OPTIMAL PENALTIES ON INFORMAL FIRMS[1]

APRIL 2017

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

I study a much discussed and highly policy relevant question: what, if anything, should the government do about informal production? In a simple capital accumulation model of informal firm growth and potential formalization, I show analytically that optimal penalties depend on the productivity level: the least productive firms should be left alone, the more productive ones should always face positive penalties, which increase in the productivity level. This holds for two different government objectives, i.e. maximizing formal sector tax revenue, and speeding up formalization, respectively, where the latter objective results in higher average penalties. This theoretical result provides a direct policy advice in an area where some countries may have followed such a policy, without a rigorous foundation, whereas other countries have simply let all informal firms be, independent of productivity or size. The result is in line with an emerging consensus about distinct types of informal entrepreneurs.

JEL classification: E26, H26, K42, O17

Keywords: Informality, penalties, enforcement, formalization, public policy

[1] This paper was previously circulated as "Informal firms, investment incentives and formalization". I am much indebted to Paulo Furquim for discussions that helped shape this work. I also thank Love Ekenberg, Shon Ferguson, Harry Flam, Johan Gars, Gustav Hansson, Ann Harrison, Bård Harstad, John Hassler, Per Krusell, Ted Miguel, José Mauricio Prado, Jesper Stage, Jakob Svensson, Ragnar Torvik, and Fabrizio Zilibotti for helpful discussions, feedback and suggestions, as well as participants at the Berkeley development lunch, Stockholm University Economics workshop, IIES brown bag seminar, University of Gothenburg Economics seminar, the SUDSWEc conference, the Nordic Development Economics Conference (NCDE) and the Research Workshop on Institutions and Organizations (RWIO). 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 (La Porta and Shleifer, 2014; Hassan and Schneider, 2016; Loayza, 2016B). Hence the question is highly relevant.

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 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 (2014) argue for a "dual" model (inspired by Lewis, 1954), which contrasts both to de Soto´s "romantic" and Farrell´s "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 (2014), in turn, suggests a unifying modeling framework that contains the three classes of informal firms, i.e. subsistence/survival, non-compliance/parasitism and "de Soto" entrepreneurship. The author studies the firm specific and economy wide effects of different entry-, tax- and enforcement policies.[2]

Differently from Ulyssea (2014), and the literature at large, this paper studies how an enforcement policy vis-à-vis informal firms, here referred to as *penalties*, should be optimally set, as a function of the informal firm´s productivity level.[3] 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. The objective is thus more modest than the complete welfare analysis of e.g. Ulyssea (2014), in that I take as given both the formal sector tax level and positive entry costs, and only use a penalty instrument, affecting the incentives of informal entrepreneurs. 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, Ulyssea (2014) and Loayza (2016A) 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. Yet, inherent to this paper´s modeling approach, with productivity-dependent penalties, and in line with the argument of de Mel, McKenzie and Woodruff (2010), is the recognition that the growth potential of the informal firms that do resemble formal firm owners should not be neglected.

[2] The discussion here concerns the 2014 version of the Ulyssea model, which is under revision (information from author´s webpage). Loayza (2016B) is another paper that incorporates the above empirical regularities by considering a "rudimentary" as well as a "modern" informal sector.

[3] The literature typically assumes linear penalties (e.g. Loayza, 1996; Johnson et al, 1997; Ihrig and Moe, 2004; Prado, 2011; Nguimkeu, 2016) or convex penalties (e.g. Ulyssea, 2014; and papers referenced therein), where the latter approach can be rationalized by a probability of detection that increases in informal firm size.

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. (2013). 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 two objectives, which is the topic of this paper, delivers a "zero first, then increasing" penalty, as a function of informal firm productivity.

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 penalty levels needed to achieve the optimal productivity levels, for the two objectives. Section 5 briefly discusses the model implications in relation to assuming specific distributions of firm productivities. Section 6 discusses the results, with some of the derivations, and two model extensions, in the Appendix.

