Imperfect Competition and Misallocations

Yan Liang*
London School of Economics, CFM

Latest version here

This version: July 2017

Abstract

Misallocation of resources across production units lowers aggregate TFP. The factors driving misallocation can be internal or external to the market structure. Variation in markups associated with imperfect competition is an internal distortion. Taxes and subsidies on input and output markets are examples of external distortions. The study of India's manufacturing data suggests that more than 80% of the observed misallocation can be attributed to internal distortions. Removing external distortions alone has a modest effect on aggregate TFP. It raises TFP by 14-17%, which is less than one fifth the compound effect of removing both types of distortions. The implication is that fostering competition is more effective to improve allocation efficiency and aggregate TFP. Removing individual external distortions is less effective.

JEL Codes: O11, O47, O53.
Keywords: Misallocation, Markup, Distortions, Imperfect Competition.

*First version: May 2014. Department of Economics, and Centre for Macroeconomics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE (e-mail: y.liang4@lse.ac.uk). I am grateful to Francesco Caselli for his invaluable guidance and support. My gratitude extends to Gianmarco Ottaviano, Michael Peters, and seminar participants at LSE, LBS TADC for helpful comments. All errors are my own.

1 Introduction

A growing literature has pointed out the important role of resource allocation in shaping aggregate productivity. For the allocation to be efficient, it is required that marginal products be equalized across different production units.[1] Deviations from this represent misallocation and lead to a low aggregate productivity. Most of the literature emphasizes the role of particular distortions in aggregate TFP and output. For example, it is argued that government distortionary policies[2], credit frictions[3], and information frictions[4] are possible sources of misallocation. This paper adds to the literature by arguing that imperfect competition has a profound influence in determining aggregate productivity.

My point of departure is the observation that market structure governs how factor inputs are allocated across production units. To a large extent, allocative efficiency is determined by the underlying market structure. Monopolistic competition is a commonly used market structure. It may or may not lead to misallocation of factors of productions. If the price elasticity is constant, all firms will charge the same markup and their marginal products will be equalized. In this case, there is no misallocation of resources. However, when price elasticity varies, markups and marginal products will no longer be equalized. This leads to misallocation of factor inputs, and translates into sub-optimal aggregate outcomes. One needs to take a stand on the market structure. In this paper, I assume it is monopolistic competition with variable markups. Variation in markups is a specific kind of market distortion. I refer to this distortion as *internal distortion* to emphasize the fact that it is an intrinsic feature of the market structure. By contrast, *external distortion* refers to the distortions that are imposed from outside the market. Since external distortions are not an integral part of the market, in principle they can be removed without having to alter the underlying market structure. Examples of external distortions are government taxes and subsidies, product market regulations, and trade restrictions. In this paper, I focus on a special kind of external distortion whose direct effect is to create heterogeneity in the prices faced by individual producers.

The goal of this paper is to quantify the effect of internal distortions on resource misallocation and aggregate productivity. Quantifying internal distortion is difficult because internal distortions and external distortions are observationally equivalent. When one observes a high dispersion of marginal products, it is equally valid to interpret it as internal distortion or as external distortions. To overcome this difficulty, I adopt a sequential identification strategy. I first identify firm-specific markups from data on revenue and profits, and then identify external distortions and individual productivities taking as given firms' markups. This provides a starting point to unravel the effects of the two kinds of distortions. Since internal distortions naturally interact with external distortions,

[1]More precisely, the necessary condition is that revenue marginal products be equalized across production units, since the price of output must be taken into account. A stronger condition is required when the market is not perfectly competitive. Namely, the revenue TFP must be equalized across production units. This point will become clear when we discuss the accounting framework in Section 2.4.

[2]For example, Restuccia and Rogerson (2008), Hsieh and Klenow (2009).

[3]For example, Midrigan and Xu (2014).

[4]For example, David et al. (2014).

a direct measure of the effects of the two distortions is bound to be imprecise. Quantifying internal distortions requires specifying a structural model and designing a policy experiment, to which I now turn.

The model builds on Melitz and Ottaviano (2008). I extend it to a multi-sector environment. The Melitz and Ottaviano model is a dynamic industry model with heterogeneous firms. Firms are engaged in monopolistic competition and each faces a downward-sloping linear demand for its own variety. The linear demand generates an endogenous distribution of markups across firms that responds to the toughness of competition in a market. A tougher market hosts more productive firms that set lower markups (hence smaller variance of markups). Thus a tougher market exhibits a lower level of misallocation and a higher level of aggregate productivity. The Melitz and Ottaviano model provides a laboratory for policy experiment. To link the distribution of markups with resource misallocation and aggregate productivity, I adopt the accounting framework by Hsieh and Klenow (2009). The framework, however, builds on a CES demand system. With a CES demand, the markups of all firms are exogenously given and do not respond to market conditions. I replace the CES demand with a linear demand system. In short, I combine the two models to quantify the effect of internal distortions.

The policy experiment is designed as follows. In this experiment, I ask: how much of the observed misallocation would remain if the external distortions are removed? I exploit a property of the model, that the performance measures of all firms are a function of two variables: a firm's idiosyncratic cost and the threshold cost for which a firm is just indifferent about remaining in the industry. The threshold cost is endogenously determined by a free entry condition. Therefore, the key to the experiment is to determine the threshold cost after removing external distortions. This can be done in two steps. In the first step, I estimate the distributional parameters for individual productivity, internal and external distortions, which are needed for the evaluation of the free entry condition. I assume that productivity, internal and external distortions are jointly log-normally distributed. Since I only observe data of surviving firms, I estimate the distributional parameters using a truncated maximum likelihood model[5]. In the second step, I remove all external distortions, and evaluate the free-entry condition to obtain the counterfactual threshold cost. The threshold cost, together with individual costs, allows me to compute all counterfactual equilibrium variables. From these variables, I can compute the level of resource misallocation and aggregate productivity. Since external distortions no longer exist, the misallocation of resources is entirely driven by internal distortions. Comparing the levels of misallocation before and after the removal of external distortions gives a precise measure of the effect of internal distortions.

I apply the empirical methodology to plant-level data from manufacturing firms in India. I find that removing external distortions has virtually no effect in reducing misallocation of factors of productions. The counterfactual dispersion of revenue TFP is above 80% of the level of dispersion before liberalization. In other terms, suppose there were no external distortions, one still observes

[5] See, for example, Amemiya and Boskin (1974) for an excellent exposition of the maximum likelihood estimation when the variable is truncated lognormally distributed.

about 80% of misallocation of factors due to the fact that the market is not perfectly competitive. The TFP gain from removing external distortions is about 14-17%. To put the numbers in perspective, I note that if internal distortions were not present, the TFP gain would be as large as 86-89%.[6] Hence removing external distortions alone has a very modest effect on aggregate productivity, which is about one-fifth of the combined effect of suppressing both internal and external distortions. Interestingly, a naive measure of the effects for the two kinds of distortions provides misleading results. The direct measure reads that about one-third of the dispersion in marginal products comes from internal distortion. The remaining two-thirds are attributed to external distortions. It suggests that external distortions are about twice as important as internal distortions.

Why does the direct measure provide such a misleading picture? The reason is that it ignores the interaction of the two kinds of distortions. The intuition is as follows. Conditional on survival, a firm's productivity and external distortions are negatively correlated. That is, a firm survives either because it draws on high productivity, or because it receives a subsidy to compensate for its low productivity. As a result, the distribution of marginal cost, which reflects both productivity and taxes, is less dispersed under the presence of external distortions. In our framework, a firm with a lower marginal cost produces more, earns more revenue, and charges higher markup and prices. A concentrated distribution of marginal costs translates into a concentrated distribution of markups. This explains why the dispersion of markups remains large when external distortions are removed from the market. It also explains why the elimination of external distortions has a modest effect on aggregate TFP.

To explore whether the main findings are driven by model misspecification, I perform two robustness checks. In the benchmark model, there are two kinds of input: labor and capital goods. However, for simplicity, I only consider two kinds of external distortion: a tax on capital and a tax on output. It is natural to ask whether the large effect of internal distortions comes from the omission of a labor distortion (i.e., a tax on labor inputs). Thus the first robustness analysis includes a labor distortion. The second one concerns capital shares in the production function. In the benchmark model, I follow Hsieh and Klenow (2009) to adopt the US shares in corresponding sectors. This allows me to focus on distortions instead of technology differences. However, one might worry that the use of US shares affects the main findings. Hence I estimate and use the indigenous shares instead. In both cases of analysis, the results are consistent with those of the benchmark case. This suggests that the main findings are indeed driven by the nature of imperfect competition.

