\delta+\rho)}$, at which depreciation and investment offset each other.[4]

[4] A non-explosive path of investment is profit-maximizing. Other paths, that fulfill the differential equations for i_t and k_t, are ruled out for optimality reasons (and do not fulfill $Lim_{t\to\infty}(\lambda_t k_t) = 0$). Investment is constant due to the

The informal investment path, for a given T, can, in turn, be determined by solving for the path that takes the firm from k_0 to $\widetilde{k_T}$ and then maximize total profits with respect to $\widetilde{k_T}$:

Choose i_t, $\widetilde{k_T}$ to

$$\text{Max}\left[\int_0^T (A^i k_t - \frac{z}{2} i_t^2)\, e^{-\rho t} dt + V^{formal}(T, \widetilde{k_T}) e^{-\rho T}\right] \text{ s.t. } k_t' = i_t - \delta k_t,\ k(0) = k_0 \text{ and } k(T) = \widetilde{k_T}. \quad (3)$$

The solution is derived as above, the difference being the terminal constraint on k_t.[5] We get

$$i_t^{formalization} = \frac{A^i}{z(\delta+\rho)} + \frac{(A^f - A^i)e^{(\delta+\rho)(t-T)}}{z(\delta+\rho)},$$
$$k_t^{formalization} = k_0 e^{-\delta t} + \frac{A^i(1-e^{-\delta t})}{z\delta(\delta+\rho)} + \frac{(A^f - A^i)(e^{(\delta+\rho)(t-T)} - e^{-(\delta+\rho)T-\delta t})}{z(\delta+\rho)(2\delta+\rho)}. \quad (4)$$

The investment path starts close to $\frac{A^i}{z(\delta+\rho)}$, and then increases up to the level of formal investments at T, i.e. $\frac{A^f}{z(\delta+\rho)}$. Investment increases because the marginal value of capital is higher after formalization, making it optimal for the firm to decrease profits by accumulating more capital, while still informal.

Now assume T does not exist. The firm is then informal forever. The set-up is as in the above formality problem, but productivity is A^i, time runs from 0 and initial capital is k_0. The "ever-informal" problem is:

$$\text{Choose } i_t \text{ to Max } \int_0^\infty (A^i k_t - \frac{z}{2} i_t^2)\, e^{-\rho t} dt \text{ s.t. } k_t' = i_t - \delta k_t \text{ and } k(0) = k_0, \text{ with solution} \quad (5)$$

$$i^{informal} = \frac{A^i}{z(\delta+\rho)},\ k_t^{informal} = k_0 e^{-\delta t} + \frac{A^i}{z\delta(\delta+\rho)}\left(1 - e^{-\delta t}\right). \quad (6)$$

The investment rate is again constant and the capital stock converges to $k_\infty^{informal} = \frac{A^i}{z\delta(\delta+\rho)}$.[6]

---

convexity of costs - the firm wants to spread investment over time. Investment increases in the productivity parameter A^f and decreases in the cost of investment z, the depreciation rate of capital δ and the rate of time preference ρ.

[5] Solving for the paths of i_t and k_t as functions of $\widetilde{k_T}$ and plugging these back into the profit function, we then integrate to get the optimal value of informal profits as a function of $\widetilde{k_T}$, and then differentiate with respect to $\widetilde{k_T}$. The optimality condition, i.e. $\frac{d}{d\widetilde{k_T}}\left(\int_0^T (A^i k_t(\widetilde{k_T}) - \frac{z}{2}(i_t(\widetilde{k_T}))^2)\, e^{-\rho t} dt + V^{formal}(T, \widetilde{k_T}) e^{-\rho T}\right) = 0$, is that the loss of informal profits from increasing $\widetilde{k_T}$ should be exactly offset by a gain in formal profits.

2.2 Solving for the formalization time T

If T exists, the investment path before and after formalization is given by expressions 4 and 2, respectively. The optimal T can be derived by recognizing that, at the time of formalization, it must be that formalization is just as attractive as remaining informal. This determines the capital stock at which the firm formalizes, which, with $k_t^{formalization}$ from (4), in turn gives T. The firm thus formalizes when

$$\frac{d}{dT}\left(\int_0^T \left(A^i k_t - \frac{z}{2} i_t^2\right) e^{-\rho t} dt + \int_T^\infty \left(A^f k_t - \frac{z}{2} i_t^2\right) e^{-\rho t} dt - F e^{-\rho T}\right) = 0. \quad (7)$$

As discussed above, pre-formalization investment approaches the formal investment rate as $t \to T$. At T, these effects cancel out. Condition 7 becomes $A^i k_T - A^f k_T + \rho F = 0$. The optimal capital stock at formalization, defined as k^F, thus becomes

$$k^F \equiv \frac{\rho F}{A^f - A^i}. \quad (8)$$

We get T by equating the optimal capital accumulated at $t = T$, i.e. $k_T^{formalization}$, with k^F:

$$\frac{A^i(1-e^{-\delta T})}{z\delta(\delta+\rho)} + \frac{(A^f - A^i)(1-e^{-(2\delta+\rho)T})}{z(\delta+\rho)(2\delta+\rho)} = \frac{\rho F}{A^f - A^i}. \quad (9)$$

This equation implicitly defines T. The productivity range for which firms ever formalize is derived by setting $T = \infty$ in (9), which gives a second-order equation in A^i, with formalization for $A_1^i < A^i < A_2^i$, where

$$A_1^i = \frac{\rho A^f - (2\delta+\rho)^{\frac{1}{2}}((A^f)^2(2\delta+\rho) - 4z\delta\rho F(\delta+\rho)^2)^{1/2}}{2(\delta+\rho)}, \; A_2^i = \frac{\rho A^f + (2\delta+\rho)^{\frac{1}{2}}((A^f)^2(2\delta+\rho) - 4z\delta\rho F(\delta+\rho)^2)^{1/2}}{2(\delta+\rho)}. \quad (10)$$

Note that $A_2^i < A^f$ (if $F > 0$). With $F > F_{min}$, we also get "low-end informality", i.e. $A_1^i > 0$, where

$$F_{min} = \frac{(A^f)^2}{z\rho(\delta+\rho)(2\delta+\rho)}. \quad (11)$$

---

[6] In the analysis that follows, I will set $k_0 = 0$, in order to focus on productivity differences between firms. Given that long-run capital levels (such as $k_\infty^{informal}$) depend on A^i, a feasible constraint on initial capital would also have to be a function of A^i, which introduces a new source of heterogeneity between firms, without much additional insight.

Finally, the relation between T and A^i, derived from (9), and with $\pi \equiv \frac{A^i}{A^f}$, can be written as

$$\pi = \tilde{\pi} \pm \sqrt{\tilde{\pi}^2 + \frac{\delta\left(1-e^{-(2\delta+\rho)T}\right)-z\delta\rho(\delta+\rho)(2\delta+\rho)F/(A^f)^2}{(2\delta+\rho)\left(1-e^{-\delta T}\right)-\delta\left(1-e^{-(2\delta+\rho)T}\right)}}, \text{ where } \tilde{\pi} = \frac{(2\delta+\rho)\left(1-e^{-\delta T}\right)-2\delta\left(1-e^{-(2\delta+\rho)T}\right)}{2\left((2\delta+\rho)\left(1-e^{-\delta T}\right)-\delta\left(1-e^{-(2\delta+\rho)T}\right)\right)}. \quad (12)$$

Expressions 9-12 and Lemma 1 contain the main results and intuition on which the analysis depends. First, with the entry cost restriction in (11), there is a low-productivity range for which formalization will not occur. In addition, there is also a high-productivity range for which formalization will not occur, which is in line with empirical observations (e.g. Ulyssea, 2018). Second, the left-hand side of (9) is the informal firm´s capital accumulation, which is increasing in A^i, thus facilitating formalization. The right-hand side, however, represents the benefits of formalizing: the higher is A^i, the less the firm has to gain (for a constant A^f). Expression (10) can be interpreted in a similar manner, in that these counteracting forces result in an A^i-interval over which formalization will occur. For some mid-range productivity, between A_1^i and A_2^i, the capital accumulation and threshold effects will balance, and give the speediest formalization. It can be inferred, perhaps somewhat loosely at this stage, that penalties that lower productivity would act as a "stick" over one range of productivities, but as a "carrot" over another range. This relates directly to the discussion of e.g. Bruhn and McKenzie (2014) about two types of informal firms, and potentially two different policy stances. Third, the properties of expression (12) are crucial for the general validity of the results in the paper. In order for the reasoning about different productivity ranges to always be correct, we need to show that the formalization time, which is infinite at A_1^i and A_2^i, is first always decreasing in A^i, then always increasing. This is lemma 1, proven in the appendix.

Lemma 1. The formalization time T is minimized at a mid-range productivity level $A_{T\_minimum}^i$. For all $A^i \in \left(A_1^i, A_{T\_minimum}^i\right)$, we have $\frac{dT}{dA^i} < 0$, and for all $A^i \in \left(A_{T\_minimum}^i, A_2^i\right)$, $\frac{dT}{dA^i} > 0$.

Proof. See Appendix A1.

There is thus monotonicity in the relation between T and A^i (first *always* decreasing, then *always* increasing), giving the results of sections 2-4 general validity, independent of parameter values. I next solve for the productivity levels that minimize the formalization time and maximize tax revenue, respectively, the comparison of which is straightforward thanks to lemma 1.

3. Speeding up formalization, maximizing tax revenue

This section derives the optimality conditions, as a function of the informal sector productivity parameter, for two related policy objectives: minimize the time to formalization, and maximize present value tax revenue, respectively. Following directly from the above discussion, the first problem is straightforward: The time to formalize is minimized, i.e. $\frac{dT}{dA^i} = 0$, at some interior point in the productivity range given in (10), which is formally stated in theorem 1A.

