Firms, Misallocation, and Aggregate Productivity: A Review

Hugo A. Hopenhayn

Department of Economics, University of California, Los Angeles, California 90095; email: hopen@econ.ucla.edu

Annu. Rev. Econ. 2014. 6:735–70

First published online as a Review in Advance on May 5, 2014

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi: 10.1146/annurev-economics-082912-110223

JEL codes: O10, O11, O40, O47, E23, E26

Keywords

distortions, resource allocation, firm heterogeneity

Abstract

Firm heterogeneity and the allocation of resources across firms play a key role in determining aggregate productivity. Entry barriers and misallocation can substantially impact productivity, as evidenced in recent work. This article provides a unifying theoretical framework and a review of this literature.

1. INTRODUCTION

One of the greatest challenges in economics is explaining the disparity in income per capita across countries. There is a large literature documenting this development gap. As an example, Caselli (2005) reports a 20-fold gap (1 to 0.05) between the GDP per capita of the top and bottom deciles. After controlling for differences in resource endowments, he finds a corresponding gap in total factor productivity (TFP) on the order of 7.

There is a growing literature trying to understand how the firm microstructure can contribute to explain this gap. The key point of departure in this literature is the observed high degree of heterogeneity in size and productivity across different production units. Aggregate TFP is affected both by the underlying distribution of establishments' productivities and the allocation of resources (e.g., capital and labor) across these units and by the number of firms per capita. This article provides a unifying framework and considers some of the most important contributions to the literature.

Our basic setting is a simplified model of firm heterogeneity in perfect competition as in Lucas (1978) and Hopenhayn (1992). Firms differ according to their idiosyncratic productivity and produce according to a common production function with decreasing returns. In the absence of distortions, the equilibrium (and Pareto optimum) allocates resources (labor in the simplified model) to maximize the total output of a homogeneous good. This optimal allocation implies a constant returns to scale aggregate production as a function of the number of firms and other factors of production. Aggregate TFP equals the geometric average of individual firms' productivities. The model is equivalent to one of monopolistic competition (Dixit & Stiglitz 1977, Melitz 2003) that is often used in this literature, in which decreasing returns come from the demand side.

There are three margins that determine output per capita: the number of firms per capita, the distribution of firms' productivities, and the allocation of resources across these firms. The number of firms is important owing to decreasing returns. I explore the implications of the models on firm entry, average size, and average productivity. The analysis takes into account several sources of variation: (*a*) the role of the distribution of firm's productivity, (*b*) the role of entry costs, and (*c*) selection effects due to firm heterogeneity. A consistent finding is that average firm size increases (decreases) with aggregate productivity if the elasticity of the cost of entry with respect to TFP is less (greater) than 1. I also show that, loosely speaking, an improvement in the distribution of productivities of potential entrants has a positive effect on aggregate productivity, decreases entry, and increases average firm size.

The work of Djankov et al. (2002), followed by yearly World Bank surveys, documents very high variation in entry costs across countries. Our basic model provides a parsimonious way of assessing the impact on productivity, suggesting a negative elasticity of TFP with respect to entry costs on the order of one-fourth. With this value, an increase in the cost of entry of about one-half of the standard deviation found in the data implies a 20% reduction in TFP. The results in the papers I review are consistent with this order of magnitude, so differences in the costs of doing business as measured by entry costs are an important source in explaining the TFP gap but leave much unexplained. As a related matter, I survey papers that consider the role of noncompliance by small firms, usually referred to in the literature as informal firms. It does not appear that informality per se can help explain a large part of the TFP gap, although, as suggested above, entry costs to formalization do.

Among the factors explaining the disparity of aggregate productivity across countries, the misallocation of resources across firms has received much recent attention (Banerjee & Duflo 2005, Alfaro et al. 2009, Guner et al. 2008, Restuccia & Rogerson 2008, Hsieh & Klenow 2009,

Buera et al. 2011, Bartelsman et al. 2013). The basic idea is that institutions and policies might prevent the equalization of the marginal value of inputs across firms, thus resulting in aggregate productivity losses. One of the first papers to consider the impact of wedges in the allocation of resources across firms on aggregate productivity is Hopenhayn & Rogerson (1993). As a consequence of firing costs, firms fail to fully adjust their labor force in response to shocks, especially temporary ones. As shown in this article, this can give rise to large differences in the marginal product of labor: Some firms hold workers in excess of the optimal amount, whereas others show a shortage. Such gaps can be identified with wedges or implicit taxes. Firing costs lead to a very rich set of wedges, some of which are correlated with productivity and some of which are not. In a simple calibration, I provide orders of magnitude of these wedges and their impact on aggregate productivity, which turns out to be small compared to the development gap.

The more recent literature on misallocation abstracts from the origin of distortions, while treating distortions as primitives themselves (Guner et al. 2008, Restuccia & Rogerson 2008, Hsieh & Klenow 2009, Bartelsman et al. 2013). The main emerging message seems to be that correlated distortions—those that implicitly tax more productive firms and subsidize less productive ones—can be more damaging to aggregate TFP. I review this literature and provide a characterization of the mapping between the structure of distortions and aggregate TFP. My analysis shows that the key element is the concentration of distortions and that correlation with size/productivity does not matter per se, only to the extent that it gives rise to more concentrated distortions.

Hsieh & Klenow (2009) provide a methodology to recover wedges from firm-level data. I review their paper and derive simple formulas to calculate TFP losses. Additionally, I review two other contributions (Bartelsman et al. 2013, Midrigan & Xu 2014). The calculations in Hsieh & Klenow (2009) indicate that misallocation can explain half of the TFP gap between China or India and the United States.

Although this literature has been very useful in evaluating the potential impact of misallocation, it provides less guidance to understand the major forces behind this large misallocation. Recent papers give lots of attention to the role of financial constraints, which is suggested by the very strong correlation between financial development and TFP. These papers use structural models to establish and measure the impact of the causal link (Jeong & Townsend 2007, Amaral & Quintin 2010, Buera & Shin 2013, Caselli & Gennaioli 2013, Midrigan & Xu 2014). This is a particularly interesting literature both because it has considerable success in explaining a sizable part of the development gap and because it helps identify the role of several different forces. More specifically, borrowing constraints have an impact on total capital accumulation, the allocation of capital across firms (misallocation of resources), and the distribution of firm qualities. The overall impact on productivity varies according to the papers considered. To understand these differences, I examine a simplified version of Moll (2014). In a dynamic setting, firms have an incentive to save out of these constraints. How effectively they can accomplish this depends on the persistence of productivity shocks and expected lifetime of the firm. The papers I review make very different assumptions that bear on this question, and this mostly explains the disparity of magnitudes reported. The overall impression that emerges from these quantitative models is that although borrowing constraints are a source of misallocation, the quantitative effect is moderate, whereas the effect of entrepreneurial mismatch is likely to be larger.

The recent literature on misallocation, as in Hsieh & Klenow (2009), relies heavily on the structure of technology and demand to identify the distribution of firm-level productivities/quality and misallocation. I review three sources of misspecification: measurement error, the degree of decreasing returns, and adjustment costs. As analyzed by Hsieh & Klenow (2009), classical measurement does not seem to account for the observed misallocation. As for decreasing returns, I

prove that estimates of the impact of measured distortions on TFP decrease with curvature. Consistent with this result, Hsieh & Klenow (2009) find that their estimated gains to eliminating distortions increase substantially as the elasticity of demand is increased. Finally, I consider the role of adjustment cost, emphasizing a recent paper that argues, quite convincingly, that the wedge introduced by adjustment costs can account for an important part of misallocation (Asker et al. 2014).

I conclude my review returning to the question of the kind of distortions that might be more relevant in explaining the TFP gap. Alfaro et al. (2009) and Hopenhayn (2012) derive lower bounds on distortions that are consistent with differences in the size distribution of firms across countries. The lower bounds obtained in Hopenhayn (2012) seem to explain a fairly negligible portion of the TFP gap. The bounds are obtained by imputing distortions that preserve the rank in the size of firms, at least weakly. Many size-dependent policies discussed in the literature have this property, as is the case for exemptions to taxation or labor costs that are granted to firms under a certain size threshold. I conclude by arguing that this type of policy is unlikely to be a good candidate for distortions explaining a large TFP gap. Strong reversals of ranking, as those implied by Restuccia & Rogerson (2008) or those found by Hsieh & Klenow (2009), seem necessary to accomplish this.

2. THE BASIC MODEL

This section describes a simple baseline model used throughout the article. The model is a simplified version of Hopenhayn & Rogerson (1993) that builds on Hopenhayn (1992) but without entry and exit and is closely related to Lucas (1978) and Jovanovic (1982). There is a set (measure) of firms $i = 1, \ldots, M$ with production functions

$$y_i = z_i n_i^{\eta},$$

where z_i is an idiosyncratic productivity shock for firm/establishment $i = 1, \ldots n$.[1] Production displays decreasing returns ($\eta < 1$) in the only input (labor), of which there is a total endowment N in the economy that is supplied inelastically. Firms behave competitively, taking prices as given. This economy has a unique competitive equilibrium ($\{n_i\}, w$), where n_i is the profit-maximizing input choice for firm i, and the labor market clears. The competitive equilibrium is also the solution to the output-maximization problem:

$$\max_{n_i} \sum_i z_i n_i^{\eta},$$
$$\text{subject to } \sum n_i \leq N.$$

The efficient allocation equates marginal products across all firms or equivalently, following from the homogeneity of the production function, to equate average products: $y_i/n_i = y/n \equiv a$ for all i.[2] Finally, using the aggregate resource constraint to substitute for a, it follows that the aggregate output

$$y = \left(\sum_i z_i^{\frac{1}{1-\eta}} \right)^{1-\eta} N^{\eta}. \tag{1}$$

This is an aggregate production function of the same class as the underlying firm-level production function (a similar aggregation is given in Melitz 2003). It exhibits decreasing returns,

[1] I do not make any distinctions here, although quantitative analysis is usually done by taking establishments as units of production.

[2] This obviously does not hold with a fixed cost in terms of overhead labor, as used in Bartelsman et al. (2013).

when firms are treated as a fixed factor. This can be more clearly seen by dividing the first term by $M^{1-\eta}$,

$$y = \left(E z_i^{\frac{1}{1-\eta}}\right)^{1-\eta} M^{1-\eta} N^{\eta}, \tag{2}$$

or grouping firms by productivity,

$$y = \left(\int z^{\frac{1}{1-\eta}} dG(z)\right)^{1-\eta} M^{1-\eta} N^{\eta}, \tag{3}$$

where G corresponds to the distribution of productivity across firms.

2.1. Firms Versus Managers and Aggregate Productivity

When firms are treated as a reproducible asset, as in Hopenhayn (1992), the aggregate production function given in Equation 2 has constant returns to scale in M and N and a level of aggregate TFP that is given by the geometric mean of firm-level productivity. This interpretation assumes that an expansion in the number of firms is obtained by replication (i.e., preserving the distribution of productivity and thus aggregate TFP).

An alternative model used in the misallocation literature is the Lucas model of span of control (Lucas 1978), in which firm productivities z_i correspond to managerial talent. Here G stands for the distribution of managerial talent conditional on participation, as only the most efficient producers will be active (those above some threshold z_0). This threshold obviously decreases while the number of active producers increases. Although Equation 2 remains valid under this interpretation, aggregate TFP is now a decreasing function of M. For the special case in which $(1 - G(z)) = z^{-\alpha}$ is Pareto, the production function has a simple expression. Letting N stand for the size of the population and $N_e = N - M$ the number of workers, one finds that

$$y = A_0 \left(\frac{M}{N}\right)^{-\frac{1}{\alpha}} M^{(1-\eta)} N_e^{\eta}, \tag{4}$$

or in per capita terms,

$$y/N = A_0 m^{\left(1-\eta-\frac{1}{\alpha}\right)} n_e^{\eta}$$

for some constant A_0, where $m = M/N$ and $n_e = N_e/N$.[3]

As shown in Equations 2 and 3, output per capita in this economy is determined by (*a*) the distribution of firm-level productivities and (*b*) the total number of firms, M. In this article, I refer to the first component as TFP to emphasize that the latter corresponds to a factor of production, a form of (human) capital. This distinction is often not clear, as several papers in the literature identify aggregate TFP with the term in parentheses in Equation 1, confounding these two sources.[4]

[3]The function still exhibits constant returns to scale under the Lucas-model interpretation, as the mass of producers and workers is scaled proportionately with the population without the need of changing the manager threshold. For simplicity, I am assuming that managers and workers are two separate groups, in contrast to the Lucas model of span of control. This allows one to treat N as a fixed endowment independent of M in the analysis to focus on the most relevant margins. The trade-off is considered explicitly below.

[4]In matching to data, it is a question of what fraction of this capital M is measured either in the capital or in the human capital stock. Some papers treat M as intangible capital (see Atkeson & Kehoe 2005).

It is still the case that output per capita, $y/N = Am^{1-\eta}$, is increasing in the number of firms per capita, m. Following the literature (see Banerjee & Moll 2010, Buera et al. 2011, Midrigan & Xu 2014), I refer to changes in M or the distribution of firm productivities as extensive margins.

The benchmark model implies that output per capita increases with the number of firms per capita and thus decreases with the average size of firms.[5] This is in contrast to widespread evidence indicating that the average size of firms is lower in less developed economies. We return to this question below when considering entry decisions.

2.2. A More General Production Function

Our analysis considers the case of a single input for convenience. The aggregation results presented above and most of what follows are easily generalizable. The key assumptions are that all firm production functions are homogeneous of the same degree in all inputs and that productivity shocks are multiplicative (Lucas 1978). This can be represented as follows:

$$y_i = z_i\big(f(\mathbf{x})\big)^{\eta},$$

where f is a constant returns to scale production function on a vector of inputs $\mathbf{x}$, giving rise to the aggregate production function:

$$y = \left(E z_i^{\frac{1}{1-\eta}}\right)^{1-\eta} M^{1-\eta} f(\mathbf{X})^{\eta},$$

where $\mathbf{X}$ is the aggregate vector of inputs. The parallel with the above analysis can be easily seen, by interpreting $f(\mathbf{x})$ as a technology for producing an intermediate input that is transformed into output by firms according to the homogeneous production function used in the previous section.