This paper relates to several existing branches of literature. The studies of misallocation and aggregate productivity is pioneered by Restuccia and Rogerson (2008) and Hsieh and Klenow (2009). Restuccia and Rogerson (2008) study policies which create heterogeneity in the prices faced by producers. They find sizeable decreases in output and measured TFP in the range of 30-50%. In

[6] Hsieh and Klenow (2009) find that a full liberalization of the India economy observes a TFP gain of about 100-127%. Their statistics are computed directly from the data. I study the data through the lens of the model. That is, all equilibrium variables except productivity are computed by the model. That being said, it is reassuring that my model can closely replicate the TFP gains from data.

the same vein, Hsieh and Klenow (2009) study the impact of misallocation across establishments in explaining productivity in manufacturing in China and India. They find that moving to U.S. efficiency would increase TFP by 30-50% in China and 40-60% in India. The sizable TFP losses inspires a large body of literature studying the factors driving misallocation.

A more recent body of literature studies the role of particular distortions and performs theoretical comparative static analyses to quantify the effect of misallocation. For example, Peters (2013) studies the connection between entry intensity, misallocation and long-run growth. He finds that an increase in entry reduces misallocation by fostering competition. Since a higher entry intensity also encourages innovation and growth, misallocation is negatively correlated with growth. Hence misallocation has both static and dynamic welfare consequences. He also finds that the competition channel explains about 20% of the observed dispersion of marginal products. Midrigan and Xu (2014) find that the credit frictions cannot explain the large aggregate TFP losses from misallocation of factors in Colombia and South Korea. They find that although financial frictions are estimated to be large, efficient establishments are, nonetheless, able to accumulate internal funds and quickly grow out of their borrowing constraints. Their model predicts the TFP losses from misallocation are in the order of 2% for Korea, and only 1% for Colombia. David et al. (2014) study the effect of imperfect information on resource misallocation and aggregate productivity. They estimate the TFP losses range from 7-10% for productivity in China and India, and are smaller, though still significant, in the US. They argue that private learning plays a much more significant role in mitigating uncertainty and improving aggregate outcomes than learning from financial markets. This paper is closer to Peters (2013) in the sense that both papers examine the role of variable markups in explaining misallocation. While Peters (2013) focuses on the link between static misallocation and long-run growth, this paper aims to quantify the effects of internal and external distortions on misallocation.

This paper also relates to a branch of the new trade theory. For example, Melitz (2003) develops a dynamic industry model with heterogeneous firms to analyze the intra-industry effects of international trade. Melitz and Ottaviano (2008) develop another dynamic industry model with endogenous differences in the toughness of competition across markets. Aggregate productivity and the distribution of markups respond to the size of a market and the toughness of competition in that market. Mayer et al. (2014) extend the previous model to study the behavior of multi-product firms. They show that tougher competition in an export market induces the firm to reduce the set of exported products, and to skew the export sales towards their better-performing products. The reallocation of labor towards the core products suggests a new channel of productivity gains. The current paper emphasizes that the toughness of competition matters for resource misallocation, and for aggregate productivity.

The remainder of the paper is structured as follows. Section 2 outlines the model and its sectoral equilibrium. Section 3 discusses the data and identification issues. Section 4 presents the main results regarding the impact of imperfect competition on resource misallocation. Section 5 performs robustness checks. Section 6 concludes.

2 Model

In this section, I first lay down a model with linear demand structure and variable markups, which builds on Melitz and Ottaviano (2008). Next, I extend the accounting framework by Hsieh and Klenow (2009) to this model. I will then discuss efficient allocations. Let us start with the model.

2.1 Preferences

There is a unit mass of consumers in the economy, whose preference is given by

$$U = Y_0 + \beta \sum_{s=1}^{S} \sum_{i=1}^{N_s} Y_{si} - \frac{1}{2}\gamma \sum_{s=1}^{S} \sum_{i=1}^{N_s} Y_{si}^2 - \frac{1}{2}\delta \sum_{s=1}^{S} \left(\sum_{i=1}^{N_s} Y_{si} \right)^2 - \frac{1}{2}\eta \left(\sum_{s=1}^{S} \sum_{i=1}^{N_s} Y_{si} \right)^2, \tag{1}$$

where Y_{si} represents consumption levels of the numeraire good and each variety i in sector s. The demand parameters β, γ, δ and η are all positive. The parameters β and η govern the substitution pattern between varieties and the numeraire. An increase of β and a decrease of η shift out the demand for varieties relative to the numeraire. The parameter γ indexes the degree of product differentiation between the varieties, and δ the degree of product differentiation between the sectors. When γ and δ equal zero, consumers only care about their consumption level over all varieties. The varieties are perfect substitutes. The degree of product differentiation increases with γ and δ as consumers give increasing weight to the distribution of consumption levels across varieties and sectors.

In an equilibrium in which demands for all varieties are all positive, the inverse demand function for a particular variety is given by

$$P_{si} = \beta - \gamma Y_{si} - \delta Y_s - \eta Y, \tag{2}$$

where $Y_s \equiv \sum_{i=1}^{N_s} Y_{si}$ is the aggregate of sectoral goods, and $Y \equiv \sum_{s=1}^{S} Y_s$ the aggregate of all variety goods. This equation can be inverted to yield the linear demand function for variety goods,

$$Y_{si} = \frac{\beta}{\delta N_s + \gamma} + \frac{\delta N_s}{\delta N_s + \gamma} \frac{1}{\gamma} \overline{P}_s - \frac{\eta}{\delta N_s + \gamma} Y - \frac{1}{\gamma} P_{si}, \tag{3}$$

where N_s is the number of varieties in sector s, and $\overline{P}_s \equiv (1/N_s) \sum_{i=1}^{N_s} P_{si}$ is the price index for sector s. Given the linear demand system, there will be one choke price associated with each sector. The choke price for sector s is given by

$$P_{max,s} \equiv \frac{\beta\gamma}{\delta N_s + \gamma} + \frac{\delta N_s}{\delta N_s + \gamma} \overline{P}_s - \frac{\eta\gamma}{\delta N_s + \gamma} Y. \tag{4}$$

At any price equal to or above the choke price, consumer demand for the variety is driven to 0.

Combining equations (3) and (4) yields another expression for the demand function

$$Y_{si} = \frac{1}{\gamma}(P_{max,s} - P_{si}). \tag{5}$$

This equation clearly shows that each sector faces a different downward-sloping demand system. The price elasticity of a variety is $\varepsilon_{si} \equiv [-(\partial Y_{si}/\partial P_{si})(P_{si}/Y_{si})] = (P_{max,s}/P_{si} - 1)^{-1}$, and is decreasing as one move down along the demand curve.

Given the demand system, it is simple to write down the aggregate variables. Define $N \equiv \sum_{s=1}^{S} N_s$ as the total number of varieties in the economy. Then aggregate outputs can be expressed as $Y_s = N_s(P_{max,s} - \overline{P}_s)/\gamma$ for sector s, and $Y = N(\overline{P}_{max} - \overline{P})/\gamma$ for the whole economy, where $\overline{P}_{max} \equiv (1/N)\sum_{s=1}^{S} N_s P_{max,s}$ and $\overline{P} \equiv (1/N)\sum_{s=1}^{S} N_s \overline{P}_s$.

2.2 Production

Turning to technology and production. The numeraire good is produced under constant returns to scale at unit cost. One can think of it as being produced using a bundle of capital and labor such that the bundle cost of inputs is one. Entry in the differentiated product sector is costly as each firm incurs a start-up cost. Subsequent production exhibits constant returns to scale. Specifically, the production function takes the Cobb-Douglas form $Y_{si} = A_{si}K_{si}^{\alpha_s}L_{si}^{1-\alpha_s}$, and capital's share α_s is allowed to vary across sectors. As such, each firm faces a constant marginal cost. Firms that can cover their marginal cost survive and produce. All other firms exit the industry. Surviving firms compete in a monopolistically competitive market. Each firm maximizes its profit using the residual demand (5), while taking as given all aggregate variables.

In addition, each firm faces idiosyncratic external distortions: distortion that affects both the marginal product of capital and labor by the same proportion is $\tau_{Y_{si}}$, and distortion that affects only the marginal product of capital is $\tau_{K_{si}}$. Profit is then given by

$$\pi_{si} = (1 - \tau_{Y_{si}})P_{si}Y_{si} - wL_{si} - (1 + \tau_{K_{si}})RK_{si}. \tag{6}$$

Note capital and labor are assumed to be freely mobile across sectors. As a result, all firms face the same wage and interest rate. Profit maximization yields the optimal capital-labor ratio

$$\frac{K_{si}}{L_{si}} = \frac{\alpha_s}{1-\alpha_s}\frac{w}{R}\frac{1}{1+\tau_{K_{si}}}, \tag{7}$$

and the standard pricing rule for monopolistic producers

$$P_{si} = \mu_{si}\left(\frac{R}{\alpha_s}\right)^{\alpha_s}\left(\frac{w}{1-\alpha_s}\right)^{1-\alpha_s}\frac{(1+\tau_{K_{si}})^{\alpha_s}}{A_{si}(1-\tau_{Y_{si}})}, \tag{8}$$

where $\mu_{si} \equiv \varepsilon_{si}/(\varepsilon_{si} - 1) = P_{si}/(2P_{si} - P_{max,s})$ is the markup charged by the firm. The rest of the terms on the right-hand side constitute the marginal cost for the firm. For notational convenience, denote the marginal cost by C_{si}.

