Endogenous Returns to Scale[*]

Alexandr Kopytov
University of Rochester

Mathieu Taschereau-Dumouchel
Cornell University

Zebang Xu
Cornell University

December 22, 2025

Abstract

We develop a general equilibrium model in which firms choose how scalable their production technologies are. More scalable technologies make it easier for firms to expand output but are less effective at small scale. In equilibrium, more productive firms adopt more scalable technologies and grow disproportionately large. As a result, the tail of the size distribution becomes thicker and, as resources reallocate to the most productive producers, GDP increases. Over the long-run, as aggregate productivity rises, firms adopt more scalable technologies, which lowers input prices, leading to further increases in scalability. Through this supply-chain amplification process, endogenous returns to scale raise the growth rate of GDP. A calibrated version of the model shows that these effects are quantitatively significant. We also document support for the model's predictions in firm-level data.

JEL Classifications: E23, E01, D24, O40, L11

[*] We thank Guangbin Hong, Ron Kaniel, Narayana Kocherlakota, Ezra Oberfield, Christian Opp, Tommaso Porzio, Markus Poschke and participants at various seminars and conferences for helpful suggestions.

1 Introduction

At the turn of the twentieth century, the automobile was a luxury good, carefully assembled by skilled craftsmen. State-of-the-art models like the 1909 Cadillac Model Thirty were already built with interchangeable parts, yet their production remained artisanal and their price prohibitively high. Just a few years later, Henry Ford's Highland Park plant was producing a Model T every 93 minutes. By adopting the moving assembly line, Ford had made the decision to reorganize production to increase its returns to scale. The benefit was a large decline in production costs that transformed the automobile into a mass-market good.

Ford's story is not unique. Indeed, in his seminal work, Chandler (1990) argues that the adoption of processes, organizational structures and technologies designed to achieve higher economies of scale was a key driver of modern economic growth. This historical perspective suggests that scalability is not a fixed constraint but a dimension of technology that firms actively manage. Building on this insight, we explore the idea that returns to scale are endogenous equilibrium objects driven by incentives, and study their implications for the firm size distribution, the response of the economy to shocks, and long-run growth.

To do so, we develop a multi-sector general equilibrium model with endogenous returns to scale. Within each sector, a continuum of firms with heterogenous productivity produce a common good using labor and intermediate inputs. Importantly, firms in our setup are free to choose their returns to scale subject to a technological trade-off: achieving larger scale economies comes at a cost in terms of productivity. As a result, while more scalable technologies make it easier for firms to expand output, they are less effective at small scale.

The existence of a trade-off between scale and productivity is well documented. Chandler (1977) describes how the historical shift from small, artisanal producers to large, integrated enterprises depended on costly investment in a new "technology of organization." Achieving and managing large scale required the creation of professional hierarchies and complex administrative systems. While this "visible hand" of management enabled firms to coordinate large-scale production, it introduced administrative overhead and reduced the operational flexibility that smaller firms enjoyed.

Our analysis of the model begins with the individual firm's decisions in partial equilibrium. Following McKenzie (1959), we interpret decreasing returns to scale as arising from a fixed entrepreneurial factor. A firm's choice of returns to scale therefore reflects a trade-off between this constrained in-house factor and the variable bundle of labor and intermediate inputs. We show that the optimal degree of scalability is reached when the marginal productivity loss from expanding scalability exactly offsets the cost savings from relying less on the fixed factor. This condition dictates how firms adapt to their environment: any change that pushes the firm to expand puts pressure on the fixed factor and encourages the adoption of a technology with higher returns to scale. Consequently, higher productivity, higher output prices, or cheaper intermediate inputs all induce the firm to adopt a more scalable, input-intensive production function.

This mechanism generates a "double blessing" for the most productive firms. Their intrinsic productivity (the first blessing) naturally leads to larger size and higher profits. This expansion, in turn, tightens the constraint imposed by the fixed entrepreneurial factor, creating an incentive to adopt more scalable technologies (the second blessing), which lead to a further increase in size. This disproportionate growth creates superstar firms and a thick Pareto tail for the firm-size distribution.

Despite heterogeneity in returns to scale, firms in a sector can be aggregated in a tractable way. Because of free entry, production exhibits constant returns to scale at the sector level, with firm-level scalability decisions affecting the importance of labor and intermediate inputs in sectoral production. Returns to scale decisions also manifest themselves in sectoral productivity. We show that by allowing high-productivity firms to grow larger, endogenous returns to scale increase sectoral productivity and, through that channel, the level of GDP.

The model admits a unique and efficient competitive equilibrium, which can be characterized as the solution to a social planner's problem. We use this characterization to study the determinants of returns to scale in general equilibrium. We find that any shock that lowers the relative cost of intermediate inputs induces firms to adopt more scalable technologies. For instance, a productivity improvement in an upstream sector reduces input costs for downstream customers, encouraging them to increase their scalability. This shift toward greater scalability implies a heavier reliance on intermediate inputs, which raises the Domar weights of upstream suppliers. Consequently, sectors experiencing productivity gains become more central to the economy, amplifying their impact on GDP. Through this channel, endogenous returns to scale magnify the benefits of positive shocks. Symmetrically, the same mechanism dampens the adverse impact of negative shocks, as firms substitute away from intermediate inputs and reduce the importance of the affected sectors.

Endogenous returns to scale also matter for long-run growth. With recurrent productivity improvements, firms continuously adopt more scalable technologies. This leads to an increase in Domar weights, making subsequent productivity gains even more impactful. As this process unfolds, the economy enters an acceleration phase where the GDP growth rate rises over time, eventually converging to a long-run rate strictly higher than in a fixed-technology economy. In our model, growth is therefore driven by the interaction between exogenous innovation and endogenous scaling decisions. A back-of-the-envelope calculation suggests that this mechanism can have a significant impact on long-run growth.

To study how policy interventions and market frictions affect returns to scale decisions, we extend our baseline model to include wedges, such as sales taxes or tariffs on intermediate inputs. We find that such distortions, by incentivizing firms to shrink, lead to the adoption of inefficiently low returns to scale. In this distorted equilibrium, productivity shocks have a first-order effect on welfare by altering the economy's structure. A positive productivity shock, for instance, not only increases output directly but also acts as a corrective force: by lowering input costs, it encourages firms to increase their scalability, moving the economy's production technology closer to the efficient benchmark.

We use detailed data covering the near-universe of firms in Spain to test the core predictions of the model. Consistent with our theory, we document a strong positive correlation between firm productivity, size, and returns to scale. We also find evidence supporting the model's input-cost mechanism. By exploiting variation in import tariffs, we show that firms more exposed to costlier intermediate inputs tend to reduce their returns to scale, in line with the model's prediction. Finally, cross-country patterns corroborate these mechanisms at the aggregate level. Countries where returns to scale are more responsive to productivity—indicating a stronger endogenous scalability mechanism—exhibit higher income per capita. This suggests that endogenous scalability decisions may play a role in long-run economic development.

Finally, to quantify the importance of our mechanism, we calibrate the model to the Spanish economy. We find that endogenous returns to scale are a first-order determinant of economic performance. Eliminating the ability of firms to adjust their scalability reduces the level of GDP by nearly 12% and lowers the long-run growth rate of the economy by 0.8 percentage points. Crucially, these gains are driven by the capacity of high-productivity firms to adopt more scalable technologies. To illustrate this, we examine the impact of size-dependent distortions that are particularly detrimental to large firms. Constructing wedges as in Hsieh and Klenow (2009), we confirm that larger firms face higher effective distortions in the data. We find that removing these wedges yields welfare gains that are more than twice as large in our model compared to a standard framework. This highlights that policies burdening large firms are particularly costly when they stifle the adoption of high-scale technologies.

Literature review

Some early work emphasizes the importance of changing returns to scale for economic outcomes. Kuznets (1973) argues that the rise of large-scale enterprises reflected an adaptive process through which firms learned to coordinate production and distribution over expanding markets. Chandler (1977) documents how managerial hierarchies and integrated production systems enabled firms to realize "economies of scale and scope." These classic accounts emphasized the importance of changes in returns to scale for growth. Our work formalizes that idea in a general equilibrium framework.[1]

Since we focus on the aggregate impact of endogenous scalability we adopt a holistic approach and do not take a stance on the underlying margins that firms use to adjust their returns to scale (several are likely at work). In contrast, some recent studies have focused on specific mechanisms. Argente et al. (2025) propose a model of multi-product firms in which standardization can increase

[1] Our paper also relates to classic models of firm heterogeneity and dynamics such as Lucas (1978) and Hopenhayn (1992). Typical work in this literature assumes that returns to scale are exogenous. We also relate to a literature that studies the internal organization of firms as in Garicano (2000) and Garicano and Rossi-Hansberg (2006).

a firm's returns to scale. Like us, they find that endogenous scalability can lead to fat-tailed firm-size distributions. Engbom et al. (2025) build a model in which entrepreneurs can professionalize administrative tasks by hiring white-collar workers, thereby relaxing constraints on their returns to scale. They find that the scarcity of skilled labor in developing countries limits this reorganization, and that increasing the aggregate supply of skills can explain two-thirds of the shift into large firms observed during development. Also in the development literature, Gottlieb et al. (2025) propose a model in which firms can choose between a high and a low returns to scale technologies. They use the model to explain empirical patterns related to the effect of skill endowments on the firm size distribution.

Smirnyagin (2023) proposes a business cycle model with financial frictions in which firms can choose between two returns to scale levels. Focusing on long-run patterns, Lashkari et al. (2024) document that the decline in IT prices led to an increase in returns to scale in France. To explore the implications of this finding, they propose a model with non-homothetic production in which returns to scale can vary with input factors. Hubmer et al. (2025) use administrative data from Canada and the United States to document that larger firms operate technologies with higher returns to scale—a finding that is consistent with our model. They explore the implications of these patterns in an entrepreneurial model with fixed heterogenous returns to scale and financial frictions.[2]

A distinguishing feature of our work is that we study the impact of input-output linkages on scalability decisions. In doing so, we identify a novel channel through which adjustments in returns to scale propagate through supply chains, reshaping Domar weights throughout the economy. This channel has important implications for the macroeconomic impact of endogenous scalability and is essential for our long-run growth results.