2.3. Connection to Monopolistic Competition

In a monopolistically competitive economy (Dixit & Stiglitz 1977), output y is produced by aggregating a continuum of intermediate inputs y_i with the production function

$$y = \left(\int y_i^{\eta} di\right)^{\frac{1}{\eta}},$$

and each intermediate good is produced with a linear function of labor,

$$y_i = \tilde{z}_i n_i,$$

where $\tilde{z}_i$ is the productivity of an intermediate producer i. As it is well known, in an equilibrium, firms choose a constant markup over the marginal cost such that $p_i = (1/\eta)(w/\tilde{z}_i)$. Letting $z_i = \tilde{z}_i^{\eta}$, one can easily show that $y_i^{\eta} \propto z_i^{1/(1-\eta)}$ and $n_i \propto z_i^{1/(1-\eta)}$. Comparing these to the derivations in the first section, it follows that output under monopolistic competition is $y^m = y^{1/\eta}$. Hence, we can rewrite the expression for Equation 2 as

[5] The effect is mitigated in the Lucas model because of negative selection but does not disappear. For example, under the Pareto distribution, one finds that $y/N = A_0 m^{1-\eta-\frac{1}{\alpha}}$, which in an interior solution increases with m but at a lower rate.

$$y = \left(E z_i^{\frac{1}{1-\eta}}\right)^{\frac{1-\eta}{\eta}} M^{\frac{1-\eta}{\eta}} N$$
$$= \left(E \tilde{z}_i^{\frac{\eta}{1-\eta}}\right)^{\frac{1-\eta}{\eta}} M^{\frac{1-\eta}{\eta}} N. \quad (5)$$

These are the familiar equations for the monopolistically competitive case (see Melitz 2003). Note that for fixed M, aggregate TFP is the same (given the transformation of the z_i's)[6] as in the perfectly competitive case, and the only difference remains in the increasing returns to scale in M, N. As for the determination of entry, it is well known that in the monopolistic competition setting, the equilibrium zero-profit condition for entrants implies that the number of firms is efficient, subject to pricing at the markup rule. That is, the equilibrium number of firms solves the problem of efficiently assigning labor to the creation of new firms to maximize the total output. It must also maximize $y^\eta = \left(E\tilde{z}_i^{\eta/(1-\eta)}\right)^{1-\eta} M^{1-\eta} N^\eta$, as this is a monotone transformation of output y. It follows that the optimal and equilibrium number of entrants analyzed below applies also to this case.

3. ENTRY AND THE EXTENSIVE MARGIN

This section considers the relationship between TFP and firm size and considers the impact of barriers to entry. We first extend the model to incorporate entry decisions. Suppose that c_e workers are needed to create a firm with productivity that is randomly drawn from cumulative distribution function (CDF) G, independently for all entrants. A competitive equilibrium is defined as follows. In the first stage, a large mass of identical potential entrants decides whether to enter. An entrant must pay the cost of entry given by c_e units of labor and then draws its productivity e according to a CDF G. Assuming there is a large number of entrants and that draws of potential entrants are independent, the distribution of realizations is approximately given by G. Entry decisions are driven by the expected profits of a firm $E\pi(w) = \int \pi(z, w) G(dz)$, where $\pi(z, w) = \max_n z n^\eta - wn$. In an equilibrium, we find that $E\pi(w) = c_e w$.

In this simple economy, the welfare theorems hold, so equilibrium and Pareto optimal allocations—those that maximize total output—coincide. For a fixed number of firms, we construct an equilibrium and optimal allocation in the previous section. The optimal choice of the number of firms solves

$$y = \max_{M, N_e} A M^{1-\eta} N_e^\eta, \quad \text{subject to } c_e M + N_e \leq N. \quad (6)$$

The solution to this problem is $N_e = \eta N$, the number of firms is $M = \big((1-\eta)N\big)/c_e$, and the equilibrium wage is the multiplier of the constraint. The corresponding production function in terms of the total labor endowment N is given by

$$y = A_0 c_e^{-(1-\eta)} N, \quad (7)$$

where A_0 is a constant multiplicative in A. Given the aggregate production function above, the total output will be split between wages and firm profits with shares η and $(1-\eta)$, respectively, and

[6]This transformation of the z_i's is exactly what would be needed to match firm-level employment (e.g., the size distribution of firms measured by employment).

the equilibrium entry condition $E\pi(w) = \int \pi(z,w)G(dz) = wc_e$ is verified. Equation 7 implies a constant elasticity of output per capita with respect to the cost of entry equal to $(1-\eta)$.[7] A similar expression for output is obtained in the Lucas-style model[8] when managerial talent has a Pareto distribution with parameter α, in which the elasticity of output per capita to the cost of entry is $-(1-\eta-1/\alpha)$.[9] Equation 7 displays constant returns to scale in population size N and an aggregate TFP that is the product of two terms: (*a*) the geometric mean of productivities of participating producers and (*b*) the cost of entry. A similar expression with a slightly different constant term is obtained in a steady state when firms last more than one period and die exponentially at a rate δ, as is commonly assumed in the literature.[10]

3.1. Average Firm Size and Total Factor Productivity

The evidence on firm size and development is mixed. According to Alfaro et al. (2009), using Dun & Bradstreet data, it is significantly negative; it is mildly negative in Bollard et al. (2014), using UNIDO data; and it is significantly positive (with an elasticity of 0.45) in Poschke (2014), using GEM and Amadeus data. I now examine in more detail the implications of the above models on this relationship.

Consider first the effect of exogenous changes in the number of firms on aggregate productivity. According to Equation 3, a higher number of firms contributes positively to GNP per capita, and if not fully accounted for by measured inputs (e.g., the case of intangible capital), it will show up in the national accounts as higher TFP. So holding employment fixed, this channel implies a negative relationship between average firm size and aggregate productivity.

In our simple model, the distribution of firm productivity is independent of the number of firms, so the total effect of the number of firms on TFP is at worst neutral and is positive when firm capital is not properly accounted for. In contrast, the Lucas model provides a negative selection effect.

Although results in general depend on the details of the distribution, there are interesting implications that can be explored in the Pareto case. If the smaller average size in less developed countries is the result of excessive entry, Equation 4 can be used to derive implications for TFP. For example, the average firm size in India is one-third of what it is in the United States, so there are roughly three times more firms per capita in India. The value of $1/\alpha$ is bounded by $1-\eta$; otherwise, there is no equilibrium with heterogeneous firms. Conservatively, I take $\eta = 3/4$, giving an upper bound of $1/\alpha = 1/4$.[11] The selection effect accounts for 25% lower TFP in India relative to the United States. The impact on output per capita is lower because, as explained above, firms are a productive factor.

Consider now the case of endogenous entry. If the cost of entry (measured in goods) increases in the same proportion as aggregate TFP (e.g., when it is denominated in units of labor),

[7]These results easily generalize to the formulation given in Section 2 assuming that the technology for producing firms is linear in the constant returns to scale aggregator $f(\cdot)$. For example, if $f(k, n) = (k^\alpha n^{1-\alpha})^\eta$, the aggregate production function is $y = A_0 c_e^{-(1-\eta)} K^\alpha N^{1-\alpha}$.

[8]The Lucas model of span of control is exactly the model with an endogenous distribution and $c_e = 1$.

[9]However, in the latter case, the number of firms given by $M = \big((1-\eta-1/\alpha)/(1-1/\alpha)\big)/c_e$ is lower and increasing in α, so it is not independent of the distribution of firms' productivities.

[10]With discount factor β and survival rate $1-\delta$, the term c_e needs to be substituted by $(\eta(1-\beta)(1-\delta)+\delta)c_e$. If there also are fixed costs, the present value of these fixed costs (appropriately discounted) must be added to the entry costs in the above calculation.

[11]Our model abstracts from capital, so η cannot be taken as the labor share; rather, it indicates the degree of decreasing returns. It is quite standard in the literature to set it equal to 0.85. For the monopolistic competition model, one finds that $\eta = (\sigma-1)/\sigma$, where σ is the elasticity of substitution. In Hsieh & Klenow (2009), $\sigma = 3$ is used, implying $\eta = 2/3$. Others argue for considerably higher elasticity.

total entry—and thus average firm size—remains unchanged. It easily follows that average firm size will increase, stay constant, or decrease with development depending on whether the cost of entry increases, stays constant, or decreases more than TFP. The effect is reinforced in the Lucas model due to selection.

A change driven by differences in the distribution of firm productivity has additional implications in the Lucas model. Assuming a Pareto distribution with parameter α, lower TFP corresponds to higher α (smaller tail of the distribution). It is then easy to show that as TFP falls, the number of firms increases and the average size decreases (see footnote 9). To get an idea of how large this effect can be, I perform a simple calculation setting $c_e = 1$—the value that corresponds exactly to the Lucas model—and, conservatively, $\eta = 3/4$. I vary the parameter $\alpha \in \{5, 5.1, \ldots, 10\}$ to generate a range of firm sizes from 6 to 16, which is consistent with variation in cross-country data. The implied size distributions of firms are Pareto, with coefficients ranging from 1.25 to 2.5, which seem fairly reasonable.

These parameter values give a ratio of highest to lowest TFP of 1.65 and an elasticity of average firm size to TFP of almost 2. This implies an elasticity of average firm size to income per capita of 2/3, close to the elasticity of 0.45 found in Poschke (2014). The implied range of TFP variation is well within the cross-country variation (which is 5 to 1). This simple back-of-the-envelope calculation shows that large variation in average firm size across countries can be easily accommodated given observed differences in aggregate TFP.

A related question is, What drives differences in entrepreneurship across different economies? Guiso & Schivardi (2011) provide a very insightful empirical analysis. Using data for Italian manufacturing firms at the regional level, they show that the number of entrepreneurs is positively associated with the distribution of entrepreneurial productivity. The result seems somewhat counterintuitive to the above analysis, but there is an important difference. In Guiso & Schivardi's (2011) model, all regions form part of an integrated economy, with relatively free mobility of labor and capital but much less mobility of entrepreneurs. Because of factor mobility, entrepreneurs in a given region do not compete with each other for resources. The absence of this general equilibrium effect implies that, contrary to our model, the higher incidence of entrepreneurship (as a ratio of people born in the same region) is consistent with larger average firm size.

3.2. Cost of Entry and Total Factor Productivity

Equation 7 implies an elasticity of aggregate TFP to costs of entry equal to $-(1 - \eta)$. The value for $\eta = 3/4$ used above implies a moderate elasticity of TFP to costs of entry equal to one-fourth. Several recent papers emphasize the differences in the costs of doing business as a potential source of cross-country disparities in income per capita (see, e.g., Barseghyan 2008, Poschke 2010, Barseghyan & DiCecio 2011, D'Erasmo & Moscoso Boedo 2012, Moscoso Boedo & Mukoyama 2012). These papers exploit recent cross-country evidence documenting the wide variation on the costs (time and fees) involved in creating a business. [Measurements of the cost of doing businesses come from Djankov et al. (2002) and yearly follow-ups by the World Bank (e.g., Doing Business 2009).]

Based on World Bank data for 2007, Barseghyan (2008) finds a standard deviation in log costs of entry of 1.61. Using instrumental variables, he shows that an increase in the cost of entry of 80% (half this figure) accounts for a 22% reduction in TFP and a 29% reduction in output per worker, implying an elasticity between one-fourth and three-eighths. Barseghyan & DiCecio (2011) calibrate a sophisticated model of firm dynamics using a coefficient $\eta = 0.85$ and get an elasticity of -0.14. Their model generates 45% productivity differences between countries in the top and bottom deciles of the entry cost distribution. Similarly, in Moscoso Boedo & Mukoyama (2012), differences between the entry costs of lower-income countries [with 2% gross national income (GNI) of the United States] and median–entry cost countries explain

a TFP gap of 21%. These results suggest that cross-country variation in the cost of entry is an important part of the story, but they are still far from explaining the TFP gap (see also Poschke 2010 for a similar exercise).

3.3. Cost of Entry and Informality

Higher costs of entry help explain lower TFP in less developed economies but have potentially one counterfactual implication, as they lead to less entry and thus to higher average firm size. The effect of selection is not clear and depends on whether selection operates prior to or after entry. In the former case (as in the Lucas-style model), selection will be positive. In the latter case (as in Hopenhayn 1992), the higher cost of entry lowers the threshold for exit, giving rise to negative selection.

The existence of a large informal or noncomplying sector is often associated with high entry costs and underdevelopment (Rauch 1991, Loayza 1996, Amaral & Quintin 2006, Pratap & Quintin 2008, Quintin 2008a, D'Erasmo & Moscoso Boedo 2012).[12] Informal firms are defined in many different ways, but the most conventional and easiest to interpret is that of noncompliance with many forms of regulation. Informal firms are much smaller than their formal counterpart[13] and obviously contribute to reducing the average size of firms in less developed economies, where they account for most employment. There is also a very strong negative correlation (−0.897) between the share of the informal labor force and output per capita, as shown in **Figure 1**. D'Erasmo & Moscoso Boedo (2012) consider the impact of a series of distortions—including higher costs of entry—on aggregate TFP in a model in which firms have the option to operate in the informal sector to avoid these and other costs. Their model also includes borrowing constraints, as described in Section 6. According to these results, the joint effect of these distortions explains a 25% gap between low-income countries (those between 2% and 8% of the US GNI per capita) and the United States, or roughly 36% of the actual gap. The lion's share goes to differences in the cost of entry, accounting for 29% of this gap (borrowing constraints explain approximately 6%). Their numerical exercise also shows that informality per se does not play a substantial role, as the impact of all these distortions is quite similar in an economy in which firms cannot avoid compliance.