2. A model of informal firm formalization

Consider a dynamic model of firm investment, growth and possible formalization.[4] Starting out as informal, the question in this section is if, when and at what size the firm will become formal, and how this decision depends on the firm's productivity level. The derivation arrives at expressions (9-12), which are then analyzed in the sections that follow. The production function is simple: output is linear in the capital stock, k_t.[5] As informal, the firm produces $A^i k_t$, if it has formalized, production is $A^f k_t$, with

[4] Some parts of the modeling is inspired by the framework in Harstad and Svensson (2011).
[5] The final section of the appendix discusses a production function also containing labor, with results unchanged.

$A^f > A^i$. Thus, formality is desirable.[6] The firm can grow by investing (i_t) in its capital stock. The cost of investing is convex, $\frac{z}{2} i_t^2$. This gives a profit flow, $\pi_t = A^i k_t - \frac{z}{2} i_t^2$, in case the firm is informal. The capital stock depreciates at the rate δ. Capital 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 the rate ρ. Starting with a capital stock of k_0, the informal firm chooses an investment path, whether it should become formal and the time of formalization (T). 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.

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:

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

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

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.[7]

[6] I return to the specification of A^f in section 3, when discussing tax revenue. A^f can be thought of as the after-tax productivity in the formal sector.

[7] 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 informal investment path, for a given T, can, in turn, be determined by solving for the investment 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

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

The solution is derived as above, the difference being the terminal constraint on k_t.[8] 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:

Choose i_t to Max $\int_0^\infty (A^i k_t - \frac{z}{2} i_t^2)\, e^{-\rho t} dt$ s.t. $k_t' = i_t - \delta k_t$ and $k(0) = k_0$, with solution (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 $k_\infty^{informal} = \frac{A^i}{z\delta(\delta+\rho)}$.[9]

Solving for the formalization time *T*

If T exists, the investment path before/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

the convexity of costs - the firm wants to spread it 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.

[8] Solving for the i_t- and k_t-paths 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.

[9] 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.

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

As discussed above, pre-formalization investment approaches the formal investment rate as $t \rightarrow T$. 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

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

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

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

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

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

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

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

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

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

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 of firms that will not formalize. In addition, there is a high-productivity range of firms that also do not formalize, which is in line with empirical observations. 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 $A^i_{T\_minimum}$. For all $A^i \in \left(A^i_1, A^i_{T\_minimum}\right)$, we have $\frac{dT}{dA^i} < 0$, and for all $A^i \in \left(A^i_{T\_minimum}, A^i_2\right)$, $\frac{dT}{dA^i} > 0$.

Proof. See appendix A1.

There is thus monotonicity in the relation between T and A^i (first *always* decreasing, then *always* increasing), which gives the results general validity, independent of the parametrization of the model. 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 return to the formal sector productivity, A^f. It can be thought of as the after-tax productivity parameter, once a revenue/output tax (specified as in e.g. Prado, 2011 or Ulyssea, 2014) has been levied on A^F, the "baseline" formal productivity parameter. We thus have $A^f \equiv A^F(1-\tau)$, where τ is the tax rate. A^f is the relevant parameter for formalization incentives, hence A^F was not introduced until this point. 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

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

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

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

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

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 $\frac{dT}{dA^i} > 0$. 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 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 the interval $\left(A^i_1, A^i_2\right)$, and is larger than $A^i_{T\_minimum}$.

Proof. The results follow from lemma 1 and the above tax derivation. Appendix A2 shows that the inflection points are minimum and maximum points, respectively.

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 an optimal tax rate and specifying a full welfare function, I 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 which 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

Penalties vis-à-vis the informal sector is a policy instrument the government can use in order to affect formalization incentives. I assume that such monitoring/enforcement is costless for the government. This is unrealistic, but positive monitoring costs would strengthen the argument of the paper, I therefore postpone a discussion until section 6. One way of thinking about the effects of enforcement is that it makes informal firms divert time from production, in order to avoid the authorities.[10] I will instead discuss penalties as the de facto reduction of the informal productivity parameter needed to achieve the two policy objectives, recognizing there may be several channels than can produce such a reduction. An additional issue, discussed only in some models of informality/formality, is what can actually be monitored by the authorities. In the present context, where I assume that penalties affect productivity, it means productivity must be observable. Given the high correlation of manager characteristics and firm productivity in the data, cited by e.g. La Porta and Shleifer (2014), one (imperfect) measure of productivity could be observation of owner characteristics, from audits. Another method is to back out an (imperfect) measure of productivity from an observation of 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.

