Firms, Informality and Welfare*

Gabriel Ulyssea
University of Chicago and IPEA

Job Market Paper

November 25, 2013

Abstract

This paper shows the empirical and theoretical importance of distinguishing between two margins of informality: (i) when firms do not register and pay entry fees – the *extensive margin*; and (ii) when firms hire workers "off the books" – the *intensive margin*. I build an equilibrium entry model where heterogeneous firms can exploit both. I show that the proposed framework encompasses the leading views of informality and integrates them in a unified setting. I estimate the model with firm-level data and use it to infer the empirical relevance of these views. I then assess the effects of reducing regulatory costs and increasing enforcement in a general equilibrium setting. The simulation results indicate that: (a) partial equilibrium LATE estimates tend to overestimate policies' impacts; (b) worker and firm informality show very different policy responses once the intensive margin is accounted for; and (c) informality reductions are not necessarily associated to welfare gains.

JEL Codes: O17, C54, O12.

*I am indebted to James Heckman, Steven Durlauf and Chang-Tai Hsieh for their guidance and constant encouragement. I would like to thank Azeem Shaik, Claudio Ferraz, Rafael Lopes de Melo, Ben Moll, Carlos Henrique Corseuil and Miguel Foguel for their comments and helpful discussions. Special thanks to Stephane Wolton and Dimitri Szerman, who provided me with detailed and extremely helpful comments and feedback. I am also thankful to seminar participants at various universities and conferences for helpful comments. Finally, I would like to thank Ricardo Paes de Barros for first encouraging me to pursue this project. Of course, all errors are mine. Financial support from CAPES, IPEA and The University of Chicago is gratefully acknowledged.

1 Introduction

The informal sector is a central feature of most developing economies. It constitutes the majority of firms and accounts for between a third and half of GDP in these countries [e.g. Schneider (2005)]. The aim of this paper is to improve our knowledge about informality and its potential consequences for economic development.

This paper distinguishes two margins of informality: (i) whether firms register and pay entry fees to achieve a formal status – the *extensive margin*; and (ii) whether firms that are formal in the first sense hire workers "off the books" – the *intensive margin*. I show that the intensive margin is empirically very important and accounting for it clarifies the role of informality in developing countries. I build an equilibrium model to analyze firms' choices regarding entry, production and both margins of informality. The existing literature focuses on the extensive margin alone, which implies that being informal is a binary decision to comply or not with taxes and regulations.[1] In this context, firm (in)formality implies worker (in)formality. However, this is highly at odds with the available evidence, since a large share of informal employment comes from formal firms.[2] The intensive margin breaks the direct association between firm and worker informality, and implies that formal and informal are no longer disjoint states, as registered firms may hire a share or all of their labor force "off the books".

The framework developed in this paper encompasses the leading views about informal firms [La Porta and Shleifer (2008)], and is able to integrate them in a unified setting. These views constitute the main approaches to the study of informality, but their actual empirical relevance remains as an open question in the literature [Arias et al. (2010)]. This is an important gap, as they offer very different interpretations of informal firms' role in economic development and the potential harms and benefits of informality. I show that despite their differences, these views simply reflect heterogeneous firms choosing whether to comply given the institutional framework they face. I argue that the central distinction, however, lies in their predictions about informal firms' behavior in face of policy changes. I exploit these differences to define a taxonomy of informal firms based on these views, which provides a natural setting to infer their empirical relevance.

I estimate the model with the simulated method of moments and data of formal and informal firms in Brazil. I use the estimated model to analyze firms' choices in the counterfactual scenario where formal sector's entry costs are removed, and apply the proposed taxonomy to infer the relative size of each view in the data. The results

[1] This is an important and extensive literature that I build upon. It includes, among others, Rauch (1991), Fortin et al. (1997), Amaral and Quintin (2006), de Paula and Scheinkman (2011), and Galiani and Weinschelbaum (2012).

[2] Section 2 presents evidence to support this claim.

show that the potentially productive entrepreneurs that formalize and thrive once formal sector's entry costs are removed represent a small fraction of all informal firms (12.3%). The largest group corresponds to those that are productive enough to survive in the formal sector but choose to remain informal to earn higher profits from the cost advantages of non-compliance (47.8%).[3] The remaining firms correspond to those too unproductive to ever become formal, which are only able to survive because they avoid taxes and regulations.

Based on the estimated model, I proceed to assess the effects of interventions that are typically analyzed in the empirical literature: reducing formal sector's entry costs or payroll taxes; and increasing enforcement of existing institutions.[4] These are large scale interventions that have impacts both on firms and aggregate outcomes, which have been analyzed by completely separate literature streams [a notable exception is the recent work by Meghir et al. (2012)].[5] The present model embeds firm behavior into aggregate relationships, which allows me to dissect policies impacts into their firm-level and aggregate effects.[6]

Firm-level results indicate that partial-equilibrium Local Average Treatment Effect (LATE) estimates tend to overestimate the impacts of policies that reduce entry costs and payroll taxes. The general equilibrium effects go in the direction of increasing wages, especially in the payroll tax experiment, and thus partially revert some of the benefits from firms' standpoint. The results also put into question the common interpretation of a causal effect of formality, which is typically estimated from exogenous policy variations [e.g. McKenzie and Sakho (2010)]. I find that the LATE estimates of reducing regulatory costs (either entry costs or payroll taxes), and of increasing enforcement on the extensive margin are of large magnitude but have opposite signs. In the first case, the interpretation would be that the formal status has a positive causal impact on firms' performance, while in the latter the interpretation would be the reverse. These results thus illustrate the difficulty of interpreting as causal the effects from changes in formality status that are

[3]The first view dates back to De Soto (1989) and has been recently expanded by, for example, Djankov et al. (2002). The second view has been put forward by Farrell (2004) and Levy (2008), among others.

[4]Monteiro and Assuncao (2011) and Fajnzylber et al. (2011)] analyze tax reduction and simplification in Brazil. Bruhn (2011), Kaplan et al. (2011) and de Mel et al. (2013) analyze the effects of policies that reduce bureaucratic entry costs. Almeida and Carneiro (2009) and Almeida and Carneiro (2012) analyze the impacts of higher government auditing at the municipality level in Brazil.

[5] Meghir et al. (2012) focus on the analysis of labor markets (as opposed to firms) and their approach can be seen as complementary to the present one. Recent studies that analyze aggregate effects include Ulyssea (2010), Prado (2011), Charlot et al. (2011), Aruoba (2010), D'Erasmo and Boedo (2012), and Leal-Ordonez (2013), among others.

[6]This is a common approach in the international trade literature [see for example Eaton et al. (2010), Cosar et al. (2010), Dix-Carneiro (2013), and Donaldson (2013)].

policy-induced. Additionally, I find a large degree of effect heterogeneity across firms, even within the group of firms that formalize because of a given policy.

The aggregate results show that the share of informal workers and the share of informal firms – the two most commonly used measures of informal sector's size – have very different policy responses once the intensive margin is accounted for. Reducing entry costs, for example, has strong impacts on the share of informal firms but barely reduces the share of informal workers. This is observed because newly formalized firms are less productive and thus hire a large fraction of their labor force informally. The results also indicate that aggregate TFP and GDP do not necessarily move in the same direction in the present context. Increasing enforcement on the extensive margin sharply reduces informal sector's size, which has positive impacts on TFP through composition effects. However, firms displacement is too strong, and in net GDP remains basically unaltered.[7] When entry costs are reduced, TFP decreases because there is greater entry of low-productivity firms. However, this higher entry causes GDP to increase because the economy is expanding on the extensive margin of production.

Looking at welfare effects, reducing formal sector's entry costs and increasing enforcement on informal firms are the two best performing ones, with a 2.3% increase. However, the two policies operate through very different channels. In the first case, welfare increases because there is a substantial reduction of deadweight losses from wasteful barriers to entry, which increases competition, GDP and wages. In the second case, welfare improves because of substantially higher tax revenues. This effect is in line with the mechanisms highlighted by the literature on fiscal capacity [e.g. Besley and Persson (2010)].[8] Finally, although the different policy instruments always succeed in decreasing at least one dimension of informality, they do not always lead to welfare improvements. Hence, lower informality does not necessarily imply greater welfare.

The remaining of the paper is organized as follows. Section 2 presents the data and some key stylized facts. Section 3 presents the model, while Section 4 discusses informal firms taxonomy. Section 5 contains the estimation method and results. Section 6 presents the quantitative results and Section 7 concludes.

[7] Leal-Ordonez (2013) provides an interesting discussion of the effects of enforcement within the leading models in the literature of resource misallocation.

[8] It is worth noting, however, that this effect depends on the assumption that all tax revenues are rebated to the households and that there is no government waste, resource dissipation and so on. Thus, this should be seen as an upper bound for this effect.

2 Some Facts about Firm Informality

2.1 Definitions and Data

Throughout this paper, I define as *informal workers* those employees who do not hold a formal labor contract (*sem carteira de trabalho*),[9] and as *informal firms* those not registered with the tax authorities. Registering a worker implies several variable costs that come from labor regulation (e.g. payroll taxes). To register a firm in Brazil is a lengthy and costly process (Table 8), and in practice constitutes an entry cost into the formal sector, but also implies additional variable costs due to tax regulations. These definitions are used both in the data and theory. The Brazilian data sets provide a precise way of measuring both types of informality, as firms are directly asked whether they are registered with the tax authorities and whether their workers have a formal labor contract.[10]

I use four data sources to conduct the empirical analysis. The two main data sets used are those that contain information of formal and informal firms in Brazil. The first is the ECINF survey (*Pesquisa de Economia Informal Urbana*), a repeated cross-section of small Brazilian firms (up to five employees), which was collected by the Brazilian Bureau of Statistics (IBGE) in 1997 and 2003. This is a matched employer-employee data set that contains information on entrepreneurs, their business and employees. The ECINF is designed to be representative at the national level for firms with at most five employees. However, the effective sample includes firms with up to 10 employees or more, although the information for larger firms becomes sparser and it is not statistically reliable.[11]

Although ECINF's sample size cap is not likely to be a problem for its ability to capture informal firms (which are predominantly small scale enterprises) it certainly is for formal firms. I thus use the RAIS data set to complement the information on formal firms. This is a matched employer-employee, administrative data set collected by the Brazilian Ministry of Labor. It is an annual panel of both workers and firms and it contains the universe of formal firms and workers. Table 1 compares the main moments from both data sets for 2003.[12] Finally, I also use two surveys collected by the Brazilian

[9] In Brazil, all formal workers are required to have their labor contracts registered in a booklet (*carteira de trabalho*).

[10] These are self-reported variables and naturally raise measurement error concerns. Nonetheless, the National Bureau of Statistics (IBGE) has a long tradition in measuring labor informality with high accuracy, and it has very strict confidentiality clauses, so the information cannot be used for auditing purposes (which could incentivize respondents to misreport). These features, associated to the actual high levels of informality observed in the data, increase the confidence that respondents are not deliberately underreporting their informality status.

[11] See de Paula and Scheinkman (2010) for a more detailed description of the ECINF data set.

[12] Appendix B describes the details of the construction of the data sets used

Bureau of Statistics to compute some aggregate labor market statistics (such as the share of informal workers). The first is the National Household survey (PNAD), a repeated cross section that is representative at the national level. The second is the Monthly Employment Survey (PME), which is a rotating panel of workers that has the same design as the U.S. Current Population Survey and covers the 6 main metropolitan areas.

Table 1: Comparing RAIS and ECINF

	Formal – RAIS	Formal – ECINF	Informal – ECINF
Sector composition (%)			
Services	41.9	42.5	53.7
Manufacturing	12.6	7.9	8.9
Commerce	45.5	49.6	37.4
Size Distribution (# workers)			
Pc. 25	1	1	1
Pc. 50	3	2	1
Pc. 75	7	3	1
Pc. 95	31	5	3
Mean	10.8	2.1	1.3
Obs.	476,299	2,600	18,736

Source: Author's own tabulations from RAIS and ECINF, 2003. In 2003, RAIS had information on over 2.1 million firms; I work with a 25% random sample from the original data set.

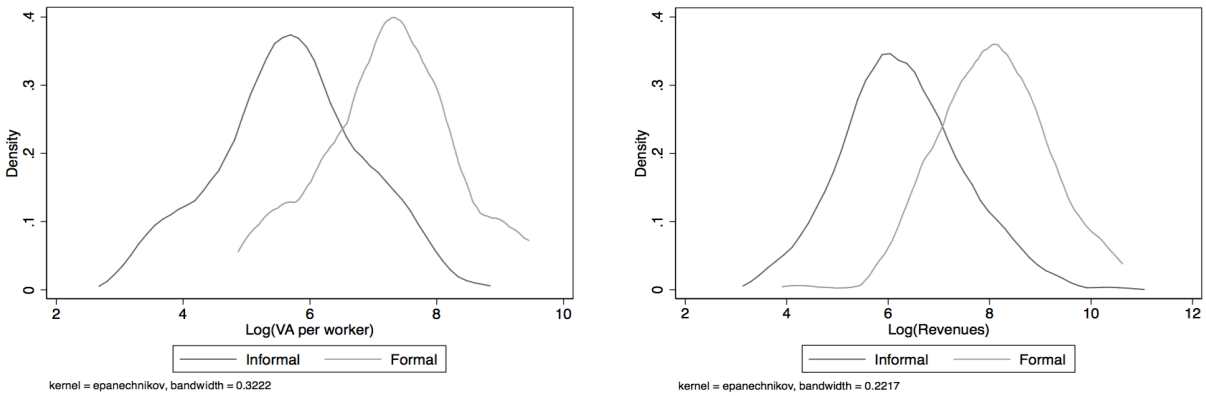
2.2 Facts

There exist some well-stablished facts about informal firms in the literature [e.g. Perry et al. (2007) and La Porta and Shleifer (2008)]: on average they have less educated entrepreneurs, are smaller both in terms of employees and revenues, pay lower wages and earn lower profits relatively to formal firms. These facts are also present in the Brazilian data [see de Paula and Scheinkman (2011) and Table 9 in the Appendix]. Despite these stark differences between formal and informal firms, the evidence indicates that they coexist even within narrowly defined industries (see Figure A.3 in the appendix), which contradicts the notion that formal and informal firms operate in completely different markets.