4. MISALLOCATION

The aggregation results given in Section 2 assume that resources (e.g., labor) are allocated efficiently across firms. In our analysis, that means equating marginal products across firms or, similarly, equating average products [total factor productivity revenue (TFPR)].[14] There are many policies and institutional constraints that prevent such equalization, and this has been the focus of the literature on misallocation.

As a motivating example, consider the effect of labor tax avoidance, prevalent among many firms in developing countries. With τ representing the tax per worker, firms that comply equate the marginal product of labor to $w(1+\tau)$, whereas those that do not equate it to w. Equivalently, if τ were a tax on output, revenues for complying firms would be $(1-\tau)y_i$, whereas it would be y_i for

[12]There is such a large literature on this, starting with De Soto (1989), that I am forced to leave many important references aside.

[13]Pratap & Quintin (2008) document that the informal sector mainly comprises small-scale, unskilled intensive activities.

[14]As explained above, the two conditions are equivalent for homogeneous production functions.

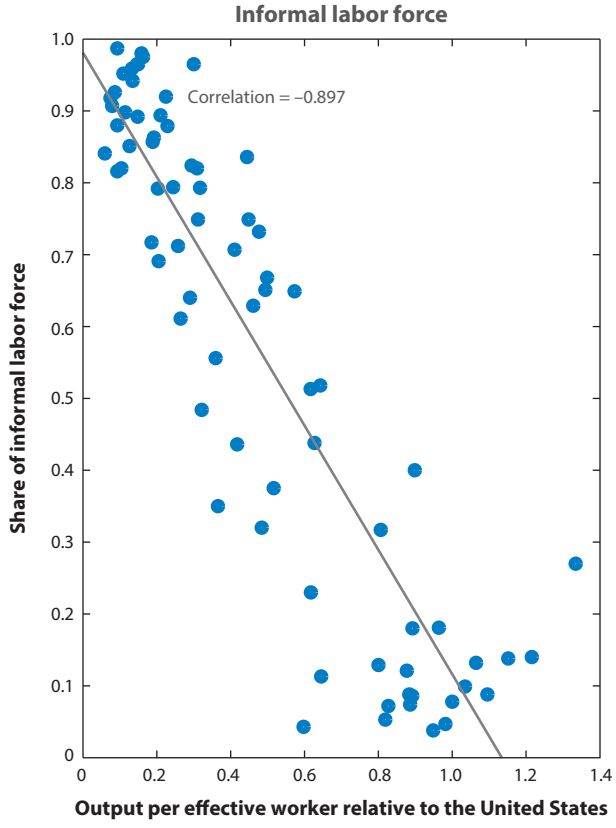

Figure 1

Informal labor force and output per worker. Figure adapted from D'Erasmo & Moscoso Boedo (2012), with permission from Elsevier.

those that do not comply. A new formula for aggregate productivity incorporating these distortions can be easily derived and is given by

$$y = \frac{E z_i^{\frac{1}{1-\eta}} (1+\tau_i)^{-\frac{\eta}{1-\eta}}}{\left(E\left[z_i/(1+\tau_i)\right]^{\frac{1}{1-\eta}}\right)^{\eta}} m^{1-\eta} n^{\eta},$$

where τ_i corresponds in the example to firm-specific labor taxes.[15] Similar effects follow from tax exemptions to certain classes of firms, regional subsidies, etc. More generally, these firm-specific taxes represent wedges on firms' choices that result from policies or other constraints in the economy.

In the simple form in the above aggregation, a term with weighted distortions replaces the geometric average of firm productivities. In particular, it preserves the same functional form and multiplicative separability on firms and labor. One implication of this property is that in our benchmark model with entry costs denominated in labor, distortions are neutral on the equilibrium number of firms and on the average size. When entry costs are partly denominated in output,

[15]The formula for the sales tax is identical with $(1 + \tau_i)$ replaced by $(1 - \tau_i)^{-1}$.

the resulting lower productivity will have a negative impact on entry and result in higher average size, as seen above.[16]

4.1. Total Factor Productivity Revenue

Misallocation results in a difference in the marginal value of inputs. For the one-input case, this means dispersion in marginal products or equally in the average output per worker. With multiple inputs under a Cobb-Douglas aggregator, there is also a convenient representation. Assuming the firm's production function is $z_i(k^\alpha n^{(1-\alpha)})^\eta$, an efficient allocation requires equalizing $y_i/k_i^\alpha n^{1-\alpha}$ across firms. This term corresponds to what has been called TFPR in the literature.[17] The variance of the log of TFPR has been taken as an aggregate measure of distortions, as shown below.

4.2. From Policies to Wedges

Hopenhayn & Rogerson (1993) were perhaps the first to consider the impact of misallocation on aggregate productivity by examining the impact of layoff costs. They also develop the general equilibrium framework described above with the addition of time-varying idiosyncratic shocks and provide a method to calibrate the model using data on firm dynamics that is widely used. Firing costs are a direct tax on the downward adjustment of employment. This by itself will prevent firms from adjusting to their static optimal employment levels. In addition, firms might choose to delay adjustment even further if a current bad shock could revert in the near future. A similar real options problem occurs when making hiring decisions in which, in anticipation of a possible reversal of fortunes, firms will hire less than the static optimal employment. These two effects will obviously be stronger the less persistent are the shocks.

In the presence of firing/hiring costs, past employment becomes a state variable affecting choices in the present. Two firms that arrived at the same level of productivity but from very different histories will have different stocks of employed workers and may decide to adjust differently. The solution to this problem is an *sS* policy characterized as follows: For every current productivity level z, there are two values, $n_l(z) < n_h(z)$, that are increasing in z, and three regions of adjustment: (*a*) If a firm starts with employment below $n_l(z)$, it hires up to that level; (*b*) if it starts with employment above $n_h(z)$, it fires workers up to that level; and (*c*) in the intermediate region (zone of inaction), the firm does not change its employment. The gap between these two values implies a maximum wedge between the two firms. This gap increases with firing costs and decreases with the persistence of productivity shocks. **Figure 2** is an example of an *sS* policy.

4.2.1. Firing costs and implicit wedges.[18] The history of productivity shocks of a firm, through repeated application of this employment policy, determines the current employment of the firm. A stationary equilibrium implies a joint distribution of productivity/employment levels and consequently an aggregate level of TFP, as discussed in the previous sections. Firms with employment below the undistorted level (the one given by the middle line in **Figure 2**) have

[16]The neutrality result also does not hold when firm productivity changes over time. As shown in Fattal-Jaef (2014) and Fattal-Jaef & Hopenhayn (2012), correlated distortions lead to more entry of firms and thus lower average size.

[17]The term comes from the fact that Hsieh & Klenow (2009) instead use the (equivalent) model of monopolistic competition in which curvature is on the demand side, and in that case, TFPR $= py_i/k_i^\alpha n_i^{1-\alpha}$.

[18]This section relies heavily on Hopenhayn & Neumeyer (2008).

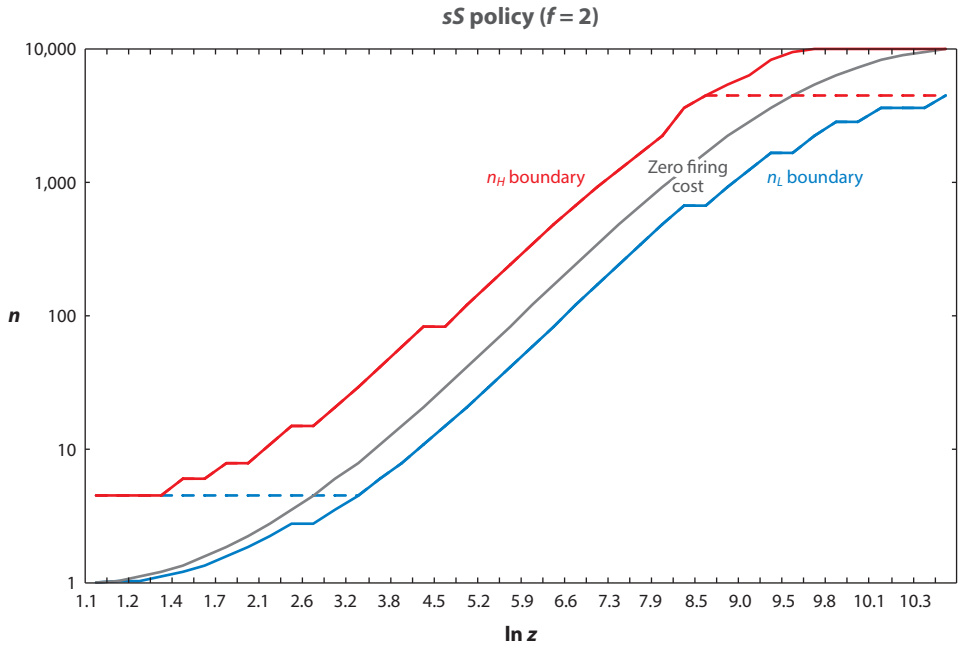

Figure 2

An example of an *sS* policy. The lower line corresponds to $n_L(z)$, the upper line to $n_H(z)$, and the middle line to the optimal employment with no distortions, which is always between the other two. The dashed lines show the restriction of the boundary decision rules to the support of the stationary distribution of productivity/employment states.

a positive implicit τ, whereas those with employment above the unconstrained level have negative τ.

To get better insight on the nature of distortions generated from layoff costs, consider the following hypothetical example. Let there be three levels of productivity, $z_1 < z_2 < z_3$, and suppose that the corresponding employment thresholds are $n_L = \{5, 8, 14\}$ and $n_H = \{9, 12, 20\}$ and the unconstrained employment levels are $n^* = \{7, 10, 18\}$. Suppose the Markov process governing firms' productivities has the property that any level of productivity can be reached after some time with positive probability from any history. This process generates a long-run distribution with the following support: $\{z_1, 9\}$, $\{z_2, 9\}$, $\{z_2, 12\}$, and $\{z_3, 14\}$.

The explanation is quite simple. (*a*) Eventually state z_3 is reached, and employment increases to 14. Once it reaches this level, it can only approach state z_1 from above (n_H), so only $n = 9$ will be observed for firms with productivity z_1. (*b*) Because state z_3 is approached only from below, employment $n = 14$ is the only compatible level in the long run for z_3. (*c*) State z_2 can be reached from state z_1 or z_3. In the first case, employment will be 9, and in the second case, it will be 12.

This implies that employment will be above the optimal level for z_1 and below the optimal level for z_2, as if firms in z_1 faced a subsidy and those in z_3 faced a tax, hence a positive correlation of wedges. Conversely, there will be firms with two implicit tax levels in state z_2 (one positive and the other negative) and correspondingly a positive variance of wedges.

To get an order of magnitude for these wedges, consider the following variant of Hopenhayn & Rogerson (1993), with no entry and exit. Let

$$y_i = z_i n_i^{0.85},$$

$$\ln z_{it} = \bar{z}(1-\rho) + \rho \ln z_{it-1} + \varepsilon_{it},$$

where ε_{it} is a lognormal random variable, independently and identically distributed (i.i.d.) across firms and time with mean zero and variance σ^2; ρ is 0.92, consistent with firm-level employment in the United States and a time unit of 5 years; σ^2 is chosen to match residuals of employment regressions in the United States (from Hsieh & Klenow 2009); and $\bar{z}$ generates an average size of firms with 50 workers in the undistorted economy similar to the US level for manufacturing (although it does not affect the analysis much). Take a firing cost f equal to two years of wages, which is not an unreasonable mean value for countries with such type of regulations. **Figure 2** shows precisely the *sS* policy corresponding to the equilibrium of this economy. The solid lower and upper lines are the n_L and n_H boundaries. The middle line corresponds to the zero–firing cost optimal employment level. As can be readily seen, there is a wide range of employment values between the two boundaries.

Table 1 illustrates the range for a few productivity values. The ratio of n_H/n_L is extremely high, reaching almost a sixfold value. This employment range can be rationalized by taxes and subsidies to employment, as described above. The tax gap, reflecting the difference between implicit subsidy and tax rates for a given level of productivity, ranges from 15% to 27%.

These gaps might seem substantial, but what are their implications for aggregate TFP? **Table 2** provides an answer to this question: The level of firing costs that we consider ($f = 2$ years) results in a 2.8% reduction in TFP. If $f = 5$ years, the TFP loss is 7.5%, and if $f = 25$ years, it is 24.3%. This level of firing costs is close to prohibitive, and firms respond by not adjusting employment at all. With such a degree of firing costs, it seems very likely that firms would negotiate and bargain with workers and induce quits to get around these barriers to adjustment. Hence, a moderate range of firing costs seems a more plausible, realistic scenario for employment rigidity and consequent TFP losses.

Firing costs are an example of the very subtle connection that might appear between policies and implicit wedges. **Figure 3** presents a scatter plot of the values of log TFPR in the support of the joint distribution of productivity and employment when firing costs equal five years of wages. Two types of distortions can be observed. On the one hand, there is a positive correlation between TFPR and productivity, indicating that less productive (more productive) firms tend to overemploy (underemploy) labor. This is what the literature has called a correlated distortion. The correlation is 0.76, as can be observed in **Table 3**, explaining 58% of the total variation. On the other hand, there is variation in TFPR for fixed levels of productivity, or what has been called uncorrelated distortions. The variance is larger for intermediate levels of productivity. This can be understood by observing **Figure 2**. Whereas the width of the bands does not seem to change much with productivity, the boundaries for the support of the long-run joint distribution

Table 1 Firing costs and tax gaps

$\ln z$	n_L	n	n_H	n_H/n_L	Tax gap
2.9	4.5	6	20.7	4.6	25%
5.2	29	59	170	5.9	27%
7.3	243	483	1,257	5.2	25%
9.3	1,677	3,607	4,454	2.7	15%

Table 2 Firing costs and total factor productivity (TFP) losses

Firing costs	TFP loss	Gap (range of equivalent labor taxes)
2 years	−2.8%	32.9
5 years	−7.5%	56.0
25 years	−24.3%	97.6

of productivity/employment (given by the dashed lines) get narrower for low and high shocks. This partly results from the bounded support for productivity shocks used in the calculations but also reflects the effect of mean reversion (recall that the AR1 coefficient used is 0.92).

