Abstract

A number of recent papers argue that the misallocation of resources can explain large cross-country TFP differences. This argument is underpinned by empirical evidence documenting substantial dispersion in the marginal products of resources, particularly capital, in developing countries. But why does misallocation persist? That is, why don’t distortions disappear on their own? This is particularly true for capital misallocation, a point we illustrate in a simple model of capital accumulation with credit constraints: misallocation implies high marginal products for constrained firms and therefore a strong pressure for accumulation and to eliminate the distortion. We distinguish between misallocation on the intensive and the extensive margin, and show that the former should disappear asymptotically under fairly general conditions while the latter may persist. We conclude by discussing possible theories of persistent misallocation.

Introduction: Misallocation and Underdevelopment

There is growing interest in the view that underdevelopment may not just be a matter of lack of resources like capital, skilled labor, entrepreneurship or ideas but also a consequence of the misallocation or misuse of available resources. In particular Banerjee and Duflo (2005), Jeong and Townsend (2007), Restuccia and Rogerson (2008), Hsieh and Klenow (2009), Bartelsmann,
Haltiwanger and Scarpetta (2008), Alfaro, Charlton and Kanczuk (2008) and Buera, Kaboski and Shin (2008) all argue that the extent of misallocation of resources in poor countries is large enough to explain a very large part of the TFP gap between rich and poor countries.

Evidence on the misallocation of resources takes many forms. There is evidence on interest rates suggesting that many smaller firms in developing countries borrow at interest rates of 50%, 80% or even higher.\(^1\) This suggests that these firms must have marginal returns on capital that are even higher. These high estimates of the marginal returns are consistent with the direct evidence on the return on capital in small to medium sized firms in developing countries that we get from randomized experiments, natural experiments and other sources.\(^2\)

However, we see a very different picture when we look at the evidence on the aggregate marginal product of capital for various developing countries. Caselli and Feyrer (2007) find that the marginal product of capital is the same in poor and rich countries and is in fact below 10% everywhere. Swan (2008), using a different series for the prices of capital goods does find substantially higher estimates for developing countries, but even his high estimates are much lower than a lot of the firm level estimates for the marginal product of capital. Bai, Hsieh and Qian (2006) come up with 20% as the aggregate marginal product of capital in China, down from 25% in the earlier period.

The obvious way to reconcile these two sets of facts is to assume that marginal products, contrary to what efficiency would require, are not equalized across firms – some firms have very high marginal products but a lot of other firms do not.

A second source of evidence involves fitting a production function to firm level data and directly estimating the distribution of marginal products or something related within an industry. Hsieh and Klenow (2009) estimate the distribution of TFPR, which turns out to be a geometric average of the marginal products of capital and labor and calculate that the ratios of 90th to 10th percentiles of TFPR are 5.0 in India, 4.9 in China and 3.3 in the U.S. Moreover the most productive firms (firms where the conventional measure of TFP is the highest) tend to be the most distorted in the direction of being too small in both countries, which amplifies the effect of the TFPR not being equalized.

A less structural version of the same exercise involves comparing the distribution of firm

---

\(^1\)Banerjee (2003) describes the evidence on this point and emphasizes that default is rare – so that these interest rates should be thought of as the rates that firms actually pay.

\(^2\)See, for example, de Mel, McKenzie and Woodruff (2008), Banerjee and Duflo (2008), Udry and Anagol (2006).
sizes across countries to argue that the distribution of firm sizes in most developing countries looks different from the presumed efficient US distribution (Alfaro, Charlton and Kanczuk, 2008). An alternative approach uses the correlation between firm size and the average product of labor as a measure of allocative efficiency, under the theory that the most productive firms should be the biggest (Bartelsmann, Haltiwanger and Scarpetta, 2008). Both these exercises yield some evidence that less developed countries have a joint distribution of firm size and productivity quite unlike the US.

Finally one could do a pure calibration exercise using some plausible parameter values and a model, to put some magnitudes on the potential extent of output loss due to misallocation (as in Jeong and Townsend (2007), Restuccia and Rogerson (2008), Buera, Kaboski and Shin (2008), Banerjee and Duflo (2005)). All of these papers find that 50% or more of the difference between rich and poor countries (or in the case of Jeong and Townsend, 73% of the increase in TFP in Thailand) can be explained by the effects of misallocation under reasonable assumptions about parameter values.

While each of these pieces of evidence has its limitation, taken together they strongly suggest that misallocation is quantitatively important as an empirical phenomenon.

I. Theorizing Misallocation

One very natural explanation of why there is so much misallocation, especially given the evidence presented above about the high rates of interest, is to blame asset markets: the inefficiency in the functioning of asset markets makes it harder for successful firms to acquire the assets they need to expand and simultaneously allows failed firms to survive (because the alternative of downsizing and putting the rest of the money in asset markets is unattractive). As a result high productivity firms underinvest in what they need – be it management ideas, new technology, marketing advice, reputation building or just new plants or machinery.