As for maximizing a firm´s present value tax payments, to which the firm contributes once formal, I first discuss the formal productivity parameter, A^f. It can be thought of as the after-tax productivity, once a revenue/output tax (specified as in e.g. Prado, 2011 or Ulyssea, 2018) has been levied on A^F, the before-tax formal productivity parameter. We thus have $A^f \equiv A^F(1-\tau)$, where τ is the tax rate (which is assumed to be constant, throughout the paper). A^f is the relevant parameter for formalization incentives. With k_t^{formal} from (2), a tax rate τ on output $A^F k_t^{formal}$, and initial capital given by (8), we get a per-period tax revenue, from $t = T$ onwards, of

$$\tau A^F k_t^{formal} = \tau A^F \left(\frac{\rho F}{A^f - A^i} e^{-\delta(t-T)} + \frac{A^f}{z\delta(\delta+\rho)}\left(1 - e^{-\delta(t-T)}\right)\right). \quad (13)$$

Integrating this expression, discounted at the rate ρ, gives a (time zero) net present value of

$$PV(TAX) = \tau A^F e^{-\rho T}\left(\frac{\rho F}{(\delta+\rho)(A^f - A^i)} + \frac{A^f}{z(\delta+\rho)^2\rho}\right), \text{ with derivative} \quad (14)$$

$$\frac{dPV(TAX)}{dA^i} = \tau A^F \rho e^{-\rho T}\left(-\frac{dT}{dA^i}\left(\frac{\rho F}{(\delta+\rho)(A^f - A^i)} + \frac{A^f}{z(\delta+\rho)^2\rho}\right) + \frac{F}{(\delta+\rho)\left(A^f - A^i\right)^2}\right). \quad (15)$$

Setting $\frac{dPV(TAX)}{dA^i} = 0$ gives $\frac{dT}{dA^i} = \frac{F}{(\delta+\rho)\left(A^f - A^i\right)^2}\left(\frac{\rho F}{(\delta+\rho)(A^f - A^i)} + \frac{A^f}{z(\delta+\rho)^2\rho}\right)^{-1}$. Tax revenue is thus maximized when $\frac{dT}{dA^i} > 0$.[7] The above condition for speeding up formalization, $\frac{dT}{dA^i} = 0$, does not consider the size of the firm at formalization, which is instead incorporated in (15). Because $\frac{dT}{dA^i}$ solving (15) is less than infinity, we can infer that also tax revenue is maximized at an interior point in the productivity range. In addition, Lemma 1 gives that the productivity level for which tax revenue is maximized is larger than the productivity level that speeds up formalization the

[7] Expression 15 is also zero at $A^i = A_1^i$ and $A^i = A_2^i$, as we then get $e^{-\rho T} = 0$. Both points correspond to zero revenue, as is the case for $A^i < A_1^i$ and $A^i > A_2^i$, for which expression 15 does not exist.

most. Theorem 1 summarizes these results (where, as customary, the parenthesis notation indicates an interval not including the endpoints).

Theorem 1A. The informal sector firm productivity for which formalization is fastest, $A^i_{T\_minimum}$, lies in the interval $\left(A^i_1, A^i_2\right)$.

Theorem 1B. The informal sector firm productivity for which the present value of the firm´s tax contributions is maximized, $A^i_{TAX\_maximum}$, lies in $\left(A^i_1, A^i_2\right)$, and is larger than $A^i_{T\_minimum}$.

Proof. The results follow from lemma 1 and the above tax derivation. Appendix A3 shows that the stationary points are minimum and maximum points, respectively, making use of the fact that $T(A^i)$ is convex, which is shown in Appendix A2 (the proof concerns the range where $\frac{dT}{dA^i} > 0$).

In the above specification of tax revenue, and elsewhere in the paper, the focus is on the properties of the model as a function of the informal sector productivity parameter A^i. Rather than choosing optimal penalties *and* tax rates, I thus study a more modest and partial question, which still is in line with much of the policy discussion: what to do about informality? In addition, and in line with the arguments of e.g. Arruñada (2007), the paper implicitly assumes that there is some level of entry control that is socially desirable. This motivates a model in which F has a minimum level, and I treat F as fixed, and instead focus on A^i. I next derive optimal penalties for the two objectives.

4. Optimal penalty policies to speed up formalization and to maximize tax revenue

Penalties vis-à-vis the informal sector is a policy instrument the government can use in order to affect formalization incentives. The monitoring/enforcement of informal firms can be that the authorities, e.g. tax inspectors or police, carry out controls of manufacturers, businesses, shops, vendors etc., in geographical areas suspected to contain much informal economic activity, and then penalize firms that are found to operate informally. I first assume that monitoring/ enforcement of informal firms and observation of firm productivity is costless for the government. Although unrealistic, positive monitoring costs strengthen the arguments of the paper, and I therefore postpone a discussion until the end of section 4 (section 4.1).

One way of thinking about the effects of enforcement is that it makes firms divert time from production, in order to avoid inspections, which results in a reduction of productivity and output.[8] I take a more straightforward approach and discuss penalties as the de facto reduction of the informal productivity parameter needed to achieve the above discussed policy objectives, while recognizing that there may be several channels that can bring about such a reduction. In the present context of penalties that affect firm productivity, and as discussed in the preceding paragraph, I assume that firm productivity is observable. Given the high correlation between manager characteristics and firm productivity in the data, cited by e.g. La Porta and Shleifer (2008, 2014), one means through which productivity can be assessed is through observation of firm owner characteristics. Another method is to back out a measure of productivity from observing the firm´s capital stock and output.

From theorem 1, it is straightforward to derive the optimal penalties/productivity reductions needed to speed up formalization or maximize a firm´s contribution to tax revenue. Theorem 2 summarizes this result, which is displayed in figure 1.

Theorem 2A. The penalty level that minimizes the formalization time is first zero, until $A^i = A^i_{T\_minimum}$, then increases one-to-one with A^i.

Theorem 2B. The penalty level that maximizes the present value of the firm´s tax contributions is first zero, until $A^i = A^i_{TAX\_maximum}$, then increases one-to-one with A^i.

Proof. Follows directly from Theorem 1.

In figure 1, there is first a range of productivity levels, up to A^i_1 (from expression 10), for which a firm will never formalize. These firms should always be left alone. Next follows a productivity range $A^i_1 < A^i < A^i_{T\_minimum}$. Applying penalties on these firms would slow down formalization,

[8] Informal firm production could be specified as $p(h)A^i k_t(1-h) + \big(1-p(h)\big)A^i k_t(1-h)(1-\psi)$, where $h \in (0,1)$, the fraction of time allocated to "hiding" (instead of production), is a new choice variable (its introduction would not affect the dynamic problem). The probability of non-detection, $p(h)$, can be specified as $p(h) = \sqrt{h}$ and ψ is the fraction of output confiscated if the firm is detected. It is then straightforward to derive that, in the optimal allocation, hiding increases and output decreases in ψ, through the time diversion mechanism. In real life, hiding could be activities such as temporarily closing when inspectors (are suspected to) arrive, locating in less visible and remote locations, changing locations, etc.

hurting both policy objectives. Hence firms in this range should also always face zero penalties. If the objective is to speed up formalization, all firms with productivity above $A^i_{T\_minimum}$ should face penalties that bring down their productivity to $A^i_{T\_minimum}$, as this strengthens the incentive to formalize, without hurting capital accumulation too much. This is the upward-sloping 45-degree line in the upper panel. If the objective is instead to maximize tax revenue from the informal firm, it is optimal to also leave firms in the range $A^i_{T\_minimum} < A^i < A^i_{TAX\_maximum}$ alone. The reason is that the size of the capital stock matters for tax revenue, and, over this productivity range, penalties that speed up formalization leave formal firms of initially small sizes, reducing initial tax collection. Finally, above $A^i_{TAX\_maximum}$, up to the highest informal sector productivity level, penalties should be applied to bring down productivity to $A^i_{TAX\_maximum}$. This is the 45-degree line in the lower panel.

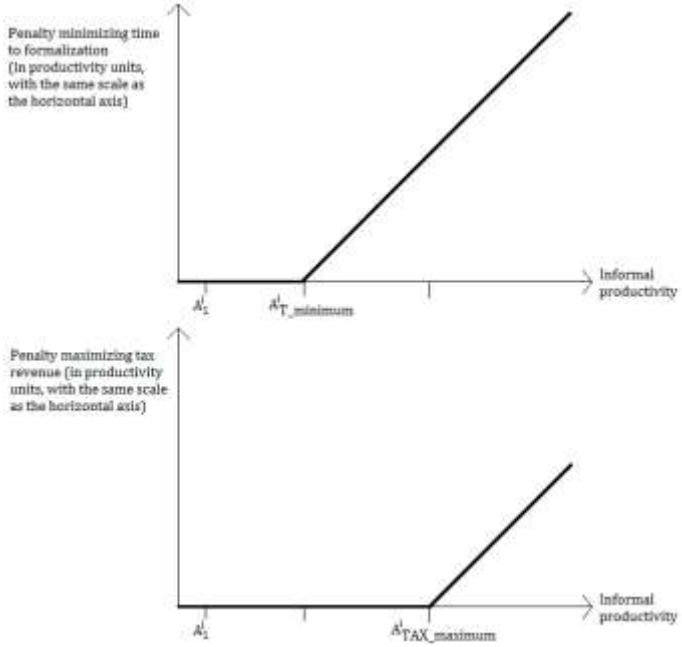

Figure 1. Optimal penalties, for each level of informal firm productivity, in order to minimize the time to formalization, and maximize the firm´s contribution to tax revenue once formal, respectively. The scale of all axes is the same, and the optimal penalty is first zero, then increases one-to-one with informal productivity, once the respective threshold has been reached.