[10] Informal firm production could be specified as $p(h)A^i k_t(1-h) + \left(1-p(h)\right)A^i k_t(1-h)(1-\psi)$, 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.

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.

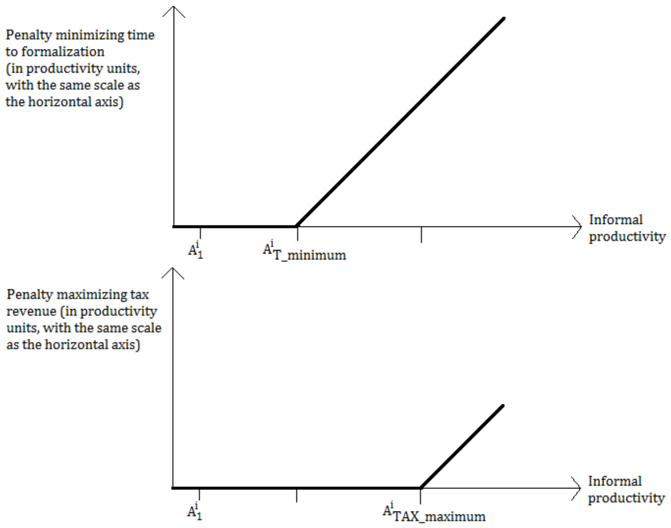

Figure 1. Optimal penalties, for each level of informal firm productivity, in order to minimize the time to formalization, and maximize the firm´s contribution to tax revenue once formal, respectively. The scale of all axes is the same, and the optimal penalty is first zero, then increases one-to-one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

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, 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 level 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.

5. Distributions of firm productivities

The model above assumes that A^f is constant. As I discuss in Appendix A3, however, the results are valid also for a specification in which the only requirement is that, as A^i increases, it increases faster than A^f. The results thus carry over to a model where "entrepreneurial ability" increases a firm´s productivity in both sectors, as long as the gap $A^f - A^i$ does not widen when A^i increases. This assumption is plausible, given the paper´s objective of studying if informal firms of different productivity should be treated differently.

The results in the paper are derived on a "per firm" basis, and no assumption is made about the distribution of informal sector productivities. The above discussed empirical evidence 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. Discussion

This paper derives optimal penalties on informal firms, under two different objectives, i.e. speeding up formalization, and maximizing the firm´s contribution to tax revenue once formal, respectively. For both objectives, I show that low-productivity firms should be left alone. High-productivity informal firms should instead face penalties that increase in productivity, with fewer penalized firms and lower penalties in case of the tax revenue objective. These results are summarized in figure 1 and relate directly to the analysis of e.g. Bruhn and McKenzie (2014), who discuss how a differentiated treatment vis-à-vis different types of informal firms may be the optimal policy stance. Figure 1 can also be fed back into models of informal-formal linkages, where typically linear penalties (e.g. Prado, 2011) or quadratic penalties are assumed. A penalty specification such as figure 1, instead of a quadratic assumption, could probably revert some negative welfare results related to increased enforcement in models of the informal sector (such as in Ulyssea, 2014).[11]

The paper does not model or study welfare linkages between the formal and informal sectors, and assumes taxes and entry costs to be constant, instead focusing on the informal sector productivity and penalties. It is quite likely however, that the intuition about two different policies carry over to a welfare setting. Assume taxes contribute to public goods, which in turn increase formal sector productivity, but with decreasing returns. Assume a certain percentage of firms (the formal ones) pay taxes. If the marginal productivity of public spending is high, then bringing in more firms from informality would almost certainly be socially beneficial. If this happened, we would subsequently get a decrease in the