Focusing on asset markets would also be consistent with Hsieh and Klenow’s result that most of the gain in both India and China would come from reallocating capital across firms. From the point of view of reallocating resources the key asset markets are the markets for land, financing and opportunities for risk diversification (which we will call risk capital). Of the primary assets of a firm, land (and what gets built on it) is the one where the physical adjustment costs are high everywhere in world. However while the acquisition of land has been
a major issue in both India and China, this is less of a problem at the firm level (unless it is a very large firm) than at the regional level, whereas the reallocation that Hsieh and Klenow are emphasizing is mostly between firms within the same region. On the other hand, there may well be constraints on selling land (though this seems unlikely since both India and China had a boom in residential real estate in this period, which made it extremely lucrative to sell existing land-holdings which were often in prime locations) and getting building permits and infrastructure connections are quite likely to have been a problem. At this stage we know too little, descriptively, about the workings of the urban "land" market in developing countries to say anything useful about this.

Finance and risk capital are of course the other two key assets that any firm needs (and the availability of which constrains its ability to then acquire machines, ideas, consulting, etc.) and there we know that the US financial infrastructure is much better. The banking sector in both India and China continues to be dominated by slow-moving and badly managed public sector banks and the system as a whole is notoriously ineffective in the enforcement of credit contracts, so that even the private sector is often unwilling to lend. The stock markets are not known for their effective regulations (in India things are said to have improved a lot, but only after 2000, while the data ends in 1995). And venture capital as an institutional form is more or less in its infancy in both countries.

However an alternative to acquiring these assets on the market is to accumulate them. The high rates of return faced by firms that are underinvesting, create a strong pressure for accumulation. Similarly, the lack of risk capital generates a precautionary savings motive which may drive firms to accumulate so much capital that they can self-insure. Both these forces generate forces towards eliminating the distortions across firms.

This does not mean, as we will emphasize later in the paper, that these asset market failures do not have aggregate consequences. But it does pose a challenge to explaining why distortions would not go away on their own; since underdevelopment is a persistent phenomenon, we need a theory that explains the persistence of misallocation.\(^3\)

The next section sets up a simple model which helps us understand this challenge. We focus

\(^3\)There is some evidence on the rate of change in the extent of misallocation: Hsieh and Klenow (2009) report that for China, hypothetically moving to “U.S. efficiency” might have boosted TFP by 50% in 1998, and 30% in 2005. But for India, hypothetically moving to U.S. efficiency might have raised TFP around 40% in 1987 or 1991, and 59% in 1994, notwithstanding the fact that in this period India liberalized substantially. However there is some danger of overinterpreting this evidence, since taking short term changes seriously asks a lot of the data.
on the role of credit constraints, for simplicity suppressing the issue of risk capital by assuming away all risk.

II. A Simple Model of Capital Accumulation with Credit Constraints

II.A. Preferences and Technology

Time is discrete. There is a continuum of households that are indexed by their ability \( z \) and their wealth \( a \). At each point in time \( t \), the state of the economy is some joint distribution \( G_t(a, z) \). The marginal distribution of ability is denoted by \( \mu(z) \) and its support by \([\underline{z}, \bar{z}]\).

Agents have preferences

\[
\sum_{t=0}^{\infty} \beta^t u(c_t),
\]

where \( u \) is strictly increasing, strictly concave and satisfies standard Inada conditions. Each household owns a private firm which uses \( k_t \) units of capital to produce \( f(k_t, z) \) units of output. We assume that the function \( f \) is strictly increasing but not necessarily concave. Capital depreciates at the rate \( \delta \).

II.B. Market Structure and Equilibrium

Denote by \( a_t \) an agent’s wealth and by \( r_t \) the (endogenous) interest rate. Agents can rent capital \( k_t \) in a rental market at a rental rate \( R_t = r_t + \delta \).\(^4\) Then an agent’s wealth evolves according to

\[
a_{t+1} = f(k_t, z) - (r_t + \delta)k_t + (1 + r_t)a_t - c_t.
\]

Agents face borrowing constraints:

\[
k_t \leq \lambda(z, r_t)a_t,
\]

\(^4\)Here the capital is accumulated by some intermediary who then rents it out to entrepreneurs. That the rental rate equals \( r_t + \delta \) can be shown by a standard arbitrage argument. This way of stating the problem avoids carrying \( k_t \) as a state variable in the agent’s problem.
where $\lambda$ is continuous and non-increasing in its second argument.\textsuperscript{5} A collateral constraint as in (3) is a particularly simple form of credit constraint. As will become clear later, our main result (Proposition 1) does not depend on the particular form of credit constraints we assume – the underlying logic is very simple and robust so that the particular form of credit constraint is not likely to matter. Moreover, the specification (3) nests the extreme case of no borrowing, $\lambda = 1$. This is a useful benchmark delivering an upper bound on the effects of credit constraints. In our model, there are also no intermediation costs so that the borrowing rate equals the deposit rate. We could incorporate a spread between the two rates as in Erosa (2001) at the expense of some extra notation. All our results still hold in the presence of intermediation costs.