The first question to be examined refers to whether firms sort between sectors based on productivity right upon entry. If true, this would imply that formal sector's distribution is shifted to the right relatively to informal sector's, even amongst young firms. A second and related question is whether there is some overlap between productivity distributions in both sectors. Since the model implies a one-to-one relationship between productivity and size (measured as employment or revenues), I also examine whether these facts hold for firm size distribution as well. Measuring productivity is a difficult task and there is

an entire body of literature devoted to it. Due to data limitations, I use a crude but common proxy for productivity, which is simply value-added per worker. As for the size measure, I use log-revenues. Both measures are computed among firms with at most 12 months to proxy for entrants.[13] Figure 1 indicates that indeed firms sort based on productivity right upon entry, as both productivity and size distributions in the formal sector are shifted to the right. Moreover, there is a substantial overlap between them.

Figure 1: Productivity and size distributions among entrants



(a) Productivity: Log(VA/Worker)

(b) Size: Log(Revenues)

The second question to be examined refers to the patterns and empirical relevance of the intensive margin. It is worth highlighting that a well-known fact in the literature is that the probability of firms being informal strongly decreases with firms' size, usually measured as number of employees [e.g. Perry et al. (2007)], which also holds in the Brazilian data [see Figure A.4 in the appendix and de Paula and Scheinkman (2011)]. One possible rationale behind this fact is simply that larger firms are too visible to the government and thus more likely to be audited. Given this argument, it is likely that the same is holds true for the intensive margin: larger formal firms (in number of employees) are likely to have a lower share of informal employees. The Brazilian data suggests that this is true, as the intensive margin of informality is also decreasing in firm's size (Figure A.4 in the appendix).

Finally, I assess the empirical relevance of the intensive margin, which can only be done indirectly with the data available. In Table 2, I use data from the Monthly Employment Survey (PME) to show that 52% of all informal workers are employed in firms with 11 employees or more. As already discussed, the likelihood of a firm with 11 employees or

[13]To obtain cleaner measures, I regress the log of value-added and log-revenues on a set of industry dummies to purge inter-industry variation. The computed log-residuals are the productivity and size measures used.

more to be informal is very low. These two pieces of evidence combined thus suggest that there is a large fraction of informal workers who are employed in formal firms. Hence, if one does not include this additional margin, one is likely to be missing a considerable share of informal labor.

Table 2: Formal and informal employment composition by firm size

	Informal Workers (in %)	Formal workers (in %)
Firm size (# employees)		
0–5	35.8	6.6
6–10	11.7	7.2
11 or more	52.5	86.2

Source: Author's own tabulations from the Monthly Employment Survey (PME) 2003.

3 An Equilibrium Model of Entry and the Two Margins of Informality

3.1 Set up

There is a continuum of firms that are indexed by their individual productivity, θ. Firms produce a homogeneous good using labor as their only input. Product and labor markets are competitive, and formal and informal firms face the same prices.[14] Workers are assumed to be homogenous and formal and informal employees perform the exact same task within the firm.[15] Output of a given firm θ is given by

$$y(\theta, \ell) = \theta q(\ell)$$

where the function $q(\cdot)$ is a assumed to be increasing, concave, and twice continuously differentiable.

[14] As argued in Section 2, formal and informal firms coexist even within narrowly defined markets, so the assumption that firms face the same output price seems like a reasonable approximation. Nevertheless, the model can be readily modified to a monopolistic competition setting where firms produce different varieties.

[15] This is of course a strong assumption, but since the goal is to characterize firms' behavior, I choose to use the simplest model possible.

Incumbents

Informal incumbents are able to avoid taxes and labor costs, but face a probability of detection by government officials. If caught, they must pay a penalty that can assume the form of a bribe or fine. This expected cost assumes a very general form of a labor distortion term denoted by $\tau_i(\ell)$.[16] Informal firms' profit function is given by:

$$\Pi_i(\theta, w) = \max_{\ell} \{\theta q(\ell) - \tau_i(\ell) w\ell\} \tag{1}$$

where the price of the final good is normalized to one.

This profit function is a standard one except for the term $1 \leq \tau_i(\cdot) < \infty$, which is the labor distortion mentioned above. It is assumed to be increasing and convex in firm's size ($\tau_i', \tau_i'' > 0$). Regardless of how one motivates the existence of this labor distortion, the central assumption is that informality costs are increasing in firm' size, which is a common assumption in the literature [e.g. Fortin et al. (1997), de Paula and Scheinkman (2011) and Leal-Ordonez (2013)].

Formal incumbents must comply with taxes and regulations, but they can hire informal workers to avoid the costs implied by the labor legislation. Even though workers are homogeneous, hiring costs of formal and informal workers differ due to institutional reasons: formal firms have to pay a constant payroll tax on formal workers, while they face an increasing and convex expected cost to hire informal workers, which is summarized by the function $\tau_{fi}(\cdot)$, $\tau_{fi}', \tau_{fi}'' > 0$ and $1 \leq \tau_{fi}(\cdot) < \infty$. The cost for formal firms of hiring informal workers is thus given by $\tau_{fi}(\ell) w\ell$, while the cost of hiring formally is $(1+\tau_w) w\ell$, where τ_w is the labor tax. Since formal and informal workers are perfect substitutes, on the margin firms hire the cheapest one, and hence there is an unique threshold $\tilde{\ell}$ above which formal firms only hire formal workers (on the margin).[17] Formal firms' profit function can be written as follows:

$$\Pi_f(\theta, w) = \max_{\ell} \{(1-\tau_y)\theta q(\ell) - C(\ell)\} \tag{2}$$

[16] This general cost function $\tau_i(\cdot)$ can be directly obtained from a formulation that explicitly accounts for a detection probability. The appendix D.1 shows this correspondence.

[17] The marginal cost of hiring informal workers is strictly increasing $\left[C_{fi}'(\ell) = w\left(\tau_{fi}'(\ell)\ell + \tau_{fi}(\ell)\right)\right]$ and the marginal cost of hiring formal workers is constant $\left[C_{ff}' = (1+\tau_w) w\right]$. Hence, there is an unique value of ℓ such that $C_{fi}'(\tilde{\ell}) = C_{ff}'$. If the labor quantity that maximizes formal firm's profit is such that $\ell^* \leq \tilde{\ell}$, then the formal firm will only hire informal workers. If $\ell^* > \tilde{\ell}$, then the firm will hire $\tilde{\ell}$ informal workers and $\ell^* - \tilde{\ell}$ formal workers.

where τ_y denotes the revenue tax and

$$C(\ell) = \begin{cases} \tau_{fi}(\ell)\, w\ell, & \text{for } \ell \leq \tilde{\ell} \\ \tau_{fi}\left(\tilde{\ell}\right) w\tilde{\ell} + (1+\tau_w)\, w\left(\ell - \tilde{\ell}\right), & \text{for } \ell > \tilde{\ell} \end{cases} \tag{3}$$

The two margins of informality thus introduce a size-dependent distortion in the economy, as lower productivity firms face *de facto* lower marginal costs.[18] Incumbents in both sectors must pay a per-period, fixed cost of operation, which is denoted by $\bar{c}_s$, $s = i, f$. This a standard formulation in the literature and can be interpreted as the opportunity cost of operating in sector s. The profit function net of this fixed cost of operation is denoted by $\pi_s(\theta, w) = \Pi_s(\theta, w) - \bar{c}_s$.

Entry

Every period there is a large mass of potential entrants of size M. Potential entrants only observe a pre-entry productivity parameter, $\nu \sim G$,[19] which can be interpreted as a noisy signal of their effective productivity. G is the same for all firms and independent across periods (i.e. ν is i.i.d.). Hence, the mass of entrants in one period does not affect the composition of potential entrants in the following period. To enter either sector, firms must pay a fixed cost that is assumed to be higher in the formal sector: $c_f >> c_i$.[20] After entry occurs, firms draw their actual productivity from the conditional c.d.f. $F(\theta|\nu)$, which is the same in both sectors and independent across firms. $F(\theta|\nu)$ is assumed to be continuous in θ and ν, and strictly decreasing in ν. Hence, a higher ν implies a higher probability of a good productivity draw after entry occurs.

Because firms face uncertainty about their effective post-entry productivity, the model allows for the possibility of overlap between formal and informal productivity distributions. Even though the sorting behavior in the present model is commonly found in the literature, the fully static models without uncertainty imply perfect sorting and no overlap between formal and informal firms' productivity and size distributions,[21] which is at odds with the data (as shown in Section 2).

If firms are surprised with a low productivity draw $\theta < \bar{\theta}$, such that $\pi_s\left(\bar{\theta}, w\right) =$

[18] Size-dependent frictions have been increasingly analyzed in the literature, see for example Guner et al. (2008), Garicano et al. (2013) and Adamopoulos and Restuccia (2013).

[19] G is assumed to be absolutely continuous with support $(0, \infty)$ and finite moments.

[20] The difference between these entry costs is interpreted here as a consequence of the regulation of entry into the formal sector (e.g. red tape bureaucracy). Under this interpretation, the entry cost into the informal sector can be seen as the initial investment or minimum scale required to operate in the given industry.

[21] See, for example, Rauch (1991), Fortin et al. (1997), de Paula and Scheinkman (2011), Prado (2011), and Galiani and Weinschelbaum (2012).

0, they decide to exit immediately without producing. If firms decide to stay, their productivity remains constant forever and they face a constant exit probability denoted by κ_s, $s = i, f$.[22] Note that this exit probability could also be interpreted as sector-specific discount rates, which could reflect for example differential borrowing rates. Aggregate prices remain constant in steady state equilibria and since firms' productivity also remains constant, firm's value function assumes a very simple form:

$$V_s(\theta, w) = \max\left\{0, \frac{\pi_s(\theta, w)}{\kappa_s}\right\}$$

where for notational simplicity I assume that the discount rate is incorporated in the exit rate. The expected value of entry for a firm with pre-entry signal ν is thus given by

$$V_s^e(\nu, w) = \int V_s(\theta, w)\, dF(\theta|\nu), \qquad s = i, f \tag{4}$$

Entry into the formal sector occurs if $V_f^e(\nu, w) - c_f \geq V_i^e(\nu, w) - c_i$, while entry into the informal sector occurs if $V_i^e(\nu, w) - c_i > \max\left\{V_f^e(\nu, w) - c_f, 0\right\}$. If entry in both sectors is positive the following entry-conditions hold:

$$\begin{aligned} V_i^e(\overline{\nu}_i, w) &= c_i \\ V_f^e(\overline{\nu}_f, w) &= V_i^e(\overline{\nu}_f, w) - (c_i - c_f) \end{aligned}$$

where $\overline{\nu}_s$ is the pre-entry productivity of the last firm to enter sector $s = i, f$. The appendix D.2 show the effective, post-entry productivity distributions in both sectors can be derived as functions of these thresholds.

Demand

Finally, the demand side of the model is kept extremely simple. It is characterized by a representative household that inelastically supplies $\overline{L}$ units of labor and that derives utility solely from consuming the final good:

$$U = \sum_{t=0}^{\infty} \beta^t u(c_t)$$

[22] The assumption of constant post-entry productivity is a reasonable approximation to firms' behavior as long as the post-entry productivity process shows a large degree of persistence. In the Appendix F I use the panel of formal firms to present evidence that indicates that the productivity process is indeed persistent in Brazil. Additionally, La Porta and Shleifer (2008) show indirect evidence that only a small fraction of firms formalize their business after entry occurs.

3.2 Equilibrium

The focus lies on stationary equilibria, where all aggregate variables remain constant. In particular, the size of the formal and informal sectors must remain constant over time, which implies the following flow condition:

$$\mu_s = \frac{1 - F_{\theta_s}\left(\overline{\theta}_s\right)}{\kappa_s} M_s \tag{5}$$

where μ_s denotes the mass of active firms in sector s, $M_i = \left[G\left(\overline{\nu}_f\right) - G\left(\overline{\nu}_i\right)\right] M$ and $M_f = \left[1 - G\left(\overline{\nu}_f\right)\right] M$ denote the mass of entrants into the informal and formal sectors, respectively. In words, the above condition simply states that the mass of successful entrants in both sectors must be equal to the mass of incumbents that exit.

As for the consumer problem, in a stationary equilibrium with constant prices it simplifies to a very simple static optimization problem:

$$\max_c u\left(c\right) \qquad \text{s.t.} \quad c \leq w\overline{L} + \Pi + T \tag{6}$$

where Π is profits and T tax revenues that the government transfers back to households, while entry costs are assumed to be dissipated.