Our work also relates to different strands of the production network literature (Long and Plosser, 1983; Acemoglu et al., 2012). As in Baqaee and Farhi (2019a), Hulten's (1978) theorem only provides a first-order approximation to the economy's response to shocks in our model. We also build on previous work that studies the role of wedges in network economies (Jones, 2011; Baqaee and Farhi, 2019b; Liu, 2019; Bigio and La'O, 2020). Finally, we relate to a literature on production networks that treats the production function as endogenous (Oberfield, 2018; Acemoglu and Azar, 2020; Kopytov et al., 2024a,b). We share with this literature the assumption that firms have control over their production technologies. However, in this literature, returns to scale are always constant. Endogenizing returns to scale yields novel predictions for the firm size distribution and has important aggregate implications.

[2] In many models, firms must pay a fixed cost to operate, and this fixed cost therefore influences the *average* returns to scale of the firm. In contrast, our setup is interested in *marginal* returns to scale, which capture how a marginal increase in size affects the marginal cost of production.

2 A model of endogenous returns to scale

We introduce endogenous returns to scale into an otherwise standard multisector economy. Each sector produces a differentiated good that can be used for final consumption and as an intermediate input. Within each sector, there is a continuum of firms that differ in terms of their productivity. A representative household supplies labor, owns all firms, and consumes the final good. Importantly, firms choose their returns to scale in order to maximize profits. Changes in the environment can therefore affect individual returns to scale and, through that channel, macroeconomic aggregates.

2.1 Production technology

There are N goods, each produced by a different sector. Each sector i consists of a continuum of competitive firms whose mass M_i is determined by a free-entry condition. Upon paying $\kappa_i > 0$ units of labor to enter, a firm l draws a random productivity level $\varepsilon_{il} \sim \text{iid } \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ from a normal distribution. The firm can then produce using a Cobb–Douglas technology but, crucially, it can choose how scalable that technology is. Specifically, if it selects returns to scale $0 < \eta_{il} < 1$, firm l's output is given by

$$F_i\left(L_{il}, X_{il}, \eta_{il}\right) := e^{\varepsilon_{il}} A_i\left(\eta_{il}\right) \zeta\left(\eta_{il}\right) \left(L_{il}^{1-\sum_{j=1}^{N} \alpha_{ij}} \prod_{j=1}^{N} X_{ij,l}^{\alpha_{ij}} \right)^{\eta_{il}}, \tag{1}$$

where L_{il} is labor, $X_{il} = \left(X_{i1,l}, \ldots, X_{iN,l}\right)$ is a vector of intermediate inputs, and $\zeta\left(\eta_{il}\right)$ is a normalization term to simplify subsequent expressions.[3] We assume that operating technologies with higher returns to scale is costly, so that the productivity shifter $A_i\left(\eta_{il}\right)$ is strictly decreasing. This captures the idea that achieving greater scalability often requires more complex processes, incurs significant coordination and communication costs, and demands more managerial attention (Chandler, 1977). For tractability, we impose that A_i is smooth, strictly log-concave and that $A_i\left(\eta_{il}\right) \to 0$ as $\eta_{il} \to 1$.

2.2 Firm problem

To explore what drives a firm's returns to scale decision, we first analyze its problem in partial equilibrium. A firm l in sector i simultaneously chooses its returns to scale η_{il} and variable inputs (labor and intermediates) to maximize profits:

$$\Pi_{il} := \max_{\eta_{il}, L_{il}, X_{il}} P_i F_i\left(L_{il}, X_{il}, \eta_{il}\right) - W L_{il} - \sum_{j=1}^{N} P_j X_{ij,l}. \tag{2}$$

[3] We set $\left(\zeta\left(\eta\right)\right)^{-1} := \left(\left(1 - \sum_j \alpha_{ij}\right)\eta\right)^{\eta\left(1-\sum_j \alpha_{ij}\right)} \prod_j \left(\eta \alpha_{ij}\right)^{\eta \alpha_{ij}} \left(1-\eta\right)^{1-\eta}$ to simplify the unit cost expression below. Here, this term can be subsumed in $A_i\left(\eta_{il}\right)$.

where W is the wage and P_j is the price of good j. Solving this problem, we derive the firm's marginal cost of production as a function of its output Q_{il} and its chosen technology η_{il}.

Lemma 1. *The firm's marginal cost of production λ_{il} is given by*

$$\lambda_{il} := \frac{1}{e^{\varepsilon_{il}} A_i\left(\eta_{il}\right)} H_i^{\eta_{il}} \Pi_{il}^{1-\eta_{il}}, \tag{3}$$

where $H_i := W^{1-\sum_{j=1}^{N} \alpha_{ij}} \prod_{j=1}^{N} P_j^{\alpha_{ij}}$ is the price of the variable input bundle used by firms in sector i, and

$$\Pi_{il} = \left(1 - \eta_{il}\right) \lambda_i Q_{il} \tag{4}$$

is profits.[4]

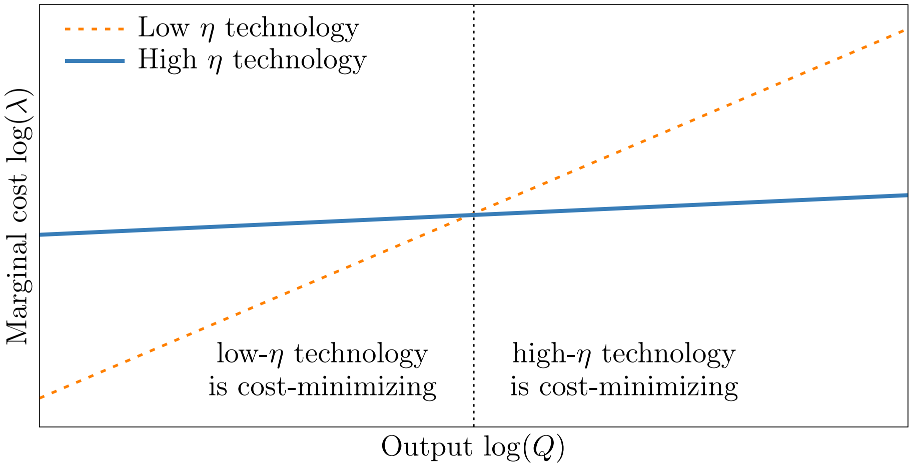
As usual, the firm's marginal cost λ_{il} decreases with productivity and increases with input prices. The wage and the price of intermediate inputs, for instance, affect λ_{il} through the variable input bundle price H_i. Crucially, the firm's profits Π_{il} also show up as an input price in (3). To understand why, recall that we can interpret a decreasing-returns production function as a constant-returns technology with an additional fixed entrepreneurial factor in unit supply (McKenzie, 1959). Under this interpretation, profits Π_{il} are simply the payment to that input. As output Q_{il} increases, the pressure on this fixed factor rises, increasing its shadow cost Π_{il} and, in turn, driving up the marginal cost λ_{il}.

Returns to scale η_{il} play a dual role in shaping marginal costs. Higher returns to scale lower baseline productivity through $A_i\left(\eta_{il}\right)$, but they also increase the weight of the variable bundle (labor and intermediate) relative to the fixed factor. Through this second channel, η_{il} determines how steeply the marginal cost rises with output. Figure 1 illustrates this trade-off by plotting λ_{il} as a function of Q_{il} for high and low values of η_{il}. The high-η technology offers greater scalability and thus a flatter marginal cost curve, allowing the firm to increase its size with only a small increase in its marginal cost. This makes it particularly effective for large firms. However, because it incurs a large productivity penalty A_i, this technology is inefficient at small scales. In contrast, the low-η technology benefits from high baseline productivity $A\left(\eta_{il}\right)$, making it the preferred choice for small-scale production.

This leads to the key sorting mechanism of our model. While adopting a technology with higher returns to scale is costly in terms of baseline productivity, firms that choose these technologies are, in equilibrium, more productive overall. This is because only firms with a sufficiently high idiosyncratic productivity draw ε_{il} find it optimal to operate at the large scale necessary to make a high-η technology worthwhile. In Section 6, we will show that this positive correlation between productivity and returns to scale is supported by the data. Finally, profit maximization implies that the firm selects output Q_{il} so that its marginal cost λ_{il} equals the price of its good P_i.

[4] All proofs are in Appendix C.

Figure 1: The trade-off between baseline productivity and returns to scale.



2.3 Choosing returns to scale

Building on these insights, we can characterize the optimal returns to scale decision of the firm.

Lemma 2. *At an interior solution, the firm chooses its returns to scale $\eta_{il} \in (0,1)$ according to*

$$\frac{da_i\left(\eta_{il}\right)}{d\eta_{il}} = \log H_i - \log \Pi_{il}, \tag{5}$$

where $a_i\left(\eta_{il}\right) := \log A_i\left(\eta_{il}\right)$.

This expression describes the core trade-off behind the returns to scale choice. It is better understood as the derivative of the log of λ_{il}, given by (3), with respect to η_{il}. When increasing η_{il} at the margin, the firm shifts its input mix away from the fixed entrepreneurial factor, whose cost is Π_{il}, toward the variable input bundle, whose cost is H_i. The right-hand side of (5) captures the marginal change in cost associated with that shift. The firm balances that change in cost with any loss in TFP associated with the higher returns to scale, as reflected by the left-hand side of (5).