The production and savings/consumption decisions separate in a convenient way. Define the profit function

\[
\pi_t(a, z) = \max_k \{ f(k, z) - (r_t + \delta)k + (1 + r_t)a \text{ s.t. } k \leq \lambda(z, r_t)a \}. \tag{4}
\]

It is easy to see that this profit function is increasing in both its arguments. Also denote the optimal capital choice from this profit maximization problem by

\[
k_t(a, z) = \min \{ \lambda(z, r_t)a, k^u(z, r_t) \},
\]

where

\[
k^u(z, r_t) \equiv \arg \max_k \{ f(k, z) - (r_t + \delta)k \}
\]

is unconstrained capital demand.

Summarizing, at each point in time $t$, each household solves

\[
v_t(a, z) = \max_{\{c_s, a_s\}_{s=t}^\infty} \sum_{s=t}^\infty \beta^{s-t}u(c_s) \text{ s.t. } a_{s+1} = \pi_s(a_s, z) - c_s, \forall s \geq t, \quad a_t = a. \tag{6}
\]

\textsuperscript{5}For example, suppose that agents can avoid the payment for rented capital $(r + \delta)k$ by incurring a cost proportional to capital usage, $\phi k$ where we assume $\phi < r + \delta$. If they default they also lose their savings $(1 + r)a$. The enforcement constraint is $f(k, z) - (r + \delta)k + (1 + r)a \geq f(k, z) - \phi k$ so that $\lambda(z, r) = (1 + r)/(r + \delta - \phi)$. More generally $\lambda$ also depends on ability $z$.\n
The problem for each household can be written in recursive form:

$$v_t(a, z) = \max_{a'} u[\pi_t(a, z) - a'] + \beta v_{t+1}(a', z). \tag{7}$$

Note that the value function is indexed by $t$. This is because $r_t$ varies over time, albeit exogenously from the point of view of the household. Denote the optimal choice of savings $a'$ by $s_t(a, z)$. This is the policy function of a household with assets $a$ and productivity $z$.

This paper studies capital misallocation. We find it useful to distinguish between two forms of misallocation.

**Definition 1** (i) We say there is capital misallocation on the intensive margin at time $t$ if marginal products of capital $f_k(k_t(a, z), z)$ are not equalized across all agents who have positive levels of capital usage, $k_t(a, z) > 0$.

(ii) We say there is capital misallocation on the extensive margin if it is possible to redistribute capital from one agent to another individual with either an equal marginal product or zero capital and raise the sum of their outputs.

Misallocation on the intensive margin is misallocation in the conventional sense and is what the empirical evidence has been principally about (Hsieh and Klenow, 2009; Restuccia and Rogerson, 2008). On the other hand, misallocation at the extensive margin exists only if there are non-convexities in production or if some individuals have zero capital and therefore are not be picked up at all by the current methodologies for measuring misallocation, which focuses on the equalization of the marginal products. Therefore there may be much more misallocation than the data on marginal products suggests – in particular because there are talented people who never have enough money to set up a business and therefore we do not even see them in the data. Jeong and Townsend (2007) and Buera, Kaboski and Shin (2008) attempt to get at this by making assumptions about the underlying distribution of talent.

A competitive equilibrium in this economy is defined in the usual way. That is, (i) individuals solve (6) taking as given the equilibrium time path for the interest rate $\{r_t\}_{t=0}^{\infty}$, and (ii) the capital market (which is the only market in this economy) clears at every point in time

$$\int k_t(a, z)dG_t(a, z) = \int adG_t(a, z), \quad \text{all } t \geq 0. \tag{8}$$
The main question that we are interested in is whether misallocation disappears over time. The answer turns out to depend on the shape of the production function (in particular whether it exhibits diminishing returns or not) and on whether we are looking for misallocation at the intensive or the extensive margin – in particular, misallocation at the intensive margin tends to disappear in the long run even when the misallocation at the extensive margin does not. This is what we now turn to.

II.C. Diminishing Returns

In this sub-section, we make the following assumption.

Assumption 1 The function \( f(k, z) \) is concave in \( k \) for any \( z \) and satisfies standard Inada conditions.

As already noted, with diminishing returns there is no role for misallocation on the extensive margin except in the case where some individuals start out with zero wealth. The Euler equation corresponding to (6) is

\[
\frac{u'(c_t)}{u'(c_{t+1})} = \beta \left( \frac{1 + r_{t+1} + \lambda(z, r_{t+1})\psi_{t+1}}{1 + r_t + \delta} \right),
\]

where

\[
\psi_t = \max\{f_k(\lambda(z, r_t)a_t, z) - (r_t + \delta), 0\}
\]

is the Lagrange multiplier on the borrowing constraint (2). If there is no misallocation on the intensive margin at time \( t \), marginal products are equalized across individuals and all multipliers are zero. In this case the unconstrained Euler equation

\[
\frac{u'(c_t)}{u'(c_{t+1})}(1 + r_{t+1})
\]

holds for all agents.