Since consumers do not derive any disutility from work and cannot save, they simply consume all their income. Hence, the natural measure of welfare in this model is total consumption, which in equilibrium is given by $C = w\overline{L} + \Pi + T$

That said, the equilibrium conditions are given by the following: (i) markets clear, $L_i + L_f = \overline{L}$; (ii) The zero profit cutoff (ZPC) condition holds in both sectors, $\pi_s\left(\overline{\theta}_s, w\right) = 0$; (iii) the free entry condition holds in both sectors, with equality if $M_s > 0$; and (iv) the equilibrium flow condition holds in both sectors (expression 5). These equilibrium conditions are straightforward and need no further discussion. In the appendix D.3 I show that the equilibrium exists and it is unique.

4 A Taxonomy of Informal Firms

In this section I propose a simple taxonomy of informal firms based on the three main competing views of their role in economic development [see La Porta and Shleifer (2008) for a discussion]. The starting point of the analysis is to establish a precise definition of each type, which comes directly from these views:

- Type 1: Informal firms that are too unproductive to ever become formal, even if entry costs were removed. These are entrepreneurs with low human capital, who are

only able to survive in the informal sector because they avoid taxes and regulations.

- Type 2: Informal firms that are productive enough to survive as formal firms once entry barriers are removed, but *choose* not to do it because it is more profitable for them to remain informal.

- Type 3: Higher productivity informal firms that are kept out of formality by high entry costs. If these were removed, they would become formal and improve their performance, as they would no longer have the size constraints imposed by the informality status.

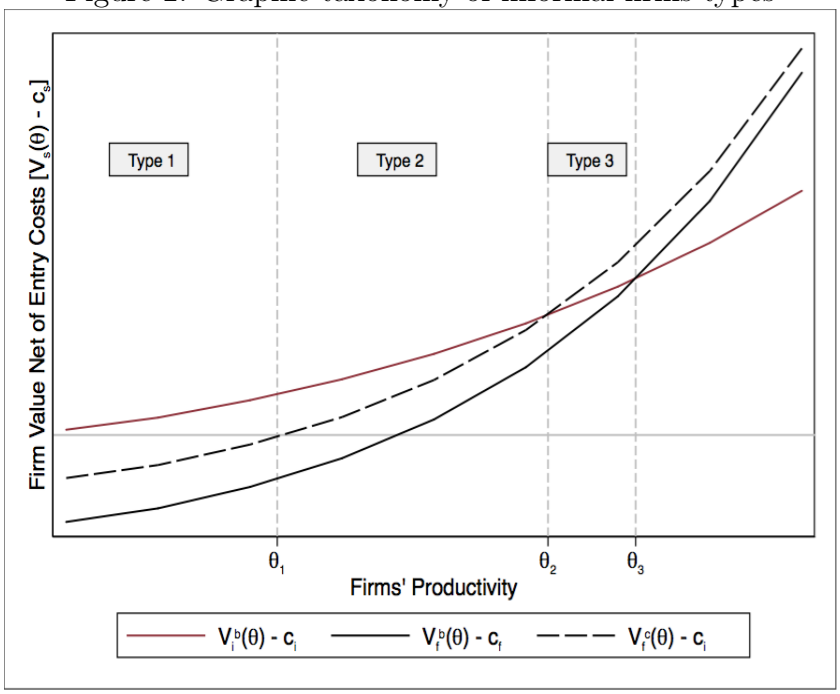
The first crucial difference between these types is how they would respond to a policy that eliminates entry costs into the formal sector. Type 3 firms would formalize their business and would be better off in this counterfactual scenario, as they are no longer constrained by the growth limitations imposed by informality. Hence, any model that does not account for entry costs into the formal sector cannot account for this view, as there would be no bunching of informal firms near the transition threshold. However, the other two types are no so easily distinguishable, as both are predicted to remain informal in the absence of entry costs. Any model that has firms sorting between sectors, even without entry costs and productivity uncertainty, would be able to account for these two types. The crucial differentiation between Types 1 and 2 is the reason why they remain informal. Type 2 firms are productive enough to survive in the formal sector (once entry barriers are removed), but choose to remain informal to receive higher profits. Type 1 are simply not productive enough to be formal, and are only able to survive due to the cost advantages of non-compliance.

Given this reasoning, it is straightforward to represent the different types using the model just discussed. The thought experiment is to ask how the different informal firms would respond to an intervention that equalizes entry costs in the formal and informal sectors ($c_f = c_i$). To disentangle types 1 and 2 it is necessary to ask a somewhat harder question: Which firms could actually become formal in the counterfactual scenario but choose not to do it? Figure 2 summarizes this thought experiment. For each productivity level (θ), it plots firm's baseline life time value net of entry costs in both sectors, $V_s^b(\theta) - c_s$, and the value of being formal under the counterfactual scenario where formal and informal sectors' entry costs are equalized, $V_f^c(\theta) - c_i$.[23]

The baseline curves for the formal and informal sectors intersect each other at $\theta = \theta_3$, and all firms with $\theta \geq \theta_3$ will always choose to be formal, as their life time value is

[23]To obtain these curves it is necessary to specialize the model to specific functional forms and parameter values. I postpone this discussion to the following section, where I present the estimation procedure.

Figure 2: Graphic taxonomy of informal firms types



Note: The figure shows, for each productivity level, firms' baseline life time value net of entry costs in the formal and informal sectors, $V_f^b(\theta) - c_f$ and $(V_i^b(\theta) - c_i)$, respectively. The third curve displays the net value of being formal in a scenario in the counterfactual scenario where formal sector's entry costs are eliminated, so that $c_f = c_i$: $V_f^c(\theta) - c_i$.

higher than being informal. Firms with productivity $\theta \in [\theta_2, \theta_3)$ are the Type 3 firms: once entry barriers into the formal sector are removed, they migrate to the formal sector, improve their performance and achieve higher life-time profits. Firms with productivity $\theta \in [\theta_1, \theta_2)$ correspond to Type 2 firms: They are productive enough to produce in the formal sector – their life-time value in the formal sector is everywhere above zero – but choose not to do it to obtain higher returns in the informal sector. Finally, firms with $\theta < \theta_1$ are the Type 1 firms, which are not productive enough to go to the formal sector even without entry costs, and use informality as a survival strategy.

Figure 2 thus shows that the different views are not necessarily competing, as the informal sector may contain all types of firms. The crucial question is therefore to infer the relative importance of each view in the data. For that, it is necessary to estimate the model and use it to back out the mass of firms in each of the θ intervals just described. The following section describes the estimation procedure, while Section 6 describes the quantitative results.

5 Estimation

The model presented in Section 3 describes firms' decisions regarding entry, production and compliance with regulations in an equilibrium setting. To perform counterfactual analysis and make quantitative statements about firms types and how they would respond to policy changes, it is necessary to estimate all objects in the model's structure. I estimate the model using a two-stage simulated method of moments (SMM). This approach combines direct estimation and calibration from micro and macro data in the first stage, with the SMM estimator itself in the second stage.[24]

To proceed with the estimation, it is first necessary to complete the model's parameterization and assume functional forms for the different objects in the model.[25] These are extra-theoretic assumptions, which naturally raise identification concerns. In the Appendix E.1 I argue that non-parametric estimation of the different objects in the model is either not feasible (given the goals of this paper and the data available), or the assumptions needed are not attainable. The fully parametric approach adopted here is crucial in order to overcome some data limitations and to provide a rich counterfactual analysis.

5.1 Parameterization

Up to this point, the initial productivity distribution, G_ν, the productivity process, $F(\theta|\nu)$, the production function, $q(\cdot)$, and the cost functions, $\tau_s(\cdot)$, were left unspecified. This section completes the model's parameterization by assuming specific functional forms for these objects. Starting with the pre-entry productivity distribution, it is assumed that it has a Pareto distribution:[26]

$$F_\nu\left(\nu \geq x\right) = \begin{cases} \left(\frac{\nu_0}{x}\right)^\xi & \text{for } x \geq \nu_0 \\ 1 & \text{for } x < \nu_0 \end{cases} \tag{7}$$

Firms' actual productivity is only determined after entry occurs. I assume a very simple log-additive form for the post-entry productivity process, which is determined as follows: $\theta = \varepsilon\nu$, where the unexpected shock ϵ is i.i.d. and has a log-normal distribution

[24] For examples of this two-stage approach, see for example Gourinchas and Parker (2002) and Cosar et al. (2010).

[25] It is not always the case that one needs to identify and parameterize all the objects in the model's structure in order to answer specific policy questions [see Marschak (1953) for an early discussion and more recently Heckman (2001) and Ichimura and Taber (2002)]. However, *ex ante* policy evaluations typically require the full specification of a behavioral model in order to perform the counterfactual analysis [e.g. Keane et al. (2011)].

[26] A well documented fact in the literature is that the Pareto distribution fits firms' size distribution remarkably well. This empirical regularity was first documented by see Simon and Bonini (1958), and more recently by Luttmer (2007), among others.

with mean zero and variance σ^2. As for production, I assume that firms use a Cobb-Douglas technology: $y(\theta, l) = \theta l^{\alpha}$, $\alpha < 1$.

The cost functions of both margins take a very simple functional form: $\tau_s(\ell) = \left(1 + \frac{\ell}{b_s}\right)$, where $b_s > 0$ and $s = i, f$. Finally, that the per-period, fixed costs of operation are a function of the equilibrium wage, which makes the exit margin more meaningful since it now responds to market conditions. The fixed costs are determined as follows: $\bar{c}_s = \gamma_s w$, $0 < \gamma_s \leq 1$. Note that I do not impose that these costs are different, this will be determined in the estimation procedure.

The parameter vector to be estimated is thus given by

$$\Gamma = \{\tau, \kappa_f, \kappa_i, \alpha, \sigma, \gamma_f, \gamma_i, b_i, b_f, \xi, \nu_0, C_f, C_i\}$$

which is partitioned into two sub-vectors, $\Gamma = \{\psi, \varphi\}$, the first and second stage parameters, respectively. The MSM estimator used in the second stage takes the parameters in ψ as given in order estimate the vector φ. The next subsection describes the estimation steps.

5.2 Fitting the model to the data

The vector of parameters determined in the first stage is given by $\psi = \{\tau, \kappa_f, \nu_0, \alpha, \sigma, \gamma_f\}$. The tax rates are set to their statutory values: $\tau_w = 0.375$ and $\tau_y = 0.293$.[27] The α is set at the value of the labor share in the economy, which is computed using total labor income from the 2003 Brazilian Household Expenditure Survey (POF).[28] This gives a labor share of 66.4%, which is within the 65-80% range found by Gollin (2002) in his cross-country analysis. The exit probability in the formal sector is set to $\kappa_f = 0.129$ that is estimated using the panel structure in the RAIS data set. This estimate is obtained using the predicted exit probability for the average firm in the sample. As for the remaining parameters, the Pareto distribution scale parameter (ν_0) is set so that the minimum size of a potential entrant in the informal sector is one employee. The standard error of the post-entry shock ($\sigma^2 = 0.15$) and the formal sector's fixed cost of operation ($\gamma_f = 0.5$) are calibrated using as reference different moments of the size distributions in both sectors.

[27] The value of τ_w corresponds to the main payroll taxes, namely, employer's social security contribution (20%), direct payroll tax (9%), and severance contributions (FGTS), 8.5%. The value of τ_y includes two VAT-like taxes: the IPI (20%) and PIS/COFINS (9.25%). These values can be easily obtained in the compilation by the World Bank's Doing Business initiative (www.doingbusiness.or).

[28] In principle, this share could be readily computed from the National Accounts, but as shown in Gollin (2002) this approach can lead to a substantial underestimation, specially in countries where self-employment represents a large share of the working force (as it is the case of Brazil). Indeed, Barros et al. (2007) show that total labor income in the POF is 33% higher than in the National Accounts.

Second stage: The SMM estimator

For a given parameter vector (Γ), wage (w) and individual productivity shocks (ν_j and ε_j), I can completely characterize firms' behavior. The SMM estimator proceeds by using the model to generate simulated data sets of formal and informal firms and computing a set of moments that are also computed from real data. The estimate is obtained as the parameter vector that best approximates the moments computed from the simulated data to the ones computed from real data.[29]

Let $\hat{m}$ denote the vector of moments computed from data, and let $m^s(\varphi;\psi)$ denote the vector of the same moments computed from the simulated data. Define $g(\varphi;\psi) = \hat{m} - m^s(\varphi;\psi)$; the MSM estimation is based on the moment condition $E[g(\varphi_0;\psi_0)] = 0$, where φ_0 and ψ_0 denote the true values of φ and ψ, respectively. The second-stage, MSM estimator is then given by

$$\hat{\varphi} \quad = \quad \arg\min_{\varphi} Q(\varphi;\psi) = \left\{ g(\varphi;\psi)' \, \hat{\mathbf{W}} \, g(\varphi;\psi) \right\} \tag{8}$$

where $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, and $\mathbf{W}$ is a positive semi-definite weighting matrix. Under the suitable regularity conditions (which are discussed in the appendix), the MSM estimator is consistent and asymptotically normal. The Appendix E.2 describes the derivation of the asymptotic variance-covariance matrix and the computation of the optimal $\hat{\mathbf{W}}$, while the appendix E.3 describes the optimization algorithm.