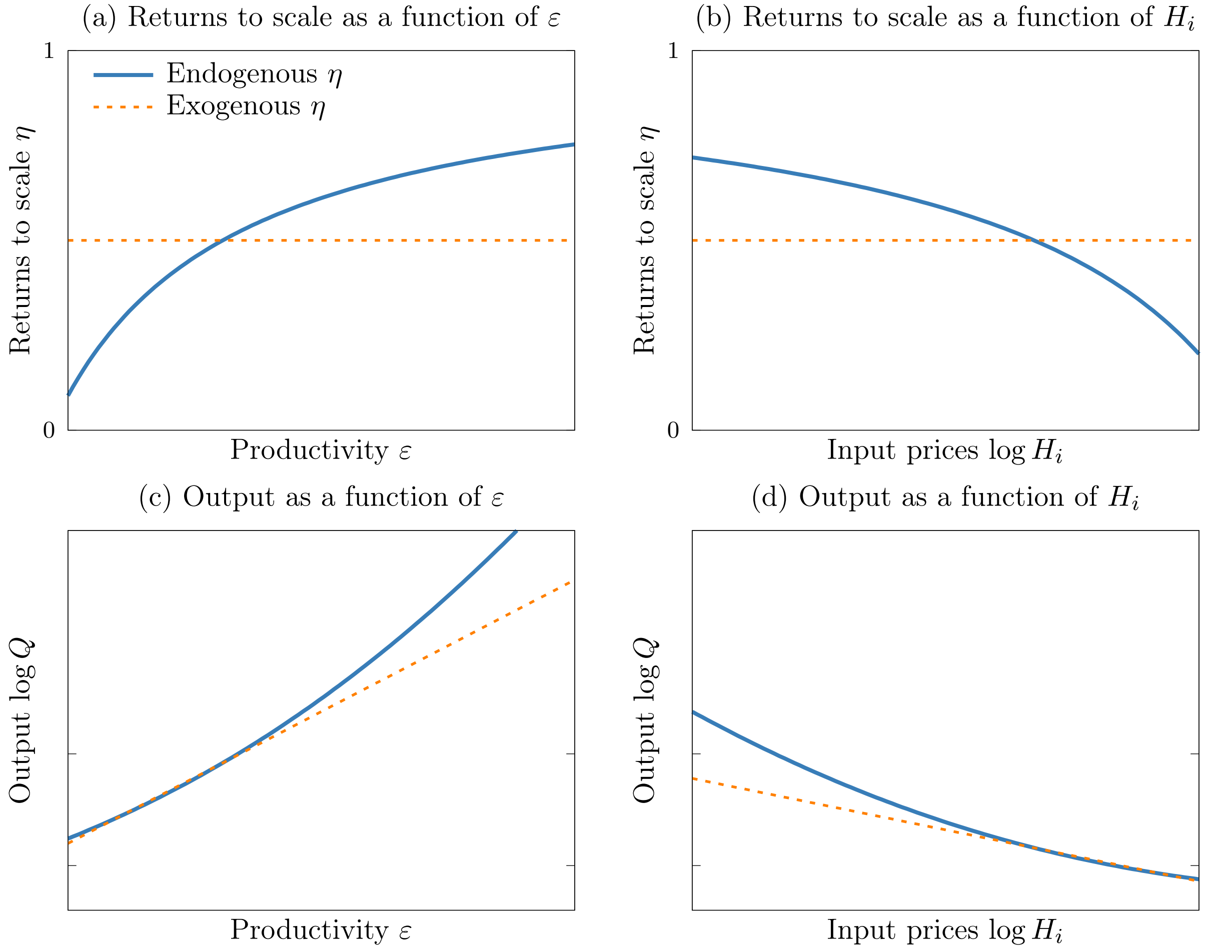
From (5), we can determine how η_{il} responds to changes in the economic environment. The top two panels of Figure 2 illustrate the forces involved. Since a_i is concave, its derivative is decreasing. Therefore, any change that increases profits Π_{il}—such as a higher output price P_i or a better productivity draw ε_{il}—makes the fixed factor more expensive, pushing the firm to adopt a higher η_{il}. Conversely, an increase in the variable input cost H_i, as it incentivizes the firm to rely more on its fixed factor, lowers the optimal η_{il}. The following lemma formalizes this intuition.

Lemma 3. *At an interior solution, the returns to scale parameter η_{il} satisfies*[5]

$$\frac{d\eta_{il}}{d\varepsilon_{il}} = \frac{d\eta_{il}}{d\log P_i} = -\left[(1-\eta_{il})\,\frac{d^2 a_i}{d\eta_{il}^2}\right]^{-1} > 0, \qquad \textit{and} \qquad \frac{d\eta_{il}}{d\log H_i} = \left[(1-\eta_{il})\,\frac{d^2 a_i}{d\eta_{il}^2}\right]^{-1} < 0.$$

This result highlights that the elasticities of the returns to scale with respect to prices and productivity depend on η_{il} itself and on the concavity of a_i.

Figure 2: Impact of productivity ε_{il} and input prices H_i on the firm


(a) Returns to scale as a function of ε

(b) Returns to scale as a function of H_i

(c) Output as a function of ε

(d) Output as a function of H_i

The endogenous choice of scalability also has crucial implications for the output Q_{il} of the firm.

Lemma 4. *At an interior solution, the elasticity of output Q_{il} with respect to productivity ε_{il} is given by*

$$\frac{d\log Q_{il}}{d\varepsilon_{il}} = \underbrace{\frac{1}{1-\eta_{il}}}_{\textit{Fixed } \eta \textit{ effect}} + \frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\varepsilon_{il}} > 0.$$

[5] When increasing P_i, we keep the price of the variable input bundle constant to distinguish the two channels that affect η_{il}.

In addition, the elasticities of output Q_i with respect to prices are given by

$$\frac{d\log Q_{il}}{d\log P_i} = \underbrace{\frac{\eta_{il}}{1-\eta_{il}}}_{\textit{Fixed } \eta \textit{ effect}} + \frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\log P_i} > 0, \ \textit{and} \ \frac{d\log Q_{il}}{d\log H_i} = \underbrace{-\frac{\eta_{il}}{1-\eta_{il}}}_{\textit{Fixed } \eta \textit{ effect}} + \frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\log H_i} < 0.$$

Furthermore, the impact of a change in ε_{il}, $\log P_i$ or $\log H_i$ on $\log Q_i$ is amplified because of the endogenous response of η_{il}.

With fixed returns to scale, productivity and prices affect output Q_{il} through standard channels, captured by the first terms in the expressions of Lemma 4. Higher productivity ε_{il}, for instance, allows the firm to produce larger quantities before its marginal cost reaches the price P_i. The magnitude of this response depends on returns to scale: a high-η firm is more sensitive to productivity and prices than a low-η firm.

In addition to this fixed-η mechanism, Lemma 4 reveals an additional mechanism at work when returns to scale are endogenous. Following an increase in productivity ε_{il}, the firm not only expands to exploit its lower marginal cost but also increases its returns to scale to better accommodate the higher production volume. This amplification mechanism creates a superstar effect, causing high-productivity firms to grow disproportionately large. A similar mechanism operates in response to price changes.

The bottom two panels of Figure 2 illustrate these forces. With exogenous returns to scale (dashed orange lines), $\log Q_{il}$ varies linearly with ε_{il} and $\log H_i$, as in standard models. In contrast, with endogenous returns to scale (blue lines), the response is convex: productivity and input prices have an outsized impact on output.

This amplification mechanism has important implications for the firm distribution. To explore them transparently, it helps to specialize the returns to scale cost function a_i.[6]

Assumption 1. *The TFP shifter function A_i takes the form*

$$a_i(\eta_{il}) = -\frac{\gamma_i}{1-\eta_{il}}, \tag{6}$$

where the parameter $\gamma_i > \sigma_i^2/2$ governs the productivity cost of increasing η_{il} in sector i.

We also define the *effective productivity dispersion* $\varphi_i := \sigma_i^2/(2\gamma_i)$ as a measure of the dispersion in sector i relative to the cost of adjusting η_{il}. This parameter plays an important role in our analysis.

[6] Assumption 1 imposes that $A_i(\eta_{il})$ satisfies an Inada condition as $\eta_{il} \to 1$, but not as $\eta_{il} \to 0$. Since productivity shocks ε_{il} are unbounded, some firms with very low ε_{il} might choose $\eta_{il} \notin (0,1)$. In any reasonable calibrations of the model, these firms are very small and their mass is negligible. It is straightforward to truncate the distribution of ε_{il} to guarantee that $0 < \eta_{il} < 1$ for all firms, but this makes the analysis burdensome without any new interesting insights. In the main text, we therefore do not impose such a truncation but we do explore a version of the model with truncated productivity in Appendix D.1. We show that aggregate quantities in this alternative model converge to their main-text counterparts as the mass of firms picking $\eta_{il} \notin (0,1)$ shrinks. In the quantitative section of the paper, we verify that the mass of such firms is indeed small.

The constraint $\gamma_i > \sigma_i^2/2$ in Assumption 1 implies that $0 < \varphi_i < 1$.[7]

With that assumption, we can describe the impact of endogenous returns to scale on the tail of the firm-size distribution.