2.3 Industrial Equilibrium

Consider the free-entry industrial equilibrium. It is convenient to characterize a firm by its marginal cost. Conditional on survival, a firm chooses its output level by equating its marginal revenue to its marginal cost, $P_{max,s} - 2\gamma Y_{si} = C_{si}$. Combining this with the demand function (5) yields

$$Y_{si} = \frac{1}{\gamma}(P_{si} - C_{si}). \tag{9}$$

This equation characterizes the firm's behavior. As the marginal cost rises, the firm charges a lower price and operates at a lower scale. Let C_{sd} reference the cost of the marginal firm who is indifferent about exiting or remaining in the industry. The marginal firm earns zero profit as its price is driven down to equalize its marginal cost, $P_{sd} = C_{sd} = P_{max,s}$. As such, it faces zero demand and produces zero output. Firms with lower marginal cost earn positive profit and remain in the industry. The threshold C_{sd} summarizes the performance of all firms. Specifically, the firms' price, output level, revenue and profit are given by

$$P_{si} = \frac{1}{2}(C_{sd} + C_{si}), \tag{10}$$

$$Y_{si} = \frac{1}{2\gamma}(C_{sd} - C_{si}), \tag{11}$$

$$P_{si}Y_{si} = \frac{1}{4\gamma}(C_{sd}^2 - C_{si}^2), \tag{12}$$

and

$$\pi_{si} = \frac{1}{4\gamma}(C_{sd} - C_{si})^2. \tag{13}$$

Assume free entry for all sectors. An entrant pays an industry-specific fixed cost $f_{E,s}$ before learning its marginal cost[7]. Prior to entry, the expected firm profit is $\int_0^{C_{sd}} \pi_{si} dG_s(C_{si}) - f_{E,s}$. Unrestricted entry of new firms drives the expected profit to zero for all sectors. We have the free-entry condition as follows:

$$\int_0^{C_{sd}} \pi_{si} dG_s(C_{si}) = \frac{1}{4\gamma}\int_0^{C_{sd}} (C_{sd} - C_{si})^2 dG_s(C_{si}) = f_{E,s}, \tag{14}$$

where $G_s(C_{si})$ is the distribution of marginal cost for sector s, and $f_{E,s}$ the entry cost for the same sector. This equation defines the cutoff C_{sd} for each sector given the parameter $f_{E,s}$. The cutoff C_{sd} in turn determines the number of firms producing in sector s, N_s. Since the cutoff cost satisfies $P_{sd} = C_{sd} = P_{max,s}$, substituting this into the expression for choke price (4) gives the zero cutoff

[7] If we assume a firm reveals its marginal cost prior to entry, then the decision of entry is a cutoff rule. A firm enters if and only if its actual profit exceeds the entry cost $f_{E,s}$. The cutoff rule characterizes a static equilibrium. The assumption that entry cost is sunk before an entrant reveals its marginal cost is needed for modeling a dynamic equilibrium. The dynamic equilibrium is characterized by two conditions: a free-entry condition and a zero cutoff condition. In a stationary equilibrium, the mass of entrants times the probability of surviving equals the mass of firms in production.

profit condition,

$$N_s = \frac{2\gamma}{\delta}\frac{\beta - C_{sd} - \eta Y}{C_{sd} - \overline{C}_s}, \tag{15}$$

where $\overline{C}_s = [\int_0^{C_{sd}} C_{si} dG_s(C_{si})]/G_s(C_{sd})$ is the average cost of surviving firms. The number of entrants is then given by $N_{E,s} = N_s/G(C_{sd})$. This completes the description of industrial equilibrium.[8]

2.4 Accounting Framework

The allocation of resources across firms depends on firm-specific productivities, markups and direct distortions. In particular, the marginal revenue products of capital and labor are given by

$$MRPL_{si} \equiv \frac{1}{\mu_{si}}(1-\alpha_s)\frac{P_{si}Y_{si}}{L_{si}} = w\frac{1}{1-\tau_{Y_{si}}}, \tag{16}$$

$$MRPK_{si} \equiv \frac{1}{\mu_{si}}\alpha_s\frac{P_{si}Y_{si}}{K_{si}} = R\frac{1+\tau_{K_{si}}}{1-\tau_{Y_{si}}}. \tag{17}$$

And revenue TFP is given by

$$TFPR_{si} \equiv \frac{P_{si}Y_{si}}{K_{si}^{\alpha_{si}}L_{si}^{1-\alpha_{si}}} = \mu_{si}\left(\frac{MRPK_{si}}{\alpha_s}\right)^{\alpha_s}\left(\frac{MRPL_{si}}{1-\alpha_s}\right)^{1-\alpha_s}. \tag{18}$$

From these expressions, it is clear that external distortions ($\tau_{K_{si}}, \tau_{Y_{si}}$) and internal distortions (μ_{si}) are observationally equivalent. In general, it is difficult to tell from the data whether the dispersion in revenue TFP is driven by internal or external distortions. This paper makes a modest attempt to unravel the two sources.

It is now ready to consider the aggregate economy. One thing that is worth notice is that due to non-CES preference $(1/N_s)\sum_{i=1}^{N_s} P_{si}Y_{si} = \overline{P}_s\overline{Y}_s - (1/\gamma)\sigma_{p,s}^2$, i.e., there is an additional term associated with price dispersion in sector s. For this reason, I will measure sectoral output as $\sum_{i=1}^{N_s} P_{si}Y_{si}$ throughout, and carefully avoid the notation $\overline{P}_s\overline{Y}_s$. Denote sectoral capital and labor by $K_s \equiv \sum_{i=1}^{N_s} K_{si}$ and $L_s \equiv \sum_{i=1}^{N_s} L_{si}$. Then sectoral TFPR is the geometric mean of sectoral revenue products

$$\overline{TFPR}_s \equiv \frac{\sum_{i=1}^{N_s} P_{si}Y_{si}}{K_s^{\alpha_s}L_s^{1-\alpha_s}} = \overline{\mu_s}\left(\frac{\overline{MRPK_s}}{\alpha_s}\right)^{\alpha_s}\left(\frac{\overline{MRPL_s}}{1-\alpha_s}\right)^{1-\alpha_s}, \tag{19}$$

[8]The general equilibrium can be solved numerically by exploiting the recursive structure of the model. Let l denote the l-th iteration. Given the aggregate output $Y^{(l)}$, we can solve for sectoral cutoff C_{sd} and the number of surviving firms N_s using the free-entry condition (14) and zero cutoff condition (15). We then sum up quantities to obtain aggregate output $Y^{(l+1)} = \sum_{s=1}^{S} N_s\overline{Y}_s$, where $\overline{Y}_s = (C_{sd} - \overline{C}_s)/2\gamma$ is the average output of surviving firms. We find the equilibrium when we find a fixed point for Y, i.e., $Y^{(l)} = Y^{(l+1)}$.

where revenue products are given by

$$\overline{MRPK}_s = R \Big/ \left(\sum_{i=1}^{N_s} \frac{(1/\mu_{si}) P_{si} Y_{si}}{\sum_{j=1}^{N_s} (1/\mu_{sj}) P_{sj} Y_{sj}} \frac{1-\tau_{Y_{si}}}{1+\tau_{K_{si}}} \right), \tag{20}$$

$$\overline{MRPL}_s = w \Big/ \left(\sum_{i=1}^{N_s} \frac{(1/\mu_{si}) P_{si} Y_{si}}{\sum_{j=1}^{N_s} (1/\mu_{sj}) P_{sj} Y_{sj}} 1-\tau_{Y_{si}} \right), \tag{21}$$

and sectoral markups given by

$$\overline{\mu}_s = \left(\sum_{i=1}^{N_s} \frac{P_{si} Y_{si}}{\sum_{j=1}^{N_s} P_{sj} Y_{sj}} \frac{1}{\mu_{si}} \right)^{-1}. \tag{22}$$

It is clear that $\overline{MRPK}_s$, $\overline{MRPL}_s$ and $\overline{\mu}_s$ are weighted harmonic means of their respective firm measures. In addition, sectoral marginal revenue products are weighted by firms' variable costs, while sectoral markup is weighted by their revenues. Finally, define sectoral physical TFP as $\overline{TFPQ}_s \equiv Y_s/(K_s^{\alpha_s} L_s^{1-\alpha_s})$. It can be expressed as

$$\overline{TFPQ}_s = \sum_{i=1}^{N_s} \frac{P_{si} Y_{si}}{\sum_{j=1}^{N_s} P_{sj} Y_{sj}} \frac{\overline{TFPR}_s}{TFPR_{si}} TFPQ_{si}. \tag{23}$$

This is the key equation that illustrates how dispersions of revenue TFP would affect aggregate physical TFP. For later use, let the sectoral TFP also be denoted by $\overline{A}_s^{DE}$, where DE stands for the decentralized economy with market distortions.