Moments and identification

There are seven parameters to be estimated in the second stage: $\{\kappa_i, \gamma_i, b_i, b_f, \xi, C_f, C_i\}$. I use 18 moments from the data to form the vector $\hat{m}$, which are the following: (i) share of informal workers; (ii) overall share of informal firms and by firms' size, with $n = 1, ..., 5$ (where n is the number of employees); (iii) average share of informal workers in formal firms with size $n = 2, ..., 5$; (iv) the 75^{th}, 95^{th} and 99^{th} percentiles of informal firms' size distribution; and (v) the 25^{th}, 50^{th}, 75^{th} and 95^{th} percentiles formal firms' size distribution.

Even though simulation-based methods typically do not allow for a formal identification arguments, it is possible to discuss the role that some moments play in the identification of given parameters [for recent examples, see Cosar et al. (2010) and Dix-Carneiro (2010)]. Starting with the shape parameter of the Pareto distribution, ξ, it is completely determined by firm size distribution, specially of formal firms, which is very

[29] The technical details are discussed in the Appendix E.2. The interested reader can find an in-depth and systematized discussion in Gourieroux and Monfort (1996) and Adda and Cooper (2003).

precisely estimated and thus highly weighted by the matrix $\hat{\mathbf{W}}$ (see appendix). In fact, given the one-to-one relationship between productivity and firms' size in the model, the estimation of the productivity distribution crucially relies on observed moments of firm size distribution (measured by number of employees).

As for the parameters that govern the cost functions of both margins of informality, the share of informal firms by firms size plays a crucial role in identifying b_i, while the share of informal workers in formal firms by firms' size identifies b_f. Informal sector's exit (or discount) rate κ_i determines the overall disadvantage of being informal relatively to being formal, as the difference between κ_i and κ_f represents an overall downward shift in life time value for informal firms. Thus, given formal sector's entry cost, the κ_i is disciplined by the overall share of informal firms and workers. Given σ and γ_f (determined in the first stage), formal sector's entry cost, c_f, is largely determined by the degree of overlap between formal and informal firm size distributions, the minimum size of entrants in the formal sector and how shifted to the right firm size distribution is relatively to informal sector's. Similarly, informal sector's entry cost is likely to be determined by entrants' minimum scale, which comes from the lower percentiles of firm size distribution in the informal sector.

Estimates and Model Fit

Table 3 shows the values of both first and second stage parameters, ψ and φ respectively. The estimates show that formal sector's entry cost is more than twice informal sector's. Exit (discount) rate in the informal sector is also more than twice as high as formal sector's, which confirms the anecdotal knowledge that informal firms have higher turnover rates than their formal counterparts. The estimates for b_f and b_i show that informal firms have more room for growth than informal employment in the formal sector. This also agrees with the evidence presented by Almeida and Carneiro (2009) regarding government's inspection technology. The authors argue that the government is much more likely to inspect formal firms simply because they are more visible than informal ones. Finally, the estimate of the Pareto's shape parameter, ξ, indicates a skewed (to the left) productivity distribution, which is indeed observed in the data for firm size.

As for the fit of the model, Table 4 shows how the model performs compared to the observed moments in the data. The model matches the share of informal firms and the share of informal workers well. However, it highly understates the share of informal firms with only one employee (the 75% percentile of the actual size distribution in the informal sector) and consequently overstates the share of firms with at most two employees. The same does not happen with the size distribution in the formal sector, which the model is

Table 3: Parameter Values

Parameter	Source	Value	Std. Errors
First Stage			
τ_w	Statutory values	0.375	–
τ_y	Statutory values	0.293	–
α	Labor share	0.664	–
κ_f	Estimated (Micro data)	0.129	–
ν_0	Calibrated	7.7	–
σ	Calibrated	0.15	–
γ_f	Calibrated	0.5	–
Second Stage (SMM)			
b_f	Estimated	3.724	0.353
b_i	Estimated	4.197	0.043
κ_i	Estimated	0.308	0.005
γ_i	Estimated	0.258	0.024
ξ	Estimated	4.205	0.030
c_f †	Estimated	5,298.2	259.3
c_i †	Estimated	2,564.4	62.8

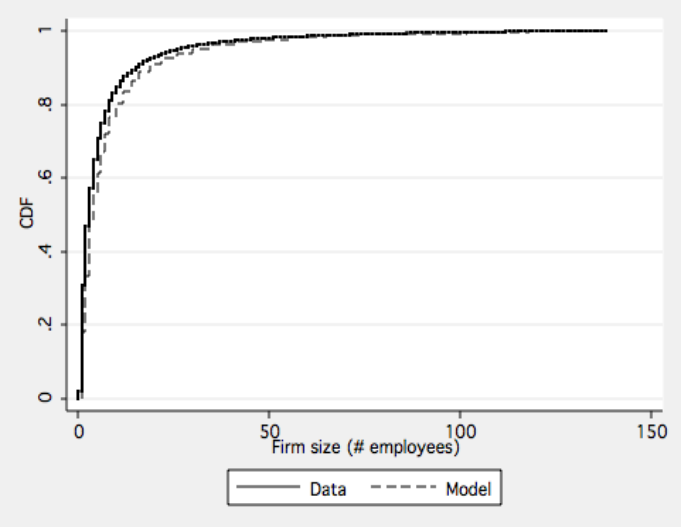
† Estimates and SD expressed in R$ of 2003. The point estimates correspond to R$8,441.2 and R$4,085.7 in 2012 values, which roughly correspond to 4,020 and 1,945 US dollars, respectively.

Table 4: Model fit

	Model	Data
Share inf. workers	0.360	0.354
Share inf. Firms	0.683	0.686
Size distribution: Informal sector		
≤ 1 *employee*	0.314	0.849
≤ 2 *employees*	0.906	0.958
≤ 4 *employees*	0.990	0.993
Size distribution: Formal sector		
≤ 1 *employee*	0.302	0.295
≤ 3 *employees*	0.545	0.563
≤ 7 *employees*	0.786	0.774
≤ 31 *employees*	0.963	0.953

able to replicate well. Figure 3 shows the same information for the entire size distribution in the formal sector.

Figure 3: Model fit: Formal sector employment distribution



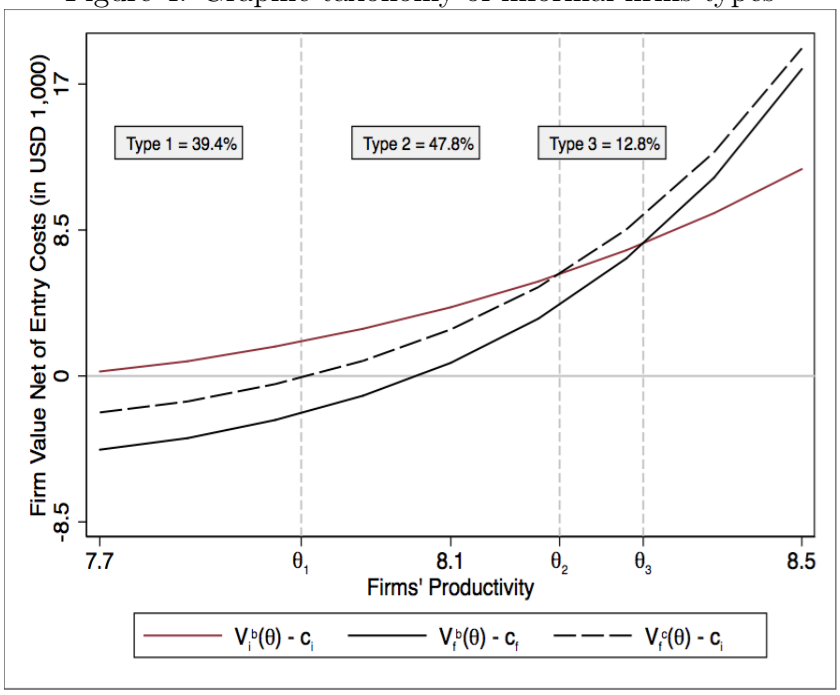
Note: The data source is a 25% random sample of the RAIS data set, which contains the universe of formal firms and their formal workers. The graph plots the CDF of formal firms size distribution (measured in numbered of employees) from both the real and simulated data sets.

6 Quantitative results

6.1 The distribution of informal firms' types in the data

In this section I use the estimation results and the taxonomy discussed in Section 4 to back out from the data the distribution of informal firms types. For that, I use the estimated model to obtain, for each firm θ, the baseline net life time values of being formal and informal that are implied by the data: $V_f^b(\theta) - c_f$ and $V_i^b(\theta) - c_i$, respectively. I then simulate the counterfactual scenario where entry costs into the formal sector are equalized to informal sector's ($c_f = c_i$), and compute for all firms the counterfactual value of being formal once entry costs are removed, which gives me the curve $V_f^c(\theta) - c_i$. Figure 4 revisits Figure 2 by displaying the corresponding empirical curves, which are drawn from the data using the estimated model.

Figure 4: Graphic taxonomy of informal firms types



Note: The figure shows, for each productivity level, firms' life time value of net of entry costs in the formal and informal sectors, $V_f^b(\theta) - c_f$ and $(V_i^b(\theta) - c_i)$, respectively. The third curve displays the net value of being formal in a scenario where the bureaucratic costs of entry into the formal sector are eliminated, so that $c_f = c_i$: $V_f^c(\theta) - c_i$.

The relative sizes are obtained by computing the mass of firms within each of the three intervals. This computation depends on the effective productivity distribution in the informal sector, which is determined by three main elements: (i) the underlying pre-entry productivity distribution F_ν; (ii) the determinants of firms' sorting between sectors; and (iii) the selection mechanism after entry occurs, which selects out the least

productive firms. The Appendix D.2 contains the derivation of post-entry productivity distributions. As discussed in the previous section, elements (i) and (iii) are pinned down by firm size distributions (measured as number of employees) in both sectors, and the degree of overlap between them.[30] Element (ii) is determined by the interplay between the pre-entry productivity distribution, entry costs, and the institutional factors that determine expected profitability in both sectors, such as taxes and the costs of both margins of informality.

The data indicates that the potentially productive informal firms that are kept out of formality by high entry costs (Type 3) are the minority, corresponding to less than 13% of all informal firms. Type 2 firms, those that could survive as formal firms once entry costs are removed but choose to remain informal to enjoy the cost advantages of non-compliance correspond to the largest group, 47.8% of all informal firms. The remaining firms, 39.4%, correspond to Type 1, which are too unproductive to survive in the formal sector. Hence, these results indicate that there is limited scope for policies that reduce formal sector's entry costs to promote substantial formalization. Indeed, the available empirical evidence on the effects of this type of intervention is at best mixed, with some studies indicating very limited impacts on formalization [e.g. Kaplan et al. (2011)]. In the following section I examine policies impacts in a general equilibrium context.

6.2 Policy Impacts: Reducing regulatory costs *versus* increasing enforcement

In this section I analyze two different policies that seek to reduce regulatory costs: (i) eliminating formal sector's bureaucratic entry costs (equalizing both sectors' entry costs); and (ii) a 20 p.p. cut in the payroll tax, which corresponds to eliminating social security contribution. For greater enforcement, I consider the effects of increasing the costs of the extensive and intensive margins separately, which could be achieved through greater monitoring efforts by the government. In the model, this translates into lower values of the parameter b_s, $s = i, f$.

In what follows I analyze the effects at the firm and aggregate levels separately.

[30] Since it is not possible to directly estimate productivity from the data, I use the model to infer it from firm size distributions. If there was a panel of formal and informal firms with information on production and inputs, I could directly estimate the production function and thus firm-level productivity.

Firm level effects: Heterogeneity, LATE and the "causal" impact of formalization

Part of the existing empirical studies aim at identifying the effects of a given policy on formalization *per se* and other outcomes such as firms' employment [e.g. Kaplan et al. (2011) and Almeida and Carneiro (2012)]. One potential difficulty in this case is that these are large-scale interventions and are thus likely to have substantial general equilibrium effects, which are not accounted for in the available partial equilibrium estimates. Another stream of studies uses policy variations to try to identify the *causal* effect of formality on firms' outcomes [e.g. McKenzie and Sakho (2010), Fajnzylber et al. (2011) and de Mel et al. (2013)].[31] However, as discussed by Arias et al. (2010): "self-selection in static models implies that the observed differences in productivity between formal and informal firms are the result of a selection process, leaving no place for a causal impact of the formal status on productivity or firms".[32]

To shed some light on these issues, I compute the equivalent of the LATE parameter, which is given by $E\left[\Delta^{mte}(\theta)\middle|D_c(\theta)=1, D_b(\theta)=0\right]$, where $D_j(\theta)$ equals one if the firm is formal in the baseline ($j=b$) or counterfactual ($j=c$) scenarios. The $\Delta^{mte}(\theta)$ denotes the marginal treatment effect (MTE) for firm with productivity θ [e.g. Heckman and Vytlacil (2005)]. The outcome variable is firm's life time value, and thus for the compliers – informal firms that go to the formal sector as the result of a given policy – the MTE is given by $\Delta^{MTE}(\theta) \equiv \log\left(V_f^c(\theta)-\tilde{c}\right) - \log\left(V_i^b(\theta)\right)$, where $\tilde{c}=c_f-c_i$.[33] To grasp the importance of general equilibrium effects, I compute the LATE parameter allowing for both general and partial equilibrium effects. The latter are computed holding wages constant, analogously to Heckman et al. (1998). Table 5 shows the results.