Table 3 gives descriptive statistics for the log TFPR resulting from different firing costs. The results are pretty transparent. Higher firing costs (which result in a wider band of inaction) lead to higher overall variance in TFPR and an increasing importance in the role of correlated distortions.

The model calibrated in this section has no entry and exit. This might affect the results somewhat. Although most of the employment adjustment takes place for incumbent firms, young firms are the ones that exhibit the highest variance of innovations. Including entry in the model would change firm demographics, generating in a steady state a cross section of different aged firms, as described above. In addition, firms start small and tend to grow over time. This should imply that for younger vintages, employment levels near the lower boundaries are more likely. Moreover, if we were to include a higher variance of growth rates for younger firms, as occurs in the real world, the width of the bands would widen for young firms, generating

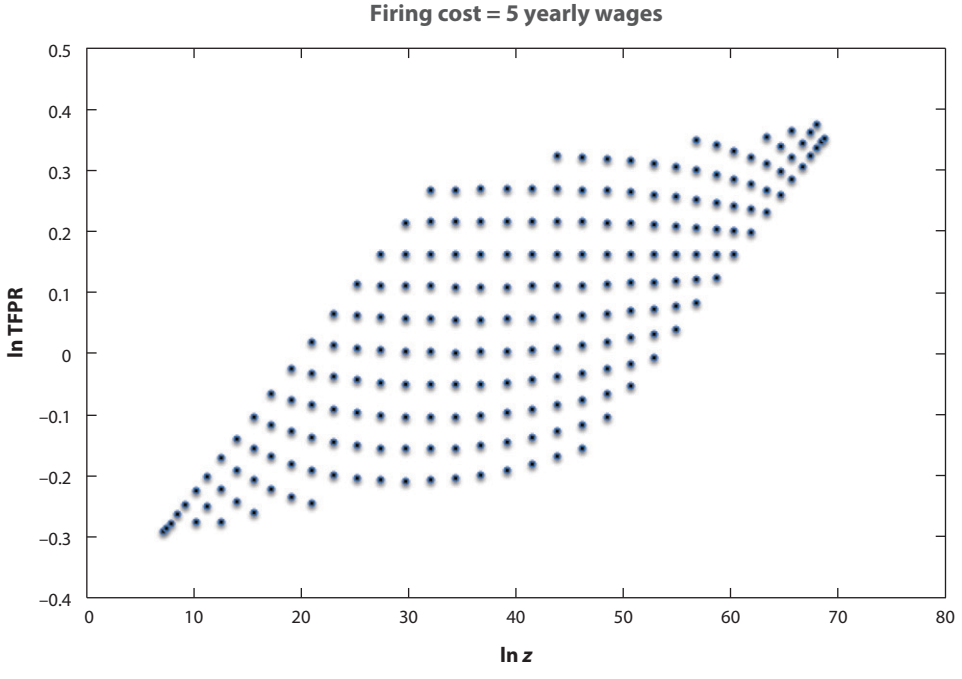

Figure 3

Firing costs: total factor productivity revenue (TFPR) and productivity.

Table 3 Firing costs and ln total factor productivity revenue

Firing costs	Standard deviation	Covariance	Correlation
2 years	0.118	1.01	0.57
5 years	0.190	2.19	0.76
25 years	0.338	5.08	1.00

higher implicit wedges. This would appear as implicit taxation to employment, as younger firms concentrate in the lower section of the *sS* band.

4.2.2. Labor taxes and exemptions. Many countries exempt small firms from certain forms of taxation. There are potentially two distortions arising from this policy. On the one hand, firms that are close to the threshold determining the exemption might choose to downsize in order to qualify. On the other hand, the exemption implies a positively correlated wedge, reducing the size of firms that comply and increasing the size of those that do not. Gourio & Roys (2014) provide very interesting evidence on this kind of policy in France and a calculation of its incidence. There is a series of exemptions to firms that have 50 employees or fewer. The size distribution of firms in **Figure 4** very clearly shows the incidence of the first effect described above, as evidenced by the rise in the number of firms right before the threshold and the large drop beyond.

Although the distortions are quite revealing, their impact on aggregate productivity appears to be relatively small. According to the calculations provided by Gourio & Roys (2014), eliminating this distortion results in a 0.3% rise in output per worker, which also disappears when considering general equilibrium effects on entry.

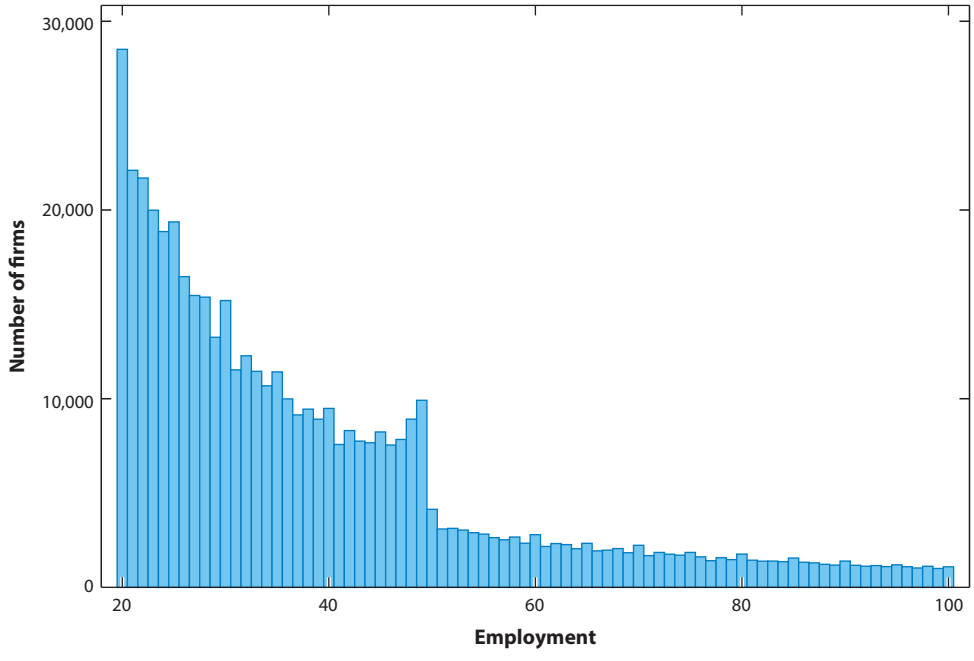

Figure 4

Firm size distribution in France. Figure adapted with permission from Gourio & Roys (2014).

Another back-of-the-envelope calculation is provided by Hopenhayn & Neumeyer (2008). This calculation considers a 30% tax on labor and noncompliance by smaller firms accounting for 50% of employment. The levels of taxes and compliance are in the right order of magnitude for many Latin American countries. The effect of this policy is to shift a considerable amount of employment away from the most efficient firms, reducing aggregate TFP by 2.4%.

In summary, the effects of the policies leading to misallocation examined above on the behavior of firms are quite noticeable, and their impact on aggregate TFP is not negligible. However, they are very far from explaining the gap in aggregate productivity. This suggests either a weakness in the models used for performing the counterfactuals or a limited scope for this type of policy in accounting for underdevelopment.

4.3. Distortions as a Primitive: A Valuable Diagnostic?

In the previous section, I explicitly consider policies that lead to misallocation and their impact on aggregate productivity; that is where the literature started. More recently, the emphasis in the literature has turned to the study and measurement of misallocation without direct reference to policies. In this section, I review two papers that motivated this literature (Guner et al. 2008, Restuccia & Rogerson 2008). Both these papers consider distortions as a primitive and explore their impact on aggregate productivity in a general equilibrium setting.

Restuccia & Rogerson (2008) use a model similar to the one described in Section 2, with the addition of capital accumulation and exogenous exit of firms. This benchmark model is calibrated to the US economy following a similar procedure as in Hopenhayn & Rogerson (1993). A series of numerical experiments is conducted that consists of introducing different firm-specific wedges on firm sales/output. In all these exercises, firms are divided in two groups, with one group taxed and the other subsidized. Firm i's profits are given by $\pi_i = (1 - \tau_i)y_i - wl_i - rk_i$, where $\tau_i \in \{\tau_T > 0, \ \tau_S < 0\}$, representing the tax and subsidy rates, respectively. Given the set of firms that are taxed and the level of τ_T, the subsidy τ_S is chosen so that there is no change in the steady-state capital stock. The exercises vary along three dimensions: (*a*) the level of the tax, (*b*) the fraction of establishments taxed, and (*c*) the productivity of the establishments taxed. For the third dimension, two cases are considered: uncorrelated distortions (taxes and subsidies chosen at random) and correlated distortions (higher quartiles of productivity taxed and the rest subsidized). These are two polar cases according to the taxonomy of distortions discussed above.

Table 4 provides some results from these experiments. The numbers represent the level of TFP in the corresponding distorted economy relative to the benchmark. For example, when 90% of the firms picked at random face $\tau = 0.2$, aggregate productivity falls by 16%. The following observations can be made: Larger distortions (higher τ) have a higher impact on TFP; the same occurs when a larger fraction of firms are taxed and when these distortions are correlated. The last column of the table considers a different exercise, in which the top 90% are taxed at 40%, but there is no

Table 4 Impact of misallocation: total factor productivity relative to no-distortion case

	Uncorrelated τ_T		Correlated τ_T		Exempt
Establishments taxed	0.2	0.4	0.2	0.4	0.4
90%	0.84	0.74	0.66	0.51	0.85
50%	0.96	0.92	0.80	0.69	0.78
10%	0.99	0.99	0.92	0.86	0.85

Table 5 Aggregate and productivity effects (undistorted = 100)

Statistic	Tax on capital		Tax on labor	
Fall in average size	10%	20%	10%	20%
Implicit tax	13.4	34.4	5.87	13.8
Aggregate output	96.1	91.9	99.9	99.5
Capital	88.8	78.7	99.9	99.5
Average managerial ability[a]	98.3	96.4	98.3	96.4
Consumption	97.8	94.8	99.9	99.5
Welfare cost (%)	0.30	1.52	0.08	0.43

[a]This number differs from the one in Guner et al. (2008) because they use $z^{1-\eta}$ in their production function instead of z.

subsidy to the rest. Compared to the preceding column, this appears to be an important difference, which is examined in Section 4.4.

Although it is somewhat hard to gauge how realistic are the distortions considered by Restuccia & Rogerson (2008), the numerical results suggest that their impact could potentially be large. The main message the literature seems to have taken from these results is that correlated distortions are the more damaging ones to aggregate productivity. This might seem intuitive at first sight, given that these distortions imply taking resources from more to less productive firms. There is a caveat to this logic, as what matters when changing the resources available to a firm is its marginal product (or TFPR) and not its TFP; furthermore, marginal products are identical in the optimal allocation. I return to this question and the explanation of the above observations in Section 4.4.

Guner et al. (2008) conduct a related exercise by considering the impact of what they call size-dependent policies, which are indeed correlated distortions. Examples of such kinds of policies are widespread, such as tax exemptions, as the ones in France discussed above; direct subsidies to small firms; and restrictions to the operation of large retail stores. To analyze the impact of these policies, Guner et al. calibrate a Lucas-style model to the US benchmark and separately consider taxes on capital and those on labor. These apply only to firms that are above the average use of the respective input. There are three margins that adjust in their exercise: (*a*) Some firms downsize to avoid taxes (as documented in Gourio & Roys 2014), (*b*) there is a correlated distortion leading to the downsizing of firms above the threshold and overexpansion of those below, and (*c*) there is additional entry. The last margin is explained by a reduction in the marginal product of labor (wage) and also by the fact that the marginal entrant, being the smallest firm in the economy in the Lucas model, is favored by this policy.[19] **Table 5** gives a summary of the results. Taxes on capital and on labor are chosen to give a 10% and 20% reduction in the average firm size, consistent with

[19]As an example, take the model with a Pareto distribution with parameter α. The average productivity of a firm is $(\alpha z_0)/(\alpha - 1)$, where z_0 is the productivity of the marginal firm, and letting m denote the firms per capita, one then finds $z_0 = m^{-1/\alpha}$. This gives an elasticity of the marginal firm's productivity with respect to m equal to $-1/\alpha$ and an elasticity of the average firm's productivity with respect to m equal to $1/(\alpha - 1)$. Because $\alpha \geq 1/(1 - \eta)$ for an equilibrium to exist, an upper bound on the absolute value of these elasticities is $(1 - \eta)$ and $(1 - \eta)/\eta$, respectively. As argued above, reasonable values of η are above two-thirds, giving bounds of one-third and one-half, respectively. Weighted by employment, this obviously would give a much smaller elasticity of average productivity, so the impact of this selection effect on output per capita is considerably smaller.

data from the OECD (Organization for Economic Co-operation and Development). When applied to capital, distortions need to be larger because of the substitution effect. In the worst scenario, in which there is 34.4% tax on capital, output falls by 8%. The large reduction in the capital stock as a result of taxation explains approximately 85% of this drop, taking into account that their parameter values imply an elasticity of output to capital close to one-third. The large expansion in the number of firms is associated with negative selection and a drop in managerial ability of 3.6% (given the elasticity of employment to z, the new marginal firm employs 22% less workers than the original one). Consumption falls by almost 5% in this exercise, which is sizable but far from the observed TFP gaps. Welfare falls by less when taking into account the adjustment path, given the drop in investment. The impact of lower distortions is obviously smaller, and labor distortions also have a smaller impact. The latter hides the fact that the level of taxes needed to induce the desired falls in average size is much smaller when the taxes apply to labor, as seen in the second row of the table.