2.5 Efficient Allocations

Consider the first-best allocation of the resource. The social planner would choose to equalize revenue TFP across firms within each sector, i.e., $TFPR_{si} = \overline{TFPR}_s$ for all s. The first-best sectoral TFP is

$$\overline{A}_s^{FB} = \sum_{i=1}^{N_s} \frac{P_{si} Y_{si}}{\sum_{j=1}^{N_s} P_{sj} Y_{sj}} TFPQ_{si}. \tag{24}$$

However, in a decentralized economy with monopolistic competition, the first-best allocation cannot be achieved.

The constrained optimal allocation, i.e., the second-best, would be to equalize the marginal revenue of products across firms within the same sector. That is, the planner finds it optimal to eliminate all external distortions. In this case, $MRPK_{si} = \overline{MRPK}_s = R$, $MRPL_{si} = \overline{MRPL}_s = w$, and $TFPR_{si}/\overline{TFPR}_s = \mu_{si}/\overline{\mu}_s$ for all s. The second-best sectoral TFP can be expressed as

$$\overline{A}_s^{SB} = \sum_{i=1}^{N_s} \frac{(1/\mu_{si}) P_{si} Y_{si}}{\sum_{j=1}^{N_s} (1/\mu_{sj}) P_{sj} Y_{sj}} TFPQ_{si}. \tag{25}$$

It can be verified that $\overline{A}_s^{FB} \geq \overline{A}_s^{SB} \geq \overline{A}_s^{DE}$. Intuitively, sectoral TFP is high when more weights are assigned to more productive firms. The first-best allocation weighs by revenues, the second-best weighs by production costs (i.e., $(1/\mu_{si})P_{si}Y_{si} = C_{si}Y_{si}$), and the decentralized economy with distortions weighs by physical inputs (i.e., $P_{si}Y_{si}/TFPR_{si} = K_{si}^{\alpha_s}L_{si}^{1-\alpha_s}$). Since more productive firms earn relative higher revenues, weighting by revenues results in a higher sectoral TFP than weighting by production costs. Hence, the first-best sectoral TFP is higher than the second-best one. Without distortions, weighing by production costs is equivalent to weighing by physical inputs, since the bundle costs are the same within each sector. With distortions, however, productive firms operate on a lower scale and use fewer physical inputs than they would do without distortions. Therefore, the decentralized economy with distortions has a lower sectoral TFP than the second-best economy.

3 Data

The primary source of data is India's Annual Survey of Industries (ASI). India's ASI is conducted annually by its Central Statistical Organization. The basic survey unit is plant, also known as establishment. I analyze three years of data: 2001, 2004 and 2007. The three years correspond to the first year, the middle year and the last year of the data available. The raw data of 2001 contains about 26 thousand plants, and the data of 2004 and 2007 each contains about 36 thousand plants.

The variables used in the analysis include firms' value added, profits, fixed capitals, mandays, wages, bonuses and benefits. To make the study comparable with that of Hsieh and Klenow (2009), I take the average of the book values of capital at the beginning of the year and at the end of the year as the plant's capital stock, and take the sum of wages, bonuses and benefits as its employees' compensation. For capital share in each industry, α_s, I use the US shares in the NBER Productivity Database. In particular, I use the 2007 US data for it corresponds to the last year of the ASI data. The use of US shares underlies the assumption that all countries have access to the world technology frontier, and allows me to focus on India's allocative efficiency instead of its technology.[9] Finally, to alleviate the effects of outliers, I trim 1% data of TFP and TFPR dispersions and recompute all performance variables as in Hsieh and Klenow (2009).

To separate the effects of internal and external distortions, it is necessary to identify firms' markups and external distortions from the data. In general, identification is difficult because internal distortions and external distortions are observationally equivalent. When one observes a high dispersion of revenue productivity, it is equally valid to interpret it as internal distortions or as external distortions. This point is clearly illustrated in equation (18). To circumvent this problem, I adopt a sequential identification strategy. I first identify firm-specific markups from data on revenue and profits, and then identify external distortions and individual productivities taking

[9]The use of exogenous capital shares also eases the exposition. As a robustness check, section 5.2 estimates and uses the native capital shares instead of their US counterparts, and shows that switching to native shares does not change the main results.

as given firms' markups. The key identification equations are given below. We have

$$\mu_{si} = \frac{(1-\alpha_s)\,\pi_{si}}{wL_{si}} + 1, \tag{26}$$

$$1 + \tau_{K_{si}} = \frac{\alpha_s}{1-\alpha_s}\frac{wL_{si}}{RK_{si}}, \tag{27}$$

$$1 - \tau_{Y_{si}} = \mu_{si}\frac{wL_{si}}{(1-\alpha_s)\,P_{si}Y_{si}}, \tag{28}$$

and

$$A_{si} = \kappa_s \frac{\left(1+\frac{1}{\varepsilon_{si}}\right)P_{si}Y_{si}}{K_{si}^{\alpha_s}L_{si}^{1-\alpha_s}}, \tag{29}$$

where $\kappa_s = 1/P_{max,s}$ can be set to one for all sectors without affecting the dispersion of TFPR. To derive the last equation, start from the definition of price elasticity, $\varepsilon_{si} = (P_{max,s}/P_{si} - 1)^{-1}$. Rearrange to get $Y_{si} = (1/P_{max,s})(1+1/\varepsilon_{si})P_{si}Y_{si}$ and plug into the expression $A_{si} = Y_{si}/(K_{si}^{\alpha_s}L_{si}^{1-\alpha_s})$. The variables in the right-hand side of each equation above are either observable, or can be identified using preceding equations. This makes it possible to identify the key variables in a sequential manner.

4 Quantifying Internal Distortions

This section examines the central question of the paper: How much of the dispersions in revenue TFP can be explained by variable markups? In other terms, suppose that all external distortions were removed, how much of the dispersion of TFPR would remain? To answer this question, I perform a counterfactual experiment. The experiment relies on a sufficient statistic of the model: the threshold cost for which a firm is indifferent about remaining in the industry. The threshold cost is endogenously determined by the free-entry condition. The experiment proceeds in two steps. In the first step, I estimate distributional parameters that are needed for evaluation of the free-entry condition. In the second step, I remove external distortions and compute threshold costs by evaluating the free-entry condition. I then recompute the equilibrium variables and determine the level of misallocation and aggregate productivity.

The remainder of the section is structured as follows. Subsection 4.1 presents the methodology. Subsections 4.2 and 4.3 detail the two steps and present the results. Subsection 4.4 describes a naive direct measure of the effects of internal and external distortions. It shows that the naive measure provides a misleading figure and substantiates the need for our well-designed policy experiment.

4.1 Methodology

The idea for a proper policy experiment is simple. We can take advantage of the key property of the model: the performance of all firms is summarized by the threshold C_{sd}. Suppose we know the parameters of the model, we can compute price, output, revenue, profit and other equilibrium variables starting from equation (10)-(13). Using those equilibrium variables, we could compute TFPR, markups and marginal products, and the levels of TFPR dispersion and aggregate productivity. This records the level of misallocation prior to the removal of external distortions.

We then liberalize the economy by removing all external distortions. Suppose the demand parameter γ and entry cost $f_{E,s}$ will not change, then from the free-entry condition (14) we have

$$\int_0^{C_{sd}} (C_{sd} - C_{si})^2 dG_s(C_{si}) = \int_0^{\tilde{C}_{sd}} (\tilde{C}_{sd} - C_{si})^2 d\tilde{G}_s(C_{si}), \tag{30}$$

where $\tilde{C}_{sd}$ and $\tilde{G}_s(\cdot)$ are the counterfactual cutoff and distribution function for marginal costs. If we know the distribution functions G and $\tilde{G}$, we could estimate C_{sd} from data and compute the counterfactual $\tilde{C}_{sd}$ from the equation above. With the $\tilde{C}_{sd}$ at hand, we could then compute firms' counterfactual price, output level, revenue and profit again using equations (10)-(13). It is also possible to allow for endogenous exit. If a firm finds it unprofitable to produce after the liberalization, it could choose to exit rather than to produce. Finally, we could calculate the dispersion of TFPR and contrast it to the one we have computed before, and conclude how much of the dispersion of TFPR can be attributed to internal distortions.