As one could expect, there are no compliers under the intervention that increases enforcement on the intensive margin. This policy increases the costs of formality for small firms and thus one should not expect firms switching to the formal sector once it is implemented. However, for the other three interventions, Table 5 shows that the partial equilibrium estimates overestimate the full impact that takes into account general equilibrium effects. Take the experiments that reduce entry costs and payroll taxes, which are the ones usually considered in the literature [e.g. Monteiro and Assuncao (2011) and

[31] The regression of interest is usually of the form $Y_i = \gamma D_i + X_i'\beta + \varepsilon_i$, where Y_i is the outcome of interest (e.g. firm's revenue or profits), D_i is a dummy that equals one if the firm is formal and X_i is a vector of covariates. Exogenous variations in access to a given formalization policy would provide an instrument for D_i, potentially allowing to achieve identification.

[32] This is of course no longer true in a dynamic setting where the productivity processes in both sectors is different due to fundamental differences in technology [e.g. D'Erasmo and Boedo (2012)].

[33] For firms that remain in the same sector s in the baseline and counterfactual scenarios we have that $\Delta^{MTE}(\theta) \equiv \log\left(V_s^c(\theta)\right) - \log\left(V_s^b(\theta)\right)$.

Table 5: LATE – General vs. Partial Equilibrium Effects

	Partial Equilibrium	General Equilibrium
Reducing regulatory costs		
Entry Costs	0.236	0.180
Payroll Tax	0.294	0.079
Increasing Enforcement		
Extensive Mg.	-0.210	-0.250
Intensive Mg.	–	–

Note: For each counterfactual scenario c, the LATE parameter is computed as $E\left[\Delta^{mte}(\theta)\middle|D_c(\theta)=1, D_b(\theta)=0\right]$, where $D_j(\theta)$ is an indicator function that equals one if firm θ is formal and zero if informal in the baseline ($j=b$) or counterfactual ($j=c$) scenarios.

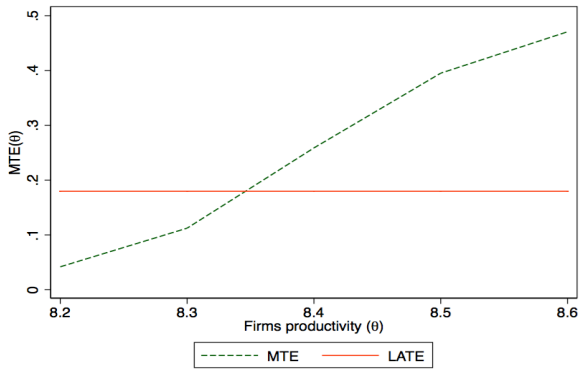
de Mel et al. (2013)]. These policies generate wage increases – and particularly strong in the payroll experiment – which partially undo some of the benefits perceived by the firms that formalize when these policies are implemented. Another interesting result is the difference in the "causal" effect of informality under policies that reduce regulatory costs and under the intervention that increases enforcement in the extensive margin. They are of similar magnitude but with opposite signs. This is not surprising, as in the latter firms are being pushed to formalize, although they had already revealed that it is not optimal for them to do so with the existing regulatory framework. This result thus reinforces the above argument that it is hard to interpret as causal the effect from changes in formality status that are policy-induced.

Looking only at the group of compliers, Figure 5 plots the MTE for each level of firm's productivity and the LATE parameter. It shows that there is substantial degree of heterogeneity even within this restricted group of firms. Examining the distribution of MTEs among all firms, the effect heterogeneity is even more substantial, as shown in Table 6. Perhaps more importantly, the LATE parameter represents the highest percentiles in the distribution of MTEs. Most firms are actually harmed even by policies that reduce regulatory costs. The reason for these negative effects is again the presence of GE effects through higher wages. Take the case of reducing entry costs. Except for firms that make the transition into the formal sector, all other firms are worse off as they are now exposed to higher competition.[34] Nevertheless, when compared to the effects of higher enforcement on the extensive margin, the effects of reducing regulatory costs are much better. The reason is that increasing enforcement on informal firms has a very strong

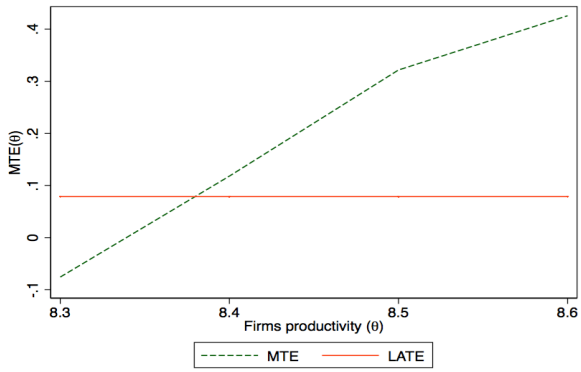
[34]However, the economy as a whole is in better situation, as discussed in the following section.

negative effect on these firms, which constitute the majority of firms, without producing any sizable positive (or negative) general equilibrium effect (from the firms' standpoint).

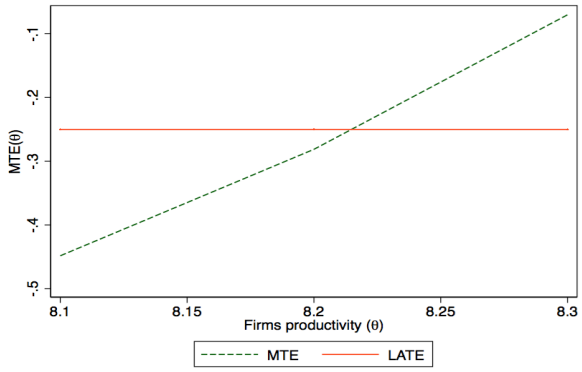
Figure 5: MTE Profiles amongst Compliers



(a) ↓↓ Entry costs



(b) ↓↓ Payroll tax



(c) ↑↑ Enforcement - Extensive Mg.

Table 6: The Distribution of MTEs – All Firms

	Reducing Regulatory Costs		Increasing Enforcement	
	Entry Costs	Payroll	Extensive Mg.	Intensive Mg.
Mean	-0.014	-0.169	-0.502	-0.013
Pctile 25	-0.037	-0.224	-0.678	0.000
Pctile 50	-0.032	-0.202	-0.675	0.000
Pctile 75	-0.028	-0.171	-0.327	0.000
Pctile 95	0.195	0.019	0.000	0.000

Macro Effects: Margins of Informality, Productivity vs. GDP, and Welfare

Reducing formal sector's entry cost leads to a substantial reduction in the share of informal firms, of nearly 18 p.p. (Table 7). The effect on the share of informal workers is however very limited, which highlights the importance of accounting for the intensive margin of informality. Although the share of informal firms decreases substantially, the incentives to hire informal workers in the formal sector remain unaltered. Moreover, the share of formal firms increases by the addition of low productivity firms to the formal sector, which are more likely to hire a large share of their labor force informally. The opposite is true when the payroll tax is reduced: informal employment is more substantially reduced than the share of informal firms, which decreases nearly 5 p.p.. Also interestingly, increasing enforcement on the intensive margin leads to greater informality levels among firms. This is observed because the effective cost of being formal increases for less productive firms, thus increasing their incentives to be informal. Nevertheless, the share of informal workers has a net reduction, which is the result of a substantial decline in the share of informal workers hired by formal firms.

Table 7: Main aggregate outcomes

Reducing Regulatory Costs			
	Baseline	Entry Cost	Payroll tax
Informal workers (share)	0.344	0.327	0.239
Informal firms (share)	0.687	0.510	0.630
GDP	1.000	1.022	1.005
Avg. Log(TFP)	1.000	0.962	1.041
Wages	1.000	1.023	1.151
Welfare	**1.000**	**1.023**	**0.993**
Increasing Enforcement			
	Baseline	Extensive Mg.	Intensive Mg.
Informal workers (share)	0.344	0.160	0.299
Informal firms (share)	0.687	0.302	0.720
GDP	1.000	1.007	0.985
Avg. Log(TFP)	1.000	1.050	1.003
Wages	1.000	1.000	1.000
Welfare	**1.000**	**1.024**	**0.992**

Notes: The variation in average log-TFP is measured as $\exp\left\{\overline{\log\left(TFP\right)}_c - \overline{\log\left(TFP_i\right)}_b\right\}$, where $\exp\overline{\log\left(TFP\right)}_c$ and $\overline{\log\left(TFP\right)}_b$ denote the average log-TFP in the counterfactual and baseline scenarios, respectively.

Reducing entry costs also reduces deadweight losses from wasteful barriers to entry and increases competition, which leads to an increase in wages. The intervention has a negative effect on aggregate TFP because more low-productivity firms enter the economy (there is a 13.6% increase in the mass of active firms). However, because there is higher entry the economy gains on the extensive margin of production and GDP increases. Hence, lower TFP does not necessarily imply lower aggregate production in the context of high informality and high entry costs. Increasing enforcement on the extensive margin is highly effective in reducing both the share of informal firms and workers. Because it drives production in the informal sector very close to zero, there is a large effect on aggregate TFP due to composition effects but with no effect on GDP, due to the displacement of a large number of firms.

The welfare analysis shows that reducing entry costs and increasing enforcement on the extensive margin show almost identical welfare improvements, of 2.3% and 2.4%, respectively. The determinants of their good performance are, however, completely different. In the entry cost intervention, welfare increases because there is a substantial reduction in deadweight losses from wasteful entry costs. Moreover, wages and to a lesser extent tax revenues also increase, which contributes to the result. As for higher enforcement on the extensive margin, the improvement in welfare is almost entirely driven by the substantial increase in tax revenues, which more than compensates the losses of informal and small formal firms. This effect can thus be interpreted as a stylized version of the mechanisms highlighted by the literature on fiscal capacity [e.g. Besley and Persson (2010)]. It is worth highlighting, however, that these results should be seen as an upper bound for the welfare effects from greater enforcement, as the exercises assume that all tax revenue are directly rebated to households, with no resources lost.

Finally, even though the different interventions always manage to reduce at least one measure of informality, they do not always lead to welfare improvements. This is the case of higher enforcement on the intensive margin and lower payroll tax. In the first case, the policy reduces the share of informal workers but negatively affects small formal firms and actually leads to an increase in the share of informal firms. These effects, combined with a slight decline in tax revenues, cause welfare to decrease. As for the lower payroll tax intervention, its general equilibrium effects (i.e. wage increases) undo some its firm-level benefits. Additionally, there is a substantial reduction in tax revenues, which direct and negatively impacts welfare.

7 Final remarks

In this paper I show the empirical and theoretical importance of distinguishing between two margins of informality: (i) when firms do not register and pay entry fees (extensive margin); and (ii) when firms pay workers "off the books" (intensive margin). The latter is a central innovation, as it is empirically important and has direct implications to the understanding of informality. The presence of the intensive margin breaks the direct association between worker and firm informality, and implies that formal and informal are no longer disjoint states for firms, as formal firms may hire part or all of their labor force informally.

The framework developed here integrates the leading views of informality in an unified setting, and provides a natural taxonomy of informal firms based on these views. I take the model to the data on formal and informal firms in Brazil to back out the empirical relevance of these competing views. The results show that firms that are potentially productive and which formalize and succeed when formal sector's entry costs are removed constitute a small fraction of all informal firms (12.3%). The view that argues that informal firms choose informality to exploit the cost advantages of non-compliance even though they are productive enough to survive in the formal sector corresponds to the largest group – 47.8% of all informal firms. The remaining firms correspond to those too unproductive to ever become formal.

This paper also estimates the effects of reducing formal sector's entry costs, reducing payroll taxes, and increasing enforcement to reduce non-compliance in both the extensive and intensive margins. These are large-scale interventions that are likely to produce sizable general equilibrium effects. Indeed, firm-level results indicate that not accounting for general equilibrium effects leads to an overestimation of the LATE parameter. The results also put into question the common interpretation of a causal effect from the formal status, as the magnitude and direction of the effects of formalization fundamentally depend on the policy that generated the change in status.

At the aggregate level, the results highlight the importance of accounting for the intensive margin. Once it is included in the analysis, the aggregate shares of informal firms and workers – the two most common measures of informal sector's size – do not respond in the same way to different policies. The same holds for TFP and GDP, which do not necessarily move in the same direction in response to policy changes. Finally, results from the welfare analysis show that the two best interventions are to reduce entry costs and increase enforcement on informal firms. Both generate sizable welfare gains, of around 2.3%. As for reductions in the payroll tax and increasing enforcement on the intensive margin, both manage to reduce labor informality but they also lead to welfare

losses (of less than 1%). Thus, informality reductions are not necessarily associated to welfare improvements.

References

Abbring, J. H. (2010). Identification of dynamic discrete choice models. *Annual Review of Economics 2*(1), 367–394.

Ackerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007). Econometric tools for analyzing market outcomes. *Handbook of econometrics 6*, 4171–4276.

Adamopoulos, T. and D. Restuccia (2013). The size distribution of farms and international productivity differences. *American Economic Review*. Forthcoming.

Adda, J. and R. Cooper (2003). *Dynamic Economics: Quantitative methods and applications*. The MIT Press.

Aguirregabiria, V. and P. Mira (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics 156*(1), 38–67.

Almeida, R. and P. Carneiro (2009). Enforcement of labor regulation and firm size. *Journal of Comparative Economics 37*(1), 28 – 46.

Almeida, R. and P. Carneiro (2012). Enforcement of labor regulation and informality. *American Economic Journal: Applied Economics 4*(3).

Amaral, P. S. and E. Quintin (2006). A competitive model of the informal sector. *Journal of Monetary Economics 53*(7), 1541–1553.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies 58*(2), pp. 277–297.