The following Lemma about the optimal savings policy function \( s_t(a, z) \) is an adaptation of a result by Dechert and Nishimura (1983) and will be useful below. Note that it applies regardless of the assumptions on technologies \( f \), for instance also for the case of (local) increasing returns.

Lemma 1 The policy function \( s_t(\cdot, z) \) is strictly increasing for all \( z \).
(All proofs appear in the Appendix.) This Lemma implies

**Corollary 1** Consider individuals with the same ability $z$. Their wealth trajectories never intersect, that is if $a_0 > \hat{a}_0$ then $a_t > \hat{a}_t$ for all $t$.

We are now in the position to prove our main result about the asymptotic behavior of misallocation with diminishing returns.

**Proposition 1** Under Assumption 1 there is no misallocation on the intensive margin asymptotically, that is (10) holds for all agents as $t \to \infty$.

While there is no misallocation on the intensive margin, there may be misallocation on the extensive margin. If there are some individuals that have zero wealth at $t = 0$, they produce zero output and will never accumulate any wealth which, according to our definition, is extensive margin misallocation. This knife-edge case can be ruled out by

**Assumption 2** The initial distribution of wealth $G_0(a,z)$ places no mass at $a = 0$.

Adding this assumption gives the following corollary to Proposition 1.

**Corollary 2** Under Assumptions 1 and 2 there is no misallocation (either on the intensive or extensive margins) asymptotically.

Informal versions of this claim have appeared in the literature in the past (for example in Banerjee and Duflo (2005)) and it is entirely intuitive. Indeed this result is rather general in many ways. We allowed for arbitrary correlations between productivity and access to credit. Moreover while we do not formally deal with that case, the result holds in the presence of intermediation costs. The result does not depend on the particular form of credit constraints (3) either: it would hold in models of credit constraints in which access to credit depends on the present as well as the future profitability of the firm. For example, Moll (2009a) analyzes a similar environment with the main exception that credit constraints take the form of endogenous and forward-looking limited enforcement constraints. Misallocation also disappears with this form of credit constraints. This makes clear that the logic behind the proof of Proposition 1 is very general.\(^6\)

\(^6\)Note also what this proposition does and does not say. It says that misallocation disappears asymptotically as time $t \to \infty$. It does not say that this will happen in finite time. In fact, Moll (2009a) finds that credit constraints always bind in finite time with slightly different modeling of credit constraints.
The one assumption that is critical for this result is the assumption that all agents are equally patient. If we drop this assumption and allow for heterogeneity in discount rates, the result does not hold any longer. Below we will demonstrate the existence of a steady state where there are two types of agents with differing levels of patience where the less patient agent is permanently credit constrained. This is a counterexample to Proposition 1 because the proposition ought to hold for any initial distribution of wealth and ability.\footnote{Moll (2009a) shows that if agents’ discount factor differ, misallocation also persists asymptotically with endogenous and forward-looking limited enforcement constraints.}

More can be said about steady states of this economy.

**Definition 2** A steady state is a competitive equilibrium with a constant interest rate \( r_t = r^* \) and a constant aggregate capital stock \( K_t = K^* \) for all \( t \).\footnote{The definition of the aggregate capital stock is the obvious one, \( K_t \equiv \int adG_t(a,z) \).}

In line with Proposition 1, we can show that there will be no capital misallocation in steady state. Furthermore the interest rate equals the rate of time preference, \( r^* = \rho \) where \( \rho \) is defined by \( \beta \equiv (1 + \rho)^{-1} \). To see this, first note that an interest rate greater than the rate of time preference, \( r^* > \rho \), is inconsistent with a steady state: from the Euler equation this would imply positive consumption growth \( c_{t+1} > c_t \) for all agents which can only be true if the aggregate capital stock grows. We can also rule out an interest rate \( r^* < \rho \). In the absence of credit constraints agents would dissave until they reach zero wealth. With credit constraints, this is not true anymore; instead individuals dissave until their wealth reaches a level satisfying

\[
1 + \rho = 1 + r^* + \lambda(z, r^*)\psi^*(a, z) = 1 + r^* + \lambda(z, r^*)[f_k(\lambda(z, r^*)a, z) - r^* - \delta].
\]  

That is, unconstrained agents only stop dissaving once they become constrained. But there cannot be only constrained agents in equilibrium; there have to be some lenders as well. This tells us that the interest rate must equal the rate of time preference.