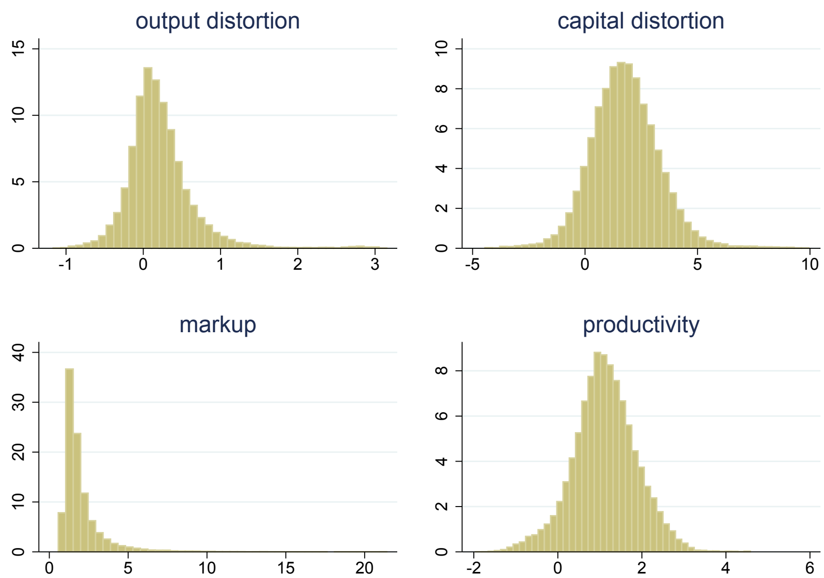
To perform the above experiment, it is necessary to estimate model parameters. Fortunately, I only need to estimate the distributional parameters for productivity and external distortions, to which now I turn.

4.2 Step 1: Estimation of Distributional Parameters

To determine G and $\tilde{G}$, I need to impose distributional assumptions on productivity and distortions. To motivate the distributional assumptions, it is instructive to inspect the data. Figure 1 illustrates the log of firm-specific technology, distortions and markups for year 2004. There are two commonly used distribution assumptions for technology, the Pareto distribution and the lognormal distribution. The data suggests that a lognormal distribution centered on one would be a reasonable approximation. By the same token, the output and capital distortions can be approximated by lognormal distributions. For surviving firms, both distortions tend to have positive means, with the latter much more dispersed than the former. A positive mean for output distortion suggests that surviving firms tend to be those which experience positive shocks. On the other hand, a positive mean for capital distortion indicates that on average firms in India operate in more labor-intensive technologies than their US counterparts.

Building on intuitions from the data, I assume that $(\ln A_{si}, \ln(1-\tau_{Y_{si}}), \ln(1+\tau_{K_{si}}))$ is multivariate normal. Let the mean and variance of $\ln A_{si}$ be denoted by μ_A and σ_A^2, the mean and variance of $\ln(1-\tau_{Y_{si}})$ and $\ln(1+\tau_{K_{si}})$ by μ_Y, μ_K, σ_Y^2, and σ_K^2, respectively, and their covariance by σ_{KY}. These assumptions are the same with Hsieh and Klenow (2009). For convenience, I reproduce the

Figure 1: Histograms of productivity, distortions and markup



expression for marginal cost below:

$$C_{si} \equiv \left(\frac{R}{\alpha_s}\right)^{\alpha_s} \left(\frac{w}{1-\alpha_s}\right)^{1-\alpha_s} \frac{(1+\tau_{K_{si}})^{\alpha_s}}{A_{si}(1-\tau_{Y_{si}})}. \tag{31}$$

From this expression, it is easy to see that C_{si} is lognormally distributed with mean $-\mu_s \equiv -\mu_A - \mu_{\tau s}$ up to a constant, where $\mu_{\tau s} \equiv \mu_Y - \alpha_s \mu_K$, and variance $\sigma_s^2 \equiv \sigma_A^2 + \sigma_{\tau s}^2$, where $\sigma_{\tau s}^2 \equiv \sigma_Y^2 + \alpha_s^2 \sigma_K^2 - 2\alpha_s \sigma_{KY}$. And $\tilde{C}_{si}$ is also lognormally distributed but with mean $-\mu_A$ and variance σ_A^2. Therefore, I need to estimate these parameters together with the cutoffs C_{sd} from the data.

Let the parameters be denoted by $\theta = (\mu_A, \mu_Y, \mu_K, \sigma_A^2, \sigma_Y^2, \sigma_K^2, \sigma_{KY})$. To identify these parameters, I perform a maximum likelihood estimation on the incidentally truncated data. For this purpose, it is more convenient to work with productivity rather than marginal cost, as the mean of the latter involves a sector-specific constant. The data selection rule is given by

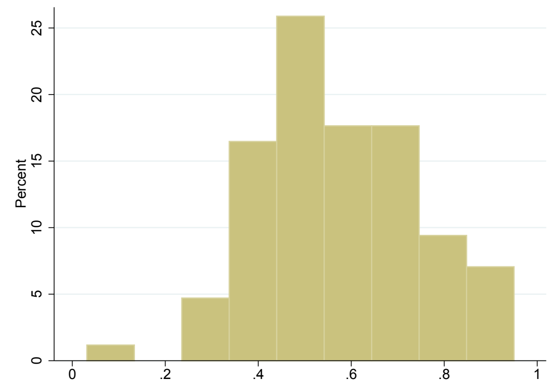
$$z_{si} \equiv \frac{A_{si}(1-\tau_{Y_{si}})}{(1+\tau_{K_{si}})^{\alpha_s}} \geq z_{sd}. \tag{32}$$

That is, only those firms with *de facto* productivity greater or equal to the industrial threshold choose to produce. Notice that productivity z_{si} is lognormally distributed with mean μ_s and variance σ_s^2. Also, note there is a one-to-one mapping from z_{sd} to C_{sd}. For notational convenience, let us define effective distortion as $\ln \tau_{si} = \ln(1-\tau_{Y_{si}}) - \alpha_s \ln(1+\tau_{K_{si}})$.

The conditional density of observing a particular data point is given by

$$f(\ln A_{si},\ \ln \tau_{si} | \ln z_{si} \geq \ln z_{sd}) = \frac{f(\ln A_{si},\ \ln \tau_{si})}{\Pr(\ln z_{si} \geq \ln z_{sd})}, \tag{33}$$

Figure 2: Distribution of capital shares for matched sectors



provided $\ln z_{si} \geq \ln z_{sd}$ holds. By noting that $\ln A_{si}$ and $\ln \tau_{si}$ are independent, I could rewrite the conditional density as

$$f(\ln A_{si}, \ln \tau_{si} | \ln z_{si} \geq \ln z_{sd}) = \frac{1}{\sigma_A} \phi\left(\frac{\ln A_{si} - \mu_A}{\sigma_A}\right) \frac{1}{\sigma_{\tau s}} \phi\left(\frac{\ln \tau_{si} - \mu_{\tau s}}{\sigma_{\tau s}}\right) \Big/ \Phi\left(\frac{\mu_s - \ln z_{sd}}{\sigma_s}\right), \tag{34}$$

where Φ is the distribution function of $N(0,1)$ and ϕ the associated density function.

Since the maximum likelihood estimation is conventional, details are relegated to the appendix. Instead, it would be interesting to discuss some intuitions about the estimation strategy. The estimated model has been constructed in a parsimonious and effective way. The conditional density function is on productivities and effective distortions only. The identification of parameters of productivity is straightforward, since we have assumed that productivity is independent of distortions. The identification of parameters associated with distortions is subtler. To estimate σ_Y^2, σ_K^2 and σ_{KY} from a single sequence of effective distortions, I rely on the fact that 1, α_s, and α_s^2 are independent polynomials. To see how this is relevant, recall that $\sigma_{\tau s} \equiv \sigma_Y^2 + \alpha_s^2 \sigma_K^2 - 2\alpha_s \sigma_{KY}$. While $\sigma_{\tau s}$ can be determined from data of effective distortions, it can be uniquely decomposed into three components when 1, α_s, and α_s^2 are independent polynomials. This explains why the simple strategy is effective in action.

The strategy of the estimation relies on the variability of sector-specific capital share. Fortunately, there is a great deal of variation in the data. Table 1 reports the descriptive statistics of capital shares for those sectors that have a counterpart in the US. The data is consistent in all three sampling years. There are more than 80 matched sectors in each year. Capital share centers on 0.57, and has a standard deviation of about 0.17. The minimum capital share is 0.031 and the maximum is 0.950. The large variation in capital share reassures us that the estimation will be effective. Figure 2 illustrates the distribution of capital share for 2004. The diagrams for the other two years are virtually identical.

The estimates of distributional parameters are reported in Table 2. The estimated parameters confirm the patterns of data. In particular, the center for (log) productivity is slightly negative.