Arellano, M. and B. Honore (2001). Panel data models: some recent developments. In J. Heckman and E. Leamer (Eds.), *Handbook of econometrics*, Volume 5 of *Handbook of Econometrics*, pp. 3229 – 3296. Elsevier.

Arias, J., O. Azuara, P. Bernal, J. Heckman, and C. Villarreal (2010). Policies to promote growth and economic efficiency in mexico. NBER Working Paper 16554.

Aruoba, S. B. (2010). Informal sector, government policy and institutions. In *2010 Meeting Papers*, Number 324. Society for Economic Dynamics.

Barros, R., S. Cury, and G. Ulyssea (2007). A desigualdade de renda no brasil encontra-se subestimada? uma analise comparativa usando pnad, pof e contas nacionais. In P. Barros, M. Foguel, and G. Ulyssea (Eds.), *Desigualdade de renda no Brasil: Uma analise da queda recente*. DFID.

Besley, T. and T. Persson (2010). State capacity, conflict, and development. *Econometrica 78*(1), 1–34.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*, 115–143.

Blundell, R. and S. Bond (2000). Gmm estimation with persistent panel data: An application to production functions. *Econometric Reviews 19*.

Bond, S. (2002). Dynamic panel data models: A guide to micro data methods and practice. *IFS Working Paper CWP09/02*.

Bruhn, M. (2011). License to sell: The effect of business registration reform on entrepreneurial activity in mexico. *Review of Economics and Statistics 93*(1), 382–386.

Charlot, O., F. Malherbet, and C. Terra (2011). Product market regulation, firm size, unemployment and informality in developing economies. IZA Discussion Papers No.5519.

Cosar, A., N. Guner, and J. Tybout (2010). Firm dynamics, job turnover, and wage distributions in an open economy. NBER Working Paper 16326.

de Mel, S., D. McKenzie, and C. Woodruff (2013). The demand for, and consequences of, formalization among informal firms in sri lank. *American Economic Journal: Applied Economics*. Forthcoming.

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

de Paula, A. and J. A. Scheinkman (2011). The informal sector: An equilibrium model and some empirical evidence. *Review of Income and Wealt'h 57*, S8–S26.

De Soto, H. (1989). *The Other Path*. Harper e Row, New York.

D'Erasmo, P. and H. Boedo (2012). Financial structure, informality and development. *Journal of Monetary Economics 59*(3), pp. 286–302.

Dix-Carneiro, R. (2010). Trade liberalization and labor market dynamics. Job Market Paper.

Dix-Carneiro, R. (2013). Trade liberalization and labor market dynamics.

Djankov, S., R. L. Porta, F. Lopez-De-Silanes, and A. Shleifer (2002, February). The regulation of entry. *The Quarterly Journal of Economics 117*(1), 1–37.

Donaldson, D. (2013). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*. Forthcoming.

Duffie, D. and K. Singleton (1993). Simulated moments estimation of markov models of asset prices. *Econometrica 61*(4), 929–952.

Eaton, J., S. Kortum, and F. Kramarz (2010). An anatomy of international trade: Evidence from french firms. *Econometrica Forthcoming*.

Fajnzylber, P., W. F. Maloney, and G. V. Montes-Rojas (2011). Does formality improve micro-firm performance? evidence from the brazilian simples program. *Journal of Development Economics 94*(2), 262 – 276.

Farrell, D. (2004). The hidden dangers of the informal economy. *The McKinsey Quarterly* (3).

Fortin, B., N. Marceau, and L. Savard (1997). Taxation, wage controls and the informal sector. *Journal of Public Economics 66*(2), 293–312.

Galiani, S. and F. Weinschelbaum (2012). Modeling informality formally: households and firms. *Economic Inquiry 50*(3), 821–838.

Garicano, L., C. Lelarge, and J. Van Reenen (2013). Firm size distortions and the productivity distribution: Evidence from france. Technical report, NBER Working Paper No. 18133.

Gollin, D. (2002). Getting income shares right. *Journal of Political Economy 110*(2), pp. 458–474.

Gourieroux, C. and A. Monfort (1996). *Simulation-Based Econometric Methods*. Oxford University Press.

Gourinchas, P.-O. and J. Parker (2002). Consumption over the life cycle. *Econometrica 70*(1), 47–89.

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

Heckman, J. and S. Navarro (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*, 1–22.

Heckman, J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrics 73*, 669–738.

Heckman, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy 109*(4), pp. 673–748.

Heckman, J. J., L. Lochner, and C. Taber (1998). General equilibrium treatment effects: A study of tuition policy. *American Economic Review, Papers and Proceedings 88*(2), 381–386.

Ichimura, H. and C. Taber (2002). Semiparametric reduced-form estimation of tuition subsidies. *The American Economic Review 92*(2), pp. 286–292.

Kaplan, D. S., E. Piedra, and E. Seira (2011). Entry regulation and business start-ups: Evidence from mexico. *Journal of Public Economics*.

Keane, M. P., P. E. Todd, and K. Wolpin (2011). The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications. Volume 4A of *Handbook of Labor Economics*, Chapter 04, pp. 331–461. Elsevier.

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

Leal-Ordonez, J. C. (2013). Tax collection, the informal sector, and productivity. *Review of Economic Dynamics In Press*.

Levy, S. (2008). *Good Intentions, Bad Outcomes: Social Policy, Informality, and Economic Growth in Mexico*. Brookings Institution Press.

Luttmer, E. (2007). Selection, growth and the size distribution of firms. *The Quarterly Journal of Economics 122*(3), 1103–1144.

Marschak, J. (1953). Economic measurements for policy predictions. pp. 1–26. New York: John Wiley.

Matzkin, R. L. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*(2), pp. 239–270.

Matzkin, R. L. (2003, 09). Nonparametric estimation of nonadditive random functions. *Econometrica 71*(5), 1339–1375.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*(5), 995–1026.

McKenzie, D. and Y. S. Sakho (2010). Does it pay firms to register for taxes? the impact of formality on firm profitability. *Journal of Development Economics 91*, 15 –24.

Meghir, C., R. Narita, and J. Robin (2012). Wages and informality in developing countries. NBER Working Paper 18347.

Monteiro, J. C. and J. J. Assuncao (2011). Coming out of the shadows? estimating the impact of bureaucracy simplification and tax cut on formality in brazilian microenterprises. *Journal of Development Economics* (0), –.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume 4 of *Handbook of Econometrics*, pp. 2111 – 2245. Elsevier.

Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–1297.

Pakes, A. and D. Polard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*(5), 1027–1057.

Perry, G., W. Maloney, O. Arias, P. Fajnzylber, A. Mason, and J. Saavedra-Chanduvi (2007). *Informality: Exit or exclusion*. The World Bank. Washington, D.C.

Prado, M. (2011). Government policy in the formal and informal sectors. *European Economic Review 55*, 1120–1136.

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

Rust, J. (1994). Structural estimation of markov decision processes. Volume 4 of *Handbook of Econometrics*, pp. 3081 – 3143. Elsevier.

Schneider, F. (2005). Shadow economies around the world: what do we really know? *European Journal of Political Economy 21*(3), 598 – 642.

Simon, H. A. and C. P. Bonini (1958). The size distribution of business firms. *The American Economic Review 48*(4), 607–617.

Taber, C. R. (2000, June). Semiparametric identification and heterogeneity in discrete choice dynamic programming models. *Journal of Econometrics 96*(2).

Tauchen, G. (1986). Finite state markov-chain approximation to univariate and vector autoregressions. *Economic Letters 20*, 177–81.

Ulyssea, G. (2010, January). Regulation of entry, labor market institutions and the informal sector. *Journal of Development Economics 91*, 87–99.

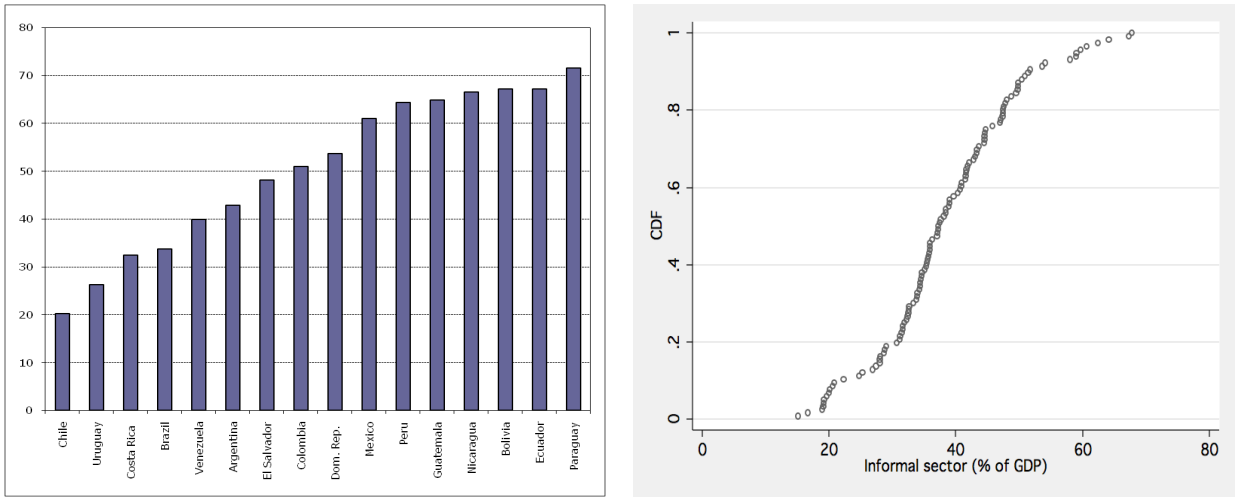
Ulyssea, G. (2012). Formal and informal firms ' dynamics. Unpublished manuscript, University of Chicago, Department of Economics.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step gmm estimators. *Journal of Econometrics 126*(1), 25 – 51.

A Cross country evidence

Figure A.1's left panel displays informal sector's size in Latin American countries, which is measured as the share of employees not covered by social security. [35] The right panel shows the c.d.f. plot of informal sector's size for 116 countries that have a GDP per capita that is less or equal to half of USA's. The size measure used in this graph is informal sector's share of GDP, which comes from La Porta and Shleifer (2008)'s data set that uses Schneider (2005)'s methodology. Figure A.2 uses the same data set to plot the size of the informal sector (as a share of GDP) and the GDP per capita, with a linear fitted line. The cost of regulation data comes from the Doing Business Initiative (*www.doingbusiness.org*).

Figure A.1: Informal sector's size



(a) Labor informality (Latin America)

(b) CDF of size measure (developing countries)

[35] The data come from the Socio-Economic Data Base for Latin America and the Caribbean (SEDLAC), a joint initiative by the World Bank and Universidad Nacional de La Plata (available at *http://sedlac.econo.unlp.edu.ar/esp/*).

Figure A.2: Informal sector's size and GDP per capita

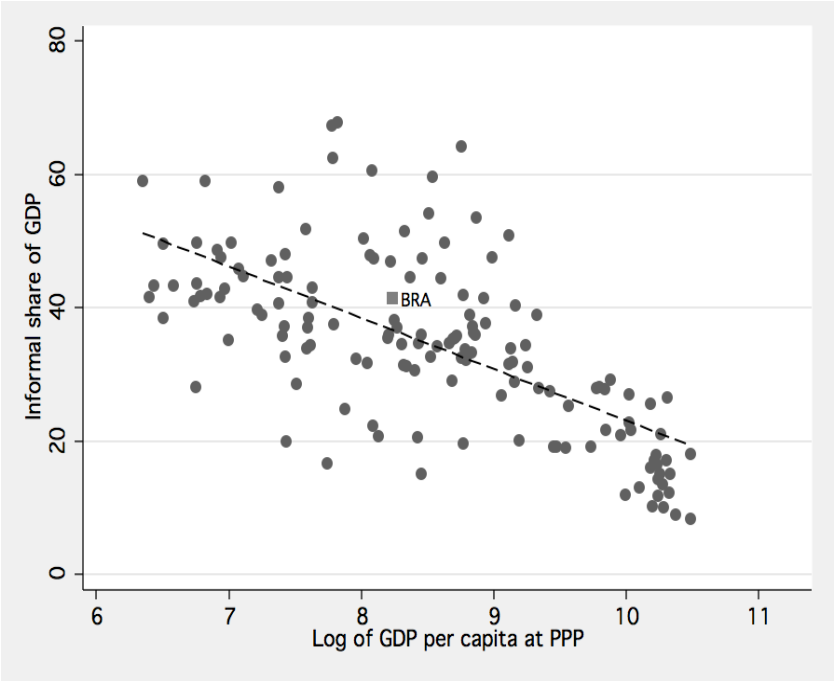


Table 8: Regulatory costs

	Entry Costs		Labor tax (%)
	# Procedures	# Days	
E. Asia & Pacific	7	37	10.7
E. Europe & C. Asia	6	16	21.7
L. A.C.	9	54	14.6
Mid. East & N. Africa	8	20	16.9
OECD high income	5	12	24
South Asia	7	23	7.7
Sub-Saharan Africa	8	37	13.5
Brazil	**13**	**119**	**40.9**

Source: Doing Business Initiative, 2010 (*www.doingbusiness.org*).

B Data appendix

As described in Section 2, the two main data sets used in this paper are the ECINF survey (*Pesquisa de Economia Informal Urbana*) and the *Relacao Anual de Informacoes Sociais* (RAIS), an administrative data set from the Brazilian Ministry of Labor. In this section I simply describe the filters used in both data sets in order to reach the final sample used in the estimations.