Given this, no one is decumulating capital in the steady state. Now if some individuals were credit constrained in steady state they would have an incentive to accumulate wealth because the right-hand side of (11) would be greater than the left-hand side. Therefore the capital stock must be going up, contradicting the definition of a steady state. Summarizing, in any steady state of the economy with diminishing returns all agents must be unconstrained and the interest rate equals the rate of time preference.
The allocation of capital is then first-best and can be described by an aggregate production function. Individual capital usage \(k^*(a, z)\) is therefore simply the same as in the standard neoclassical growth model

\[
f_k(k^*(a, z), z) = \rho + \delta, \quad \text{all} \ (a, z).
\]

Recalling the definition of unconstrained capital demand \(k^u(z, r)\) in (5), this implies that individual steady state capital stocks \(k^*(a, z)\) are equal to the unconstrained capital demands \(k^u(z, \rho)\). The unique aggregate steady state capital stock is first-best and equals

\[
K^* = \int k^u(z, \rho)\mu(z)dz.
\]

Once again this is something that holds with some generality. The nature of the credit constraint does not matter and nor does the presence of intermediation costs. Since in steady state relative prices are constant, the argument that no firm can be credit constrained also immediately extends to the case where there are multiple goods. The one assumption that is essential for the claim that the steady state capital stock is first best efficient is the assumption of identical discount factors. To see how there can be a steady state where agents are credit constrained and output is less than first-best efficient, suppose that there are two types of agents with the same ability \(z\) (so that we can drop \(z\) from the \(f\) and \(\lambda\) functions), but different discount factors \(\beta_2 < \beta_1\). We show now, that it is possible to construct a steady state in which type 2 agents are credit constrained. If type 2 agents are constrained it must be that type 1’s are lending and therefore from their Euler equation the interest rate is given by \(r^* = \rho_1\). Similarly from type 2’s Euler equation, it must be that

\[
1 + \rho_2 = 1 + \rho_1 + \lambda(\rho_1)[f'(\lambda(\rho_1)a^*_2) - \rho_1 - \delta].
\]

This determines the steady state wealth of a type 2 agent. Since \(f\) satisfies an Inada condition, this only has a solution if \(a^*_2 > 0\). Therefore the constrained type must own positive wealth. To complete the construction, note that the steady state wealth of a type 1 agent must be such that he is able and willing to meet the demand for loans from type 1 agents. Assuming that
there are equal numbers of type 1 and type 2 agents, this requires that
\[ a_1^* - k^u(\rho_1) = (\lambda(\rho_1) - 1)a_2^*. \]

Finally observe that
\[ \lambda(\rho_1)[f'(\lambda(\rho_1)a_2^*) - f'(k^u(\rho_1))] = \rho_2 - \rho_1 > 0. \]

Interestingly, the wedge in the marginal products of capital of the two agents exactly equals the difference in their discount rates.

Given that misallocation disappears asymptotically as \( t \to \infty \) and the steady state is first best efficient, a natural question is: how long does this convergence of marginal products take? To get a feel for the length of time involved, we have conducted some numerical experiments.\(^9\) We compare two individuals, one rich and one poor, with the same \( z \). The borrowing constraint of the poor agent binds at time zero, so much so that the poor agent has a three times higher marginal product than the rich agent, \( f_k(k^p, z)/f_k(k^r, z) = 3 \). Figure 1 plots the ratio of the marginal products over time under the assumption that \( r \) remains fixed over time, for different values of \( \lambda(z, r) \) The speed at which the gap between the two marginal products narrows is quite striking: Even if credit markets are completely shut down, \( \lambda = 1 \) the gap has almost disappeared after seven years. For better functioning credit markets convergence is even faster. With \( \lambda = 3 \) for example, the gap has essentially disappeared after five years. From a theoretical perspective, the result becomes less surprising if one keeps in mind that the model is a variant of the standard neoclassical growth model. With no credit markets, \( \lambda = 1 \), the models are in fact identical. It is well known that the speed of convergence of the neoclassical growth model is relatively high (King and Rebelo, 1993). With better working credit markets, this speed is only increased.

II.D. Local Increasing Returns

The assumption of global diminishing returns is not always a great description of how actual firms function – there is often a set-up cost involved in starting a business and more generally,

---

\(^9\)Production takes the Cobb-Douglas form \( f(k, z) = k^\alpha \). Utility is of the CRRA form with parameter \( \sigma \). Borrowing constraints just take the simple form \( k_t \leq \lambda a_t \) where \( \lambda \geq 1 \) is a constant. We choose the following very conventional parameter values: \( \alpha = 0.3, \delta = 0.05, \beta = 0.95, \sigma = 2. \).
non-convexities arise naturally from the fact that machines come in fixed sizes and ideas tend to be indivisible. Following Skiba (1978) and Dechert and Nishimura (1983) we now allow for increasing returns over some range.

**Assumption 3** $f(k, z)$ is convex over some range of $k$, but there is a $\hat{k} < \infty$ such that $f(k, z)$ is concave for $k > \hat{k}$.

$$f(k, z) = z(k - \check{k})^\alpha, \text{ for } k \geq \check{k} \text{ and zero otherwise, is an example of this kind of production function.}$$

In this section, we only look at steady states (as in Def. 2) – we are not able to guarantee that the economy converges to a steady state, though in our simulations it always does. As a result of the steady state assumption, in contrast with the earlier analysis, the problem of an agent is now stationary. Nevertheless, because the production function is non-convex, so is the agent’s maximization problem. The result is that people at different wealth levels may exhibit radically different behaviors – those who are not too far below a particular non-convexity will save up and “cross” the non-convex region to get to the high returns available at high levels of investment, while those with only slightly less assets will prefer to dissave because the climb to get to the high returns is too far for them. Therefore where you converge to depends on where you started and there are multiple individual steady state wealth levels with different levels.
of output associated with them (these results are not formally demonstrated here but follow directly from the logic of such problems spelt out in Skiba (1978) and Dechert and Nishimura (1983)).