Table 1: Descriptive statistics for matched sectoral-specific capital share

Year	Mean	Median	S.D.	Min	Max	#obs
2001	0.571	0.548	0.174	0.031	0.950	82
2004	0.572	0.550	0.173	0.031	0.950	85
2007	0.569	0.546	0.177	0.031	0.950	81

Table 2: Maximum likelihood estimation of parameters

Year	μ_A	μ_Y	μ_K	σ_A^2	σ_Y^2	σ_K^2	σ_{KY}	#obs
2001	-0.412***	0.539***	3.038***	0.689***	0.361***	3.237***	0.899***	17887
	(0.0119)	(0.0226)	(0.0448)	(0.0112)	(0.0238)	(0.137)	(0.0580)	
2004	-0.344***	0.671***	3.212***	0.613***	0.450***	3.901***	1.171***	27609
	(0.00818)	(0.0156)	(0.0333)	(0.00765)	(0.0190)	(0.109)	(0.0456)	
2007	-0.321***	0.740***	3.314***	0.593***	0.438***	3.982***	1.172***	24475
	(0.00827)	(0.0175)	(0.0381)	(0.00773)	(0.0192)	(0.123)	(0.0486)	

Notes: ** 5%, *** 1% levels of significance.

The variance of capital distortion is much larger than that of output distortion. Interestingly, output distortion is positively correlated with capital distortion. This implies that, on average, firms receiving subsidies tend to be those which operate in labor-intensive technologies. It thus helps us understand why removing distortions would improve India's aggregate productivity.

4.3 Step 2: Quantifying Effects of Imperfect Competition

Now that I have the distributional parameters, I could perform the policy experiment explained in section 4.1. To proceed, I first compute equilibrium variables using equations (26)-(29), and calculate the dispersion of revenue TFP. Next, I remove all external distortions and compute the counterfactual cutoff prices $\tilde{C}_{si}$. At this stage, endogenous exits are allowed. If a firm finds itself making a loss after liberalization, it can choose not to produce. I then recompute equilibrium variables using equations (26)-(29), and recalculate the TFPR dispersion around its sectoral means. The ratio of TFPR dispersion before and after removing distortions is interpreted as the remaining fraction of TFPR dispersion after the liberalization. Table 3 reports the three numbers for each of the sampling years.

The experiment shows that in 2001, if market distortions were removed, TFPR dispersion would drop from 0.362 to 0.301. The removal of distortions is not very effective in reducing TFPR dispersion. Indeed, about 83.1% of the TFPR dispersion still remain due to variable markups. The percentages remaining for the other two years are very similar, with both of them exceeding

Table 3: TFPR dispersions before and after liberalization

TFP Dispersion	2001	2004	2007
Before Liberalization	0.362	0.338	0.330
After Liberalization	0.301	0.284	0.281
% Remained	83.1	83.9	85.2

Table 4: OLS regression of productivity on effective distortion

Dependent variable: $\ln TFPQ_{si}$

	(1) 2001	(2) 2004	(3) 2007
$\ln \tau_{si}$	-0.709***	-0.698***	-0.718***
	(0.037)	(0.036)	(0.040)
Industry FE	yes	yes	yes
N	17887	27609	24475
R^2	0.339	0.330	0.335

Notes: ** 5%, *** 1% levels of significance. Standard errors are clustered at the industry level.

Table 5: First-best and second-best TFP gains

TFP Gain	2001	2004	2007
First-best	88.7	87.3	86.1
Second-best	16.6	14.4	13.7

80%. Interestingly, despite markups only accounting for about 30% of the total variations in TFPR before liberalization (see section 4.4), after the removal of distortions, more than 80% of the TFPR dispersion still remains. This suggests that distortions somehow reduce the variation in markups.

To understand the factors that drive the results, it is instructive to inspect the data. The sample correlation between the effective distortions and firms' productivities is negative. To further confirm this negative relationship, I run OLS regressions of firms' productivities on their effective distortions, controlling for industrial fixed effects. The OLS estimates are reported in Table 4. These regressions confirm the robust negative correlation between productivity and effective distortions.

The negative correlation is consistent with the predictions of our model. Recall, a firm chooses to produce only if $\ln A_{si} + \ln \tau_{si} \geq \ln z_{sd}$. This implies that, conditional on survival, a firm that draws a high productivity is likely to face an unfavorable distortion, and vice versa. Since marginal cost depends on a firm's productivity as well as effective distortion, the negative correlation tends to reduce the dispersion of marginal costs. In our settings, a firm with lower marginal cost produces more, earns more revenue, and charges higher markup and price. A low dispersion of marginal costs translates into a low dispersion of markups. This is the reason why distortions reduce the variations in markups.

It is also interesting to see the effects of variable markups on aggregate productivity. In the section 2.5, I briefly discussed the first-best and second-best allocations. I now compute the TFP gains of efficient allocations relative to the actual economy. Denote the first-best, second-best and the actual level of output by Y^{FB}, Y^{SB} and Y, respectively. All the three variables are evaluated using our model. Given these variables, the TFP gain of moving to the first-best is given by $100 \times (Y^{FB}/Y - 1)$. The definition for TFP gain of moving to the second-best is similar. The results are reported in Table 5 below.

The data in 2001 reveals that the TFP gains of moving to efficient allocations are: 88.7% to the first-best, and 16.6% to the second-best. The first-best gain is much higher than the second-best one, as the former eliminates both market distortions and markup variations, whilst the latter only eliminates market distortions. This pattern repeats itself in the other two sampling years. Given the results for TFPR dispersion, it is not surprising that second-best TFP gains are very modest. When the market is liberalized, the dispersion of markups blows up, negating the effect of removing market distortions. As a result, second-best TFP gains are small. The first-best TFP gains, however, reflecting the effects of removing market distortions and variable markups, are much larger than the second-best ones. Nevertheless, the first-best is not implementable given the imperfectly competitive market structure. The second-best can be achieved by removing all market distortions.

4.4 A Naive TFPR Decomposition

This subsection describes a naive decomposition of TFPR dispersions. It shows that the naive measure provides misleading results and justifies the need for our policy experiment. Consider the deviation of TFPR from the industrial mean

$$\ln\left(\frac{TFPR_{si}}{\overline{TFPR}_s}\right) = \ln\left(\frac{\mu_{si}}{\overline{\mu}_s}\right) + \ln\left[\left(\frac{MRPK_{si}}{\overline{MRPK}_s}\right)^{\alpha_s}\left(\frac{MRPL_{si}}{\overline{MRPL}_s}\right)^{1-\alpha_s}\right]. \tag{35}$$

For notational convenience, denote the three terms in the preceding equation by $TF\hat{P}R_{si}$, $\hat{\mu}_{si}$, and $M\hat{R}PF_{si}$ (F means factors). The variance of TFPR can be decomposed as

$$\mathrm{Var}\left(TF\hat{P}R_{si}\right) = \beta_1\,\mathrm{Var}\left(\hat{\mu}_{si}\right) + \beta_2\,\mathrm{Var}\left(M\hat{R}PF_{si}\right), \tag{36}$$

where the coefficients are given by $\beta_1 = \frac{\mathrm{Cov}(TF\hat{P}R_{si},\hat{\mu}_{si})}{\mathrm{Var}(\hat{\mu}_{si})}$ and $\beta_2 = \frac{\mathrm{Cov}(TF\hat{P}R_{si},M\hat{R}RF_{si})}{\mathrm{Var}(M\hat{R}PF_{si})}$.

The coefficient β_1 captures how much the volatility of TFPR would increase if the volatility of the markup increases by one unit, holding the volatility of revenue marginal products constant. A similar interpretation applies to β_2. Both coefficients can be evaluated using corresponding sample moments. Moreover, the two terms in the right-hand side constitute the variance decomposition of the left-hand side variable. In particular, the first term divided by the variance of TFPR gives the fraction of variance that comes from markup variations. The second term divided by the variance of TFPR yields the fraction of variance that results in marginal product variations.

The results are presented in Table 6. All three years show consistent results. The variance decomposition shows that about one-third of the variance in TFPR dispersions is attributed to variable markups. The remaining two-thirds are attributed to revenue marginal products. Moreover, the direct contribution of markups to TFPR dispersions, as captured in β_1, stays roughly constant (at 0.63) over the three sampling years. In all three years, the direct contributions of markups β_1 are slightly smaller than those of market distortions β_2. Overall, the direct measure suggests that external distortions are about twice as important as internal distortions.

Why does the direct measure provide such a misleading picture? The reason is that it ignores the interaction of the two kinds of distortions. As explained in the previous subsection, conditional on survival, the productivities and external distortions of the firms are negatively correlated.