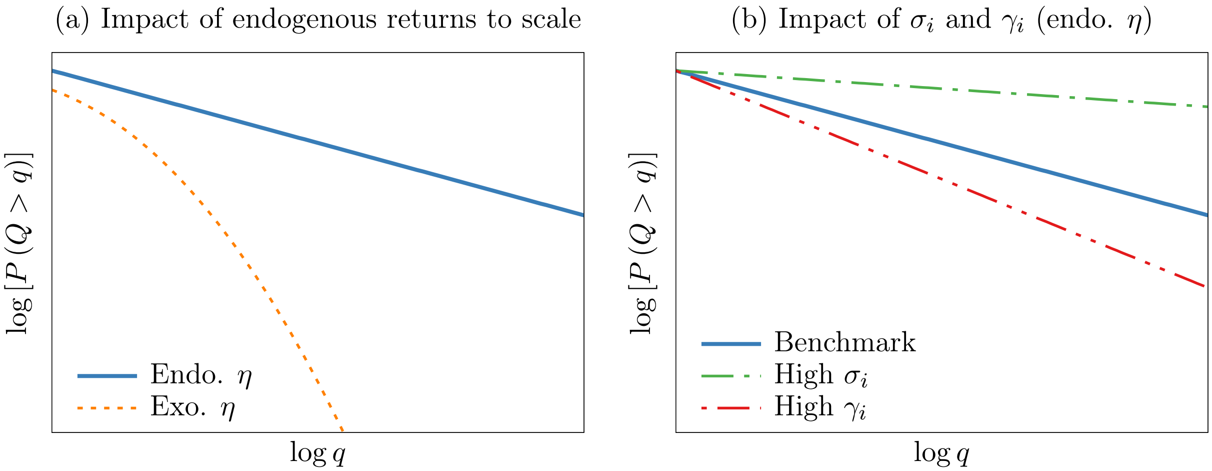
Proposition 1. *Suppose that Assumption 1 holds. Without endogenous returns to scale, the distribution of Q_{il} in sector i is log-normal. With endogenous returns to scale, the right tail of the distribution of Q_{il} behaves like a Pareto distribution with tail index $1/\varphi_i$, in the sense that*

$$\log\left(\mathbb{P}\left(Q_{il} > q\right)\right) \sim -\frac{1}{\varphi_i}\log q, \ \ \textit{as } q \to \infty.$$

In the absence of endogenous scalability, all firms within a sector operate with identical returns to scale. Consequently, the distribution of firm output simply mirrors that of the underlying productivity distribution and is log-normal. In contrast, when returns to scale are endogenous, the most productive firms choose higher returns to scale. This stretches the right tail of the distribution, making it thick and Pareto-like. Proposition 1 also shows that the thickness of the tail depends on the effective productivity dispersion φ_i. When productivity shocks are highly dispersed (large σ_i^2) or scalability is cheap (low γ_i), the firm-size distribution is thicker. Figure 3 illustrates these mechanisms.

Proposition 1 is reassuring, as the thick tail of the firm size distribution is well-documented empirically (Axtell, 2001). Our model generates this property endogenously from fundamental productivity shocks that are not themselves fat-tailed, with superstar firms emerging from the decisions of high-productivity producers to become more scalable.

Figure 3: Tail of the distribution of firm-level output Q_{il}



[7] Without the constraint $\gamma_i > \sigma_i^2/2$, returns to scale increase so rapidly with ε_{il} that sectoral output becomes infinite.

2.4 Household

A representative household owns the firms, supplies $\bar{L} > 0$ units of labor inelastically, and consumes a bundle $Y := \prod_{i=1}^{N} \left(\beta_i^{-1} C_i\right)^{\beta_i}$ of the different consumption goods, where $\sum_{i=1}^{N} \beta_i = 1$. Since Y measures aggregate value added in this economy, we refer to it as (real) GDP.

The household maximizes Y subject to the budget constraint[8]

$$\sum_{i=1}^{N} P_i C_i \leq W\bar{L}, \tag{7}$$

Because of the free-entry condition, all profits from the firms are dissipated through the entry cost, and the household's only income comes from labor.

Maximization by the household implies that spending on good i amounts to a fraction β_i of total expenditure, so that $P_i C_i = \beta_i \bar{P} Y$, where $\bar{P} := \prod_{i=1}^{N} P_i^{\beta_i}$ is the ideal price index which we adopt as numeraire, so that $\bar{P} = 1$. Consequently, nominal and real GDP are equal, and the budget constraint simplifies to $Y = W\bar{L}$.

2.5 Equilibrium conditions

For any firm-level quantity B_{il}, we denote by $B_i = \int_0^{M_i} B_{il} dl$ the sum of that quantity across all firms in sector i. We also use brackets $\{B_{il}\}$ to denote the set of that quantity over all sectors and firms.

We define an equilibrium as an allocation in which the optimality conditions of the firms and the household hold simultaneously, and all markets clear.

Definition 1. An *equilibrium* is a set of prices (P^*, W^*), a choice of returns to scale $\{\eta_{il}^*\}$, a tuple of quantities $\{C_i^*, L_{il}^*, X_{il}^*, Q_{il}^*\}$, and a mass of firms M^* in each sector such that

1. (Optimal returns to scale choice) For each $i \in \{1, \ldots, N\}$ and $l \in [0, M_i]$, the returns to scale decision η_{il}^* solves (2) given prices (P^*, W^*).

2. (Optimal input choice) For each $i \in \{1, \ldots, N\}$ and $l \in [0, M_i]$, factor demands L_{il}^* and X_{il}^* solve (2) given prices (P^*, W^*).

3. (Consumer optimization) The consumption vector C^* maximizes GDP Y subject to (7) given prices (P^*, W^*).

4. (Free entry) For each $i \in \{1, \ldots, N\}$, the expected profit of a potential entrant in sector i solves

$$\mathrm{E}_i \left[\Pi_i \left(\varepsilon_{il}, P^*, W^*\right)\right] = \kappa_i W^*, \tag{8}$$

[8] Because the model is static and there is no aggregate uncertainty, the household could instead maximize a strictly increasing function of Y without affecting the results.

where Π_{il} is given by (4), and where the expectation E_i is taken over ε_{il}.

5. (Market clearing) For each $i \in \{1, \ldots, N\}$,

$$C_i^* + \sum_{j=1}^{N} X_{ji}^* = Q_i^* = \int_0^{M_i} F_i\left(L_{il}^*, X_{il}^*, \eta_{il}^*\right) dl, \text{ and } \sum_{i=1}^{N} L_i^* + \sum_{i=1}^{N} M_i \kappa_i = \bar{L}. \tag{9}$$

Conditions 2 to 5 are standard and imply that the household and the firms maximize their objective functions, that all markets clear, and that the free-entry condition holds. Condition 1 states that firms pick their returns to scale to maximize profits.

3 Aggregation

In this section, we aggregate the economy and derive equations for equilibrium prices and GDP. While most of our partial equilibrium results hold under general A_i's, we need to impose additional restrictions to aggregate the economy in a tractable way. We therefore assume that Assumption 1 holds from now on. Under that assumption, we can derive a tractable mapping between a firm's productivity ε_{il} and its returns to scale η_{il} in equilibrium. Equation (5) implies a simple mapping between prices, productivity, and returns to scale:

$$\frac{1}{1-\eta_{il}} = \frac{1}{2\gamma_i}\left(\varepsilon_{il} + \log P_i - \log H_i\right). \tag{10}$$

In the remainder of this section, we take advantage of (10) by first aggregating the firms in each sector. We then build on that characterization to derive equations for equilibrium prices and GDP.

3.1 Sectoral aggregation

To aggregate the economy, we define the Domar weight of a production unit (a firm or a sector) as the share of its sales in nominal GDP. For a firm l in sector i and for the sector as a whole, those are given by

$$\omega_{il} := \frac{P_i Q_{il}}{\bar{P}Y} \text{ and } \omega_i := \frac{P_i Q_i}{\bar{P}Y}.$$

We also introduce the *effective returns to scale* $\hat{\eta}_i$ of a sector, defined as the sales-weighted average of firm-level returns to scale:

$$\hat{\eta}_i := \int_0^{M_i} \frac{P_i Q_{il}}{P_i Q_i} \eta_{il} dl. \tag{11}$$

This quantity will play an important role in our analysis. One can show, for instance, that the sectoral cost shares depend on $\hat{\eta}_i$:

$$\frac{WL_i}{P_iQ_i} = \hat{\eta}_i \left(1 - \sum_{j=1}^{N} \alpha_{ij}\right), \quad \frac{P_jX_{ij}}{P_iQ_i} = \hat{\eta}_i \alpha_{ij}, \text{ and } \frac{\Pi_i}{P_iQ_i} = 1 - \hat{\eta}_i.$$

In addition, we can characterize the returns to scale of any firm in sector i using $\hat{\eta}_i$.

Lemma 5. *The returns to scale η_{il} of firm l in sector i is given by*

$$\frac{1}{1-\eta_{il}} = \frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\varepsilon_{il} - \mu_i}{2\gamma_i}. \tag{12}$$

Furthermore, the moments of the firm-level returns to scale distribution in sector i are given by

$$\mathrm{E}_i\left[\frac{1}{1-\eta_{il}}\right] = \frac{1-\varphi_i}{1-\hat{\eta}_i}, \quad \mathrm{V}_i\left[\frac{1}{1-\eta_{il}}\right] = \frac{\varphi_i}{2\gamma_i}, \quad \textit{and} \quad \mathrm{Cov}_i\left[\frac{1}{1-\eta_{il}}, \varepsilon_{il}\right] = \varphi_i > 0. \tag{13}$$

Equation (12) links a firm's own returns to scale η_{il} to its productivity ε_{il} and the effective sectoral returns to scale $\hat{\eta}_i$. For the firm with the median productivity ($\varepsilon_{il} = \mu_i$), this equation simplifies to

$$\hat{\eta}_i = \eta_i(\mu_i) + \varphi_i(1 - \eta_i(\mu_i)). \tag{14}$$

Since $\varphi_i > 0$, it follows that the effective returns to scale $\hat{\eta}_i$ of sector i is larger than that of its median firm. This is because high-productivity firms, which have higher returns to scale and are larger (Lemmas 3 and 4), are weighted more heavily in the calculation of $\hat{\eta}_i$.

Equation (14) also shows that the gap between $\hat{\eta}_i$ and $\eta_i(\mu_i)$ increases with $\varphi_i = \sigma_i^2/(2\gamma_i)$. Intuitively, greater productivity dispersion σ_i^2 implies that there are relatively more high-productivity firms. Lower adjustment costs γ_i also allow these high-productivity firms to adopt more scalable technologies and grow more aggressively, leading to a higher sectoral returns to scale $\hat{\eta}_i$.

The second part of Lemma 5 describes the cross-sectional moments of returns to scale within a sector. The first moment shows again that φ_i controls the gap between the expected and effective returns to scale. The second moment shows that a higher productivity dispersion σ_i^2 and a lower adjustment cost γ_i both contribute to greater cross-sectional dispersion in η_{il}. The third moment confirms that high ε_{il} firms choose higher η_{il}. As the endogenous returns to scale mechanism shuts down ($\varphi_i \to 0$), the covariance between productivity and returns to scale goes to zero. Later on, we will rely on that covariance to measure the strength of the mechanism in the data.