In both data sets, I only kept the firms that were in manufacturing, services or com-

merce, dropping all firms in other industries and agriculture. I also dropped firms in the public sector and those that showed a total wage bill equal to zero. Since the RAIS data set contains the universe of formal firms, I took a 25% random sample from the original data set to decrease the computational burden of the moments estimation.

As for the ECINF, some additional filters were applied. Many of the observations regard self employed individuals, street vendors and other economic enterprises that do not correspond to the standard definition of a firm. In order to obtain the most comparable unit of analysis with the formal firms covered by the RAIS data set, I dropped the entrepreneurs who declared to have another job, and who do not have a specific physical location outside their household where their activity takes place. Moreover, in order to avoid outliers, I dropped firms that had an unusually low level of total revenues (lower than the first percentile), and unusually low or high age (less than the first or higher than the 99th percentiles).

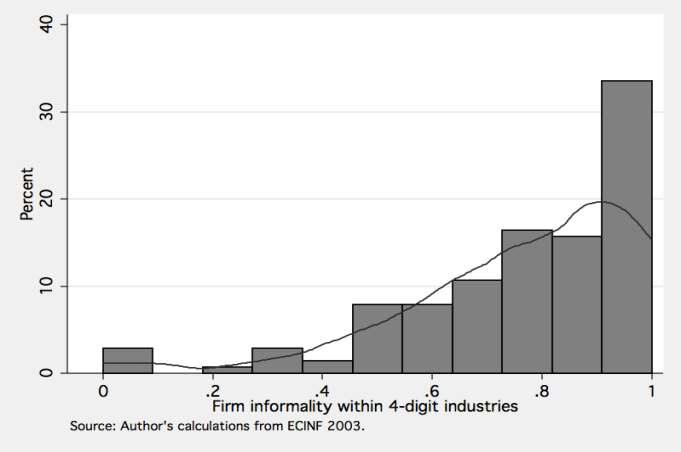
C Additional Stylized facts

Table 9: Descriptive Statistics from ECINF

	Formal		Informal	
	Mean	Sd.Dev.	Mean	Sd.Dev.
Owner's schooling (%)				
0 to 8	0.287	–	0.614	–
9 to 11	0.391	–	0.292	–
12+	0.322	–	0.094	–
Sector composition				
Services	0.394	–	0.402	–
Industry	0.078	–	0.110	–
Commerce	0.439	–	0.281	–
Construction	0.049	–	0.160	–
Wages†	0.777	1.232	0.594	0.925
Revenue†	9.119	16.679	1.363	3.323
Profit†	2.53	12.82	0.66	2.61
Firm's age (months)	110.01	98.53	106.04	105.68
# workers	2.72	1.73	1.28	0.72
Obs.	6,632		42,032	

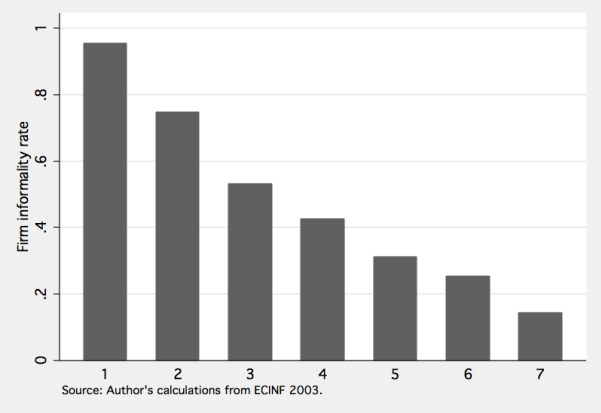
Source: ECINF 2003.
† Normalized by the (country-wide) average wage of prime-age workers who are heads of their household and work in the formal sector.

Figure A.3: Share of informal firms at the 4-digit industry level: Histogram

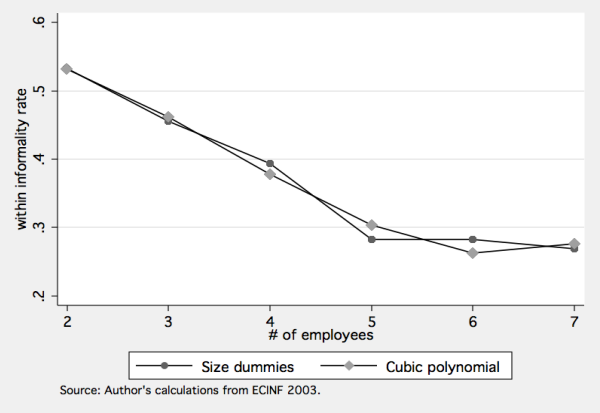


Note: The variable used is the share of informal firms measured at the 5-digit industry level. The figure shows the histogram of this industry-specific measure of firm informality.

Figure A.4: Informality margins and firms' size



(a) Extensive margin



(b) Intensive margin

Note: The panel on the left shows the share of informal firms among firms with size $n = 1, \ldots, 7$ (where size is measured as number of employees). The panel on the right shows the average share of informal workers within formal firms, among firms with size $n = 2, \ldots, 7$.

D Model Appendix

D.1 Different formulations for the costs of informality

This discussion heavily relies on Ulyssea (2012). For the cost of the extensive margin of informality, instead of the profit function described in Section 3, consider the following representation:

$$\Pi_i(\theta) = \max_{l_i} \left\{ \left[1 - p(l_i)\right] \theta l_i^{\alpha} - w l_i \right\}$$

where $0 < p(l_i) \leq 1$ is strictly increasing and convex.

The $0 \leq p(l_i) \leq 1$ can be interpreted as the probability of being caught by government's officials, in which case the informal firm looses all of its production. Assume that $p', p'' > 0$. This probability thus imposes a size limit to informal firms. To obtain a direct correspondence between this formulation and $\tau_i(\cdot)$, one could parametrize it as $\tau_i(l_i) = \frac{1}{1-p(l_i)}$, so that as $\lim_{p(l_i)\to 1} \tau_i(l_i) = \infty$ and $\lim_{p(l_i)\to 0} \tau_i(l_i) = 1$.

For formal firms, the unit cost of labor can be expressed as:

$$c(s_i, l_i) = 1 + \tau_w(1 - s_i) + [\tau_{fi}(l_{fi}) - 1] s_i$$

such that if the share of informal workers $s_i \to 0$ then $c(s_i, l_i) \to 1 + \tau_w$, which would be the unit cost of labor if no formal firm could hire informally. One can thus simply define the labor distortion in the formal sector as

$$\tau_f(s_i, l_i) = \frac{c(s_i, l_i)}{1 - \tau_y}$$

and therefore as $s_i \to 0$, $\tau_f(s_i) \to \frac{1+\tau_w}{1-\tau_y}$; that is, as the share of informal workers within a formal firm goes to zero, the labor wedge that it faces goes to the official level inplied by government's regulations and taxes.

D.2 Productivity distributions in both sectors

The post-entry, unconditional productivity distribution in the informal and formal sectors, respectively, is given by the following expressions:

$$\begin{aligned} f_{\theta_i}(x) &= \frac{1}{G(\overline{\nu}_f) - G(\overline{\nu}_i)} \int_{\overline{\nu}_i}^{\overline{\nu}_f} f(x|\nu)\, dG(\nu) \\ f_{\theta_f}(x) &= \frac{1}{1 - G(\overline{\nu}_f)} \int_{\overline{\nu}_f}^{\infty} f(x|\nu)\, dG(\nu) \end{aligned} \tag{9}$$

where f_{θ_s} is absolutely continuous and $F_{\theta_s}(\cdot)$ denotes the corresponding c.d.f..

As mentioned above, firms can be surprised with a bad productivity draw. Those with a $\theta < \overline{\theta}_s$, where $\overline{\theta}_s$ is such that $\pi_s\left(\overline{\theta}_s, w\right) = 0$, will not produce and will leave immediately. Hence, the effective productivity distribution among successful entrants is given by the following expressions:

$$\tilde{f}_{\theta_s}(x) = \begin{cases} \frac{f_{\theta_s}(x)}{1 - F_{\theta_s}\left(\overline{\theta}_s\right)} & \text{if} \quad x \geq \overline{\theta}_s \\ 0 & \text{if} \quad \theta < \overline{\theta}_s \end{cases} \tag{10}$$

where $s = i, f$.

D.3 Uniqueness of Equilibrium

This section contains a simple argument to prove the uniqueness of equilibrium. The key equilibrium conditions are given by the zero profit conditions, the free entry conditions and the market clearing condition, respectively:

$$\begin{aligned} \pi_s\left(\bar{\theta}_s, w\right) &\equiv \Pi_s\left(\bar{\theta}_s, w\right) - \bar{c} = 0 && (11) \\ V_i^e\left(\overline{\nu}_i, w\right) &= c_i && (12) \\ V_f^e\left(\overline{\nu}_f, w\right) &= V_i^e\left(\overline{\nu}_f, w\right) - \left(c_i - c_f\right) && (13) \\ L_i + L_f &= \bar{L} && (14) \end{aligned}$$

where $s = i, f$ and the free entry conditions assume that entry is positive in both sectors.

Fix a given wage. Given the assumptions made for the cost and production functions, the functions $\pi_s(\theta, w)$ are strictly increasing in θ and decreasing in w. Moreover, as $\bar{c} > 0$, there is a $\theta > 0$ such $\pi_s(\theta, w) < 0$. Thus, there is a unique $\bar{\theta}_s$ such that 11 holds. The simple form of the value functions, $V_s(\theta, w) = \max\left\{0, \frac{\pi_s(\theta,w)}{\kappa_s}\right\}$, implies that they are also continuous and strictly increasing in θ and decreasing in w. Combining this last fact with the assumptions made about $F(\theta|\nu)$, it follows that that there is an unique $\bar{\nu}_s$, $s = i, f$, such that free entry conditions hold, and that $\bar{\nu}_f > \bar{\nu}_i$. The latter follows from the assumption that $c_f > c_i$. The unique entry thresholds pin down the mass of entrants in both sectors: $M_i = \left[G\left(\overline{\nu}_f\right) - G\left(\overline{\nu}_i\right)\right] M$ and $M_f = \left[1 - G\left(\overline{\nu}_f\right)\right] M$. Given the mass of entrants in each sector, and the unique thresholds $\bar{\theta}_s$, the flow conditions in both sectors [given by (5)] pin down the mass of firms in each sector, μ_s. The last condition to close the equilibrium determination is the market clearing condition for the labor market, which determines the equilibrium wage. Of course, if $L^d \equiv L_i + L_f > \bar{L}$ there is excess demand and the wage will increase up until the point where $L^d = \bar{L}$ (the symmetric argument is true for excess supply). Because of the properties of the profit functions, the individual labor demand functions $\ell^*(\theta, w)$ are also continuous, single valued, and strictly increasing in θ and decreasing in w. Thus, there is an unique wage such that $L^d = \bar{L}$.

□

E Estimation Appendix

E.1 Discussion: Alternative methods to semi or non-parametrically identify the model

The present framework is a very simplified version of a discrete choice dynamic programming (DCDP) model.[36] Recent developments in the literature [e.g. Heckman and Navarro (2007)] have shown that semiparametric identification is possible for some classes of dynamic discrete choice models [see Abbring (2010)]. However, in what follows I argue that non-parametric estimation of the different objects in the model is either not feasible given the goals of this paper and the data available, or the assumptions needed are not attainable.

Starting by the unconditional productivity distribution, $F(\theta)$, there are a number of approaches to estimate it non-parametrically, as long as one is willing to assume a functional form for $q(\cdot)$ (Cobb-Douglas, say). One of the most well-known approaches is the one proposed by Olley and Pakes (1996), which requires access to a panel of firms that contains information on inputs, investment and revenues or, ideally, physical production [see Ackerberg et al. (2007) for a review of more recent methods]. The data requirements are however high, and are not met in the present context.

If knowledge of the profit function alone was sufficient, one could use Matzkin (2003) to identify it nonparametrically, as profit functions are homogeneous of degree one and thus satisfy Matzkin's conditions.[37] However, the profit function is a reduced-form object and not a primitive of the model as defined by the elements in ω. In the present application, it is necessary to identify the cost and revenue functions separately in order to perform the counterfactual analysis. The cost function could be nonparametrically estimated from data on variable costs (when available) and inputs, but this would only give the relationship between inputs and costs for a given structure. Once there are changes in the intensity of government inspections say, the structure that generated the estimated cost function would no longer be valid.

Finally, Heckman and Navarro (2007) extend Taber (2000) analysis to a general finite horizon model with a rich dynamics for the unobserved shocks and are able to semiparametrically identify their full structural model (including the cost and earnings functions). Their framework however, does not apply to the family of models considered here, as they

[36] The DCDP literature is quite extensive. The interested reader can refer to the well-known review of Rust (1994). More recently, Aguirregabiria and Mira (2010) and Keane et al. (2011) provide comprehensive and updated reviews of estimation methods and applications. Abbring (2010) presents a more recent discussion on identification of different DCDP models.

[37] Note, however, that even in this case one would have to assume that θ is independent of wages, which is only true when firms are truly price takers.

rely on additive separability between observable and unobservable state variables, and on the independence between both.[38] Additionally, their identification proof for the full structural model relies on "identification at infinity" type of arguments, which require strong support assumptions that are most likely not satisfied in the present application.