Table 6: Variance decomposition of TFPR dispersions

Year	Markup	MRP	β_1	β_2
2001	32.2	67.8	0.63	0.78
2004	33.0	67.0	0.62	0.77
2007	33.9	66.1	0.65	0.78

That is, a firm survives either because it draws a high productivity, or because it receives a subsidy to compensate for its low productivity. As a result, the distribution of marginal cost, which reflects both productivity and taxes, is less dispersed under the presence of external distortions. This suggests that the direct measure underestimates the importance of internal distortions, and overestimates the importance of external distortions. This exercise substantiates the need for our well-designed policy experiment.

5 Robustness Analysis

The results may seem startling – over 80% of resource misallocation can be attributed to imperfect competition. The benefit from eliminating distortions is largely negated by the cost of rising variation in markups, making the change of policy virtually ineffective. It is natural to wonder: How robust are these results? This section performs several checks and shows that these findings are fairly robust.

5.1 Labor Distortions

The model in the main text does not include a labor distortion. Is there any risk that the big role of markups comes from the misspecification of distortions? I now include a labor distortion, $1+\tau_{L_{si}}$, to the benchmark model. Profit is now given by

$$\pi_{si} = (1-\tau_{Y_{si}})P_{si}Y_{si} - (1+\tau_{L_{si}})wL_{si} - (1+\tau_{K_{si}})RK_{si}. \tag{37}$$

The methodology in the main text broadly goes through, except for a few amendments. First, I need to embody the labor distortion in the marginal products. Second, I need additional identification of labor distortion from the data. The identification strategy is simple, since I have data on compensations and man-days for all establishments. Suppose labor is mobile within the country, wage will be equalized across sectors. We have the relationship $comp_{si} = wL_{si}(1-\tau_{L_{si}})$, where $comp_{si}$ is the compensations, w the national wage, and L_{si} the man-days. Then labor distortions can be revealed as the OLS residuals of the following regression while restricting the coefficient γ_1 to be unity,

$$\ln comp_{si} = \gamma_0 + \gamma_1 \ln L_{si} + u_{si}. \tag{38}$$

Furthermore, the constant γ_0 is the log of our national wage. After recovering labor distortions, I proceed to identify markups, productivities and other distortions using equation (26)-(29), with

Table 7: TFPR dispersions before and after liberalization

TFP Dispersion	2001	2004	2007
Before Liberalization	0.421	0.399	0.395
After Liberalization	0.327	0.315	0.316
% Remained	77.6	78.8	80.1

Table 8: First-best and second-best TFP gains

TFP Gain	2001	2004	2007
First-best	106.3	102.4	106.2
Second-best	22.5	18.8	18.4

the term wL_{si} replaced with $comp_{si}$.

Third, I need to modify the estimation procedures for distributional parameters. Assume that $(\ln A_{si}, \ln(1-\tau_{Y_{si}}), \ln(1+\tau_{K_{si}}), \ln(1+\tau_{L_{si}}))$ is multivariate normal. As before, assume that distortions are independent of productivities. Let the variance of distortions be denoted by σ_Y^2, σ_K^2, and σ_L^2, and their covariance by σ_{KY}, σ_{LY} and σ_{KL}. There are 11 distributional parameters awaiting estimation. Let the parameters be denoted by $\theta = (\mu_A, \mu_Y, \mu_K, \mu_L, \sigma_A^2, \sigma_Y^2, \sigma_K^2, \sigma_L^2, \sigma_{KY}, \sigma_{LY}, \sigma_{KL})$. The conditional density function is still given by (34), except that the second term is replaced by a trivariate normal density for $\ln(1-\tau_{Y_{si}})$, $\ln(1+\tau_{K_{si}})$ and $\ln(1+\tau_{L_{si}})$. The estimation proceeds broadly as before. For conciseness, I will not elaborate on the procedures. Instead, I turn now to reporting the results.

Table 7 reports the dispersion of revenue TFP before and after removing distortions. It shows that about 80% of TFPR dispersion would remain even if all distortions are wiped out. This is consistent with the results in the benchmark model. Table 8 presents the TFP gain from eradicating market distortions. If there were no variable markups, it is expected that the aggregate TFP will increase by 106.3%. With variable markups, however, the TFP gains shrivel to 22.5%. This comes as no surprise given that about 80% of the TFPR dispersion remains in the market. These consistent results suggest that our main findings are not driven by the omission of a labor distortion, but lie in the fact that the market is imperfectly competitive.

5.2 Indigenous Production Functions

Turning to the next robustness analysis. In the benchmark identification, I follow Hsieh and Klenow (2009) and use the US capital shares for each industry. This helps to unravel the effect of distortions from that of production technologies. However, it may be argued that using the US shares might lead to an overestimation of the distortions. More importantly, do the main conclusions of this paper hinge on the choice of capital shares? To address these concerns, I estimate capital shares from India's data, and use these shares to re-evaluate the effect of imperfect competition.

To estimate the capital shares, I resort to the method developed in Olley and Pakes (1996). The basic idea is as follows. Suppose we want to estimate a Cobb-Douglas production function. If we proceed with convention OLS, the coefficients will be biased due to two reasons. First, since labor is variable input, it is likely to be correlated with firm's productivity. This would lead to an upward bias for labor's share and a downward bias for capital's share. This is a simultaneity effect. Second, there is an additional selection effect. That is, a firm that survives in the current period is likely to have experienced a favorable productivity shock in last period. Since firms with higher capital stocks can withstand worse productivity shocks, ignoring the selection effect would lead to an underestimation of the capital share. To correctly estimate a production function, Olley and Pakes (1996) propose using investment as a proxy to control for firm-specific productivity. I adopt their method, but focus on the simultaneity problem for two reasons. The first is due to data constraints. To address the selection problem, I would need a panel data. However, the data at my disposal is a cross-section dataset. Second, as shown in their paper, once the simultaneity problem has been addressed, their selection correction does not change their results. Hence I focus on the first problem. I will sketch the estimation procedure below. The interested reader is referred to the original paper.

The starting point is our Cobb-Douglas production function,

$$R_{si} = P_s(z_{si}) z_{si} K_{si}^{\beta_k} L_{si}^{\beta_l} e^{u_{si}}, \tag{39}$$

where R_{si} is value added, z_{si} is the effective productivity and u_{si} is measurement errors. Note the price $P_s(\cdot)$ is a function of the effective productivity. Taking logs, we obtain

$$r_{si} = \beta_{sk} k_{si} + \beta_{sl} l_{si} + \omega_{si} + p_s(\omega_{si}) + u_{si}, \tag{40}$$

where I use lower case to denote a variable in log terms, and $\omega_{si} \equiv \ln z_{si}$. From the preceding equation, it is clear that ω_{si} is correlated with l_{si}. This is the source of the simultaneity problem. Labor is assumed to be the only variable factor. Investment is made in the previous period and capital is predetermined. This suggests that optimal investment policy takes the form $i_{si} = i_s(\omega_{si}, k_{si})$. Provide $i_{si} > 0$, this equation is strictly increasing in ω for every k. As a result, for the subset of observations for which $i_{si} > 0$, we can invert it and write $\omega_{si} = h_s(i_{si}, k_{si})$. This function allows us to express firm's productivity as a function of observables, and hence to control for it in estimation.

Substituting this function into (40) we have

$$r_{si} = \beta_{sl} l_{si} + \phi_s(\omega_{si}, k_{si}) + u_{si}, \tag{41}$$

where

$$\phi_s(\omega_{si}, k_{si}) = \beta_{sk} k_{si} + h_s(i_{si}, k_{si}) + p_s(h_s(i_{si}, k_{si})). \tag{42}$$

The partial linear model in (41) is a semiparametric model, which can be implemented by regressing value added on labor and a polynomial in investment and capital. This identifies β_{sl} for each sector. Given that we have assumed that production is CRS, capital's share is simply $1 - \beta_{sl}$. This approach has an advantage that, for the current problem, it is no more difficult to implement than OLS. The rest of the procedure is the same as in the main text, except that the US capital shares are replaced

Table 9: TFPR dispersions before and after liberalization

TFP Dispersion	2001	2004	2007
Before Liberalization	0.306	0.284	0.297
After Liberalization	0.260	0.246	0.253
% Remained	85.0	86.1	86.0

Table 10: First-best and second-best TFP gains

TFP Gain	2001	2004	2007
First-best	77.5	75.5	79.4
Second-best	15.4	14.4	14.8

with the indigenous ones.

Table 9 reports the result on dispersions of revenue TFP. It shows that over 80% of the TFPR dispersion remains in the market despite the fact that external distortions are eradicated. The result is even stronger than that in the benchmark case. Moreover, it is not surprising that switching to native production functions reduces the levels of TFPR dispersions. When we use the US capital shares, the TFPR dispersions reflect both the direct distortions, and the indirect wedges from failing to adopt more advanced production technologies. After switching to the indigenous shares, the TFPR dispersions only reflect the direct distortions.