Aggregating firms within a sector using the free-entry condition (8) yields the following results.[9]

[9] Since all firms in a sector face the same output price, they have the same marginal cost through profit maximization. We therefore define the marginal cost λ_i of a sector i as the marginal cost of any firm in that sector, such that $\lambda_i := \lambda_{il}$ for any (or all) l. Equivalently, since, as we show later, the economy is efficient, one can write the cost minimization problem of the sector and find the same expression.

Proposition 2. *The marginal cost of sector i is given by*

$$\lambda_i = \frac{1}{Z_i(\hat{\eta}_i)} W^{1-\hat{\eta}_i \sum_{j=1}^{N} \alpha_{ij}} \prod_{j=1}^{N} P_j^{\hat{\eta}_i \alpha_{ij}}, \tag{15}$$

where sectoral total factor productivity $Z_i(\hat{\eta}_i)$ is defined as

$$\log Z_i(\hat{\eta}_i) := \underbrace{\mu_i + a_i(\hat{\eta}_i) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i}}_{\textit{Exogenous returns to scale}} + \underbrace{\frac{1}{2}(1-\hat{\eta}_i)\log\left(\frac{1}{1-\varphi_i}\right)}_{\textit{Superstar effect}} - \underbrace{(1-\hat{\eta}_i)\log\kappa_i}_{\textit{Entry cost}}. \tag{16}$$

Furthermore, the effective returns to scale $\hat{\eta}_i$ is given by

$$\frac{1}{1-\hat{\eta}_i} = \frac{1}{2\gamma_i(1-\varphi_i)}(\mu_i + \log P_i - \log H_i). \tag{17}$$

The sectoral marginal cost λ_i takes the standard form associated with a Cobb–Douglas production function with two factors: labor, which is used for production and entry costs, and intermediate inputs. The cost shares of these inputs are driven by the effective returns to scale $\hat{\eta}_i$ of the sector. As these shares sum to one, the sector as a whole exhibits *constant* returns to scale. Intuitively, free entry acts as an adjustment margin: while individual producers may operate under decreasing returns, the entry of new firms allows the sector to expand to achieve constant returns. A higher $\hat{\eta}_i$ implies that firms are larger and more scalable, so fewer of them have to enter to achieve a given amount of production. Consequently, less labor is consumed by entry costs, lowering the overall labor share, $1-\hat{\eta}_i\sum_j \alpha_{ij}$, of the sector.

Equation (16) characterizes the sector's total factor productivity $Z_i(\hat{\eta}_i)$. As expected, sectoral productivity depends on the mean firm-level productivity μ_i and the productivity cost $a_i(\hat{\eta}_i)$ associated with the effective returns to scale. However, firm heterogeneity also plays a role. The third term in (16) captures a standard *selection effect*: as more productive firms grow larger, they receive a larger share of input factor, raising sectoral productivity. These first three terms would also appear in an exogenous returns-to-scale model in which all firms share a common $\eta_{il} = \hat{\eta}_i$. The fourth term in (16), however, captures a novel amplification channel from endogenous returns to scale. In our model, high-productivity firms not only produce more but also choose more scalable technologies, which allows them to grow even larger. This *superstar effect* amplifies their contribution to sectoral productivity beyond the standard selection effect. Finally, the last term in (16) captures the role of entry costs. A higher entry cost κ_i diverts labor away from production, lowering sectoral productivity, with the magnitude of this loss determined by the importance of the fixed factor, $1-\hat{\eta}_i$.

Proposition 2 also provides an expression for the sector's effective returns to scale. This expression is analogous to the one determining firm-level returns to scale, given by (10), but it includes the adjustment term φ_i to account for firm heterogeneity. This adjustment reflects that larger, more productive firms have a disproportionate impact on the aggregate measure of returns to scale $\hat{\eta}_i$.

3.2 Prices and GDP

Having characterized sectoral production, we can now aggregate the economy to derive expressions for prices and GDP. To do so, we define the sectoral Leontief inverse matrix

$$\mathcal{L} := \left(I - \operatorname{diag}(\hat{\eta})\,\alpha\right)^{-1},$$

where $\operatorname{diag}(\hat{\eta})$ is the diagonal matrix with the effective returns to scale vector $\hat{\eta}$ on the main diagonal. An element $\mathcal{L}_{ij}$ of this matrix captures the importance of sector j in the production of good i, taking into account direct and indirect connections through the production network. For example, $\mathcal{L}_{ij}$ is large if sector i uses a large share of inputs from j (i.e., $\hat{\eta}_i \alpha_{ij}$ is large), or if i relies on another sector k that, in turn, relies heavily on j, and so on.

We show in Appendix C.1 that the Leontief inverse can be used to write the sectoral Domar weights as

$$\omega_i := \frac{P_i Q_i}{\bar{P} Y} = \beta^\top \mathcal{L} \mathbf{1}_i, \tag{18}$$

where $\mathbf{1}_i$ is the ith standard basis vector. As usual in network economies, the Domar weight ω_i provides a measure of the importance of sector ω_i as a supplier. A sector i has a large Domar weight if its output is heavily demanded, either directly by the household (high β_i), or indirectly by other sectors that the household favors (high $\mathcal{L}_{ji}$ and β_j).

Sectoral returns to scale $\hat{\eta}$ play an important role in shaping the production network. Intuitively, when a downstream producer increases its returns to scale, it effectively shifts its input mix toward intermediate goods, thereby increasing its demand for upstream suppliers. This strengthens the input-output linkages and raises the Domar weights of those suppliers. As we will show, this mechanism has important implications for the impact of endogenous returns to scale on the macroeconomy.

We can now characterize equilibrium prices and GDP.

Proposition 3. *The equilibrium price vector $P = (P_1, \ldots, P_N)$ satisfies*

$$\log \frac{P}{W} = -\mathcal{L}(\hat{\eta})\, z(\hat{\eta}), \tag{19}$$

where $z(\hat{\eta}) = (\log Z_1(\hat{\eta}_1), \ldots, \log Z_N(\hat{\eta}_N))$ is the vector of log sectoral productivities (16). *Furthermore, equilibrium log GDP $y := \log Y$ is given by*

$$y(\hat{\eta}) = \underbrace{[\omega(\hat{\eta})]^\top z(\hat{\eta})}_{\textit{Aggregate productivity}} + \underbrace{\log \bar{L}}_{\textit{Labor endowment}}. \tag{20}$$

In equilibrium, prices must equal marginal production costs ($P_i = \lambda_i$). This condition, combined with Proposition 2, allows us to solve for the vector of sectoral prices as 19. Intuitively, the price of good i is low if its key suppliers—both direct and indirect, as captured by the i-th row of $\mathcal{L}$—have high productivity z.

Equation 20 shows that the contribution of a sector's productivity z_i to GDP is proportional to its Domar weight ω_i, as in standard production network economies. One key feature of our model, however, is that both ω and z depend on the endogenous effective returns to scale $\hat{\eta}$. We will explore in Section 5 the role played by $\hat{\eta}$ in shaping GDP.

3.3 Equilibrium existence, uniqueness and efficiency

The preceding analysis describes key equilibrium objects, such as prices and GDP, as functions of the vector of effective returns to scale $\hat{\eta}$. To solve for $\hat{\eta}$ itself and characterize how it responds to changes in the environment, it is convenient to rely on the problem of a social planner. Since there is a single representative household in the economy, the planner's problem is to maximize that household's utility (GDP) subject to the physical constraints of the environment. The following result characterizes that problem and its relation to the set of equilibria.

Proposition 4. *There exists a unique equilibrium, and it is efficient. Furthermore, the equilibrium vector of effective returns to scale $\hat{\eta}$ maximizes GDP $y(\hat{\eta})$, as given by (20).*

The proof of this proposition establishes an equivalence result between the set of equilibria and the set of efficient allocations. It further shows that since there exists a unique efficient allocation, there is also a unique equilibrium. We can then use the first-order conditions of the planner to find the equilibrium $\hat{\eta}$. With that object in hand, the returns to scale of all the firms can be recovered using (12).

4 Forces shaping returns to scale decisions

In this section, we study how changes in the environment affect returns to scale in equilibrium. To do so, we rely on the fact that the equilibrium is efficient, and that the effective returns to scale vector $\hat{\eta}$ maximizes GDP. The first-order condition associated with that problem is[10]

$$\underbrace{\omega_i \alpha_i^\top \mathcal{L}\, z}_{\text{Network adjustment: } d\omega^\top / d\hat{\eta}_i} \quad + \quad \underbrace{\omega_i \left[\frac{da_i}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2} \frac{1}{(1-\hat{\eta}_i)^2} - \frac{1}{2} \log \left(\frac{1}{1-\varphi_i} \right) + \log \kappa_i \right]}_{\text{Productivity adjustment: } dz_i / d\hat{\eta}_i} = 0. \tag{21}$$

A marginal increase in $\hat{\eta}_i$ has two effects on the economy. First, it makes intermediate inputs more important production factors. As a result, the network becomes more connected and Domar weights

[10] See the proof of Proposition 6 for a derivation.

increase. Through this channel, captured by the first term in (21), the productivities z of the sectors that are upstream of i have a larger impact on GDP.

Furthermore, increasing $\hat{\eta}_i$ affects the productivity z_i of sector i directly, as captured by the term in square brackets in (21). The first term there captures the additional productivity cost a_i of selecting a higher returns to scale $\hat{\eta}_i$. In addition, when returns to scale are larger, production can move more easily to the most productive firms within the sector. This amplified selection process is captured by the second term between brackets. Higher returns to scale also affect the importance of returns to scale *dispersion* within a sector. When $\hat{\eta}_i$ is close to 1, there is less room for the most productive firms to increase their returns to scale and, through that channel, reach a larger size. This effect is captured by the third term between brackets. Finally, a higher $\hat{\eta}_i$ means that firms can scale up more freely and thus produce more. Consequently, fewer firms enter, and sectoral productivity benefits from a reduction in the total entry costs. This effectively leads to an increase in sectoral productivity, as the last term in (21) shows.