E.2 Asymptotic properties of the MSM estimator

In order to reduce the notational burden, I discuss the asymptotic properties of the one-stage MSM estimator, instead of the two-stage used in this paper. The reason for that is simply to omit the conditioning on the first-stage parameter vector, ψ. Throughout the discussion of the asymptotic properties of the estimator, the number of replications (S) per observation $i \in N$ is assumed to be fixed. The asymptotic analysis is therefore done with S fixed and $N \to \infty$.[39]

Consistency

The conditions for consistency of the MSM estimator are close to the ones for extremum estimators. In fact, the substantive assumptions are exactly the same as in the GMM case, namely, the GMM identification assumption, and the requirement that the parameter space is compact. The main difference thus lies on the type of regularity conditions required to guarantee consistency. I follow the discussion in Duffie and Singleton (1993), who provide conditions for both weak and strong consistency, which are satisfied by the present model.[40] I focus on the discussion of weak convergence for the sake of simplicity. The maintained assumption is that the state process in the underlying d.g.p. is *geometrically ergodic*. Their Lemma 1 (page 937) states the conditions under which geometric ergodicity holds, which are satisfied by as simple first-order Markov process. They also define a global modulus of continuity condition, which is that the simulator function is Lipschitz. With these conditions and the identification assumption, one can show that the MSM estimator is (weakly) consistent and thus converges in probability to φ_0.

[38] They apply results from Matzkin (1992) on nonparametric identification of static binary choice models.

[39] More generally, the simulation sample size can be defined by the function $\mathcal{S}(N)$, $\mathcal{S}: \mathbb{N} \to \mathbb{N}$, where as before N denotes the actual sample size, and $\mathcal{S}(N) \to \infty$ as $N \to \infty$. The case where replications per observation remains fixed is simply $\mathcal{S}(N) = S \times N$.

[40] Pakes and Polard (1989) provide the regularity conditions under which a broad class of simulation-based estimators are both consistent and asymptotically normal,which includes the MSM estimator proposed by McFadden (1989).

Asymptotic normality

Under the assumptions used to analyze consistency, asymptotic normality follows almost naturally. The derivations here follow closely the discussion in Newey and McFadden (1994) and Duffie and Singleton (1993).

Remember that the second-stage, SMM estimator is given by

$$\hat{\varphi} \quad = \quad \arg\min_{\varphi} Q(\varphi) = \left\{ g(\varphi)' \hat{\mathbf{W}} g(\varphi) \right\}$$

where $g(\varphi) = \hat{m} - m^s(\varphi)$ and I omit the conditioning arguments for notational convenience.

The following assumptions are made for asymptotic normality to hold:

(E.1) Assumptions required for asymptotic normality

1. φ_0 and $\hat{\varphi}$ are interior to the parameter space.
2. The simulator used to generate the simulated data is continuously differentiable w.r.t. φ in a neighborhood $\mathcal{B}$ of φ_0.
3. $\mathbf{G}_0 \equiv E\left[\nabla_\varphi g(\varphi_0)\right]$ exists, is finite and $\mathbf{G}_0'\mathbf{W}\mathbf{G}_0$ is nonsingular.

The first order condition of the MSM estimator is given by $\nabla_\varphi g_s(\hat{\varphi})' \hat{\mathbf{W}} g_s(\hat{\varphi}) = 0$, where $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, and $\mathbf{W}$ is positive semi-definite. For notational simplicity, let $g_s(\varphi) \equiv g_s(\varphi)$. Using the mean value theorem to expand $g_s(\hat{\varphi})$ around φ_0 and combining with the FOC gives

$$\sqrt{N}(\hat{\varphi} - \varphi_0) = -\left[\nabla_\varphi g_s(\hat{\varphi})' \hat{\mathbf{W}} \nabla_\varphi g_s(\overline{\varphi})\right]^{-1} \nabla_\varphi g_s(\hat{\varphi})' \hat{\mathbf{W}} \sqrt{N} g_s(\varphi_0)$$

Given that the simulator is unbiased and by the CLT, $\sqrt{N} g_s(\varphi_0)$ converges to a Normal distribution with zero mean and the following asymptotic variance

$$\Sigma_s = \left(1 + \frac{1}{S}\right)\Sigma$$

where $\Sigma = E\left[g(\varphi_0) g(\varphi_0)'\right]$ is the GMM asymptotic variance-covariance matrix.

Finally, from the WLLN both $N^{-1}\nabla_\varphi g_s(\hat{\varphi})$ and $N^{-1}\nabla_\varphi g_s(\overline{\varphi})$ converge in probability to G_0 and by the Slutzky theorem one gets

$$\sqrt{N}(\hat{\varphi} - \varphi_0) \xrightarrow{d} N\left(0, (\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1} \mathbf{G}_0'\mathbf{W}\Sigma_s\mathbf{W}\mathbf{G}_0 (\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1}\right)$$

Analogous to the GMM estimator, the optimal weighting matrix is given by $W^* = \Sigma_s^{-1}$, which therefore reduces the asymptotic variance-covariance matrix to

$$V_s\left(W^*\right) = \left(\mathbf{G}_0'\mathbf{\Sigma}_\mathbf{s}^{-\mathbf{1}}\mathbf{G}_0\right)^{-1}$$

The actual variance-covariance is computed using the empirical counterpart of Σ (estimated from real data), and the computational equivalent of G_0, which can be obtained using standard numerical differentiation methods.

E.3 Simulation Algorithm

For the simulations, I consider a mass of $M = 300,000$ potential entrants. For each potential entrant, I draw a pre-entry productivity parameter (ν) and a post entry productivity shock (ϵ). I use 77 equally spaced grid points for the productivity space. The maximum value in the productivity grid implies a firm's size of more than 18,000 employees and is not binding. Since each potential entrant has an individual pre-entry productivity parameter, it is necessary to compute a vector of transition probabilities for each point in the grid in order to compute the expected post-entry values in each sector for each potential entrant. For that, I use the method proposed by Tauchen (1986).

The stochastic components of the model are drawn only once in the beginning of the procedure and are kept fixed during the algorithm's execution. The estimation procedure is done conditional on the observed wage, which is the mean real wage for prime age males in the period 1997–2003 (pooling formal and informal employees together). I take the mean using the six years prior the baseline year to approximate the steady state wage. The steps of the estimation algorithm are the following:

1. Compute the observed wage.
2. Draw 300,000 observations of the random variables $X_1 \sim U(0,1)$ and $X_2 \sim N(0,1)$.
3. Compute the moments from the data, $\hat{m}$.
4. Initiate the optimization algorithm:
   (i) Guess φ.
   (ii) Obtain the post-entry productivity as $\log(\theta) = \log(\nu) + \epsilon$, where $\nu = G_\nu^{-1}(X_1)$ and $\log(\varepsilon) \equiv \epsilon = \sigma X_2$.
   (iii) Use the model to generate a simulated data set and compute $m^s(\varphi;\psi)$.
   (iv) Compute the loss function $Q(\varphi;\psi)$, as defined in (8).

(v) Check if the objective function is minimized (according some tolerance level). If not, return to step (i) and guess a new φ.

F Assessing the persistence of firms' productivity process

Suppose that instead of assuming that firms' productivity remains constant after entry, it is allowed to evolve over time. As it is usually assumed in the literature, suppose it follows a AR(1) process:

$$\log(\theta_{j,t}) = \tilde{\eta}_j + \rho \log(\theta_{j,t-1}) + u_{j,t} \tag{15}$$

for $t \geq 2$, where $\tilde{\eta}_j \equiv (1-\rho)\log(\nu_j)$, where ν_j still denotes firm j's pre-entry productivity parameter, ρ is the persistence parameter and the contemporary shock $u_{j,t}$ is assumed to be i.i.d. with $u_{j,t} \sim N(0, \sigma_u^2)$.

Let the log of total employment be denoted by $n_{j,t} = \ln(l_{j,t})$; using formal firms' FOC and the law of motion for the log-productivity (15), one can write

$$n_{j,t} = \gamma_0 + \gamma_1 w_t + \gamma_1 \log(\theta_{j,t}) + m_{j,t} \tag{16}$$

where w_t denotes wages, $\gamma_0 = \frac{1}{1-\alpha}\left[\log(\alpha) + \log\left(\frac{1-\tau_y}{1+\tau_w}\right)\right]$, $\gamma_1 = -\frac{1}{1-\alpha}$, and $m_{j,t}$ is measurement error.

One can then use (15) to write (16) in its dynamic representation:

$$n_{j,t} = b_0 + \gamma_1 w_t + b_2 w_{t-1} + \rho n_{j,t-1} + \eta_j + e_{j,t}$$

where $\eta_j \equiv \gamma_1 \tilde{\eta}_j$, $b_0 \equiv (1-\rho)\gamma_0$, $b_1 \equiv \gamma_1$, $b_2 \equiv \gamma_1 \rho$, and the error term is given by $e_{j,t} = \gamma_1 u_{j,t} + m_{j,t} - \rho m_{j,t-1}$.

Thus, the employment process itself can be represented as a simple AR(1) process with an MA(1) error. If there is no measurement error in (16), then the error term in the above expression, $e_{j,t}$, is serially uncorrelated. This representation is in line with a large literature on the estimation of dynamic panel models of employment evolution.[41] The final expression is the following:

$$n_{j,t} = \rho n_{j,t-1} + \beta \mathbf{x}_{j,t} + \eta_j + e_{j,t}, \tag{17}$$

[41] One of the most well-known applications is the one in Arellano and Bond (1991) [see also Blundell and Bond (2000)]. Arellano and Honore (2001) provide a comprehensive review of the literature on dynamic panel estimation.

where $\mathbf{x}_{j,t}$ denotes a vector of controls in addition of $n_{j,t-1}$ and η_j denotes firm's fixed effect. The vector $\mathbf{x}_{j,t}$ includes a set of year dummies and the current and lagged log-average wage rate calculated at the 4-digit indutry level.[42]

I start by estimating (17) using both OLS and within-groups estimators. The former is known to be upward biased while the latter is downward biased, and they can thus be used as upper and lower bounds to any consistent estimator [Bond (2002)]. The third model used is the standard first-differenced GMM estimator [e.g. Arellano and Bond (1991)], which is subject to finite sample biases (towards zero) when the lagged levels are weak instruments for the first-differenced equation. This will be the case when the series are very persistent ($\rho \to 1$) or when $\frac{\text{var}(\eta_j)}{\text{var}(e_{j,t})}$ is high [see Blundell and Bond (1998) for a detailed analysis]. The forth model is the system GMM estimator proposed by Blundell and Bond (1998), which uses lagged differences as instruments for equations in levels. [43]

For the first-differenced and system GMM models, I consider two scenarios for the error term. The first is the assumption of no measurement error, which allows for the use of lagged levels dated $t-2$ and earlier as instruments for the differenced equations, and lagged differences dated $t-1$ and earlier for the level equations. The second case assumes that there is measurement error, which implies that the error term in (17) will have a MA(1) structure. I this case one can only use lagged levels dated $t-3$ (and earlier) and lagged differences dated $t-2$ and earlier as instruments. All GMM regressions consider the log-wage as a predetermined variable and they are estimated using the two-step estimator with the correction for the variance-covariance suggested by Windmeijer (2005). Table 10 shows the results.

The results shown in Table 10 follow the pattern expected from the standard results in the literature. The OLS result already points to a very persistent process, which is confirmed by the downward bias in the first-differenced GMM model (DIFF) as compared to the system GMM estimator (SYS). Blundell and Bond (1998) find similar results when comparing these two estimators using a small sample of british firms. There is also evidence that measurement error is indeed present, as the estimates under the assumption of no measurement error (DIFF1 and SYS1) seem to be substantially downward biased. The difference-Sargan tests for the additional instruments available in the DIFF1 and SYS1 models (relatively to the DIFF2 and SYS2, respectively) strongly reject the validity of these additional instruments, reinforcing the evidence of measurement error (results not reported).

[42] This specification with current and lagged wages is standard in the empirical literature [e.g. Arellano and Bond (1991) and Blundell and Bond (1998)].

[43] The validity of this additional set of moment conditions relies on a relatively mild stationarity assumption regarding the process' initial conditions [see Blundell and Bond (1998) or Bond (2002) for a discussion].

Table 10: Productivity process estimation

	OLS	FE	DIFF1	DIFF2	SYS1	SYS2
n_{t-1}	0.944**	0.497**	0.594**	0.728**	0.713**	0.921**
	(0.000)	(0.002)	(0.007)	(0.011)	(0.009)	(0.005)
$\log(w_t)$	0.0030	0.0100	-0.0920	-0.158*	-0.210**	-0.339**
	(0.006)	(0.008)	(0.066)	(0.062)	(0.072)	(0.069)
$\log(w_{t-1})$	0.006	0.005	0.069	0.054	0.110	0.075
	(0.006)	(0.007)	(0.095)	(0.048)	(0.069)	(0.053)
Obs.	741,268	741,268	741,268	741,268	741,268	741,268

**Significant at 1% level; *Significant at 5% level.
DIFF1 and DIFF2 are the difference GMM models without and with measurement error; SYS1 and SYS2 are the system GMM models without and with measurement error, respectively.

That said, the preferred model is the system GMM under the assumption of measurement error (SYS2), which provides a reasonable value for the persistence parameter.[44] The result shows a large degree of persistence of the employment series, which under the current model implies that firms' productivity process is also extremely persistent.

[44]The Sargan test of overidentifying restrictions rejects instruments' validity in all models, with p-values of zero (not reported). However, Arellano and Bond (1991) show in their simulations exercise that the Sargan test rejects too often in the presence of heteroskedasticity, which is confirmed in their empirical application [they estimate a dynamic employment equation very similar to (17)].