The result on first-best and second-best TFP gains, as reported in Table 10, follows the same pattern. That is, the levels of gain for both efficient allocations decrease, but the fact that the first-best gain is much higher than the second-best gain remains unaltered. Overall, these results suggest that our main findings are not driven by the adoption of US production functions. Rather, they come from the fact that market is not fully competitive.

6 Concluding Remarks

Factor misallocation across productive uses undermines aggregate economic performance. The factors driving misallocation can be internal or external to the market structure. Variation in markups associated with imperfect competition is an internal distortion. Taxes and subsidies on input and output markets are examples of external distortions.

Previous literature has documented empirically that removing external distortions leads to a significant improvement of aggregate productivity. However, the conclusion hinges on the assumption that all varieties within the same sector are equally substitutable. Under this assumption, all firms have equal monopolistic power and they charge the same markup. Since firms only compete in extensive margins, there is no resource misallocation due to internal distortions. While the CES demand structure provides a nice benchmark, it might be restrictive to model markets in reality. If firms could charge different markups, there would be an endogenous distribution of markups which leads to misallocation of factors of production. This paper argues that internal distortion is an important source of misallocation. Since internal distortion is an inner feature of the underlying market structure, eliminating external distortions alone is less effective in improving aggregate productivity.

To pursue this idea, I adopt the linear demand structure proposed by Melitz and Ottaviano (2008). Combined with the accounting framework in Hsieh and Klenow (2009), I can separate the effect of internal distortions from that of external distortions. Using India's manufacturing data, I perform a policy experiment to quantify the effects of internal distortions. In this experiment, I first document the level of misallocation observed in the market. I then remove all external distortions, and allow firms to reoptimize over prices and quantities, and to choose whether to produce or to exit. Surprisingly, I find that more than 80% of the observed misallocation remains after the removal of external distortions. Elimination of external distortions is ineffective to improve market allocative efficiency and aggregate productivity. The experiment shows that aggregate productivity increases by 14-17% after the liberalization. By contrast, the removal of internal and external distortions has a fivefold effect on aggregate productivity (86-89%). The intuition is that despite the adverse effect, external distortions play a role in suppressing the dispersion of markups. When external distortions are removed, the dispersion of markups blows up, negating the gain from the elimination of external distortions. As a result, removing external distortions has only a modest effect on aggregate productivity.

The magnitude of gain certainly depends on the assumption of market structure. This paper relies on a particular structure that allows variable markups. Thus the magnitude of gain must await further investigation. The bottom line is that internal distortions may also lead to resource misallocation. Fostering market competition can be an effective means to enhance macroeconomic performance.

References

Amemiya, Takeshi and Michael Boskin, "Regression analysis when the dependent variable is truncated lognormal, with an application to the determinants of the duration of welfare dependency," *International Economic Review*, 1974, *15*, 485–496.

Caselli, Francesco, "Accounting for cross-country income differences," *Handbook of Economic Growth*, 2005, *1*, 679–741.

_ **and James Feyrer**, "The marginal product of capital," *Quarterly Journal of Economics*, 2007, *122*, 535–568.

_ **and John Coleman**, "The world technology frontier," *American Economic Review*, 2006, *96*, 499–522.

David, Joel M., Hugo A. Hopenhayn, and Venky Venkateswaran, "Information, misallocation and aggregate productivity," *NBER Working paper 20340*, 2014.

Hsieh, Chang-Tai and Peter J. Klenow, "Misallocation and manufacturing TFP in China and India," *Quarterly Journal of Economics*, 2009, *124*, 1403–1148.

Levinsohn, James and Amil Petrin, "Estimating production functions using inputs to control for unobservables," *Review of Economic Studies*, 2003, *70*, 317–341.

Mayer, Thierry, Marc J. Melitz, and Gianmarco Ottaviano, "Market Size, Competition, and the Product Mix of Exporters," *American Economic Review*, 2014, *104* (2), 495–536.

Melitz, Marc J., "The impact of trade on intra-industry reallocations and aggregate industry productivity," *Econometrica*, 2003, *71*, 1695–1725.

_ **and Gianmarco Ottaviano**, "Market Size, trade and productivity," *Review of Economic Studies*, 2008, *75*, 295–316.

Midrigan, Virgiliu and Daniel Yi Xu, "Finance and misallocation: evidence from plant-level data," *American Economic Review*, 2014, *104*, 422–458.

Olley, G. Steven and Ariel Pakes, "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 1996, *64*, 1263–1297.

Peters, Michael, "Heterogeneous mark-ups and endogenous misallocation," *Working paper*, 2013.

Restuccia, Diego and Richard Rogerson, "Policy distortions and aggregate productivity with heterogeneous establishments," *Review of Economic Dynamics*, 2008, *11*, 707–720.

A MLE of Distributional Parameters

The log likelihood function is, aside from constants,

$$\ln L = -\frac{N}{2}\ln\sigma_A^2 - \frac{1}{2\sigma_A^2}\sum_{s,i} v_{si}^2 - \frac{1}{2}\sum_{s,i}\ln\sigma_{\tau s}^2 - \sum_{s,i}\frac{1}{2\sigma_{\tau s}^2}(\ln\tau_{si})^2 - \sum_{s,i}\ln\Phi_{sd}, \tag{43}$$

where I use abbreviations $v_{si} = \ln A_{si} - \mu_A$, $v_{sd} = \mu_A - \ln z_{sd}$, $\phi_{sd} = \phi(v_{sd}/\sigma_s)$ and $\Phi_{sd} = \Phi(v_{sd}/\sigma_s)$.

Since the likelihood is increasing in z_{sd}, the maximum likelihood estimator for the cutoff is $\hat{z}_{sd} = \min_i\{z_{si}\}$. From now on, assume z_{sd} is known. The maximum likelihood equations are obtained by equating the partial derivatives of (43) with respect to θ to 0 as follows:

$$\frac{\partial \ln L}{\partial \mu_A} = \frac{1}{\sigma_A^2}\sum_{s,i} v_{si} - \sum_{s,i}\lambda_{sd} = 0, \tag{44}$$

where $\lambda_{sd} = \phi_{sd}/(\sigma_s\Phi_{sd})$ is the inverse Mill's ratio of $N(0, \sigma_s)$ evaluated at v_{sd}. We also have

$$\frac{\partial \ln L}{\partial \sigma_A^2} = \frac{N}{\sigma_A^2} - \frac{1}{\sigma_A^4}\sum_{s,i} v_{si}^2 - \sum_{s,i}\frac{\lambda_{sd}v_{sd}}{\sigma_s^2} = 0, \tag{45}$$

$$\frac{\partial \ln L}{\partial \sigma_Y^2} = \sum_{s,i}\frac{1}{\sigma_{\tau s}^2} - \sum_{s,i}\frac{1}{\sigma_{\tau s}^4}(\ln\tau_{si})^2 - \sum_{s,i}\frac{\lambda_{sd}v_{sd}}{\sigma_s^2} = 0, \tag{46}$$

$$\frac{\partial \ln L}{\partial \sigma_K^2} = \sum_{s,i}\frac{1}{\sigma_{\tau s}^2} - \sum_{s,i}\frac{1}{\sigma_{\tau s}^4}(\ln\tau_{si})^2\alpha_s^2 - \sum_{s,i}\frac{\lambda_{sd}v_{sd}}{\sigma_s^2}\alpha_s^2 = 0, \tag{47}$$

and

$$\frac{\partial \ln L}{\partial \sigma_{KY}} = \sum_{s,i}\frac{1}{\sigma_{\tau s}^2} - \sum_{s,i}\frac{1}{\sigma_{\tau s}^4}(\ln\tau_{si})^2\alpha_s - \sum_{s,i}\frac{\lambda_{sd}v_{sd}}{\sigma_s^2}\alpha_s = 0. \tag{48}$$

The maximum likelihood estimates of θ are defined as the roots of (44)-(48). The solution will be obtained by the Newton-Raphson procedure as follows: given an initial estimate $\hat{\theta}_1$, define $\hat{\theta}_2$ by

$$\hat{\theta}_2 = \hat{\theta}_1 - \left[\frac{\partial^2 \ln L(\hat{\theta}_1)}{\partial\theta\partial\theta'}\right]^{-1}\frac{\partial \ln L(\hat{\theta}_1)}{\partial\theta}. \tag{49}$$

After obtaining $\hat{\theta}_2$, one iterates again to obtain $\hat{\theta}_3$, continuing the process until $\hat{\theta}_n$ converges.

The initial estimate could be obtained by performing unconstrained maximum likelihood estimation of $(\ln A_{si}, \ln(1-\tau_{Y_{si}}), \ln(1+\tau_{K_{si}}))$. Let their joint distribution be denoted by $N(\mu, \Sigma)$, then the unconstrained maximum likelihood estimates of μ and Σ are the sample mean and sample covariance matrix, respectively. This completes the estimation procedure.