4.1 Sectoral productivity

We now analyze how changes in the environment affect equilibrium returns to scale. Two prices play a key role in this process. The first is the price H_i of the variable input bundle. At the *firm* level, more expensive inputs move firms toward technologies with lower returns to scale. The second price is the wage W, which plays an additional role at the *sector* level. When labor is expensive, entry becomes costly, reducing the mass of firms. This allows incumbents to expand and encourages the adoption of higher returns to scale. Combining these two effects, we find that the ratio H_i/W is key to determine how returns to scale evolve at the sector level.[11] To capture the role of this ratio, we define the input-price sensitivity matrix $\mathcal{K}$ with typical element

$$\mathcal{K}_{ij} := \frac{\partial}{\partial z_j} \log\left(\frac{H_i}{W}\right) \leq 0,$$

where the partial derivative holds $\hat{\eta}$ fixed. The matrix $\mathcal{K}$ summarizes how productivity shocks propagate through the network to affect input costs. Using the pricing equation (19), we can show that $\mathcal{K} = -\alpha\mathcal{L}$. Since higher productivity lowers prices, the elements of $\mathcal{K}$ are non-positive, with $\mathcal{K}_{ij} < 0$ whenever sector j is an upstream supplier to sector i (i.e., $\mathcal{L}_{ij} > 0$).[12] The matrix $\mathcal{K}$ plays an important role in our analysis and depends crucially on the input-output structure of the economy. Without network connections, $\mathcal{K} = 0$ and several of the mechanisms that we explore below disappear.

Since the sensitivity of sectoral productivity z_i to changes in $\hat{\eta}$ also influences how fundamentals

[11] One can show that H_i/W is the quantity that is raised to the power $\hat{\eta}_i$ in the marginal cost expression (15), which explains its importance for scalability decisions.

[12] We have $\mathcal{L} = (I - \operatorname{diag}(\hat{\eta})\,\alpha)^{-1} = I + \sum_{i=1}^{\infty} (\operatorname{diag}(\hat{\eta})\,\alpha)^i = I - \operatorname{diag}(\hat{\eta})\,\mathcal{K}$, so that $\mathcal{K}_{ij} < 0 \Leftrightarrow \mathcal{L}_{ij} > 0$ if $i \neq j$.

affect equilibrium returns to scale, it is convenient to define

$$\Psi_i := \frac{d^2 z_i}{d\hat{\eta}_i^2} = (1-\varphi_i)\,\frac{d^2 a_i}{d\hat{\eta}_i^2} < 0,$$

where the second equality follows directly from (16). The inverse Ψ_i^{-1} therefore captures how elastic returns to scale are in sector i. A small $|\Psi_i|$ implies that returns to scale are flexible and respond strongly to changes in the environment.

Using these definitions, we can characterize the impact of μ_j and σ_j^2 on sectoral returns to scale.

Lemma 6. *An increase in average productivity μ_j increases returns to scale in all downstream sectors, such that*

$$\frac{d\hat{\eta}_i}{d\mu_j} = \Psi_i^{-1}\mathcal{K}_{ij} \geq 0. \tag{22}$$

Furthermore, the impact of productivity dispersion σ_j^2 on $\hat{\eta}_i$ is given by

$$\frac{d\hat{\eta}_i}{d\sigma_j^2} = \Psi_i^{-1}\left(\mathcal{K}_{ij}\frac{\partial z_j}{\partial \sigma_j^2} - \mathbb{I}_{\{i=j\}}\frac{\partial^2 z_i}{\partial \sigma_i^2 \partial \hat{\eta}_i}\right), \tag{23}$$

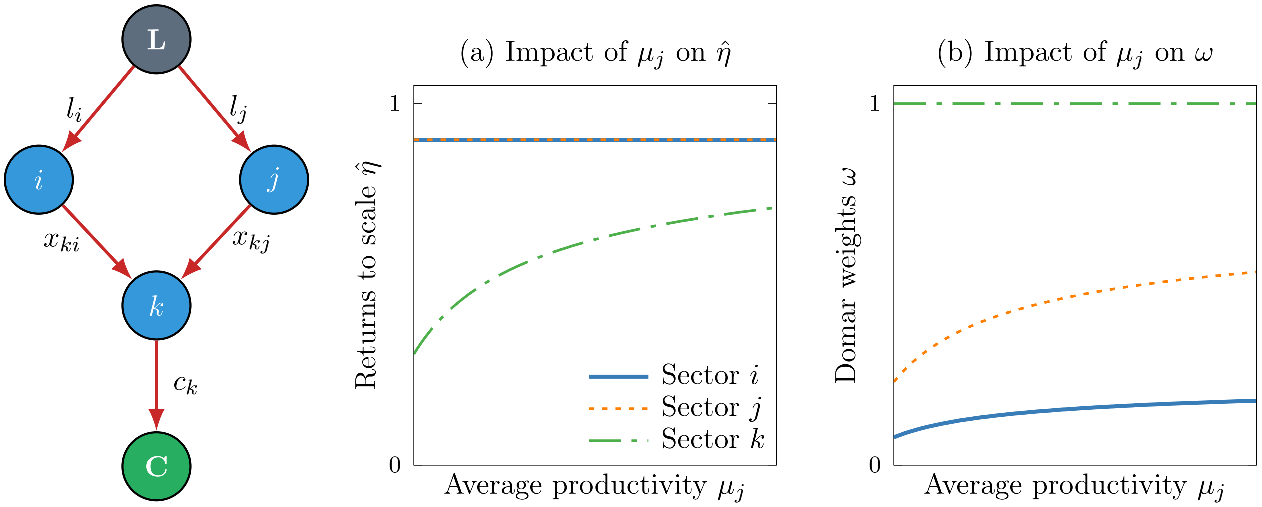
where

$$\frac{\partial z_j}{\partial \sigma_j^2} = \frac{1}{2\,(1-\hat{\eta}_j)} + \frac{1-\hat{\eta}_j}{4\gamma_j\,(1-\varphi_j)} > 0, \text{ and } \frac{\partial}{\partial \sigma_i^2}\left(\frac{\partial z_i}{\partial \hat{\eta}_i}\right) = \frac{1}{2\,(1-\hat{\eta}_i)^2} - \frac{1}{4\gamma_i\,(1-\varphi_i)}.$$

In particular, $d\hat{\eta}_i/d\sigma_j^2 \geq 0$ for $i \neq j$.

Consider (22) first. An increase in μ_j makes firms in sector j more productive, which lowers the price P_j through competition. If sector i is a downstream customer of j ($\mathcal{L}_{ij} > 0$), this lowers the price of its variable input bundle by an amount proportional to $|\mathcal{K}_{ij}|$. This in turn pushes firms in sector i to increase their returns to scale to take advantage of the cheaper intermediate inputs (Proposition 2). The magnitude of this response depends on how elastic $\hat{\eta}_i$ is, as given by Ψ_i^{-1}. Both σ_i^2 and γ_i influence this elasticity through φ_i. Specifically, if φ_i is large, which is the case if productivity is dispersed (high σ_i^2) or scalability is cheap (low γ_i), the response is stronger. Intuitively, in such sectors, there is a larger mass of high-productivity firms able to aggressively scale up in response to cheaper inputs.

Example. Consider as an example, the economy depicted in the left panel of Figure 4. Since sector k is downstream from j, an increase in μ_j reduces the price of the input bundle in sector k. In response, firms in sector k increase their returns to scale $\hat{\eta}_k$, as panel (a) shows. In contrast, since sector i is not downstream from j, input prices in sector i are unchanged and so are their returns to scale $\hat{\eta}_i$.



(a) Impact of μ_j on $\hat{\eta}$

(b) Impact of μ_j on ω

Figure 4: The impact of sectoral productivity on returns to scale and Domar weights in a vertical economy

The second part of Proposition 6shows that an increase in dispersion σ_j^2 affects returns to scale similarly to a productivity shock μ_j. Indeed, higher dispersion σ_j^2 raises sectoral productivity z_j through two channels. First, it strengthens the standard selection mechanism: even with fixed returns to scale, higher variance reallocates market share to more productive firms. Second, it amplifies the superstar effect: with a fatter right tail of productivity, the most productive firms adopt even higher returns to scale, further boosting aggregate efficiency. This increase in z_j lowers output prices, triggering a downstream increase in returns to scale $\hat{\eta}_j$ analogous to the response to a μ_j shock.

For the shocked sector j itself, however, an additional channel is at work. As shown in (21), higher dispersion σ_j^2 generally increases the marginal benefit of scalability (i.e., the sensitivity of z_j to $\hat{\eta}_j$). Intuitively, when the productivity distribution is more dispersed, the gains from allowing the best firms to scale up are larger. This direct incentive effect, captured by the term $\frac{\partial}{\partial \sigma_i^2}\left(\frac{\partial z_i}{\partial \hat{\eta}_i}\right)$, leads to a further increase in $\hat{\eta}_j$.[13]

Lemma 6 focuses on changes in the productivity process, but it is straightforward to derive similar results for changes in entry costs κ_j and the cost of scalability γ_j. We provide these results in Appendix D.2. In a nutshell, an increase in the entry cost κ_j in sector j always reduces the effective returns to scale of any other downstream sector $i \neq j$ since it leads to an increase in the price of good j. An increase in the cost of scalability γ_j naturally leads to a reduction in returns to scale in sector j. The price of good j increases as a result, leading to lower returns to scale in other sectors as well.

[13] This term is positive for reasonable calibrations of the model, but can be negative under some parametrizations.

4.2 Implications for Domar weights

The comparative statics results derived so far describe how the equilibrium returns to scale $\hat{\eta}$ respond to the environment. These movements in $\hat{\eta}$ are important, in part, because they directly alter the sectoral Domar weights ω and, therefore, GDP. The following lemma formalizes this relationship. In the next section, we use this result to analyze how endogenous returns to scale shape GDP and welfare.