The impact of distortions in Guner et al. (2008) is much lower than that in Restuccia & Rogerson (2008). There are several differences, but my impression is that the main one has to do with the orders of magnitude of distortions. The largest effects in Restuccia & Rogerson (2008) are obtained when a large fraction of the population is taxed and the remainder is subsidized, where the level of these subsidies is huge, as are the implied distortions. Guner et al. (2008) gauge distortions to match changes in the average firm size, a source of discipline from the data. Their model also has the feature that the employment rank of firms is not changed: Those that have higher employment in the undistorted economy continue (at least weakly) to have higher employment in the distorted one. As shown below, this is an important feature that limits the impact of misallocation.

4.4. A Measure of Distortions

This section considers the mapping from the distribution of distortions to productivity. It reproduces results developed in Hopenhayn (2012), which readers are referred to for more details. Distortions are represented above as a system of wedges, which indirectly focuses on firms' decisions. Instead, here I express distortions directly in terms of their implications for allocations. Letting $n(z)$ denote the optimal (and equilibrium) employment in the undistorted economy for firm i with productivity z_i, define a distortion as the ratio θ_i from actual employment to the undistorted one: $n_i = \theta_i n(z_i)$. Let $N(\theta) = \frac{1}{N}\sum_{i:\theta_i \leq \theta} n(z_i)$. This corresponds to the sum of employment that firms with $\theta_i \leq \theta$ would have chosen in the undistorted economy. Hopenhayn (2012) shows that the ratio of TFP in the distorted economy to that in the efficient one is given by

$$\frac{\text{TFP}}{\text{TFP}_{\text{eff}}} = \int \theta^{\eta} dN(\theta). \tag{8}$$

Notice that this formula is silent about the productivity of the firms underlying these distortions, so whether they are correlated is not important per se. Correlation matters for a different reason. For example, consider two groups of firms m_1 and m_2 with productivities $z_1 < z_2$ and optimal employments $n_1 < n_2$. Suppose further that optimal total employment in both groups is identical [i.e., $m_1 n(z_1) = m_2 n(z_2)$]. Then the impact on productivity from shifting a fixed number of workers from one group to the other is the same regardless of whether they are shifted from the lower- to higher-productivity group or vice versa.

The formula given in Equation 8 provides a natural order on systems of distortions as summarized by the measure $N(d\theta)$, as mean-preserving spreads of distortions result in lower TFP.[20] It is easy to show that correlated distortions are mean-preserving spreads of uncorrelated ones. Similarly, moving upward in each column of **Table 4** also involves mean-preserving spreads. This explains the larger effect of these types of distortions.

There is a close connection between θ and TFPR. In our benchmark model, TFPR equals labor productivity, y_i/n_i. Let $a = y_i^o/n_i^o = y^o/n$ denote the average labor productivity in the optimal allocation, where it is equated across firms. (Variables with no subscripts denote aggregate values.) Then $\text{TFPR}_i = y_i/n_i = \theta^{\eta-1}a$, so $\theta = (y_i/an_i)^{1/(\eta-1)} = (\text{TFPR}_i/a)^{1/(\eta-1)}$. An alternative formula that is closer to the one used in the literature (Hsieh & Klenow 2009, Bartelsman et al. 2013) can be derived, exploiting this connection between θ and TFPR. Substituting in Equation 8, one can easily show that

$$\frac{\text{TFP}^e}{\text{TFP}} = \left(\sum_i \frac{n_i}{n}\left(\frac{y_i/n_i}{y/n}\right)^{\frac{1}{1-\eta}}\right)^{1-\eta}. \tag{9}$$

The corresponding formula for the monopolistic competition case is

$$\frac{\text{TFP}^e}{\text{TFP}} = \left(\sum \frac{R_i}{R}\left(\frac{\text{TFPR}_i}{\overline{\text{TFPR}}}\right)^{\frac{\eta}{1-\eta}}\right)^{\frac{1-\eta}{\eta}}, \tag{10}$$

where R_i is the revenue of firm i, $\text{TFPR}_i = R_i/(k_i^\alpha l_i^{1-\alpha})$, and letting R be total revenue, $\overline{\text{TFPR}} = R/(K^\alpha N^{1-\alpha})$.[21]

5. MEASURING DISTORTIONS

Distortions are damaging for aggregate productivity, but how much of the development gap can they explain? The analyses in Section 4.3 suggest that the answer to this question depends on the extent and properties of misallocation. To answer this question, Hsieh & Klenow (2009) propose a method to measure distortions and apply it to three countries: China, India, and the United States. The method and results are reviewed here.

Equation 10 can be used to perform this calculation. The data needed are firm-level sales (revenues), capital, and labor. If output were observed, firm-level productivities can be obtained as the ratio $y_i/(k_i^\alpha n^{1-\alpha})$. Unfortunately, data on physical units of output are usually not available, so most empirical calculations of productivity use value-added instead of output in the denominator. In the monopolistic competition model analyzed in Section 2, revenues ($p_i y_i$) are proportional to y_i^η, so relative outputs of firms can be derived from their relative revenues. This allows one to identify the ratios of productivities across firms.

The calculations provided by Hsieh & Klenow (2009) use a similar methodology. In addition, they allow for sectoral differences in capital intensity and use wages to control for human capital differences across firms. Summary results from their computations are provided in **Table 6**. Standard deviations of TFPR are very large for all countries but are much more so for China and India. For reference, it is useful to compare these numbers to the ones in **Table 3** reporting standard deviations of TFPR of 0.118, 0.190, and 0.338 for an economy with firing costs of 2, 5, and 25

[20]This follows immediately from Jensen's inequality given that $\eta < 1$.

[21]In the case of monopolistic competition, R_i is proportional to the input aggregator (e.g., $k_i^\alpha n_i^{1-\alpha}$), so if there is only labor, R_i is proportional and can be replaced by n_i in the last formula.

Table 6 Total factor productivity revenue (TFPR) and misallocation

	Standard deviation TFPR	TFP^e/TFP	
		Value	Relative to United States
China, 1998	0.74	2.15	50.5%
China, 2005	0.63	1.87	30.5%
India, 1987	0.69	2.00	40.2%
India, 1994	0.67	2.28	59.2%
United States, 1997	0.49	1.43	

years of wages, respectively. The last number also corresponds to the TFPR from an allocation in which all firms are given equal employment, regardless of their productivities! The gains from eliminating distortions are equally large. Naturally, there are good reasons to expect that these numbers considerably overstate the true misallocation because of measurement and specification error, as discussed by Hsieh & Klenow (2009). For this reason, Hsieh & Klenow provide the alternative, preferred measure, which is calculating the gains from equalizing TFPR relative to those of the United States. These are reported in the last column of **Table 6**.

The gains from misallocation are still very high: China could have benefitted from a 50% increase in TFP in 1998 and India from a 40% increase in 1987. According to estimates by Hsieh & Klenow (2009), this would represent closing roughly 49% of the TFP gap between the United States and China and 35% of the gap between the United States and India. The observed improvement in China between 1998 and 2005, a period in which China's TFP grew at 6.2% per year, explains one-third of this growth. In contrast, India's dismal TFP growth (0.3% per year) from 1987 to 1994 could be partly explained by its increasing misallocation.

What explains misallocation in China and India? Our previous analysis suggests that size correlation (as would result from size-dependent policies) might be a good candidate. Another natural candidate is whether firms are public or private, as the literature has suggested that nonprofit motives in public firms would lead to overexpansion, and thus a lower TFPR. Theories of borrowing constraints typically predict that younger firms are more borrowing constrained than are older ones, suggesting that age might also be an important determinant of TFPR. Firm age effects are also implied by theories that stress the slow growth of firms stemming from difficulties in expanding markets or resources. Hsieh & Klenow (2009) consider all these factors and geography to explain the observed dispersion in TFPR.

Table 7 gives the cumulative percentage of the variance of TFPR (within industry-years) explained by the corresponding dummies. Ownership plays a more important role in China. Age explains less than 1%, and size explains between 2% and 2.5%, suggesting that correlation does not play an important role. Regional dummies do not add much either. Overall, the total variance explained by these variables is disappointingly low. Even if we consider the variance explained in China as a ratio to the difference between the variance of China and the United States, it is still

Table 7 Percent sources of total factor productivity revenue variation

	Ownership	Age	Size	Region
India	0.58	1.33	3.85	4.71
China	5.25	6.23	8.44	10.01

below 20%. As a point of reference, in our analysis of firing costs, size explains 57%, 76%, and 100% of the variance of TFPR for firing costs of 2, 5, and 25 years, respectively.

Hsieh & Klenow (2009) provide a very useful quantitative diagnostic suggesting the importance of misallocation in explaining the TFP gap. The next challenge in this literature is to find a useful taxonomy for understanding the major forces behind this large misallocation. Correlated distortions (according to size and age) explain only a small part.

More recently, Midrigan & Xu (2014), discussed in more detail in Section 6, provide a related calculation using panel data for Korea (for the years 1991–1999), China (1998–2007), and Colombia (1985–1990). There are a few differences with the calculations of Hsieh & Klenow (2009), making the comparison difficult. The most important one is that Midrigan & Xu's measure of misallocation considers only the dispersion in output per unit of capital, as opposed to the aggregator $k^{\alpha}l^{1-\alpha}$. Consequently, their counterfactual results should be interpreted as the gains from reallocating capital while keeping the same allocation of labor across firms fixed. Measured TFP losses are 22.4% for China, 17.7% for Colombia, and 16.2% for Korea. Note that the gap for China is approximately only one-fifth of the one reported in **Table 6**, suggesting that labor wedges play a very important role.

In Midrigan & Xu (2014), TFP losses accounted for by age dummies are relatively small: 0.2% in Korea, 0.3% in China, and approximately 2.7% in Colombia. This follows from the low dispersion in the average product of capital across producers of various ages that they find in their data.

Midrigan & Xu (2014) provide a very useful additional calculation. Taking deviations of each firm's TFP from its average, so as to focus on the temporary component, they compute the aggregate productivity loss that would result if firms were not able to adjust their capital stocks to these temporary shocks at all. This is an interesting bound, as it addresses some of the specification issues discussed below relating to the absence of frictions or costs to the adjustment of capital in the model. It also might be associated with informational frictions that prevent firms from setting their capital target correctly, as in David et al. (2014). The bound gives a TFP loss on the order of 2–3%, a modest contribution to explaining misallocation in these countries.

Bartelsman et al. (2013) also find large deviations in TFPR and labor productivity revenue (LPR), the ratio of revenue to labor, across a series of European countries and the United States (**Table 8**). Moreover, using an Olley-Pakes (OP) decomposition (Olley & Pakes 1996), they find

Table 8 Productivity dispersion and Olley-Pakes (OP) covariance

	SD in revenue labor productivity	SD in revenue total factor productivity	OP covariance term
United States	0.58	0.39	0.51
United Kingdom	0.59	0.42	0.15
Germany	0.71	NA	0.28
France	0.53	0.23	0.24
Netherlands	0.55	0.15	0.30
Hungary	1.04	0.92	0.16
Romania	1.05	0.55	−0.03
Slovenia	0.80	0.22	0.04

Averages are over 1993–2001 data. Industry-level firm-based TFP measures are not available for Germany. Abbreviations: NA, not available; SD, standard deviation. Data taken from a firm-level database in Bartelsman et al. (2013).

Table 9 Correlated distortions and consumption

Country	COV_LPR (data)	COV_LPR (model)	STD_LPR (data)	STD_LPR (model)	STD_TFPR (data)	STD_TFPR (model)	Consumption index (model)
United States	0.51	0.51	0.58	0.75	0.39	0.47	1.00
United Kingdom	0.15	0.15	0.59	0.66	0.42	0.69	0.93
Germany	0.28	0.28	0.71	0.59	NA	0.64	0.97
France	0.24	0.24	0.53	0.60	0.23	0.66	0.96
Netherlands	0.30	0.30	0.55	0.59	0.15	0.63	0.97
Hungary	0.16	0.16	1.04	0.65	0.92	0.69	0.93
Romania	−0.03	−0.03	1.05	0.72	0.55	0.70	0.88
Slovenia	0.04	0.04	0.80	0.70	0.22	0.70	0.89

Data taken from a firm-level database in Bartelsman et al. (2013). Abbreviations: LPR, labor productivity revenue; NA, not available; TFPR, total factor productivity revenue.

that the covariance between a firm's labor productivity and its employment is quite high and is even higher in the United States, where it accounts for 51% of the total variation. Bartelsman et al. (2013) interpret this fact as a misspecification in their benchmark model (similar to the ones considered in Hsieh & Klenow 2009) and provide a remedy by introducing fixed costs in the form of overhead labor. In the absence of distortions, average productive LPR measured as the ratio of revenue to production workers (total workers minus overhead) is equated across firms, while LPR will vary. Letting f denote overhead labor, we can rewrite LPR as follows:

$$\frac{p_i y_i}{n_i} = \frac{p_i y_i}{n_i - f} \frac{n_i - f}{n_i}.$$

The first term on the right-hand side is equated across firms. The second term is increasing in n_i, implying that LPR increases with firm size, measured by employment. In addition, the model has the feature that capital is adjusted prior to observing the realization of a contemporaneous productivity shock. Whether this is a technological, institutional, or informational assumption, it is meant to capture some of the variation in TFPR and LPR.

Bartelsman et al. (2013) calibrate all parameters and the overhead cost f to match similar moments as used in previous papers in addition to the moments given in **Table 8** for the US economy. Whereas the OP covariance term is matched exactly, the resulting standard deviations of LPR and TFPR are somewhat high. Overhead labor ends up representing approximately 14% of total employment. Given that the average establishment size in manufacturing was about 45 workers in the United States at this time, this implies a calibrated f of almost 7 workers. Correlated distortions are introduced for each country to match their corresponding OP covariances. Results are presented in **Table 9**. These correlated distortions give rise to fairly large reductions in consumption per capita, ranging from a 3% gap for Germany and the Netherlands to close to 12% for Romania and Slovenia.[22]

[22]Unfortunately, Bartelsman et al. (2013) do not provide information to decompose this fall in consumption into changes in the capital stock, the number of firms, and TFP.