[11] Note that also a static model can be constructed, yielding policy recommendations similar to the present paper. The productivity of the most informal firms would need to be reduced, for these firms to find formalization attractive. The less productive informal firms should be left alone, as they would not formalize anyway. However, such a model would miss out completely on the dynamic aspects, i.e. that firms can grow and then formalize. How such firms should be treated, penalty-wise, is of fundamental importance. In light of the high costs of becoming formal in many countries, compared to average- and informal sector incomes (e.g. de Soto, 1989; World Bank, 2017), a dynamic model is also justified.

marginal return to public spending. In the limit, only the least productive and smallest informal firms would remain, and would contribute very little to taxes and public goods if formalized (their disappearance would be more likely, but formalization is at least theoretically possible). In addition, this would happen in a situation where the marginal return to public spending is already very low. Leaving these firms alone, rather than forcing formalization, would most probably be socially optimal in a majority of models, unless initial conditions or assumptions about the productivity of public goods are extreme.

Now consider monitoring costs. Prado (2011) assumes both monitoring costs and government revenue from audits of informal firms. The latter would not be realistic for the least productive firms here considered, although it cannot be theoretically ruled out for the largest informal firms. Overall, penalty revenues seem less plausible for the context here imagined. As for monitoring costs, if there is a fixed cost of monitoring per firm, it would never be optimal to audit the smallest firms in the present model. If the probability of detection is increasing in informal firm productivity, it would be most beneficial to spend auditing resources on the largest informal firms. Both specifications would strengthen the argument presented in the paper.

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.

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.

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.

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.

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.

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., 2014. Informality and Development. Journal of Economic Perspectives 28 (3): 109-126.

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.

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.

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

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

Ulyssea, G., 2014. Firms, informality and development: Theory and evidence from Brazil. PUC Rio Economics Department Working Paper 632.

World Bank, 2017. The World Bank Doing Business. Measuring Business Regulation. http://www.doingbusiness.org

Appendix

A1. Proof to Lemma 1

Starting at infinity, we need to show that $T(A^i)$ first monotonously decreases in A^i, then monotonously increases and goes to infinity. Alternatively, we can show that to each T correspond exactly two solutions A^i (with the exception of the inflection point). This is the idea of writing the solution to (9) in the form of expression (12), which converges to (10), as $T \rightarrow \infty$. Starting with (12), where $\pi \equiv \frac{A^i}{A^f}$, I use the auxiliary expressions $\tilde{F} = \frac{z\delta\rho(\delta+\rho)(2\delta+\rho)F}{(A^f)^2}$ and $\widetilde{D} = (2\delta+\rho)\left(1-e^{-\delta T}\right) - \delta\left(1-e^{-(2\delta+\rho)T}\right)$, in addition to $\tilde{\pi} = \frac{(2\delta+\rho)\left(1-e^{-\delta T}\right)-2\delta(1-e^{-(2\delta+\rho)T})}{2\widetilde{D}}$. Further, let $\pi_1 = \tilde{\pi} - \sqrt{\tilde{\pi}^2 + \frac{\delta\left(1-e^{-(2\delta+\rho)T}\right)-\tilde{F}}{\widetilde{D}}}$ and $\pi_2 = \tilde{\pi} + \sqrt{..}$, where $\sqrt{..}$ denotes the square root expression. We thus need to show that

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

These derivatives can be written as follows:

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

$$\frac{d\pi_2}{dT} = \frac{1}{\sqrt{..}}\Big(\frac{d\tilde{\pi}}{dT}\pi_2 + \frac{1}{2}\Big(\frac{\delta(2\delta+\rho)e^{-(2\delta+\rho)T}}{\widetilde{D}} - \frac{\frac{d\widetilde{D}}{dT}}{\widetilde{D}^2}\big(\delta\left(1-e^{-(2\delta+\rho)T}\right) - \tilde{F}\big)\Big)\Big) \tag{A2}$$