Lemma 7. *The impact of a parameter $\chi \in \{\mu_j, \sigma_j, \kappa_j, \gamma_j\}$ on sectoral Domar weights is given by*

$$\frac{d\omega_i}{d\chi} = -\sum_{k=1}^{N} \mathcal{K}_{ki} \omega_k \frac{d\hat{\eta}_k}{d\chi}. \tag{24}$$

Proof. Follows directly from differentiating the expression for Domar weights ω given by (18). □

This result shows that if a shock to χ leads sector k to increase its returns to scale ($\hat{\eta}_k \uparrow$), the Domar weights of all sectors *upstream* of k increase. To illustrate this mechanism, consider again the example in Figure 4.

Example. Recall that an increase in μ_j raises productivity in sector j, lowering P_j through competition. Since k is a downstream customer of j, firms in k respond by increasing their returns to scale $\hat{\eta}_k$ to take advantage of the cheaper input (Lemma 6). Lemma 7 shows that this shift raises the Domar weights of all of k's suppliers (panel (b) in Figure 4). Intuitively, as sector k scales up, it becomes more input-intensive, increasing its demand for upstream goods. Consequently, the sales and the Domar weights of sectors i and j rise, reflecting their increased centrality to aggregate production. Notably, sector i's importance grows even though its own returns to scale are unaffected by the shock.

This example highlights a key feature of our model: while productivity shocks propagate *downstream* to affect returns to scale, adjustments in returns to scale propagate *upstream* to reshape the network structure and alter Domar weights. As we show next, this interaction has important implications for GDP.

5 Endogenous returns to scale and GDP

Endogenous scalability has important consequences for GDP and hence welfare. In this section, we show that it raises the level of GDP through better resource allocation and alters the economy's response to shocks. We then show that endogenous returns to scale lead to higher long-run economic growth.

Throughout our analysis, we compare the equilibrium of our model to a counterfactual economy in which returns to scale are exogenously fixed.

Definition 2 (Fixed returns-to-scale economy)**.** Let the equilibrium effective returns-to-scale vector in the baseline model be $\hat{\eta}$. The *fixed returns-to-scale economy* is an otherwise identical economy in which the returns to scale of all firms are exogenously set to $\eta_{il} = \hat{\eta}_i$ for all i and l. All other quantities are chosen optimally. Furthermore, the returns to scale $\{\eta_{il}\}$ are fixed and do not respond to any changes in the environment.

Endogenous returns to scale do two main things in our model: They create dispersion in η_{il} within a sector, and they allow η_{il} to respond to changes in parameters. The fixed returns-to-scale economy, by construction, shuts down both of these channels. By comparing our baseline model to this counterfactual, we can therefore isolate the full impact of endogenous scalability on economic outcomes.[14] Note that by construction, sectoral Domar weights are the same in both economies. In what follows, we use $\tilde{\cdot}$ to denote quantities in the fixed returns-to-scale economy.

5.1 Contribution of endogenous returns to scale to the level of GDP

We first examine the impact of endogenous returns to scale on GDP by comparing its level in the baseline and the fixed returns-to-scale economies. Since both economies share the same Domar weights, any difference in GDP must arise from differences in sectoral productivity. Comparing that quantity in the baseline model (Z) with its counterpart in the fixed returns-to-scale economy ($\tilde{Z}$), we find that[15]

$$\log Z_i\left(\hat{\eta}_i\right) - \log \tilde{Z}_i\left(\hat{\eta}_i\right) := \frac{1}{2}\left(1-\hat{\eta}_i\right)\log\left(\frac{1}{1-\varphi_i}\right) > 0. \tag{25}$$

Thus, sectoral productivity is always larger in the model with endogenous returns to scale. Intuitively, when returns to scale are fixed, high-ε_{il} firms are no longer able to adjust their scalability to take advantage of their high productivity. This limits how much they produce, and sectoral productivity falls as a result. The superstar effect is completely neutered in that case. Equation (25) shows that the difference between the two economies is particularly pronounced when the effective dispersion φ_i is large and the effective returns to scale $\hat{\eta}_i$ is low. In those circumstances, highly productive firms in the baseline model can deviate strongly from $\hat{\eta}_i$ and thus contribute more to sectoral productivity.

The following results characterize the aggregate impact of endogenous returns to scale.

Proposition 5. *The difference in log GDP between the baseline model and the fixed returns-to-scale economy is given by*

$$y - \tilde{y} = \sum_{i=1}^{N} \omega_i \frac{1}{2}\left(1-\hat{\eta}_i\right)\log\left(\frac{1}{1-\varphi_i}\right) > 0. \tag{26}$$

The gain in GDP from endogenous returns to scale is simply the Domar-weighted gain in sectoral productivity. We see from (26) that $y-\tilde{y}$ is particularly large if the sectors in which high-productivity

[14] In Section 7, we disentangle the impact of these two channels in our calibrated model.

[15] See the proof of Proposition 5 for the derivation.

firms can scale their returns to scale more easily (high φ_i and low $\hat{\eta}_i$) are also important suppliers (high ω_i). In the calibrated economy of Section 7, we will see that the impact of endogenous returns to scale on the level of GDP can be sizable.

5.2 How GDP responds to changes in the environment

In addition to the level of GDP, endogenous returns to scale also affect how GDP responds to changes in the environment.

Proposition 6. *In equilibrium, the following holds.*

1. *An increase in average productivity μ_j raises GDP:*
$$\frac{dy}{d\mu_j} = \frac{\partial y}{\partial \mu_j} = \omega_j > 0. \tag{27}$$

2. *An increase in productivity dispersion σ_j^2 raises GDP:*
$$\frac{dy}{d\sigma_j^2} = \frac{\partial y}{\partial \sigma_j^2} = \omega_j \left(\frac{1}{2} \frac{1}{1-\hat{\eta}_j} + \frac{1-\hat{\eta}_j}{4\gamma_j} \frac{1}{1-\varphi_j} \right) > 0. \tag{28}$$

3. *An increase in entry cost κ_j lowers GDP:*
$$\frac{dy}{d\log\kappa_j} = \frac{\partial y}{\partial \log\kappa_j} = -\omega_j \left(1-\hat{\eta}_j\right) < 0.$$

4. *An increase in the returns to scale productivity cost γ_j lowers GDP:*
$$\frac{dy}{d\gamma_j} = \frac{\partial y}{\partial \gamma_j} = -\omega_j \left(\frac{1}{1-\hat{\eta}_j} + \frac{1-\hat{\eta}_j}{2\gamma_j} \frac{\varphi_j}{1-\varphi_j} \right) < 0. \tag{29}$$

In these expressions, the partial derivatives indicate that returns to scale $\{\eta_{il}\}$ are taken as fixed.

Proof. The result follows directly from the envelope theorem. □

Since the equilibrium $\hat{\eta}$ maximizes GDP, any *marginal* adjustment in returns to scale must have no impact on GDP. This implies that GDP responds to marginal changes in the environment *as if* returns to scale were fixed. Consequently, Hulten's (1978) theorem applies: the first-order impact of a productivity shock $d\mu_j$ is simply the Domar weight ω_j of the affected sector. Note, however, that Domar weights themselves are endogenous in our model and depend on the incentives shaping returns to scale. Similarly, the impact of a change in entry costs $d\log\kappa_j$ is determined by its direct effect on sector j's productivity z_j, captured by $1-\hat{\eta}_j$, weighted by that sector's importance, ω_j.

The impacts of σ_j^2 and γ_j operate in a similar way, but here endogenous returns to scale feature more prominently. Consider first an increase in σ_j^2. This has two positive effects on sector j's productivity z_j. First, the higher dispersion in ε_{il} implies a larger mass of very productive firms, with positive consequence for GDP. This selection effect is captured by the term $\frac{1}{2}\frac{\omega_j}{1-\hat{\eta}_j}$ in (28) and would be at work even without within-sector dispersion in $\hat{\eta}_{il}$ (i.e., in the fixed returns-to-scale economy). Second, increasing σ_j^2 interacts with the superstar effect. Since high-ε_{il} firms already have high returns to scale, the increase in dispersion has a disproportionate impact on them. This effect is captured by the remaining term in (28). Overall, increasing σ_j^2 always has a positive effect on GDP, and the presence of endogenous returns to scale, as it creates dispersion in within sector firm-level returns to scale, make that effect larger, even to a first order.

Conversely, an increase in the cost of adjusting returns to scale γ_j has two adverse effects on GDP. The first effect is mechanical: a higher γ_j directly increases the average productivity cost $-a_i(\hat{\eta})$ in sector i, which lowers GDP. This effect is captured by the first term in (29). Second, γ_j interacts with the superstar effect. Since the biggest producers have the highest returns to scale, they suffer particularly strongly from an increase in γ_j. Indeed, recall that $a_i(\eta_{il}) = -\gamma_i/(1-\eta_{il})$ such that for $\eta_{il} \approx 1$, a marginal increase in γ_j has a particularly severe impact on productivity. This effect is captured by the second term in (29). Proposition 6 shows that increasing γ_j always hurts y, and all the more so when the superstar effect is stronger.

5.3 Second-order impact of productivity shocks

In our baseline model, the equilibrium is efficient, which implies that GDP responds to infinitesimal productivity shock as if returns to scale were held fixed. For larger shocks, however, the Domar weights themselves respond and, doing so, affect GDP.

Proposition 7. *The response of log GDP y to a shock $\Delta\mu_i$ is given by*

$$\Delta y = \omega_i \Delta\mu_i + \frac{1}{2}\frac{d\omega_i}{d\mu_i}(\Delta\mu_i)^2 + o\left((\Delta\mu_i)^2\right). \tag{30}$$

Furthermore, the second-order term is non-negative, $d\omega_i/d\mu_i \geq 0$ and given by (24).

Equation (30) provides a second-order approximation of the GDP response to a shock $\Delta\mu_i$. While the first-order term is the standard Domar-weight effect from Hulten's theorem, the second-order term captures a novel channel driven by the endogenous adjustment in returns to scale. That channel operates through the response of Domar weights to the shocks.