6. FINANCIAL CONSTRAINTS: MISALLOCATION AND SELECTION

Financial frictions have been one of the favorite candidates for explaining the development gap (Banerjee & Duflo 2005, Jeong & Townsend 2007, Banerjee & Moll 2010, Buera et al. 2011, Buera & Shin 2013, Caselli & Gennaioli 2013, Midrigan & Xu 2014). There is also considerable supportive evidence of credit constraints as a source of misallocation in developing economies, as summarized by Banerjee & Duflo (2005). **Figure 5** shows scatter plots relating development measures to the ratio of external finance to GDP. All measures of development are highly correlated with this measure of financial development: The correlation coefficient is 0.34 for GDP/worker, 0.26 for TFP, and 0.76 for the capital-output ratio. The evidence shows that financial development is associated with capital deepening, as well as improvements in the allocation of capital. These observations have motivated the above-referenced papers that attempt to establish a causal link and provide orders of magnitude about its strength. This section reviews some of the most recent contributions.

6.1. Theory

There are three main channels considered by the literature. Financial development (in the form of broader access to lending for capital financing) leads to (*a*) capital deepening as measured by higher capital/output ratios, (*b*) reduced misallocation (i.e., a better allocation of capital across firms), and (*c*) better selection of active firms. Increased capital deepness contributes to GDP per capita, but only the last two channels have a direct effect on TFP and are the focus of most of the following discussion.

I follow Moll's (2014) analysis to illustrate the impact of financial constraints on TFP. The production function is given by $y_i = z_i k_i^{\alpha} n_i^{1-\alpha}$. Given the absence of decreasing returns, only the most productive firms are active in the optimal allocation. With wages w and cost of capital r, it

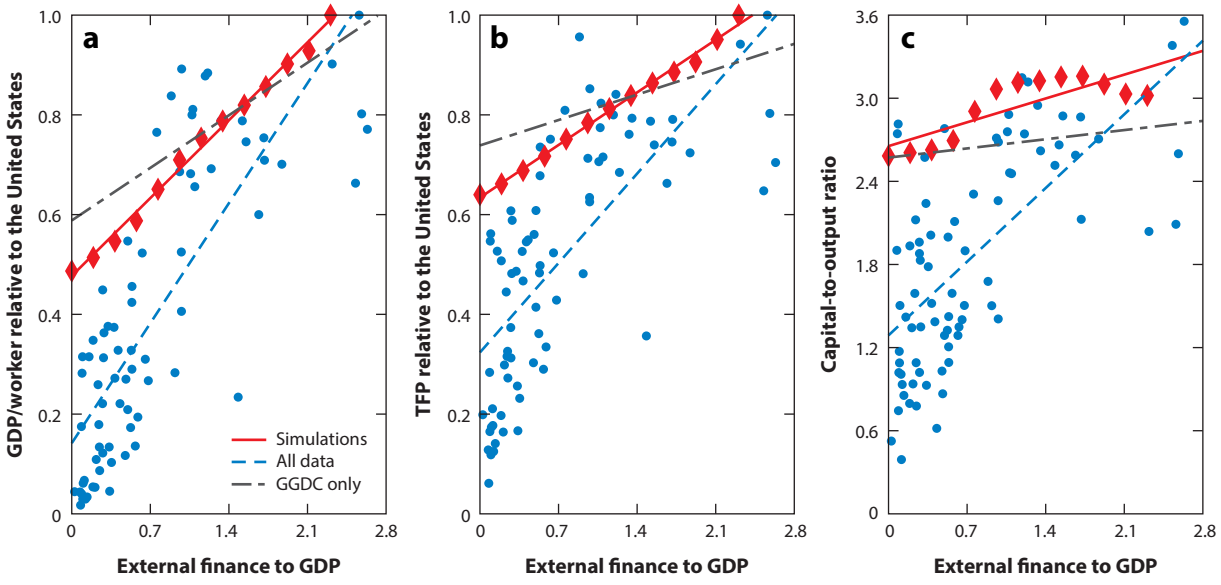

Figure 5

Financial frictions and development. Abbreviations: GGDC, Groningen Growth and Development Centre; TFP, total factor productivity. Figure adapted with permission from Buera et al. (2011).

easily follows that profits are of the form $z_i^{1/\alpha} k_i \pi(w) - r k_i$, where $\pi(w) = \alpha\big((1-\alpha)/w\big)^{(1-\alpha)/\alpha}$ is decreasing in the wage rate w.

Entrepreneurs are constrained in the choice of capital by their assets a_i according to $k_i \leq \lambda a_i$, where $\lambda \geq 1$. Here $\lambda = 1$ corresponds to a total shutdown of lending, so each entrepreneur can produce with his or her own assets only. Borrowing constraints bind for all active entrepreneurs owing to the assumption of constant returns to scale. In this simple environment, entrepreneurs do not have an alternative use of their time, so they produce if profits are positive (including the opportunity cost of assets). This implies that those above a threshold $\underline{z}$ are active. This threshold is easily derived by equating demand to supply in the capital market: $\int_{\underline{z}} \lambda a \, dG = K = \int a \, dG$, where $G(a, z)$ is the joint distribution of assets and abilities of entrepreneurs. The wage w is determined by market clearing in the labor market, and r is chosen to make the threshold entrepreneur indifferent between operating as a producer and lending his or her assets at rate r.

This economy also has a simple expression for the aggregate production function, $y = AK^{\alpha}N^{1-\alpha}$, where

$$A = \left(\int_{\underline{z}} \frac{a}{A(\underline{z})} z^{1/\alpha} \, dG \right)^{\alpha},$$

and $A(\underline{z})$ are the total assets of active entrepreneurs. TFP is thus the geometric mean of productivities but is weighted by the corresponding share of assets. An increase in the correlation of assets and productivities increases TFP by raising this mean and potentially also increasing $\underline{z}$ (more precisely, an increase in the affiliation between a and z).[23] A higher value of λ corresponds to less financial constraints and results in an increase in $\underline{z}$ and the amount of external finance, which equals $1 - A(\underline{z})$, together with an increase in TFP. This increase in TFP results both from reduced misallocation (capital is reallocated to firms with higher marginal product) and from better selection. However, these two effects are hard to disentangle because of constant returns to scale.[24] The increase in $\underline{z}$ also implies a reduction in the number of firms and consequently an increase in the average size measured by employment (also if measured by capital or value-added). Hence, less financial constraints can also contribute to explaining the higher average firm size in developed economies. This point has been raised and analyzed by Quintin (2008b).[25]

6.1.1. Endogeneity of assets. The above analysis treats assets exogenously, but in a dynamic setting, they evolve as the result of the reinvestment of past cash flows. This suggests that with sufficient incentives to save (e.g., proper values for discount factors of entrepreneurs), financial constraints might disappear in the long run. If productivity shocks are fully persistent (i.e., $z_{it+1} = z_{it}$ for all i and t), financial constraints disappear in the long run, and the economy converges

[23]This is not necessarily the case with decreasing returns, as an increase in the correlation could put too many assets in the hands of a few firms with high TFP but low marginal product, and they might have difficulties lending the excess assets when λ is low.

[24]Although the reduction in misallocation would seem to be a robust implication of a reduction in financial frictions, positive selection might not.

[25]Quintin (2008b) also calibrates a model of limited enforcement to the US economy and shows that when matched to measures of financial development in Argentina and Mexico, the model accounts well for differences between the size distributions of these countries and those of the United States.

to an efficient steady state. This follows immediately from the fact that constrained firms face higher rates of return to savings and will continue to do so and increase their capital until they become unconstrained (Banerjee & Moll 2010, Moll 2014).[26] Note also that higher-productivity firms, with higher profits, will increase their levels of assets faster, so as time goes by, the share of assets will become more correlated with productivity. In the case of constant returns to scale analyzed above, this implies both that stronger selection will take place over time and that capital will become more correlated with firm productivity, contributing to a further increase in TFP. Finally, although the steady state has an efficient allocation of capital under full persistence, the speed of convergence to this steady state could be quite slow. As shown by Moll (2014), the speed of convergence decreases with the level of persistence ρ.

At the other extreme, consider the case in which productivity shocks are i.i.d. Although firms will differ in the level of assets according to their past history of shocks, independence implies that assets and productivity are contemporaneously uncorrelated. As shown by Moll (2014), this implies that TFP will be constant in the above model, so the gap with the efficient level of TFP will never be reduced. Although this will not hold exactly in the case of decreasing returns, the result is still very useful as a benchmark.[27]

6.1.2. Limited enforcement. Most papers in the literature derive borrowing constraints from limited enforcement of lending contracts. To illustrate the mechanism, consider a one-period-lived entrepreneur with assets a who needs to incur in debt $D = k - a$ to produce with capital input k according to production function $zf(k)$. At the end of the period, the agent can choose to default and not pay the debt D but then loses a fraction ϕ of revenues. This puts a limit on lending as given by the no-default constraint,

$$(1 - \phi)zf(k) \leq zf(k) - R(k - a),$$

or, more conveniently, $R(k - a) \leq \phi zf(k)$; that is, the limits to borrowing are given by foregone revenues at default. This defines implicitly a constraint on capital $k(a, z, \phi)$ that is increasing in ϕ, z, and a. Moreover, it is easy to see that $\partial k/\partial a > 1$ owing to increased leverage.[28] The constraint ceases to bind after some $\phi < 1$ that makes it possible to choose a level of capital such that $zf'(k) = R$. For the linear production case, one finds that $k(a,z,\phi) = (Ra)/(R - \phi z)$; the positive interaction between ϕ and z implies that better enforcement (higher ϕ) increases capital more to high-productivity firms than to low-productivity ones. It also follows that better enforcement also increases the share of capital of higher-z firms, thus reducing misallocation and increasing TFP. This property also holds with decreasing returns and a homogeneous production function. Interestingly, in the latter case, TFPR $= y/k$ is increasing in z for fixed a, so better enforcement also increases TFP when all firms/entrepreneurs have the same level of assets.

[26] Again, misallocation might not improve monotonically in the case of decreasing returns. As an example, take an economy in which firms are self-financed ($\lambda = 1$) that starts with a distribution of assets in which the marginal product of capital is equated across firms but all are below the efficient level. More productive firms will save more as a result of their higher profits, and it is unlikely that marginal products of capital continue to equalize.

[27] Buera & Shin (2013) make a related point in their numerical analysis. Their results show that the speed of convergence depends on whether the economy converges from a distribution that has a similar correlation between wealth and ability, but with a lower mean, or whether it starts with a different correlation. In the former case, λ plays a less important role for the speed of convergence.

[28] Interestingly, when $f(k) = k^\alpha$ and $\alpha < 1$, TFPR $= y/k$ is also increasing in z for fixed a.

6.2. Quantifying the Effect of Financial Constraints

There are numerous papers that consider the quantitative impact of financial constraints. I refer to a very small subset of these (Amaral & Quintin 2010, Buera et al. 2011, Caselli & Gennaioli 2013, Midrigan & Xu 2014) owing to space limitations and to focus on models that translate more clearly into the above structure. I first briefly discuss the characteristics and results of these papers. **Table 10** provides a summary of these models; their distinguishing characteristics; a quantification of the impact of financial constraints on GDP per capita, TFP, and K/Y; and observations regarding the roles of misallocation and selection in explaining results. The models are quite different in several dimensions, so it is hard to make comparisons. However, in a broad sense, the forces at play are the ones discussed above: misallocation, selection, and capital deepening. Financial development reduces misallocation by contributing to more efficient resource allocation across firms. At the same time, financial development contributes to the participation of the most efficient entrepreneurs/firms, the selection effect. Finally, financial development contributes to capital deepening (an increase in K/Y), which can have important effects on GDP per capita.

6.2.1. Caselli and Gennaioli. Caselli & Gennaioli's (2013) main motivation is the prevalence of dynastic-family firms in less developed countries. Although mitigating financial constraints, this type of arrangement often puts firms in the wrong hands, and this is costly for productivity. In Caselli & Gennaioli's model, agents live for one period (generation) and bequest a fraction of their wealth and firms (in the case of entrepreneurs) to a single offspring, who in addition is born with a random endowment of wealth. The set of firms is fixed, so there is no entry. Agents differ in managerial skills (low, high), which are mildly correlated across generations. A market for corporate control in which agents buy and sell firms helps to mitigate managerial mismatch. Technology is given by a constant returns to scale production function, and limited enforcement constrains capital acquisition.

This model shares the feature that more productive entrepreneurs can borrow more and benefit relatively more from improvements in financial development. In addition, financial development boosts the market for corporate control. This happens precisely because the value of transferring firms from low- to high-productivity entrepreneurs increases with better enforcement. Finally, as the average quality of managers increases, less firms are needed, and the average firm size increases. This effect is explained in Section 6.1.

The main counterfactual exercise in Caselli & Gennaioli (2013) is comparing an economy with essentially no financial constraints (the US-calibrated economy) to one with no lending. GDP falls by 70%, to a great extent because of the fall in K/Y but also as the result of a 21% drop in TFP. According to Caselli & Gennaioli's discussion, at least half of the TFP fall is explained by selection and the rest by misallocation. Given that entrepreneurs live for only one period, the scope for saving out of borrowing constraints is very limited in the model. Increasing intergenerational persistence from the benchmark value of 0.3 to 1 reduces the TFP drop to only 6%, suggesting that a more realistic number should be somewhere in between. The market for corporate control plays a very important role: In the absence of this market and an almost perfect capital market, Caselli & Gennaioli (2013) find that TFP would be 15% lower than in the benchmark.