4.1. Monitoring costs

The preceding discussion assumed that monitoring/enforcement of informal firms and observation of firm productivity is costless for the government. Real-life monitoring costs are positive, however. In fact, the distinction between informality and formality would make little sense if the government could costlessly monitor all firms. If costs for monitoring and for assessing firm productivity were included in the model, however, the results of the paper would be further strengthened. From the above discussion, we know that the optimal penalty policy vis-à-vis low-productivity firms is to leave them alone. A costly policy to monitor such firms would therefore optimally not be implemented. In practice, a policy of non-enforcement vis-à-vis low-productivity firms could be instituted after an initial assessment of the likely productivity and size characteristics of informal firms in different geographical areas, business sectors, etc.

Informal firms of high productivity grow faster and become larger than firms of low productivity. It is often argued that such firms become more visible, an implication of which is that they are easier to monitor (e.g. de Paula and Scheinkman, 2011; de Mel, McKenzie and Woodruff, 2013). In the government's choice of monitoring and enforcement policy, lower costs for monitoring high-productivity firms is therefore an argument, in itself, that enforcement resources should be spent on this group, rather than on low-productivity firms. As we already know that high-productivity firms are the (only) ones that should face positive penalties, however, the inclusion in the model of government costs for enforcement and assessment of firm productivity would not alter this conclusion. If anything, the result of positive monitoring costs would be to further raise the productivity thresholds, in Figure 1, above which firms should be penalized.

5. A more general model

The above model assumes that A^f is larger than A^i and, furthermore, that A^f is constant. With respect to the first assumption, formality is thus assumed to result in a productivity increase. Many explanations have been given for such productivity differentials in the literature. In this paper, I assume that formality implies that the firm gets access to some public good, which results in a productivity increase. A rudimentary example, discussed by La Porta and Shleifer (2008), is that formality may imply that a manufacturing firm can apply for and get access to a

more reliable electricity connection, thus be less affected by power outages. A second example is dispute resolution mechanisms, where only formal firms have access to courts. Public credit is another example (e.g. Braun and Loayza, 1994).

The second assumption, a constant A^f, is less standard. As is outlined in the next paragraph and shown in the appendix, however, this assumption can be relaxed and a more general model of informal and formal sector productivities can be constructed, which yields the same results. Before discussing the more general model, however, a first observation is that the informal firms with the lowest productivity have the strongest incentive to formalize (right hand side of expression 9). In fact, formalization would imply the same formal productivity as for much more productive informal firms. Even with such a strong incentive, however, low productivity informal firms do not formalize. As a consequence, any specification of the model that implies a narrowing of the formal-informal productivity gap, $A^f - A^i$, for low-productivity firms, will leave the above results unchanged, that is, that such firms shall not be penalized. The interesting case to focus on is instead informal firms with higher productivity. Second, and importantly, the discussion in this section concerns productivity differences when penalties are zero. As was discussed in section 4, a common characterization of informal/formal productivity differences is that, as informal firms grow, they become more visible and can be audited by the authorities, which may lead such firms to spend resources on hiding rather than on production. As a result, the productivity, growth and size of such firms would potentially be "capped" (similar to the models in e.g. Rauch, 1991; de Paula and Scheinkman, 2011). The present model also allows for such a possibility and contains such an effect, through the government's choice of penalty policy. The analysis of penalties and the resulting decrease in informal firm productivities thus comes in addition to what is discussed in this section.

Consider each entrepreneur as endowed with a level of ability, θ, where higher ability implies higher productivity, whether operating informally or formally, i.e. $\frac{dA^i}{d\theta} > 0$, $\frac{dA^f}{d\theta} > 0$. We then have that A^f increases together with A^i, which is different from sections 2-4. In appendix A4 I show that the qualitative results of the paper remain, also for high productivity informal firms, as long as $A^f - A^i$ decreases when ability goes up. This is consistent with a framework where formality, through access to public goods, gives some benefit, but its importance decreases with the ability of the entrepreneur. Returning to the above example of electricity connections from

La Porta and Shleifer (2008), formality thus implies that a manufacturing firm has a more reliable electricity connection and is less affected by power outages. But high ability informal firms make enough profits to acquire their own generators, and would be less affected, in the case of an outage. More generally, the higher the ability of the entrepreneur, in an environment with no penalties for operating informally, the better the informal firm will develop substitutes for formality. In the second above example of productivity enhancing public goods, dispute resolution mechanisms, formality would thus give access to courts, but informal firms of high ability (which will also imply larger size) will develop capacities to settle disputes outside courts. High ability as a substitute for formality, for large informal firms, is found also by McKenzie and Sakho (2010), in a study from Bolivia (see also World Bank, 2009). The assumption of decreasing $A^f - A^i$ is plausible, given the paper´s objective of studying whether informal firms of different productivity should be treated differently.[9]

5.1 Distributions of firm productivities

The results in sections 2-4 are derived on a "per firm" basis, and no assumption is made about the distribution of informal sector productivities. The above discussed empirical evidence, e.g. La Porta and Shleifer (2008, 2014), certainly points to a situation where most informal firms are in the low-end of the productivity range of figure 1 (horizontal axis). Hence under both policy objectives penalties should probably be zero, for the overwhelming majority of informal firms. Non-zero penalties would concern a limited number of highly productive informal firms with potential for becoming formal and contributing to tax revenue.

6. Maximizing welfare

In the model presented so far, two different government objectives, and the resulting optimal penalties vis-à-vis informal firms, are compared. The formalization and tax objectives result in different levels of government (present value) tax revenue, from firms that formalize. How such revenue is used, however, was not modeled. In fact, it has no productive function in the above

[9] Note that the assumption of formality and ability as substitutes does not counter arguments presented by e.g. Levenson and Maloney (1998), that firms choose a higher degree of formality as they grow, indeed, such is the outcome for a range of firms in the present model.

model.[10] The focus of the analysis was instead on the effect of different penalty levels on an individual firm´s formalization decision, and the resulting contribution to government tax revenue from the individual firm. In the next paragraphs I build on this model, to arrive at a specification where one can analyze, within the same broad framework, how penalties should be set if the objective is to maximize welfare. For such an analysis to make sense, however, the model needs to incorporate some tradeoff, where the penalization of informal firms is compensated by gains elsewhere. I will ultimately consider, in section 6.2 (with further discussion in section 7), productivity-enhancing public goods in the formal sector, the provision of which depends on government tax revenue. Part of the revenue stems from the formalization of informal firms. In principle, the distribution of firms in the economy must now be taken into account, as it affects total tax revenue.

For the purpose of the present discussion, I define a long-run zero-penalty baseline scenario. Given a distribution of informal firm productivities, and based on the model in sections 2-4, firms with productivity $A_1^i < A^i < A_2^i$ will ultimately become formal, which will result in a certain level of present value tax revenue. The level of "per period" public goods provisioning, and the productivity parameter A^f, both of which I assume are constant, are determined by the present value of total tax revenue.[11] [12]

6.1 Preliminary analysis

Taking the baseline scenario as the point of departure, the incentive to formalize, informal firms' future tax contributions, and government tax revenue in general, can be affected by penalties. Consider first, somewhat imprecisely, a scenario where the government needs to infinitesimally increase (present value) tax revenue, above that of the baseline, using the penalty instrument. Which firm productivity level(s) should be penalized in order to raise the additional revenue, yet

[10] A specification where all such tax receipts are transferred back to households, thus not affecting formalization incentives, would yield the same results as above.

[11] One can think of the government, which earns more tax revenue over time, as borrowing from future receipts, such that the level of public goods and the resulting productivity parameter A^f (which depends on the public goods technology) are constant over time. As discussed in section 3, in order to focus the analysis on the policy stance vis-à-vis informal firms, the formal sector tax rate is assumed to be constant.

[12] Again, the results of this section can be carried over to a model such as the one discussed in section 5, where A^f varies between firms, based on an underlying ability parameter.

cause a minimal impact, i.e. lower lifetime firm profits as little as possible? From theorem 2B we can infer that penalties will never be set such that A^i is lowered below $A^i_{TAX\_maximum}$, as it would lead to both lower tax revenue and lower firm profits. Infinitesimally increasing penalties for a firm with productivity above A^i_2 will also lower firm profits, but not affect tax revenue, as the firm would remain informal. At A^i_2 itself, the tax derivative is zero (expression 15). The "first productivity" to be penalized must thus lie somewhere in the interval $(A^i_{TAX_{maximum}}, A^i_2)$. Within this range, the formalizing firm(s) least affected should be penalized. Further increases of revenue should follow the same logic: penalize those formalizing firms, for which the lifetime firm profit is least affected, for a given amount of additional (present value) tax revenue raised. Expressed more precisely, penalties should be levied on informal firms, in descending order of their $\frac{\left|\frac{d}{dA^i}PV(TAX)\right|}{\left|\frac{d}{dA^i}\pi^{formalization}\right|}$ - ratios, and this ratio should be equalized for all penalized firms. $PV(TAX)$ is given in expression 14, $\pi^{formalization}$ is the optimal profit from expression 3, and the vertical brackets indicate absolute values.[13] As long as the maximal tax revenue is not raised, per firm penalties will be (weakly) lower than those of theorem 2B.[14] So far in the analysis, the result from the previous sections is extended: the formalization objective entails higher penalties and more penalized firms than the objective to maximize tax revenue, which in turn entails higher penalties and more penalized firms than if the government aims for less than maximum tax revenue. As a welfare objective considers not only taxes raised, but also the negative effects penalties have on informal firms, it follows that a welfare objective is kindred to the "less than maximizing tax revenue" objective.

Figure 2 sketches how optimal penalty schedules may look in the case of a small, medium and large additional tax revenue increase, from formalizing firms, for the above paragraph's preliminary analysis. Further details of the shape of these curves are given in Appendix A5. The above tax-to-profit derivatives ratio will have its maximum value for a productivity level in the interval $(A^i_{TAX_{maximum}}, A^i_2)$. Penalizing this specific productivity level will bring down its

[13] For firms with productivity above A^i_2, this ratio is zero. For high enough increases in tax revenue, these firms must also be penalized, and should be so in the order from A^i_2 and up.