Note, however, that regardless of any non-convexities, it is still true that credit constrained agents have a higher intertemporal marginal rate of substitution than unconstrained agents. This allows us to prove

**Proposition 2** Consider a steady state with constant interest rate $r$ and constant aggregate capital stock $K^*$ (definition 2). As $t \to \infty$, each agent converges to his individual steady state, there is no capital misallocation on the intensive margin, and $r = \rho$.

This result extends Proposition 1 to the case of local increasing returns. Misallocation on the extensive margin is now a “robust” phenomenon: For instance, suppose the non-convexity takes the form of a fixed cost. Then individuals starting off below some threshold wealth level, will never produce any output because they cannot cover the fixed cost. Extensive margin misallocation is “robust” in the sense that this outcome is not a knife-edge result that depends on the initial distribution of wealth $G_0(a, z)$ as in Corollary 2.

While not reported here, we also carried out numerical exercises parallel to the ones reported above for the diminishing returns case, under the current assumption about the production function. Once again, except for those who are trying to “cross the non-convex region”, convergence to an unconstrained state is relatively quick, which should not surprise us since they essentially operate in a diminishing returns environment.\(^\text{10}\)

**II.E. Implications of These Results**

These results tell us that unless convergence to an aggregate steady state fails (which we have not ruled out for the local returns increasing case) and the interest rate continues to fluctuate substantially even in the long run, we would expect to see misallocation at the intensive margin to disappear relatively quickly in the world of this model. That does not have to mean that there is no misallocation in this economy. Indeed with local increasing returns we can construct examples where the extensive margin misallocation is so large that steady state

---

\(^\text{10}\)Such non-convex problems are not much harder to solve computationally than their convex counterparts. This is because standard dynamic programming does not require the assumption of a concave period return function. This fact is, for example, used extensively in Dechert and Nishimura (1983).
output as a ratio of first best steady state output is arbitrarily close to zero. But it does raise the question: Why do we see so much intensive margin misallocation in the data (remember all the misallocation that Hsieh and Klenow, for example, find, is at the intensive margin)?

III. Conclusion: Towards an Understanding of Persistent Misallocation

Persistence of misallocation is easy to explain if we are prepared to assume, as Restuccia and Rogerson (2008) do (for illustrative purposes) that there are “taxes” on the firms that are permanently fixed. However since Hsieh and Klenow in particular only compare firms within an industry (and indeed get very similar results from comparing firms within the same industry within the same region) we need to explain why these taxes vary at the firm level. Moreover, as Restuccia and Rogerson point out, to get large effects, the taxes need to be strongly positively correlated with firm level TFP.

The problem is explaining why there would be such large firm specific taxes.

In a recent paper, Gordon and Li (2005) suggest an argument for why “taxes” will systematically vary across firms even within the same industry. Their argument is that firms face a choice between joining a formal sector in which they have to pay taxes but can grow to their appropriate size (firms in the formal sector have access to efficient intermediation through the banking system) or operating in the informal sector by staying small enough to avoid detection by the tax system. What does such a choice imply for the correlation between productivity and firm size? We believe that it is likely that less productive firms have a bigger incentive to stay small. This will for example be the case if joining the formal, intermediated sector requires paying a tax on profits. Since a profit tax works like a “fixed cost”, only the most productive firms will find it worth paying the profit tax, while the least productive firms will stay under the tax net. Note that this says that the factor that distorts firm size (i.e. Restuccia and Rogerson’s “taxes”) should be negatively correlated with firm level TFP. This is the opposite of what, according to Restuccia and Rogerson, we should be looking for. Moreover as Hsieh and Klenow point out, the potential productivity gains remain almost unchanged when they

---

11 There is a long tradition of research on the effects of fixed costs combined with credit constraints on the long run performance of the economy (Galor and Zeira, 1993; Banerjee and Newman, 1993; Aghion and Bolton, 1997; Piketty, 1997). More recently, Buera, Kaboski and Shin (2008) find that introducing fixed costs leads to larger TFP losses.
only equalize TFPR between firms within the same size quartile. In other words, misallocation within the category of firms that are all large enough to be in the tax net is very sizeable.

One scenario where high TFP firms may face high “taxes”, is where some people set up firms because they have high TFP, while others do so despite having low TFP because they are politically connected and therefore are in less danger of being expropriated. Then high TFP firms would under-invest because they fear expropriation. However while expropriation risk in today’s India and China is not entirely absent, it is hard to believe that it is a serious issue for the average firm (as against the very largest and most visible firms). But most firms that are in the Hsieh and Klenow study are neither large enough (the median firm in the top quartile of the distribution has around two hundred employees) nor in another way so special (there are thousands of such firms) to attract special attention, one way or the other, from the political system in either of the two enormous countries. And while one could imagine firms getting into an especially friendly or unfriendly relationship with the local political bosses, it is not clear why the firms that are suffering could not move to a different area, or why someone without the political baggage could not buy out the firm.