6.2.2. Amaral and Quintin. In Amaral & Quintin (2010), agents live for three periods, are born with the same level of assets, and leave no bequest. They work as labor in the first period, choose occupations (worker/entrepreneur) in the second one, and retire in the third. There is a distribution of entrepreneurial productivity as in the Lucas model, and because all agents have the same wealth, those above a certain threshold become entrepreneurs. Financial constraints are in the form of limited enforcement. A storage technology is also available, which gives another use for capital.

Table 10 Selected models of financial constraints

Model	Model characteristics	Effect on GDP	Effect on TFP	Effect on *K/Y*	Observations
Caselli & Gennaioli (2013)	One-period lived with bequest, constant returns to scale, limited enforcement, market for corporate control	−70%	−21% (30% of 10th–90th gap)	75% of 10th–90th gap	With full intergenerational persistence, only a 6% fall in TFP. At least half gain from selection. Market for corporate control plays important role in mitigating misallocation.
Amaral & Quintin (2010)	OLG three periods; occupational choice in period 2; no bequest, all same initial wealth; span of control (Lucas style); productive sector and alternative storage technology	−66.6%	−50%; −30% in productive sector	Falls by 80% in productive sector	Accounts for one-third variation in output per capita in sample of 25 countries (mostly middle and high income). Number of firms increases by five in the productive sector.
Buera et al. (2011)	Infinitely lived, two sectors (service, manufacturing), managerial talent uncorrelated across sectors, 10% death rate (redrawing productivity), higher fixed costs in manufacturing	−55%	−35% overall; −26% in service; −50% in manufacturing	Falls by ~20%; explained by increase in price of capital	TFP fall in service mostly a result of misallocation. Half of TFP fall in manufacturing is explained by selection. Overall misallocation explains ~25% of the 35% drop in TFP. Accounts for two-thirds of country variation in GDP per capita and ~60% of TFP variation.
Midrigan & Xu (2014)	Infinitely lived; traditional (no capital, low productivity) and modern sectors; borrowing constraints for capital (modern sector), limited leverage; entry cost to modern sector; can issue equity when entering modern sector; permanent and transitory shocks; new entry of entrepreneurs each period (similar to exogenous death rate)	−35%	−17% in modern sector	Falls by 32%	At least half of GDP fall explained by lower *K*. TFP fall in modern sector mostly results from fall in the number of firms. Misallocation is responsible for 4.7% of TFP fall in modern sector.

Some of the numbers in the table are my own estimates based on graphs in the papers. Abbreviations: OLG, overlapping generations; TFP, total factor productivity.

The model is calibrated to the US economy, and the experiment is to compare this benchmark to the no-lending case. GDP falls by two-thirds, while productivity falls by 50%. Productivity in the productive sector falls by 30%. There is a large reduction in total capital stock, but most importantly, there is a very large shift from the productive sector to storage technology, explaining the difference between the aggregate fall in TFP and the one in the productive sector. The negative effect of financial frictions on TFP in the productive sector operates through two channels: selection (the number of firms is multiplied by five, reducing average firm productivity) and misallocation. From Amaral & Quintin's (2010) data, it is hard to disentangle the relative importance of these two sources. In addition to this exercise, the authors choose enforcement levels to replicate the level of financial development in a sample of 25 countries (mostly middle and high income). The levels of GDP per capita predicted by the model explain roughly one-third of the actual variation.

6.2.3. Buera et al. Buera et al. (2011) consider a two-sector model (services and manufacturing), extending an occupational choice model of the Lucas style. Agents are infinitely lived and endowed with abilities as entrepreneurs in both sectors. Given these abilities and their wealth levels, agents choose whether to work as labor, as entrepreneurs in service or in manufacturing. One of the key differences between the two sectors is that fixed costs are considerably higher in manufacturing. Abilities are redrawn from a fixed distribution, independent of existing values, with a constant probability, implying de facto exogenous exit rates from entrepreneurship of approximately 10%. Borrowing constraints are the result of imperfect enforcement, as in the two previous models.

The model is calibrated to the United States, considered as an economy with perfect enforcement. As in Amaral & Quintin (2010), enforcement constraints are chosen to mimic the degree of financial development of countries in their sample. Results are presented in **Figure 5**. The model accounts for two-thirds of the relationship between financial development and GDP per worker in the data and 60% of the corresponding variation of TFP. Comparing the United States to the no-lending case leads to a 55% fall in GDP and 35% fall in TFP.

Borrowing constraints affect the two sectors differently. While TFP falls by approximately 26% in service, the drop is 50% in manufacturing. In the case of service, Buera et al. (2011) show that the drop almost entirely results from misallocation, whereas in manufacturing, at least half of the gap is explained by selection into entrepreneurship. Average entrepreneurial talent falls by 20% in service and 40% in manufacturing. Overall misallocation seems to account for close to 25% of the 35% aggregate TFP drop.

6.2.4. Midrigan and Xu. In Midrigan & Xu (2014), firms produce a homogeneous consumption good with two alternative technologies: traditional and modern.[29] Debt can only be issued in the modern sector and is constrained by assets, as in Moll (2014). This affects only the modern technology, as the traditional one does not use capital. In addition, firms have to pay an entry cost to produce with the modern technology, which is more productive. Entrepreneurs are infinitely lived, but a fraction of new ones enter each period, calibrated to 8% in the benchmark.[30] As a result, the steady-state age distribution of firms is nontrivial. This can be important for misallocation as younger firms (that have not been able to accumulate assets) face more borrowing constraints than older ones. At birth, entrepreneurs draw a permanent and transitory productivity

[29]Here traditional and modern can be interpreted also as two different sectors, similar to the informal/formal or services/industry cases considered above.

[30]I believe that the implications of this assumption for misallocation are very similar to assuming an exogenous death rate and entry to maintain a constant population of firms.

shock, in which the sum corresponds to the idiosyncratic shock considered above. The transitory shock follows a Markov process that is calibrated to have fairly low persistence. All new entrants have no assets, start with the traditional technology, and can adopt the modern one after one period by paying the entry cost. In addition, when entering the modern sector, firms can sell claims to future profits up to a limit that is treated parametrically.

The model is calibrated to the Korean economy, matching a series of moments on entry, size distribution, and growth of firms in addition to more conventional parameters, such as discount factors, decreasing returns, and factor shares that are assigned the usual values. The modern technology has 20% higher productivity than the traditional one, and given the rest of the parameters, this implies that firms employ five times more workers and have 20 times higher profits. Temporary shocks have very low persistence (an AR1 coefficient of 0.25). The variance of the innovations is considerably lower than the variance of the permanent component, which accounts for 85% of the cross-sectional distribution of productivity.

There are several interesting counterfactuals considered in Midrigan & Xu (2014). I focus on the impact of shutting down lending. As a result, 83% firms become constrained, and the capital/income ratio falls by 32%. GDP per capita falls by 33.5%, at least half explained by the reduction in the capital stock. K/Y falls from 2.7 to 2.1. There is a huge decrease in the number of producers in the modern sector, from a 93% share of total firms to 35%. This reallocation explains about two-thirds of the fall in aggregate TFP, whereas misallocation only contributes to a loss of 4.7% in the TFP in the modern sector. Age-related distortions (young firms have 73% higher TFPR than old ones) explain 3.7% of this drop, and the lack of adjustment to temporary shocks accounts for very little.

6.2.5. Overall assessment. The models given above are hard to compare, as they have quite different features. As seen, the impact of borrowing constraints depends critically on the ability of firms to save out of them. This is affected both by the level of persistence and by the expected lifetime of firms. In Caselli & Gennaioli (2013), there is a very low correlation in productivity across generations and no reinvestment of savings within. In Amaral & Quintin (2010), entrepreneurs last for only one period, and there are no bequests. Not surprisingly, these are the two papers for which the effects of borrowing constraints on GDP and on capital accumulation are the strongest. Effective persistence is considerably lower in Buera & Shin (2013) and Midrigan & Xu (2014) but is similar between the two, as there is a 10% turnover rate in the former and approximately 8% entry in the latter. These two models may still underestimate the role of persistence in the data because neither of them considers the impact of selection due to the exit of firms. It is well documented that death rates are much higher among young firms than old ones. Both the latter models assume away this effect and thus tend to overrepresent younger firm cohorts in the stationary distribution.

All the papers reviewed above support the view that borrowing constraints are important factors that contribute to explain GDP and TFP gaps. The two main channels are the effects on the aggregate capital stock and those on productivity. The range of values for the impact on GDP goes from approximately 35% to 70%. As a reference, according to Caselli (2005), the corresponding lower/upper 10-percentiles ratio is 0.05 or a range of 20 to 1. The range of values for TFP impact goes from approximately 20% to 50%, with the corresponding percentile ratio on the order of 0.2–0.3. Even if we take the lower range of values corresponding to Midrigan & Xu (2014), financial constraints can account for 35% of the 95% gap in GDP per capita and 20% of 70–80% gap in TFP.

The share of misallocation in explaining the fall in TFP varies somewhat across studies, ranging from 4.7% of the TFP fall in Midrigan & Xu (2014) to closer to 25% in Buera et al. (2011). The message here seems mixed, for the first number would suggest that borrowing constraints are not

Table 11 Standard deviations of firm input and revenue growth

	Inputs	Revenue	ε_{IR}	ε_{RI}
China	0.45	1.00	0.98	0.82
India	0.28	0.70	0.96	0.90
United States	0.68	0.43	1.01	0.82

Table reproduced from Hsieh & Klenow (2009) by permission of Oxford University Press.

an important source for misallocation, whereas the latter suggests they are.[31] In contrast, all these papers support the view that selection considerations, or the mismatch of managerial talent, are a very important channel by which financial constraints impact TFP.

7. OTHER CHANNELS

There are many other channels that can explain misallocation, selection, or more broadly the impact of institutions/policies on the distribution of firm productivities. One of them involves markups that vary widely across industries and countries. Two recent papers (Epifani & Gancia 2011, Peters 2013) consider the impact of variation in markups on misallocation. Peters (2013) considers the impact of reducing trade barriers on the dispersion of markups. Although this seems to be an important source in determining markups, the effect on aggregate TFP is very small. Epifani & Gancia (2011) find slightly larger effects when considering the welfare effect of markup variation, with estimates on the order of 3.5–10%.

The analysis in Hsieh & Klenow (2009) relies on a series of structural assumptions that are critical to the inference. I briefly consider three sources: (*a*) measurement error, (*b*) curvature, and (*c*) adjustment costs.

7.1. Measurement Error

Although arbitrary forms of measurement error cannot be ruled out as explanations, Hsieh & Klenow (2009) analyze the case in which measurement error is orthogonal to the truth and to other reported variables. Measurement error of this type will have predictable effects: (*a*) It will decrease the correlation between revenues and inputs, and (*b*) it will increase the standard deviation of firm revenue and input growth. **Table 11** gives estimates of the correlations (elasticities of inputs to revenues and vice versa) and standard deviations of growth. Elasticities are similar across the three countries, suggesting that this form of measurement cannot explain much of the observed differences in the dispersion of TFPR. The analysis of standard deviations gives a mixed message: Although the United States has a larger standard deviation of inputs, it exhibits a lower standard deviation of revenues. Under the assumption of equal driving processes for the true variables, this would require different orderings across countries for the measurement error in the two variables.

Finally, if measurement error is less persistent than are true variables, instrumenting with past values would reduce TFPR dispersion and thus the gains from eliminating distortions. According

[31] In addition, borrowing constraints imply misallocation that is highly correlated with age, which, as seen above, has very little explanatory power in Hsieh & Klenow (2009).

to Hsieh & Klenow (2009), the relative decrease in TFPR obtained when doing so is relatively larger for the United States than for China and India, suggesting, if anything, that the bias from measurement error is stronger for the United States.

7.2. Curvature

An important part of the structure is the curvature η in the production/demand, a combination of the degree of decreasing returns and demand elasticity. This is hard to identify, and several values have been used, ranging from the equivalent of $\eta = 2/3$ ($\sigma = 3$) in Hsieh & Klenow (2009) to $\eta = 0.85$ in Restuccia & Rogerson (2008). I provide here a new comparative static result. To examine the effect of curvature on the TFP gap, Equation 9 can be rewritten as follows:

$$\left(\frac{\text{TFP}^e}{\text{TFP}}\right)^{\frac{1}{1-\eta}} = \left(\sum_i \frac{n_i}{n}\left(\frac{y_i/n_i}{y/n}\right)^{\frac{1}{1-\eta}}\right). \tag{11}$$

The ratio of efficient to actual TFP in this expression is a certainty equivalent of a lottery given by the ratio in brackets in the summation under CRRA (constant relative risk aversion) utility function with exponent $1/(1-\eta)$. An increase in η decreases risk aversion (it actually increases risk love), thus increasing the certainty equivalent on the left-hand side of the equation. Consequently, an increase in η will raise TFP^e/TFP, magnifying the effect of distortions. Consistent with my comparative static result, Hsieh & Klenow (2009) show that increasing σ from 3 to 5 (η from two-thirds to three-quarters) raises China's hypothetical TFP gain (from equalizing TFPR) from 87% to 184% and India's from 128% to 230%.

7.3. Adjustment Costs

Adjustment costs introduce a wedge on input choices, giving rise to differences in TFPR across firms. An example of this is given in Section 4.2. The sluggish adjustment in inputs translates into a lower variance of inputs than revenues, as evidenced for China and India in **Table 11**. In the presence of adjustment costs, a higher volatility of fundamentals (firm productivity or demand shocks) and lower persistence can lead to lower variation in input response. Hsieh & Klenow (2009) do not seem to find large differences in volatility of input choices across the three countries, suggesting that firms in India and China face greater barriers to reallocation.