[14] Depending on the amount of tax revenue raised, some firms/productivity levels may face penalties equal to those they would face in the tax maximum, other firms lower penalties.

derivatives ratio, which then equals those of adjacent productivities. These productivities should also be penalized, albeit less, such that the derivatives ratio is equalized for all penalized firms. How large an interval of productivity levels should be penalized depends on the additional revenue to be raised. The left and middle panels illustrate the extent and intensity of penalties when the additional revenue to be raised is small to moderate. For large increases in tax revenue, all productivity levels between $A^i_{TAX_{maximum}}$ and A^i_2 will face maximum penalties (as indicated by the dashed gray line), after which the highest productivity informal firms will also face positive penalties. As the drop in profits for such productivity levels is larger, but the tax revenue contribution is equal to that of firms with productivity A^i_2, penalization will only be imposed when no more contributions can be obtained from firms in the $\left(A^i_{TAX_{maximum}}, A^i_2\right)$ range. Such a schedule is depicted in the rightmost panel in figure 2.

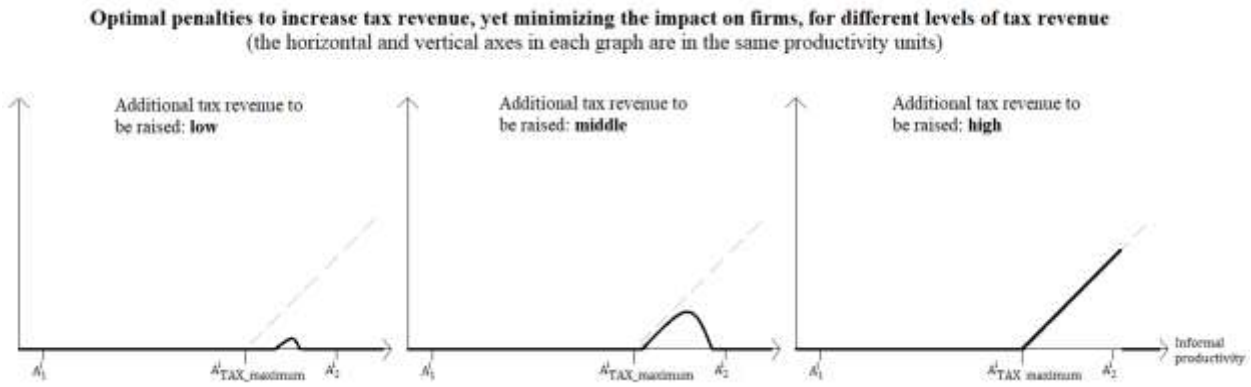

Figure 2. Optimal penalties to increase tax revenue, through the formalization of informal firms, yet minimizing the impact on profits, for different firm productivities (horizontal axis) and levels of tax revenue to be raised. The dashed gray line is the penalty schedule if the maximum possible tax revenue is to be raised.

The above preliminary analysis outlined how penalties are levied as efficiently as possible, given the objective of obtaining a certain tax revenue increase from formalizing firms. How the additional revenue is used was not specified. Such a decoupling of which productivity levels/firms should be penalized, and how the additional tax revenue is used, is valid as long as the individual firm does not internalize the possible future benefits that stem from the increased revenue. In the model of productivity-enhancing public goods, outlined next, it will mean that informal firms take the level of A^f as given, an assumption I maintain throughout.

6.2 Welfare analysis with productivity-enhancing public goods in the formal sector

Let public goods, financed by taxes, be productivity-enhancing, such that A^f is increasing in taxes collected. The government has perfect foresight and chooses A^f, and the corresponding level of additional tax revenue needed from formalizing firms, in order to maximize welfare, which is total firm profits. As discussed above, let the baseline A^f be determined by the total level of (present value) tax revenue, in the model with zero penalties. Restrict the problem such that the government indeed wants to bring informal firms into formality, in order to increase tax revenue, in line with much of the policy discussion (e.g. Bigsten et al., 2004; USAID, 2005; Prado, 2011; Bruhn and McKenzie, 2014). Let A^f be concave in total present value tax revenue. Furthermore, restrict the combination of the maximum A^f that can be obtained and entry costs F, such that $F > F_{min}$ (from 11). We thus always have some low-end informality.

Each value of A^f, above that of the baseline scenario, has a corresponding additional (present value) tax revenue that needs to be raised, from formalizing firms. As long as the revenue needed is less than the maximum possible, considering the distribution of informal firm productivities and the A^f-parameter, the government has an infinite number of choices on how to set penalties. Given that informal firms take A^f as given, however, the optimal choice is to follow the above developed criterion of descending $\frac{\left|\frac{d}{dA^i}PV(TAX)\right|}{\left|\frac{d}{dA^i}\pi^{formalization}\right|}$ – ratios. To each value of A^f corresponds such an optimal penalty schedule. Adhering to these optimal schedules, the government then compares total firm profits for all feasible values of A^f, and chooses the optimum. For the optimal A^f, the extent to which different informal firm productivities should be penalized thus follows the above developed logic, which is further discussed in section 7.

7. Discussion

This paper derives optimal penalties on informal firms, under different policy objectives. For the first two objectives analyzed, i.e. speeding up formalization and maximizing firms' contribution to tax revenue once formal, I show that low-productivity firms should always be left alone. High-productivity informal firms should instead face penalties that increase in the productivity level, with fewer penalized firms and lower penalties for the tax revenue objective. These results are

summarized in figure 1. Leaving low-productivity firms alone is intuitive: most such firms would not formalize anyway. The ones that ultimately do formalize would do so later and contribute less to tax revenue, if penalized while informal. Lower penalties for the tax revenue objective is also intuitive, as the size of firms at formalization (and hence initial tax revenue) now also matters.

The third objective, i.e. welfare, also entails zero penalties for low productivity informal firms. Before further discussing this result, I reconnect to the analysis in Ulyssea (2018). In a rich model with two margins of informality, i.e. informal firms and formal firms with informal workers, respectively, the author undertakes four separate policy experiments: reduce formal sector labor taxes, increase formal sector penalties for hiring unregistered labor, reduce formal sector entry costs, and increase enforcement vis-à-vis informal firms. The latter policy outperforms the other three in reducing informality, but also has the most negative impact, by far, on overall welfare. Informal firm profits drop significantly when harsher penalties are applied, and the negative effect thus holds both at the individual firm- and at the aggregate level. A main channel is that many informal firms shut down their business, or stay much smaller, as a result of stricter enforcement (p. 2042), which is in line with earlier arguments in the informality literature, e.g. Fafchamps (1994); Farrell (2004), La Porta and Shleifer (2008, 2014) and Charlot et al. (2015). Although the models differ, a penalty structure such as that of figure 1 would probably have improved upon the welfare result, as the many firms "going into the shadows" instead would have been unaffected (barring general equilibrium effects). More generally, the here suggested penalty structure can be fed back into models of the informal sector, where typically linear or quadratic penalties are assumed. The negative effects of such specifications, in that they penalize low productivity firms, which "may lead to poverty and destitution of informal workers and entrepreneurs" (La Porta and Shleifer, 2014, p. 125), are exacerbated by the fact that the overwhelming majority of informal firms are in the low end of the productivity spectrum.

The preceding paragraph illustrates that a reduction of informality need not imply a welfare increase, if anything, the opposite. In the present model the difference between a formalization objective, a tax revenue objective, and a welfare objective was discussed in section 6: the formalization objective entails the highest penalties, the tax revenue objective has medium penalties, and the welfare objective the lowest penalties. For the welfare objective, the fact that

losses incurred by informal firms are taken into account has a moderating effect on the penalty levels. Figure 2 shows a non-obvious penalty schedule, in that it is not the highest productivity informal firms that should face the harshest penalties. Instead, the most efficient penalties will be those that are levied on firms' whose formalization incentives are most easily manipulated.

The comparison of the welfare objective with the other two objectives is complicated, however, if the additional tax revenue, from bringing informal firms into formality, is not used for transfers.[15] In section 6.2, where I assume tax revenue is used for productivity enhancing public goods in the formal sector, the A^f-parameter will differ between the different scenarios. As a result, the productivity range of firms that formalize, from expression 10, also differs between scenarios, and would complicate the drawing of graphs corresponding to figure 2. If additional tax revenue augments formal productivity a lot, an increase in penalties on informal firms will be optimal. Given decreasing returns of additional tax revenue, however, some upper level of optimal penalties will be reached. And as long as entry costs are high enough to guarantee that the lowest productivity firms never formalize (i.e. fulfilling condition 11), there will also be a low to medium productivity range that should not be penalized, as such penalization would delay formalization, decrease future tax contributions, decrease total tax revenue and decrease welfare. One way of interpreting the model is therefore as follows: sections 3 and 4 show how individual firms should be penalized in order for the policy maker to bring the firms as fast as possible into formality, or to maximize tax revenue from such formalizing firms, respectively. A basic productivity dependent penalty structure emerges, with a slight difference between the two objectives. When we also consider, in a welfare analysis, how the benefit from increased formalization, i.e. increased tax revenue, is used, much of the logic survives. Low productivity informal firms should be left alone, and some or most of the high productivity firms should be penalized. The degree to which public goods augment productivity, and hence the amount of tax revenue needed, the overall distribution of firm productivities, together with an efficiency criterion for the order in which different productivity levels should be penalized, determine the level of penalties for each high productivity firm.

[15] As discussed above, until and including section 6.1 one can think of the tax revenue raised from formalizing firms as either not being used at all, or as transferred to households.