A third potential source of “taxes” that could be causing misallocation is an explicit policy that discriminates against large firms. It is true that India (but as far as we know, not China) has some policies – labor laws in particular – that specifically discriminate against larger firms. However Hsieh and Klenow (2009) report that their estimates of the potential productivity gain change by very little when they only equalize TFPR between firms within the same size quartile, and moreover, states in India with better labor laws do not do better in terms of TFPR dispersion.\footnote{A comment regarding the discussion of “taxes” in the preceding three paragraphs seems in order: we are not arguing against the importance of “taxes” per se; instead we are arguing against “taxes” as an explanation for persistent and large differences in marginal products between firms in both the same size category and the same four digit industry.}

The alternative approach to persistence relies on shocks. Within our framework there can be shocks to both assets ($a$) and ability ($z$) – it should be clear that any temporary shock to the profitability of the firm that occurs after investment has been chosen is isomorphic to an $a$ shock, while any shock that affects the marginal product of investment that is yet to be carried out is like a $z$ shock.

From the point of view of explaining the persistence of misallocation, shocks are important because they move people away from their steady states. Thus someone who used to have low
but now discovers that she has a high $z$ will be massively under-invested and will show a high marginal product. The same will be true if a firm loses a large part of its capital stock.

Clearly the extent to which shocks can explain the persistence of misallocation (in combination with financial constraints) depends on the frequency of the shocks: A low frequency gives firms a long time to adjust to a shock before the next shock hits and we would not expect to find very much misallocation in the stationary state.

Caselli and Gennaioli (2005) is one paper that takes a clear *a priori* position on the frequency of shocks. They look at a model where there are only $z$ shocks and these occur once in every generation, because the current owner of the business dies and is replaced by his child (because capital markets are imperfect, it does not make sense to sell the firm unless the child is especially untalented). With this they are able to explain a surprisingly large (up to 50%) fraction of the TFP gap across countries. However agents in their model have one period lives and follow a fixed bequest rule, so the possibility of undoing misallocation by accumulating resources does not arise. Buera and Moll (2009) consider a similar environment but add the possibility for individuals to accumulate wealth during their own life-time which implies smaller effects.

On the other hand there is no reason to take the idea that $z$ is ability too literally: When a firm loses a contract because its contact in the buying firm has moved on, this is also a $z$ shock, as is the introduction of a new product into the market. Interpreted in this way, the hypothesis of frequent and large $z$ shocks does not seem prima facie implausible.

Unfortunately at this point there is relatively little empirical evidence on this point. Buera and Shin (2008) use US data on the distribution of income and firm sizes, and conclude that an estimate of 0.87 for the autocorrelation of $z$ fits the data best.\footnote{They set this up using slightly different language, namely that each period individuals draw a new $z$ with probability 0.13, implying an autocorrelation of $1 - 0.13 = 0.87$.} Even more recently, Midrigan and Xu (2009) calibrate a model of firm dynamics with financing frictions to plant-level panel data from Korea with their benchmark calibration implying an autocorrelation of 0.94. Based on this correlation, they conclude that financial constraints cannot explain very much of the misallocation that they observe in the Korean data. However while 0.87 and 0.94 might seem relatively close, Moll (2009b) provides a closed-form example in which the extent of misallocation is actually highly sensitive to the autocorrelation of $z$ even in this range.\footnote{The extent of misallocation also depends on the concavity of the production function, which determines the speed at which firms adjust their capital to shifts in productivity: the more concave the technology, the higher the payoff from a small adjustment, and hence the faster this adjustment. Midrigan and Xu assume strong diminishing returns to scale (an elasticity of output with respect to scale of 0.45) so within their framework it is
However, as Buera and Shin (2008) point out, we might be missing the most important reason for the observed misallocation by focusing on stationary states. After all, even if it is true that highly autocorrelated $z$ shocks tend to generate a stationary state with limited misallocation, the transition to the stationary state from a highly distorted initial allocation (think of India or China before liberalization) can be quite slow in the presence of financial constraints and therefore in the short to medium run we will continue to observe a lot of misallocation.

Finally, is it possible that the reason why there is persistent misallocation is that there are large differences in the level of patience of different entrepreneurs? Banerjee and Mullainathan (2009), for one, argue that discount factors may be endogenous.

It is clear that what is needed at this point is empirical evidence that will help us decide between these alternative views of what lies behind the growing number of observations of large-scale misallocation: Is it financial constraints, some other form of market or government failure, some failure of patience or rationality, or are we simply reading the data wrong and there is less misallocation than we think? Answering these questions will require in part more detailed data about firms: what borrowing opportunities do they have available, what are the sources of change in their TFP, what regulations do they face, how much extortion and expropriation do small to medium firms on the ground actually face, etc? In part it would require following firms over time (i.e. panel data) to see how they make adjustments. And in part there needs to be a better understanding of how to model firm decision-making in imperfect market conditions.