This idea is pursued further in a recent paper that quantitatively examines the role of adjustment costs on misallocation (Asker et al. 2014). Their main idea is that adjustment costs can lead to quite different degrees of misallocation, depending on the variability of firm-level revenue shocks. Using panel data from several countries and following a structural approach similar to Hsieh & Klenow (2009), the authors compute TFP/revenue shocks and marginal (revenue) products of capital (MRPK) for all firms and time periods in the data.[32] Their data sets exhibit great differences in the degree of volatility of firm-level TFP growth across countries (e.g., France has twice, and Slovenia has three times, the standard deviation of the United States). Reduced-form evidence also shows a clear positive relationship across countries and across industries between the degree

[32]Asker et al. (2014) use the term TFPR to describe what Hsieh & Klenow (2009) call TFP (or revenue shocks), while measuring distortions by the MRPK. In the absence of distortions to the choice of other inputs, this measure is proportional to TFPR as defined by Hsieh & Klenow (2009).

of volatility of firm productivity and the standard deviation of MRPK across firms. Furthermore, firm-level MRPK increases with productivity shocks, with estimated elasticities ranging from 1.07 to 1.65 (1.29 for the United States). Investment responds positively to productivity shocks, but the elasticity is quite low (0.3), whereas in the absence of distortions, it should be on the order of 4 (given their calibration of demand elasticity).

Asker et al. (2014) structurally estimate a model of firm dynamics with two types of adjustment costs (fixed and convex) in addition to one month time to build. The model is estimated taking the following as identifying moments for each industry/country pair: (*a*) the fraction of firms with year-to-year growth in their capital stock of less than 5%, (*b*) the fraction adjusting over 20%, and (*c*) the standard deviation of the capital growth of firms. Interestingly, most countries in their sample exhibit larger adjustments in firm-level capital than does the United States. Their estimates give quite large adjustment costs (for the United States, fixed adjustment costs equivalent to one and one-half months of production and convex adjustment costs 8.8 as large as investment costs when doubling the capital in a period).

Asker et al. (2014) simulate the model with identical adjustment cost parameters for all countries (the US values), feeding for each country/industry pair the corresponding estimated parameters for the productivity (revenue) process,[33] comparing the limiting distribution with cross-country/industry statistics. As a measure of fit, they report

$$S^2 = 1 - \frac{(\mathbf{x} - \hat{\mathbf{x}})'(\mathbf{x} - \hat{\mathbf{x}})}{\mathbf{x}'\mathbf{x}},$$

where $\mathbf{x}$ is the actual country/industry vector for the statistic evaluated, and $\hat{\mathbf{x}}$ is the simulated one. The first statistic considered is the dispersion of MRPK across firms in the country/industry cells, with values of S^2 on the order of 0.80 (and ranging from 0.6 in Spain to 0.99 in Chile). Interestingly, repeating the exercise, but with only one month time to build (setting to zero the other adjustment costs), gives a value of S^2 on the order of 0.65 (0.5 in Spain and 0.79 in Chile). It is quite surprising that one month time to build by itself can account for such a large fraction of the variation. In contrast, whereas US level adjustment costs explain 90% of the dispersion of the growth in firms' capital, time to build alone accounts for almost none (although for some countries it is high; e.g., 0.9 for France). Although there is no final verdict as to the importance of adjustment costs as a source of misallocation, Asker et al. (2014) provide some compelling evidence and will surely impact more research to come.

8. LEARNING FROM THE SIZE DISTRIBUTIONS OF FIRMS

As discussed, one of the most striking differences across countries is the size distribution of firms. Above we examine the potential importance of the differences in average size across countries. What else can we learn from the size distributions? In particular, what can we learn about misallocation from the size distribution of firms?

Two recent papers consider this question (Alfaro et al. 2009, Hopenhayn 2012). Some identifying assumptions are necessary.[34] The identifying assumption used in these papers is to take a reference economy (the United States) and assume that it is undistorted and that other

[33]Different specifications allow country/industry-specific parameter estimates of the production function and also adjustment costs.

[34]For example, the same size distribution in an economy with no distortions can be obtained by flipping the sizes of firms above and below the median productivity.

countries share the same distribution of productivity, so all differences in size distributions can be attributed to misallocation. This is still not enough, as there could be arbitrary distortions that map the US size distribution onto that of other countries. The procedure followed is to look for the least damaging distortions that do so, thus obtaining a lower bound on their impact on aggregate TFP. The general procedure is to match quartiles of the two size distributions with the proper values for θ's, using the notation in Section 4.4 (e.g., if the number of firms is the same, assume that the largest firm in each country has the same productivity as the largest one in the United States, and so on). Equivalently, the lower bound on distortions is obtained by preserving the rank of firms: Considering two firms a and b, if the first is larger than the second one in the distorted economy, assume that it has higher productivity so that it would also be larger in the undistorted one.

In performing their accounting, Alfaro et al. (2009) consider the combined effect of the difference in the number of firms and distortions on a country's GDP per capita, taking the country-specific factor

$$D = \frac{\left[\sum_{i=1}^{N} M(1-\tau_i)^{\sigma-1} A_i^{\sigma-1}\right]^{\sigma/(\sigma-1)}}{\sum_{i=1}^{N} M(1-\tau_i)^{\sigma} A_i^{\sigma-1}},$$

where M is the number of firms per worker. According to their results, variation in D accounts for 16% of the variation in ln GDP per capita. There is no direct information given on the variation of M across these economies, but looking at the dispersion from the figures, a variance in ln M on the order of 0.5–1 does not seem unreasonable.[35] Their paper takes $\sigma = 6$, so the variation in M with an elasticity $1/(\sigma - 1) = 1/5$ seems to account for a very large part of the observed variation in D.

The exercise conducted in Hopenhayn (2012) is more limited, as it considers only four countries: China, India, Mexico, and the United States. Using a value of $\eta = 1/2$ in our benchmark model, the lower-bound misallocation accounts for a 1% fall in TFP for China and an approximately 7% fall for India and Mexico. These numbers are halved if we instead use the standard value $\eta = 0.85$.[36] Recall that the lower bound is obtained assuming that distortions preserve the rankings of firm size. It thus follows from this exercise that if distortions are to be consistent with the observed size distributions, they must involve very large rank reversals in firm size to have a large impact on aggregate TFP.[37] This observation helps explain why some of the size-dependent policies analyzed above seem to have a small impact on aggregate TFP, if they preserve (weakly) the rank of firms. Such is the case with the experiments considered in Guner et al. (2008) and tax exemptions for small firms analyzed by Gourio & Roys (2014). Policies or practices that set a limit to the size of firms, provided this limit does not vary across firms, also preserve ranking, so my analysis suggests their impact on aggregate TFP might be limited (subject, of course, to the discipline of their consistency with observed size distribution).

[35]In Alfaro et al.'s (2009) data, there is a large variation in average size across countries, and a very strong negative correlation with GDP per capita, contrary to what other data sources indicate.

[36]The results are consistent with the very little evidence of size-related distortions in Hsieh & Klenow (2009).

[37]As an example, comparing the economy with correlated distortions that result in a 50% fall in TFP to the benchmark in Restuccia & Rogerson (2008), a firm with two employees in the undistorted one has 1,000 in the distorted one, whereas a firm with 9,000 employees ends up with less than 300. These are huge rank reversals.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This article benefits greatly from earlier work and discussions with Pablo Andres Neumeyer. I would also like to acknowledge comments and discussions with Francisco Buera. I thank the EIEF and in particular the Bajola-Parisani foundation for their generous support.

LITERATURE CITED

Alfaro L, Charlton A, Kanczuk F. 2009. Plant-size distribution and cross-country income differences. In NBER International Seminar on Macroeconomics 2008, ed. JA Frankel, C Pissarides, pp. 243–72. Cambridge, MA: Natl. Bur. Econ. Res.

Amaral PS, Quintin E. 2006. A competitive model of the informal sector. *J. Monet. Econ.* 53:1541–53

Amaral PS, Quintin E. 2010. Limited enforcement, financial intermediation, and economic development: a quantitative assessment. *Int. Econ. Rev.* 51:785–811

Asker J, Collard-Wexler A, De Loecker J. 2014. Dynamic inputs and resource (mis)allocation. *J. Polit. Econ.* In press

Atkeson A, Kehoe PJ. 2005. Modeling and measuring organization capital. *J. Polit. Econ.* 113:1026–53

Banerjee AV, Duflo E. 2005. Growth theory through the lens of economic development. In *Handbook of Development Economics*, Vol. 1, ed. H Chenery, TN Srinivasan, pp. 473–552. Amsterdam: North-Holland

Banerjee AV, Moll B. 2010. Why does misallocation persist? *Am. Econ. J. Macroecon.* 2(1):189–206

Barseghyan L. 2008. Entry costs and cross-country differences in productivity and output. *J. Econ. Growth* 13:145–67

Barseghyan L, DiCecio R. 2011. Entry costs, industry structure, and cross-country income and TFP differences. *J. Econ. Theory* 146:1828–51

Bartelsman E, Haltiwanger J, Scarpetta S. 2013. Cross-country differences in productivity: the role of allocation and selection. *Am. Econ. Rev.* 103:305–34

Bollard A, Klenow PJ, Li H. 2014. *Entry costs rise with development*. Work. Pap., Stanford Univ., Stanford, CA

Buera FJ, Kaboski JP, Shin Y. 2011. Finance and development: a tale of two sectors. *Am. Econ. Rev.* 101:1964–2002

Buera FJ, Shin Y. 2013. Financial frictions and the persistence of history: a quantitative exploration. *J. Polit. Econ.* 121:221–72

Caselli F. 2005. Accounting for cross-country income differences. In *Handbook of Economic Growth*, Vol. 1, ed. P Aghion, S Durlauf, pp. 679–741. Amsterdam: North-Holland

Caselli F, Gennaioli N. 2013. Dynastic management. *Econ. Inq.* 51:971–96

David JM, Hopenhayn HA, Venkateswaran V. 2014. *Information, misallocation and aggregate productivity*. Unpublished manuscript, Stern Sch. Bus., New York Univ.

De Soto H. 1989. *The Other Path*. New York: Harper & Row

D'Erasmo PN, Moscoso Boedo HJ. 2012. Financial structure, informality and development. *J. Monet. Econ.* 59:286–302

Dixit AK, Stiglitz JE. 1977. Monopolistic competition and optimum product diversity. *Am. Econ. Rev.* 67:297–308

Djankov S, La Porta R, Lopez-de-Silanes F, Shleifer A. 2002. The regulation of entry. *Q. J. Econ.* 117:1–37

Doing Business. 2009. *Doing Business 2010*. Washington, DC: World Bank

Epifani P, Gancia G. 2011. Trade, markup heterogeneity and misallocations. *J. Int. Econ.* 83:1–13

Fattal-Jaef R. 2014. *Entry, exit and misallocation frictions*. Rep., World Bank, Washington, DC

Fattal-Jaef R, Hopenhayn HA. 2012. *Constrained optimality and the welfare effects of misallocation*. Rep., World Bank, Washington, DC
Gourio F, Roys N. 2014. Size-dependent regulations, firm size distribution, and reallocation. *Quant. Econ.* In press
Guiso L, Schivardi F. 2011. What determines entrepreneurial clusters? *J. Eur. Econ. Assoc.* 9:61–86
Guner N, Ventura G, Xu Y. 2008. Macroeconomic implications of size-dependent policies. *Rev. Econ. Dyn.* 11:721–44
Hopenhayn HA. 1992. Entry, exit, and firm dynamics in long run equilibrium. *Econometrica* 60:1127–50
Hopenhayn HA. 2012. *On the measure of distortions*. Work. Pap., Univ. Calif., Los Angeles
Hopenhayn HA, Neumeyer A. 2008. *Productivity and distortions*. Rep., Dep. Invest., Inter-Am. Dev. Bank, Washington, DC
Hopenhayn HA, Rogerson R. 1993. Job turnover and policy evaluation: a general equilibrium analysis. *J. Polit. Econ.* 101:915–38
Hsieh CT, Klenow PJ. 2009. Misallocation and manufacturing TFP in China and India. *Q. J. Econ.* 124:1403–48
Jeong H, Townsend RM. 2007. Sources of TFP growth: occupational choice and financial deepening. *Econ. Theory* 32:179–221
Jovanovic B. 1982. Selection and the evolution of industry. *Econometrica* 50:649–70
Loayza NV. 1996. The economics of the informal sector: a simple model and some empirical evidence from Latin America. *Carnegie-Rochester Conf. Ser. Public Policy* 45:129–62
Lucas RE Jr. 1978. On the size distribution of business firms. *Bell J. Econ.* 9:508–23
Melitz MJ. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71:1695–725
Midrigan V, Xu D. 2014. Finance and misallocation: evidence from plant-level data. *Am. Econ. Rev.* 104:422–58
Moll B. 2014. Productivity losses from financial frictions: Can self-financing undo capital misallocation? *Am. Econ. Rev.* In press
Moscoso Boedo HJ, Mukoyama T. 2012. Evaluating the effects of entry regulations and firing costs on international income differences. *J. Econ. Growth* 17:143–70
Olley GS, Pakes A. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64:1263–97
Peters M. 2013. *Heterogeneous mark-ups, growth and endogenous misallocation*. Unpublished manuscript, London Sch. Econ.
Poschke M. 2010. The regulation of entry and aggregate productivity. *Econ. J.* 120:1175–200
Poschke M. 2014. *The firm size distribution across countries and skill-biased change in entrepreneurial technology*. Work. Pap., McGill Univ., Montreal
Pratap S, Quintin E. 2008. The informal sector in developing countries: output, assets and employment. In *Personal Wealth from a Global Perspective*, ed. J Davies, pp. 373–94. New York: Oxford Univ. Press
Quintin E. 2008a. Contract enforcement and the size of the informal economy. *Econ. Theory* 37:395–416
Quintin E. 2008b. Limited enforcement and the organization of production. *J. Macroecon.* 30:1222–45
Rauch JE. 1991. Modelling the informal sector formally. *J. Dev. Econ.* 35:33–47
Restuccia D, Rogerson R. 2008. Policy distortions and aggregate productivity with heterogeneous establishments. *Rev. Econ. Dyn.* 11:707–20