Which of the three policy objectives is the most legitimate needs to be discussed among academics and policy makers. Prado (2011) and Bruhn and McKenzie (2014) discuss rationales for the "formalization per se" as well as the tax revenue objective. Bigsten et al. (2004) argue mainly in terms of the tax policy objective. USAID (2005) stresses all three objectives, and discuss potential conflicts between them. Ulyssea (2018) analyses the impact on informality, tax revenue and welfare, as well as other variables, from the above discussed policy experiments. La Porta and Shleifer (2008, 2014) stress the negative welfare aspects of penalizing informal firms. Although it is not primarily focused on the policy question, perhaps this latter work, which draws on the duality model of Lewis (1954), Rauch (1991) and others, most strongly argues for a "let go" attitude towards informality. Informality diminishes with development, not because informal firms formalize, but because more productive formal firms, absorbing labor, will be created. The question still remains, however, if something can and should be done about informality today. And whereas it is true, in a cross section of countries, that informality is negatively correlated with GDP/capita, most studies agree that within developing countries, the informal sector is growing rather than shrinking (Ernste and Schneider, 1998; Bigsten et al., 2004; SIDA, 2004; USAID, 2005; Chen, 2007; Tokman, 2007; World Bank, 2015). La Porta and Shleifer (2008) argue, based on data from 14 Latin American countries, that few firms transition from informality to formality, although they also hypothesize that "it is still possible of course that a minority of informal firms, and especially the most productive ones, end up joining the formal economy" (p. 35). Given that the overwhelming majority of informal firms have very low productivity levels, the observation that most such firms stay informal is not surprising. But it reflects the situation "as is" rather than answering how many firms would formalize under different policy objectives and optimal policies. As for the specific policy instrument studied, and in an empirical literature of mostly meager results from different policy initiatives, there is some evidence that increased enforcement vis-à-vis high-productivity informal firms can be a policy that works (de Andrade et al., 2014).

References

Arruñada, B., 2007. Pitfalls to avoid when measuring institutions: Is 'Doing Business' damaging business? Journal of Comparative Economics 35 (4): 729-747.

Arruñada, B., Manzanares, C., 2015. The trade-off between ex ante and ex post transaction costs: Evidence from legal opinions. Berkeley Business Law Journal 13 (1): 217-254.

Barro, R., Sala-i-Martin, X., 1999. Economic growth. Cambridge, Massachusetts: MIT Press.

Benhassine, N., McKenzie, D., Pouliquen, V., Santini, M., 2016. Can enhancing the benefits of formalization induce informal firms to become formal? Experimental evidence from Benin. World Bank Policy Research Working Paper 7900.

Bigsten, A., Kimuyu, P., Lundvall, K., 2004. What to do with the informal sector? Development Policy Review, 22 (6): 701-715.

Braun, J., Loayza, N., 1994. Taxation, public services, and the informal sector in a model of endogenous growth. World Bank Policy Research Working Paper 1334.

Bruhn, M., 2013. A tale of two species: Revisiting the effect of registration reform on informal business owners in Mexico, Journal of Development Economics 103: 275-283.

Bruhn, M., McKenzie, D., 2014. Entry regulation and the formalization of microenterprises in developing countries, The World Bank Research Observer 29: 186-201.

Charlot, O., Malherbet, F., Terra, C., 2015. Informality in developing countries: Regulation and fiscal policies. Journal of Economic Dynamics and Control 51: 1-27.

Chen, M., 2007. Rethinking the Informal Economy: Linkages with the formal economy and the formal regulatory environment. DESA Working Paper 46.

de Andrade, G., Bruhn, M., McKenzie, D., 2014. A helping hand or the long arm of the law? Experimental evidence on what governments can do to formalize firms. The World Bank Economic Review 30 (1): 24-54.

de Mel, S., McKenzie, D., Woodruff, C., 2010. Who are the microenterprise owners? Evidence from Sri Lanka on Tokman v. de Soto. In: Lerner, J., Schoar, A. (Eds.), International Differences in Entrepreneurship, pp. 63–87. Chicago: University of Chicago Press.

de Mel, S., McKenzie, D., Woodruff, C., 2013. The demand for, and consequences of, formalization among informal firms in Sri Lanka. American Economic Journal: Applied Economics 5(2): 122-150.

de Paula, A., Scheinkman, J., 2011. The informal sector: An equilibrium model and some empirical evidence from Brazil. Review of Income and Wealth, 57: S8-S26.

de Soto, H., 1989. The Other Path: The invisible revolution in the third world. New York: Harper and Row.

Dessy, S., Pallage, S., 2003. Taxes, inequality and the size of the informal sector. Journal of Development Economics 70: 225-233.

Ernste, D., Schneider, F., 1998. Increasing shadow economies all over the world – fiction or reality? IZA Discussion Paper 26.

Fafchamps, M., 1994. Industrial structure and microenterprises in Africa. The Journal of Developing Areas 29 (1): 1-30.

Farrell, D., 2004. The hidden dangers of the informal economy. McKinsey Quarterly 3: 27–37.

Harstad, B., Svensson, J., 2011. Bribes, lobbying, and development. American Political Science Review 105 (1): 46-63.

Hart, K., 1973. Informal income opportunities and urban employment in Ghana. Journal of Modern African Studies 11: 61-89.

Hassan, M., Schneider, F., 2016. Size and development of the shadow economies of 157 countries worldwide: Updated and new measures from 1999 to 2013. IZA Discussion Paper No. 10281.

Ihrig, J., Moe, K., 2004. Lurking in the shadows: the informal sector and government policy. Journal of Development Economics 73: 541-557.

ILO., 1972. Employment, incomes and equality: A strategy for increasing productive employment in Kenya. Geneva: International Labour Office.

Johnson, S., Kaufmann, D., Shleifer, A., 1997. The unofficial economy in transition. Brookings Papers on Economic Activity, Fall 1997 (2): 159-239.

La Porta, R., Shleifer, A., 2008. The unofficial economy and economic development. Brooking Papers on Economic Activity 105 (3), 473–522.

La Porta, R., Shleifer, A., 2014. Informality and development. Journal of Economic Perspectives 28 (3): 109-126.

Levenson, A., Maloney, W., 1998. The informal sector, firm dynamics and institutional participation. World Bank Policy Research Working Paper 1988.

Levy, Santiago, 2008. Good intentions, bad outcomes: social policy, informality, and economic growth in Mexico. Brookings Institution Press.

Lewis, W., 1954. Economic development with unlimited supplies of labour. Manchester School 22: 139-191.

Loayza, N., 1996. The economics of the informal sector: a simple model and some empirical evidence from Latin America. Carnegie-Rochester Conference Series on Public Policy 45: 129-162.

Loayza, N., 2016A. World Bank Policy Research Talk: Informality in the process of development and growth. June 7, 2016. http://www.worldbank.org/en/research/brief/policy-research-talks

Loayza, N., 2016B. Informality in the process of development and growth. World Bank Policy Research Working Paper 7858.

Loayza, N., Rigolini, J., 2011. Informal employment: safety net or growth engine? World Development 39 (9): 1503-1515.

McKenzie, D., Sakho, Y., 2010. Does it pay firms to register for taxes? The impact of formality on firm profitability. Journal of Development Economics 91: 15-24.

Nguimkeu, P., 2016. An estimated model of informality with constrained entrepreneurship. Manuscript, Georgia State University.

Prado, M., 2011. Government policy in the formal and informal sectors. European Economic Review 55: 1120-1136.

Ranis, G., Stewart, F., 1999. V-Goods and the role of the urban informal sector in development. Economic Development and Cultural Change 47 (2): 259-288.

Rauch, J., 1991. Modeling the informal sector formally. Journal of Development Economics 35 (1): 33-47.

Reynolds, P., Camp, S., Bygrave, W., Autio, E., Hay, M., 2001. Global Entrepreneurship Monitor, 2001 Executive Report.

SIDA, 2004. The informal economy. Fact Finding Study.

Tokman, V. 2007. Modernizing the informal sector. UN/ DESA Working Paper no. 42. United Nations, Department of Economic and Social Affairs.

Ulyssea, G., 2018. Firms, informality and development: Theory and evidence from Brazil. Manuscript. American Economic Review 108(8): 2015-2047.

USAID, 2005. Removing barriers to formalization: The case for reform and emerging best practice.

World Bank, 2009. Increasing formality and productivity of Bolivian firms. A World Bank country report. Washington DC: The International Bank for Reconstruction and Development / The World Bank.

World Bank, 2015. Informal economy and the World Bank. Policy Research Working Paper 6888.

Appendix

A1. Proof of Lemma 1

Starting at infinity, we need to show that $T(A^i)$, implicitly defined by expression 9, first monotonously decreases in A^i, then monotonously increases and goes to infinity. Alternatively, and as further motivated in the next paragraph, we can show that to each T correspond exactly two A^i (except at the stationary point, where $\frac{dT}{dA^i} = 0$). In the proof, I use the inverse function $A^i(T)$, instead of $T(A^i)$, as it simplifies the algebra. By also defining $\pi \equiv \frac{A^i}{A^f}$, the A^f-parameter is eliminated, further simplifying the expressions. This is why the solution to (9) is written as (12) (which converges to expression 10, as $T \to \infty$). Starting with (12), the proof uses the auxiliary expressions $\tilde{F} = \frac{z\delta\rho(\delta+\rho)(2\delta+\rho)F}{(A^f)^2}$, $\tilde{D} = (2\delta+\rho)\left(1-e^{-\delta T}\right) - \delta\left(1-e^{-(2\delta+\rho)T}\right)$, $\tilde{G} = \frac{\delta\left(1-e^{-(2\delta+\rho)T}\right)-\tilde{F}}{\tilde{D}}$ and $\tilde{\pi} = \frac{(2\delta+\rho)\left(1-e^{-\delta T}\right)-2\delta\left(1-e^{-(2\delta+\rho)T}\right)}{2\tilde{D}}$. Furthermore, let $\pi_1 = \tilde{\pi} - \sqrt{\tilde{\pi}^2+\tilde{G}}$ and $\pi_2 = \tilde{\pi} + \sqrt{\tilde{\pi}^2+\tilde{G}}$, and let $\sqrt{..}$ be the square root expression.