From the restriction on F in (11), we know that π_1 is always positive. In addition $\widetilde{D}$ is always positive (for $T < \infty$), as is $\frac{d\widetilde{D}}{dT} = \delta(2\delta+\rho)(e^{-\delta T} - e^{-(2\delta+\rho)T})$, 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 $\frac{\delta(2\delta+\rho)e^{-(2\delta+\rho)T}}{\widetilde{D}} - \frac{\frac{d\widetilde{D}}{dT}}{\widetilde{D}^2}\big(\delta\left(1-e^{-(2\delta+\rho)T}\right) - \tilde{F}\big)$ is positive. It remains to show that $\frac{d\tilde{\pi}}{dT} > 0$:

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

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 $X \equiv \delta$, $Y \equiv 2\delta + \rho$ and $Y > 2X$. Assuming this expression equals zero, the assumed equality can be written as

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

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, completing the proof.■

A2. 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

$$\frac{dT}{dA^i} = \frac{\frac{(1-e^{-(2\delta+\rho)T})}{2\delta+\rho} - \frac{(1-e^{-\delta T})}{\delta} + \frac{\rho Fz(\delta+\rho)}{(A^f-A^i)^2}}{A^i e^{-\delta T} + (A^f - A^i)e^{-(2\delta+\rho)T}}. \tag{A5}$$

With T implicitly defined by (9), and with the numerator in (A5) equal to 0, we get the productivity level that minimizes the formalization time. Because the numerator in (A5) and $\frac{dT}{dA^i}$ equal zero at the optimum, the second derivative at the inflection point, $\left[\frac{d^2T}{d(A^i)^2}\right]_{\frac{dT}{dA^i}=0}$, becomes

$$\left[\frac{d^2T}{d(A^i)^2}\right]_{\frac{dT}{dA^i}=0} = \frac{\frac{d}{dA^i}\left(\frac{\rho Fz(\delta+\rho)}{(A^f-A^i)^2}\right)}{A^i e^{-\delta T} + (A^f - A^i)e^{-(2\delta+\rho)T}} = \frac{\left(\frac{2\rho Fz(\delta+\rho)}{(A^f-A^i)^3}\right)}{A^i e^{-\delta T} + (A^f - A^i)e^{-(2\delta+\rho)T}}, \tag{A6}$$

which is positive , hence theorem 1A refers to a minimum. As for theorem 1B, the condition in (15) cannot represent a minimum. The optimum in theorem 1B 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. ■

A3. 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

$$\Pi = \widetilde{\Pi} \pm \sqrt{\widetilde{\Pi}^2 + \frac{\delta(1-e^{-(2\delta+\rho)T}) - z\delta\rho(\delta+\rho)(2\delta+\rho)F/(A^f)^2}{\alpha((2\delta+\rho)(1-e^{-\delta T}) - \alpha\delta(1-e^{-(2\delta+\rho)T}))}}, \text{ with } \widetilde{\Pi} = \frac{(2\delta+\rho)(1-e^{-\delta T}) - 2\alpha\delta(1-e^{-(2\delta+\rho)T})}{2\alpha((2\delta+\rho)(1-e^{-\delta T}) - \alpha\delta(1-he^{-(2\delta+\rho)T}))}. \tag{A7}$$

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.■

A4. 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 $k(T) = \widetilde{k_T}$ (A8)

In the optimal solution, 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

$$l_t = k_t (\frac{A^f(1-\eta)}{w})^{\frac{1}{\eta}}, \tag{A9}$$

where $(\frac{A^f(1-\eta)}{w})^{\frac{1}{\eta}}$ 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 $A^f \eta (\frac{l_t}{k_t})^{1-\eta} - \lambda_t \delta$. With a constant capital-to-labor ratio, from (A9), the first term in this derivative is also a constant (equaling $\eta (A^f)^{\frac{1}{\eta}} (\frac{(1-\eta)}{w})^{\frac{1-\eta}{\eta}}$). As the investment first order condition is intact from above, the only change in the dynamic equations describing i_t and k_t is in this modified production factor. All dynamic properties of the model remain. ■