Appendix

Proof of Lemma 1

The proof follows the same steps as Theorem 1 in Dechert and Nishimura (1983). By way of contradiction, let $a > \hat{a}$ but $s_t(a, z) \leq s_t(\hat{a}, z)$. By utility maximization

$$v_t(a, z) = u[\pi_t(a, z) - s_t(a, z)] + \beta v_{t+1}[s_t(a, z), z] \geq u[\pi_t(a, z) - s_t(\hat{a}, z)] + \beta v_{t+1}[s_t(\hat{a}, z), z].$$

possible that results are less sensitive to variations in the autocorrelation of shocks than it is in Moll’s example which assumes constant returns to scale. A proper assessment of firm level returns to scale clearly also has to be an important part of this research agenda.
Likewise

\[ v_t(\hat{a}, z) = u[\pi_t(\hat{a}, z) - s_t(\hat{a}, z)] + \beta v_{t+1}[s_t(\hat{a}, z), z] \geq u[\pi_t(a, z) - s_t(a, z)] + \beta v_{t+1}[s_t(a, z), z]. \]

Differencing the above two inequalities, we have

\[ u[\pi_t(a, z) - s_t(a, z)] - u[\pi_t(\hat{a}, z) - s_t(\hat{a}, z)] \geq u[\pi_t(\hat{a}, z) - s_t(\hat{a}, z)] - u[\pi_t(a, z) - s_t(a, z)]. \quad (12) \]

Note that

\[ [\pi_t(a, z) - s_t(a, z)] - [\pi_t(\hat{a}, z) - s_t(\hat{a}, z)] = [\pi_t(a, z) - s_t(\hat{a}, z)] - [\pi_t(\hat{a}, z) - s_t(\hat{a}, z)]. \]

But then the inequality in (12) contradicts the strict concavity of \( u \) (strictly decreasing differences). □

**Proof of Proposition 1**

Fix a \( z \). Consider an agent with with initial wealth \( a_0 \). Denote his wealth and consumption by \( \{a_t\} \) and \( \{c_t\} \). Consider another agent with the same ability, but initial wealth \( \hat{a}_0 < a_0 \). Denote his wealth and consumption by \( \{\hat{a}_t\} \) and \( \{\hat{c}_t\} \). Denote the ratio of their marginal utilities by

\[ \alpha_t \equiv \frac{u'(c_t)}{u'({\hat{c}}_t)}. \]

Combining the Euler equations (9), we have that

\[ \alpha_t = \alpha_{t+1} \frac{1 + r_{t+1} + \lambda(z, r_{t+1})\psi_{t+1}}{1 + r_{t+1} + \lambda(z, r_{t+1})\hat{\psi}_{t+1}} \leq \alpha_{t+1}. \]

The inequality follows because by Lemma 1 \( a_t > \hat{a}_t \), all \( t \) so that either \( \psi_t = \hat{\psi}_t = 0 \) or \( \psi_t < \hat{\psi}_t \). The sequence \( \{\alpha_t\}_{t=0}^\infty \) is nondecreasing and therefore converges on the extended real line. There are only two possible cases.

CASE 1: \( \{\alpha_t\}_{t=0}^\infty \) converges to some \( \alpha^* < \infty \). This is only possible if the sequence of multipliers \( \{\hat{\psi}_t\}_{t=0}^\infty \) converges to zero, implying the desired result.

CASE 2: \( \{\alpha_t\}_{t=0}^\infty \to \infty \). This is only possible if \( \hat{c}_t \to \infty \) or \( c_t \to 0 \). But this would imply that

\[ \lim_{t \to \infty} v_t(a_t, z) < \lim_{t \to \infty} v_t(\hat{a}_t, z). \]
Since wealth trajectories do not cross and the value function \( v_t(\cdot, z) \) is weakly increasing, this is a contradiction. □

**Proof of Proposition 2**

Consider a steady state with some interest rate \( r \) (not necessarily equal to \( \rho \)). Consider first the case where *every individual* is in steady state. Then from the Euler equation a steady state with positive wealth has to satisfy

\[
1 + \rho = 1 + r + \lambda(z, r)\psi(a, z).
\]

(13)

We next want to argue that (given that the interest rate is constant), this steady state is stable from the perspective of an individual. This can be done using the apparatus of Skiba (1978), Dechert and Nishimura (1983) and Buera (2008). We don’t include the detailed argument due to space restrictions. Suffice it to that individual wealth sequences are monotonic because policy functions are strictly increasing (Lemma 1), and that individuals either converge to a positive steady state satisfying (13) or decumulate wealth until it reaches zero. By a similar argument as in the diminishing returns case, we can rule out the case \( r \neq \rho \): with \( r > \rho \) the aggregate capital stock would grow; with \( r < \rho \) everyone would be constrained and there would be no lenders. Again as above, all multipliers \( \psi(a, z) \) must equal zero; otherwise constrained agents would be accumulating wealth. Therefore there is no misallocation on the intensive margin. □

**References**