By rewriting (9), using $\pi \equiv \frac{A^i}{A^f}$, we get $\frac{\pi\left(1-e^{-\delta T}\right)}{z\delta(\delta+\rho)} + \frac{(1-\pi)\left(1-e^{-(2\delta+\rho)T}\right)}{z(\delta+\rho)(2\delta+\rho)} = \frac{\rho F}{\left(A^f\right)^2(1-\pi)}$, where T is an implicit function of π. Differentiating with respect to π, setting $\frac{dT}{d\pi} = 0$, and rewriting, gives the condition $(1-\pi)^2 = \frac{\tilde{F}}{\tilde{D}}$. Solving for π gives $\pi = \tilde{\pi}$. That is, at the stationary point, $\frac{dT}{d\pi} = 0$ (i.e. $\frac{dT}{dA^i} = 0$), we have $\pi = \pi_1 = \pi_2$, and hence $\sqrt{..} = 0$. In order to establish that $T(A^i)$ first monotonously decreases in A^i, then monotonously increases, and working with the inverse relation $\pi(T)$, we thus need to show that

$$\frac{d\pi_1}{dT} < 0 \text{ and } \frac{d\pi_2}{dT} > 0. \qquad \text{(A1)}$$

Establishing the first part of (A1) corresponds to showing that $T(A^i)$ is decreasing for A^i below the point where $\frac{dT}{dA^i} = 0$, and the second part that $T(A^i)$ is then increasing. In conducting the proof, it is also shown that $\sqrt{..}$ is increasing in T, i.e. $\sqrt{..}$ is first imaginary (π_1, π_2 do not exist), then zero at the minimum T, then increasing.

With $\frac{d}{dT}\sqrt{\tilde{\pi}^2+\tilde{G}}=\frac{1}{2\sqrt{..}}\left(2\tilde{\pi}\frac{d\tilde{\pi}}{dT}+\frac{d\tilde{G}}{dT}\right)$ and $\frac{d\tilde{G}}{dT}=\frac{\delta(2\delta+\rho)e^{-(2\delta+\rho)T}}{\tilde{D}}-\frac{\frac{d\tilde{D}}{dT}(\delta(1-e^{-(2\delta+\rho)T})-\tilde{F}))}{\tilde{D}^2}$, the derivatives in (A1) can be written as follows:

$$\frac{d\pi_1}{dT}=-\frac{1}{\sqrt{..}}(\frac{d\tilde{\pi}}{dT}\pi_1+\frac{1}{2}(\frac{\delta(2\delta+\rho)e^{-(2\delta+\rho)T}}{\tilde{D}}-\frac{\frac{d\tilde{D}}{dT}(\delta(1-e^{-(2\delta+\rho)T})-\tilde{F}))}{\tilde{D}^2}))$$

$$\frac{d\pi_2}{dT}=\frac{1}{\sqrt{..}}(\frac{d\tilde{\pi}}{dT}\pi_2+\frac{1}{2}(\frac{\delta(2\delta+\rho)e^{-(2\delta+\rho)T}}{\tilde{D}}-\frac{\frac{d\tilde{D}}{dT}(\delta(1-e^{-(2\delta+\rho)T})-\tilde{F}))}{\tilde{D}^2})) \quad \text{(A2)}$$

From the restriction on F in (11), we know that π_1 is always positive. In addition $\tilde{D}$ is always positive (for $T<\infty$), as is $\frac{d\tilde{D}}{dT}=\delta(2\delta+\rho)(e^{-\delta T}-e^{-(2\delta+\rho)T})$, for $T>0$. The case of $T=0$ is not interesting, as it would require $F=0$. From the restriction on F (rewritten as $\tilde{F}>\delta$), we see that the entire expression $\frac{\delta(2\delta+\rho)e^{-(2\delta+\rho)T}}{\tilde{D}}-\frac{\frac{d\tilde{D}}{dT}(\delta(1-e^{-(2\delta+\rho)T})-\tilde{F})}{\tilde{D}^2}$ is positive. It remains to show that $\frac{d\tilde{\pi}}{dT}>0$:

$$\frac{d\tilde{\pi}}{dT}=\frac{((\delta(2\delta+\rho)e^{-\delta T}-2\delta(2\delta+\rho)e^{-(2\delta+\rho)T})\tilde{D}-\left((2\delta+\rho)(1-e^{-\delta T})-2\delta(1-e^{-(2\delta+\rho)T})\right)\frac{d\tilde{D}}{dT})}{2\tilde{D}^2} \quad \text{(A3)}$$

The numerator simplifies to $\delta(2\delta+\rho)\left[\delta e^{-\delta T}+(\delta+\rho)e^{-\delta T}e^{-(2\delta+\rho)T}-(2\delta+\rho)e^{-(2\delta+\rho)T}\right]$, the square bracket of which can be written as $Xe^{-XT}+(Y-X)e^{-XT}e^{-YT}-Ye^{-YT}$, where $X\equiv\delta$, $Y\equiv2\delta+\rho$ and $Y>2X$. Assuming this expression equals zero, the assumed equality can be written as

$$X+(Y-X)e^{-YT}=Ye^{(X-Y)T}. \quad \text{(A4)}$$

For $T=0$, equality holds (but is of no interest). Differentiating both sides with respect to T, with both derivatives negative, the right-hand side decreases a factor e^{XT} faster. The left-hand side is thus bigger, whenever $T>0$. The expression in square brackets and $\frac{d\tilde{\pi}}{dT}$ are hence positive, which establishes (A1). Further note that with $\sqrt{\tilde{\pi}^2+\tilde{G}}=0$ at the stationary point (a condition that determines the minimal formalization time), we get, from (A2), $\frac{d\pi_1}{dT}=-\infty$, $\frac{d\pi_2}{dT}=\infty$. To see that π_1,π_2 do not exist for smaller T, rewrite $\sqrt{..}$ to get $\frac{1}{2\sqrt{\tilde{D}}}\left[\frac{((2\delta+\rho)(1-e^{-\delta T}))^2}{(2\delta+\rho)(1-e^{-\delta T})-\delta(1-e^{-(2\delta+\rho)T})}-4\tilde{F}\right]^{1/2}$. We thus have $\sqrt{..}=0$ if the expression in square brackets equals zero, a term that increases in T if $\frac{d}{dT}\left(\frac{((2\delta+\rho)(1-e^{-\delta T}))^2}{(2\delta+\rho)(1-e^{-\delta T})-\delta(1-e^{-(2\delta+\rho)T})}\right)>0$. This derivative equals

$\frac{\delta(2\delta+\rho)^2 e^{-\delta T}\left(1-e^{-\delta T}\right)\left(\rho\left(1-e^{-(2\delta+\rho)T}\right)-(2\delta+\rho)\left(e^{-\delta T}-e^{-(\delta+\rho)T}\right)\right)}{\left((2\delta+\rho)\left(1-e^{-\delta T}\right)-\delta\left(1-e^{-(2\delta+\rho)T}\right)\right)^2}$, which is zero if $\frac{2\delta+\rho}{\rho}=\frac{1-e^{-(2\delta+\rho)T}}{e^{-\delta T}-e^{-(\delta+\rho)T}}$, a condition satisfied for $T=0$ (from l'Hôpital's rule, which gives $lim_{T\to 0}(\frac{1-e^{-(2\delta+\rho)T}}{e^{-\delta T}-e^{-(\delta+\rho)T}})$). Using the variable substitution $e^{-\delta T}=C_1, e^{-\rho T}=C_2$ (where $C_1=C_2$ if either variable equals unity), and minimizing the right hand side (i.e. $\frac{1-(C_1)^2 C_2}{C_1-C_1 C_2}$) with respect to $\{C_1, C_2\}$, there is no other solution than $C_1=C_2=1$ (i.e. $T=0$). As a consequence, the expression in square brackets increases in T and π_1, π_2 exist only for T larger than the minimal formalization time. This T must, in turn, be strictly positive for $\tilde{F}>\delta$, as, for $T=0$, the expression in square brackets is negative, which is seen by applying l'Hôpital's rule, giving $lim_{T\to 0}\left(\frac{((2\delta+\rho)\left(1-e^{-\delta T}\right))^2}{(2\delta+\rho)\left(1-e^{-\delta T}\right)-\delta\left(1-e^{-(2\delta+\rho)T}\right)}\right)=\frac{2\delta(2\delta+\rho)}{\delta+\rho}<4\delta$). Finally, the condition $\frac{d\pi_1}{dT}=-\infty$, $\frac{d\pi_2}{dT}=\infty$ is nothing but the stationary point ($\frac{dT}{dA^i}=0$), expressed in the (T,π)-plane.■

Although not strictly necessary for lemma 1, but used elsewhere in the paper, the following section proves that π_2 is concave in T, i.e. that $T(A^i)$ is convex.

A2. Proof that $T(A^i)$ is convex (the proof concerns the range where $dT/dA^i>0$)

For the results in the paper, it suffices to study the range where $\frac{dT}{dA^i}>0$, for which the proof thus establishes that $\frac{d^2T}{d(A^i)^2}>0$. Following section A1, I use the inverse relation between A^i and T, which should be concave over the relevant range, i.e. $\frac{d^2\pi_2}{dT^2}<0$. That is, if $\pi_2(T)$ is concave for all T, and hence $A^i(T)$ concave for all A^i larger than the productivity level at the stationary point, then, because $\frac{d\pi_2}{dT}$ is positive, its inverse, and hence $T\left(A^i\right)$, is convex.

I first show, by contradiction, that $\pi_2=\tilde{\pi}+\sqrt{\tilde{\pi}^2+\tilde{G}}$ must be concave if $\tilde{\pi}$ and $\tilde{G}$ are concave. Subtract $\tilde{\pi}$ and then square both sides of $\pi_2-\tilde{\pi}$ to get $\pi_2(\pi_2-2\tilde{\pi})=\tilde{G}$. Twice differentiating the left hand side gives

$$\frac{d^2(\pi_2(\pi_2-2\tilde{\pi}))}{dT^2}=2\left(\frac{d^2\pi_2}{dT^2}(\pi_2-\tilde{\pi})+\frac{d\pi_2}{dT}\left(\frac{d\pi_2}{dT}-2\frac{d\tilde{\pi}}{dT}\right)-\pi_2\frac{d^2\tilde{\pi}}{dT^2}\right) \quad \text{(A5)}$$

Let $\tilde{\pi}$ and $\tilde{G}$ be concave, i.e. $\frac{d^2\tilde{\pi}}{dT^2} < 0$ and $\frac{d^2\tilde{G}}{dT^2} < 0$. Assume π_2 is convex or linear in T, i.e. $\frac{d^2\pi_2}{dT^2} \geq 0$. The first of the three terms in the right hand side of (A5) is then non-negative, as $\pi_2 > \tilde{\pi}$. The second term is positive. From the proof to lemma 1 we know that $\frac{d\pi_2}{dT} > 0$. This derivative can be written as $\frac{d\pi_2}{dT} = \frac{d\tilde{\pi}}{dT}\left(\frac{\tilde{\pi}+\sqrt{\tilde{\pi}^2+\tilde{G}}}{\sqrt{\tilde{\pi}^2+\tilde{G}}}\right) + \frac{1}{2\sqrt{..}}\frac{d\tilde{G}}{dT}$. The restriction on F implies $\tilde{G} < 0$, hence $\frac{\tilde{\pi}+\sqrt{\tilde{\pi}^2+\tilde{G}}}{\sqrt{\tilde{\pi}^2+\tilde{G}}} > 2$. In addition, $\frac{d\tilde{G}}{dT} > 0$ was established above, hence $\frac{d\pi_2}{dT} > 2\frac{d\tilde{\pi}}{dT}$. With the assumption of $\frac{d^2\tilde{\pi}}{dT^2} < 0$, the third term, $-\pi_2\frac{d^2\tilde{\pi}}{dT^2}$, is also positive, hence expression A5 is unambiguously positive. However, concavity of $\tilde{G}$ requires (A5) to be negative, as it must equal $\frac{d^2\tilde{G}}{dT^2}$. As a result, $\frac{d^2\pi_2}{dT^2} \geq 0$ is not feasible if $\tilde{\pi}$ and $\tilde{G}$ are concave.

I next establish$\frac{d^2\tilde{\pi}}{dT^2} < 0$. Differentiate and simplify, using $X \equiv \delta$, $Y \equiv 2\delta + \rho$, as in section A1, to get

$$\frac{d^2\tilde{\pi}}{dT^2} = \frac{XY\left[X^2Ye^{-XT}\left(-1+e^{-YT}\right)\left(1+e^{-XT}-2e^{-YT}\right)+X^3e^{-XT}\left(1-e^{-YT}\right)^2+Y^3e^{-YT}\left(1-e^{-XT}\right)^2-XY^2e^{-YT}\left(1-e^{-XT}\right)\left(1+e^{-YT}-2e^{-XT}\right)\right]}{\left(Y\left(1-e^{-XT}\right)-X\left(1-e^{-YT}\right)\right)^3}.$$

The denominator is positive (equaling $\widetilde{D}^3$), as is XY in the numerator. The expression within squared brackets can be divided into two parts and rewritten as e^{-XT} times

$$\left[XY\left(e^{-XT}-e^{-YT}\right)\left(Y\left(1-e^{-XT}\right)e^{-(Y-X)T}-X\left(1-e^{-YT}\right)\right)\right]_1 + \left[(Y-X)\left(Y^2\left(1-e^{-XT}\right)^2e^{-(Y-X)T}-X^2\left(1-e^{-YT}\right)^2\right)\right]_2. \quad \text{(A6)}$$

With $e^{-XT} > e^{-YT}$, bracket 1 is negative if $Y\left(1-e^{-XT}\right)e^{-(Y-X)T} - X\left(1-e^{-YT}\right)$ is negative. This term is zero at $T = 0$, a case of no interest. Differentiating with respect to T and simplifying gives $Y(Y-X)e^{-YT}\left(1-e^{XT}\right)$, which is negative for all $T > 0$, i.e. bracket 1 is negative. Bracket 2 is negative if $Y^2\left(1-e^{-XT}\right)^2e^{-(Y-X)T} - X^2\left(1-e^{-YT}\right)^2$ is negative, i.e. if $Y^2\left(1-e^{-XT}\right)^2e^{-(Y-X)T} < X^2\left(1-e^{-YT}\right)^2$. Because both expressions contain only positive components, we can instead study the square root of each side (an order-preserving operation). Bracket 2 is thus negative if $Y\left(1-e^{-XT}\right)e^{-(Y-X)T/2} < X\left(1-e^{-YT}\right)$. Again, equality holds at $T = 0$. The left hand side has derivative $\frac{Ye^{-YT}}{2}\left((Y+X)e^{\frac{(Y-X)T}{2}} - (Y-X)e^{\frac{(Y+X)T}{2}}\right)$, which equals the right hand side derivative, XYe^{-YT}, at $T = 0$, but is smaller for $T > 0$, implying the right hand side

increases faster. Bracket 2, and the expression in (A6), are thus negative, establishing concavity of $\tilde{\pi}$.

The final part of the proof establishes $\frac{d^2\tilde{G}}{dT^2} < 0$. Differentiating $\frac{d\tilde{G}}{dT}$ (from section A1) and rewriting gives

$$\frac{d^2\tilde{G}}{dT^2} = \frac{-\delta(2\delta+\rho)^2 e^{-(2\delta+\rho)T}}{\widetilde{D}} - \frac{2\frac{d\widetilde{D}}{dT}\delta(2\delta+\rho)e^{-(2\delta+\rho)T}}{\widetilde{D}^2} - \frac{\left(\delta\left(1-e^{-(2\delta+\rho)T}\right)-\tilde{F}\right)\left(\frac{d^2\widetilde{D}}{dT^2}\widetilde{D}-2\left(\frac{d\widetilde{D}}{dT}\right)^2\right)}{\widetilde{D}^3} \quad \text{(A7)}$$

From section A1 we know that $\widetilde{D}$, $\frac{d\widetilde{D}}{dT}$ and $-\left(\delta\left(1-e^{-(2\delta+\rho)T}\right)-\tilde{F}\right)$ are all positive, hence concavity is established if $\frac{d^2\widetilde{D}}{dT^2}\widetilde{D} - 2\left(\frac{d\widetilde{D}}{dT}\right)^2$ is negative. This expression can be written as

$\frac{d^2\widetilde{D}}{dT^2}\widetilde{D} - 2\left(\frac{d\widetilde{D}}{dT}\right)^2 =$

$-\delta(2\delta+\rho)e^{-2\delta T}\left[\delta(\delta+\rho)e^{\delta T} + \delta(2\delta+\rho)e^{-2(\delta+\rho)T} + (2\delta+\rho)(\delta-(\delta+\rho)e^{-\rho T}) - (3\delta^2-\rho^2)e^{-(\delta+\rho)T}\right]$,

where the expression in square brackets, which should be > 0, using $X \equiv \delta$, $Y \equiv 2\delta+\rho$, simplifies to

$$XY(1-e^{-(Y-X)T})^2 + (Y-X)\,e^{XT}(X(1-e^{-YT}) + Ye^{-YT}(1-e^{XT})) \quad \text{(A8)}$$

The first term is positive, for $T > 0$, as is the second term, if $X(1-e^{-YT}) + Ye^{-YT}(1-e^{XT}) > 0$. Being zero at $T = 0$, this factor increases in T, as its derivative, $Y(Y-X)\,e^{-YT}(e^{XT}-1)$, is positive for $T > 0$. (A8) is thus positive, $\frac{d^2\tilde{G}}{dT^2}$ negative, $\frac{d^2\pi_2}{dT^2}$ negative, and hence $T(A^i)$ convex when $\frac{dT}{dA^i} > 0$.■

A3. Proof that theorem 1A concerns a minimum and 1B a maximum

For theorem 1A, we need $\frac{d^2T}{d(A^i)^2} > 0$ at $\frac{dT}{dA^i} = 0$. Differentiating (9) with respect to A^i and rewriting gives

$$\frac{dT}{dA^i} = \frac{\frac{(1-e^{-(2\delta+\rho)T})}{2\delta+\rho} - \frac{(1-e^{-\delta T})}{\delta} + \frac{\rho Fz(\delta+\rho)}{(A^f-A^i)^2}}{A^i e^{-\delta T} + (A^f-A^i)e^{-(2\delta+\rho)T}}. \quad \text{(A9)}$$

With T implicitly defined by (9), and with the numerator in (A9) equal to 0, we get the productivity level that minimizes the formalization time. Because the numerator in (A9) and $\frac{dT}{dA^i}$ equal zero at the optimum, the second derivative at the stationary point, $\left[\frac{d^2T}{d(A^i)^2}\right]_{\frac{dT}{dA^i}=0}$, becomes

$$\left[\frac{d^2T}{d(A^i)^2}\right]_{\frac{dT}{dA^i}=0} = \frac{\frac{d}{dA^i}\left(\frac{\rho Fz(\delta+\rho)}{(A^f-A^i)^2}\right)}{A^i e^{-\delta T}+(A^f-A^i)e^{-(2\delta+\rho)T}} = \frac{\left(\frac{2\rho Fz(\delta+\rho)}{(A^f-A^i)^3}\right)}{A^i e^{-\delta T}+(A^f-A^i)e^{-(2\delta+\rho)T}}, \quad \text{(A10)}$$

which is positive, hence theorem 1A refers to a minimum (which is global, because of lemma 1).

As for theorem 1B, the condition in (15) cannot represent a minimum. The optimum involves a higher productivity level than what gives $\frac{dT}{dA^i} = 0$. At $\frac{dT}{dA^i} = 0$, however, an infinitesimal increase in A^i produces no change in $e^{-\rho T}$ but an increase in $\frac{\rho F}{(\delta+\rho)(A^f-A^i)}$, hence tax revenue (expression 14) increases.

In order to show that (15) represents a maximum, I use the convexity of $T(A^i)$, established above. The relevant part of expression 14 is $\frac{e^{-\rho T}}{A^f-A^i}$ (where A^i is always $< A^f$, from expression 10). Its maximization is identical to minimizing $e^{\rho T}(A^f - A^i)$, where $e^{\rho T(A^i)}$ is convex because $T(A^i)$ is convex. Consider a general product of two functions, $f_1(A^i) \times f_2(A^i)$, where $f_1$is increasing and convex, and f_2 decreases at a constant rate in A^i (and is always > 0). Minimization is obtained for A^i satisfying $-\frac{f_1'}{f_2'} = \frac{f_1}{f_2}$. Due to the convexity of f_1 and linearity of f_2, increases in A^i beyond the optimum imply a more than proportional increase in f_1 than decrease in f_2, implying the product increases. Hence (15) refers to a maximum. The reasoning can be applied over the entire interval where $\frac{dT}{dA^i} > 0$, i.e. the maximum is global. An alternative expression of the proof is that an exponential function ultimately dominates any polynomial function.■

A4. A more general specification of the informal-formal productivity difference

Assume $A^i = \Pi, A^f = 1 + \Pi\theta$, $0 \leq \theta < 1$, where Π is entrepreneurial ability and θ characterizes how much A^f increases when A^i increases with one unit, and parameters are restricted (only) such that $A^i \leq A^f$. Substitute θ with $1 - \alpha$, where $0 < \alpha \leq 1$, to simplify the below expressions.

Plugging this specification of the productivity parameters into (9) to derive an expression corresponding to (12) gives

$$\Pi = \widetilde{\Pi} \pm \sqrt{\widetilde{\Pi}^2 + \frac{\delta\left(1-e^{-(2\delta+\rho)T}\right)-z\delta\rho(\delta+\rho)(2\delta+\rho)F/(A^f)^2}{\alpha((2\delta+\rho)\left(1-e^{-\delta T}\right)-\alpha\delta\left(1-e^{-(2\delta+\rho)T}\right))}}, \text{ with } \widetilde{\Pi} = \frac{(2\delta+\rho)\left(1-e^{-\delta T}\right)-2\alpha\delta\left(1-e^{-(2\delta+\rho)T}\right)}{2\alpha((2\delta+\rho)\left(1-e^{-\delta T}\right)-\alpha\delta\left(1-e^{-(2\delta+\rho)T}\right))}. \quad \text{(A11)}$$

The expression has the same structure as (12), with α=1 being the case discussed above. Differentiation, following the same steps as in Appendix A1, establishes lemma 1.■

A5. Further analyzing optimal penalties to increase tax revenue from formalizing firms

In analyzing the tax-to-profit derivatives ratio from section 6, $\frac{\left|\frac{d}{dA^i}PV(TAX)\right|}{\left|\frac{d}{dA^i}\pi^{formalization}\right|}$, I first discuss the numerator. Starting with the present value of taxes from expression 14, and using a Taylor expansion, the factor $\tau A^F(\frac{\rho F}{(\delta+\rho)(A^f-A^i)}+\frac{A^f}{z(\delta+\rho)^2\rho})$ can be approximated, arbitrarily well, with a polynomial in A^i. Given that the expression is increasing and convex in A^i, the polynomial has positive coefficients only. Next consider $e^{-\rho T}$. From above we know that $\frac{dT}{dA^i}=0$ for an $A^i>0$, and that $T(A^i)$ is convex above this point. As a result, $e^{-\rho T}$ can be approximated, above the point where $\frac{dT}{dA^i}=0$, by the exponential of a polynomial with negative coefficients only, and without a linear term. If we instead write the exponential function in the denominator of $PV(TAX)$, the minus sign goes away. We thus have an ever increasing polynomial divided by an exponential function of a positive and ever increasing polynomial Beyond $A^i = A^i_{TAX\_maximum}$, $PV(TAX)$ is decreasing, as outlined in Appendix A3. The shape of $PV(TAX)$ will be determined by the shape of $e^{-\rho T}$, which is first concave (and decreasing), then convex (and decreasing), asymptotically approaching zero, as $A^i \to A^i_2$. The absolute value of the tax derivative, in going from $A^i_{TAX\_maximum}$ to A^i_2, will therefore first be increasing and convex, concave around the peak, and decreasing and convex when approaching A^i_2.

The profit function, $\pi^{formalization}$, is convex in the productivity level. This can be seen as follows, here phrased in terms of comparing firms of different productivity levels. Consider a firm of productivity A^i. Let the optimal capital accumulation path be $k^*_t(A^i)$, and the formalization time $T^*(A^i)$, where the star indicates the optimum. Another firm, with productivity

$A^i + \Delta A^i$ has the option to choose the same investment, capital accumulation path, size at formalization, and formalization time. Profits between the two firms would then only differ by the present value of the additional production the more productive firm would have while informal, which would amount to $\Delta A^i \int_0^{T^*(A^i)} k_t^*(A^i)\, e^{-\rho t} dt$ (from expression 3). However, the firm with productivity $A^i + \Delta A^i$ does not choose this solution, instead "reoptimizes", and makes a higher total profit. Now consider a firm with productivity A^i+$2\Delta A^i$, i.e. we again increase productivity with ΔA^i. Analogously to the above, it could choose capital path $k_t^*(A^i + \Delta A^i)$ and formalization time $T^*(A^i + \Delta A^i)$, which would yield a profit difference, to the firm with productivity $A^i + \Delta A^i$, of $\Delta A^i \int_0^{T^*(A^i + \Delta A^i)} k_t^*(A^i + \Delta A^i)\, e^{-\rho t} dt$. Again, the firm chooses another solution, and earns yet higher profits. Now compare the two hypothetical profit differences, when going from A^i to $A^i + \Delta A^i$, and from $A^i + \Delta A^i$ to $A^i + 2\Delta A^i$, respectively. The latter profit difference is unambiguously larger, as $T^*\left(A^i + \Delta A^i\right) > T^*(A^i)$ over the range that we consider, and $k_t^*\left(A^i + \Delta A^i\right) > k_t^*(A^i)$, which follows from expression 4. We thus have that these "outside option" profit increments, which are not chosen by the firm as the productivity level increases, increase in A^i. This implies that the profit function is convex.

The derivatives ratio, $\frac{\left|\frac{d}{dA^i} PV(TAX)\right|}{\left|\frac{d}{dA^i}\pi^{formalization}\right|}$, will thus entail a numerator that increases from zero, peaks and then declines to zero, similar to the look of a Gaussian curve. Dividing with the increasing and convex profit derivative will impose a positive skew on the tax derivative curve, such that the peak of the ratio is to the left of the peak of the tax-derivative itself. The exact shape of $\frac{\left|\frac{d}{dA^i} PV(TAX)\right|}{\left|\frac{d}{dA^i}\pi^{formalization}\right|}$ will depend on the parametrization and figure 2 provides an illustration.

A6. Labor in the production function

Consider a production function of the Cobb-Douglas type, with capital and labor, and constant returns, i.e. $A^i k_t^\eta l_t^{1-\eta}$ as informal and $A^f k_t^\eta l_t^{1-\eta}$ as formal, with capital intensity $\eta \in (0,1)$, and where A^i and A^f are not necessarily the same constants as above. In period t, and in addition to choosing the investment level, the firm decides on how many workers, l_t, to hire, at the

exogenous wage rate w. Exemplifying with the formal profit maximization problem in (2), the problem is modified as follows:

Choose i_t, l_t to Max $\int_T^{\infty}(A^f k_t^{\eta} l_t^{1-\eta} - \frac{z}{2} i_t^2 - w l_t)\, e^{-\rho t} dt$ s.t. $k_t' = i_t - \delta k_t$ and $k(T) = \widetilde{k_T}$ (A12)

In the optimum, the firm hires a quantity of labor to maintain a constant capital to labor ratio. This can be seen through the first order condition with respect to l_t of the modified Hamiltonian, $H(t, i, l, k, \lambda) = \left(A^f k_t^{\eta} l_t^{1-\eta} - \frac{z}{2} i_t^2 - w l_t\right) e^{-\rho t} + \lambda_t (i_t - \delta k_t)$, i.e. $\frac{\partial H(..)}{\partial l} = 0$, which can be written as

$$l_t = k_t \left(\frac{A^f (1-\eta)}{w}\right)^{\frac{1}{\eta}}, \quad \text{(A13)}$$

where $\left(\frac{A^f (1-\eta)}{w}\right)^{\frac{1}{\eta}}$ is a constant. The derivative of the Hamiltonian with respect to k_t, changes from $A^f - \lambda_t \delta$, for the problem in (2), to $A^f \eta \left(\frac{l_t}{k_t}\right)^{1-\eta} - \lambda_t \delta$. With a constant capital to labor ratio, from (A13), the first term in this derivative is also a constant (equaling $\eta (A^f)^{\frac{1}{\eta}} \left(\frac{1-\eta}{w}\right)^{\frac{1-\eta}{\eta}}$). As the investment first order condition is intact from above, the only change in the dynamic equations characterizing i_t and k_t is in this modified production factor. All dynamic properties of the model remain. ■