The intuition is straightforward. Recall that a positive productivity shock in sector i propagates downstream, lowering input costs for its customer sectors. These sectors, in turn, are incentivized to increase their own returns to scale to capitalize on the cheaper inputs. This shift towards greater scalability makes the production network more reliant on sector i, which increases its Domar weight.

As a result, the derivative $d\omega_i/d\mu_i$ is positive. This implies that the second-order term is always positive, adding convexity to the GDP response. In other words, endogenous returns to scale amplify the impact of positive productivity shocks and dampen the adverse impact of negative shocks.

To identify the sources of this amplification and dampening, we can use Lemma 6 to express the second-order coefficient as

$$\frac{d\omega_i}{d\mu_i} = -\sum_{k=1}^{N} \omega_k \Psi_k^{-1} \mathcal{K}_{ki}^2 \geq 0. \tag{31}$$

The second-order term is therefore stronger when the shock hits a sector i that is a key supplier ($|\mathcal{K}_{ki}|$ large) to sectors that are large (high ω_k) and have elastic returns to scale (low Ψ_k, meaning a low cost of adjusting scalability). Those sectors more strongly increase their cost share of good i after the increase in μ_i, contributing to a larger increase in ω_i. Crucially, it is the square of $\mathcal{K}_{ki}$ that shows up in (31), implying that the convex response of GDP is disproportionately stronger for shocks to the most important suppliers.

Proposition 7 implies that whether shocks hit sectors that are upstream or downstream in the supply chain matters for their impact on GDP. The following example illustrates the mechanism.

Example. Consider the vertical economy depicted in Figure 5. An upstream sector 1 sells its entire output to a downstream sector 2, which in turn sells to the final consumer. Panel (a) shows the impact of a productivity shock μ_1 to the *upstream* sector. The solid line, representing the response of GDP, is clearly convex and lies above the linear, first-order approximation from Hulten's theorem (dashed line). This illustrates the amplification effect: as μ_1 increases, the price of good 1 falls, inducing firms in the downstream sector 2 to increase their returns to scale to capitalize on cheaper inputs. This change in production processes makes sector 1 a more important supplier, increasing its Domar weight and thus magnifying the aggregate benefit of its higher productivity. Panel (b) shows a starkly different result for a productivity shock μ_2 to the *downstream* sector. In this case, the full GDP response is linear and coincides with the Hulten's theorem approximation. Because sector 2 is at the bottom of the supply chain, a fall in its price provides no benefit to any other producing sectors. There are no downstream customers to re-optimize their scalability choices, and thus no structural amplification. This example shows a key implication of our model: the macroeconomic impact of a productivity shock depends crucially on a sector's position in the production network.

5.4 Extension: The role of wedges

So far, our analysis has focused on an efficient equilibrium where the envelope theorem holds, meaning that adjustments in $\hat{\eta}$ have only second-order effects on GDP. We now show that in the presence of frictions, markups or other distortions, this is no longer the case, and changes in returns to scale can have first-order effects on GDP.

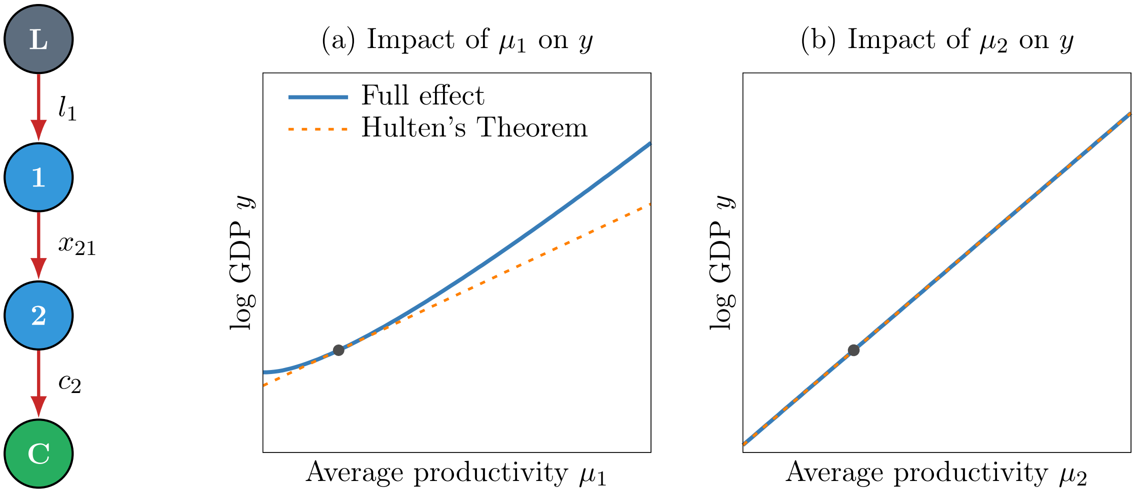


Figure 5: The impact of sectoral productivity on GDP in a vertical economy

We consider a setup with general wedges in Appendix D.5, but to illustrate the forces at work transparently, we focus here on distortionary sales wedges. Specifically, we assume that firm l in sector i retains only a fraction $1 - \tau_i^S$ of its revenue, where $\tau_i^S \in [0, 1)$ represents the wedge. From the firm's perspective, these wedges are equivalent to a reduction in productivity, leading to an inefficient adjustment in scalability. We assume that the proceeds from the wedges are fully rebated to the household lump-sum, ensuring that they act as pure distortions with no direct loss in resources.

Sales wedges directly affects the scalability choices.

Lemma 8. *An increase in the wedge τ_i^S decreases the returns to scale in all downstream sectors, such that*

$$\frac{d\hat{\eta}_i}{d\tau_j^S} = -\frac{1}{1-\tau_j^S}\Psi_i^{-1}\mathcal{K}_{ij} \leq 0. \tag{32}$$

6Intuitively, the sales wedge acts like a markup, driving a wedge between marginal cost and the market price. As τ_i^S rises, output prices increase as well, making intermediate inputs more expensive for downstream firms. Facing higher variable input costs, these firms optimally substitute away from input-intensive technologies by reducing their returns to scale. Since firms do not internalize that the revenue from wedges is rebated to the household, this adjustment leads to an equilibrium with inefficiently low returns to scale.

Wedges also have important implications for the behavior of GDP. Propositions 6 and 7 show that without wedges, changes in returns to scale only have second-order effects on GDP. This is no longer the case when wedges are present.

Proposition 8. *In the presence of sales wedges, the impact of a parameter $\chi \in \{\mu_j, \sigma_j, \kappa_j, \gamma_j\}$ on*

GDP is given by

$$\frac{dy}{d\chi} = \underbrace{\frac{\partial y}{\partial \chi}}_{\textit{Direct effect}} + \underbrace{\sum_{i=1}^{N} \frac{\partial y}{\partial \hat{\eta}_i} \frac{d\hat{\eta}_i}{d\chi}}_{\textit{Structural change effect}}, \tag{33}$$

where $\partial y/\partial \chi$ is given by Proposition 6, $d\hat{\eta}_i/\partial \chi$ is given by Lemmas 6, 9 and 10, and $\partial y/\partial \hat{\eta}_i \geq 0$.

This proposition shows that the total effect of a shock is now the sum of a direct effect (the standard Hulten's-like term) and a new structural change effect. Since wedges push $\hat{\eta}$ to be inefficiently low, any shock that incentivizes firms to increase their returns to scale ($d\hat{\eta}/d\chi > 0$) now generates an additional, first-order welfare gain.

Consider, for example, a positive productivity shock $d\mu_j > 0$. As in the efficient case, it directly raises GDP. However, it also lowers input costs, inducing firms to increase their $\hat{\eta}$. Because the economy started from an inefficiently low level of returns to scale, this structural adjustment is no longer a second-order refinement but a first-order improvement. This implies that $dy/d\mu_i \geq \omega_i$. Productivity shocks have a larger impact on GDP in this distorted economy because they not only improve productivity but also partially correct the pre-existing distorted scalability structure.

The nature of this first-order effect depends on the type of distortion introduced in the model. While sales wedges lead to inefficiently low returns to scale, other distortions can have the opposite effect. For example, as we show in Appendix D.5, a corporate profit tax can effectively raise the cost of entry, which perversely incentivizes incumbent firms to choose inefficiently high returns to scale. In such an economy, a productivity shock that further raises $\hat{\eta}$ could actually be welfare-reducing at the margin.[16] This highlights the importance of understanding the specific nature of distortions when evaluating the welfare effect of shocks in an economy with endogenous scalability.

5.5 Implications for growth

Endogenous scalability can also propel long-run economic growth. To illustrate this mechanism transparently, we simplify the model to a single-sector economy and assume that growth is driven by a constant rate of productivity improvement, $d\mu/dt = g_\mu > 0$.

We begin our analysis by characterizing the growth rate of returns to scale in this environment.

Corollary 1. *The growth of effective returns to scale $\hat{\eta}$ is given by*

$$\frac{d\hat{\eta}}{dt} = \Psi^{-1}\mathcal{K}g_\mu > 0. \tag{34}$$

Furthermore, as $t \to \infty$, effective returns to scale $\hat{\eta}$ converges to 1.

The constant improvements in sectoral productivity μ result in cheaper intermediate inputs,

[16] We show this formally in the proof of Proposition 12 in Appendix D.5. See, in particular, equation (86).

which push firms to increase their returns to scale, so that $d\hat{\eta}/dt > 0$.[17] The strength of that mechanism relies on how fast productivity improves (g_μ), but also on the importance of intermediate inputs in production, as captured by the term $-\alpha/\left(1-\hat{\eta}\alpha\right) = \mathcal{K}$. When α is large, intermediates are more important and $\hat{\eta}$ grows faster, all else equal. In addition, $d\hat{\eta}/dt$ depends on the elasticity of $\hat{\eta}$ in response to a change in input prices. As in the comparative statics exercise of Section 4, this effect is captured by $1/\Psi$. An economy that is more flexible and faces stronger incentives will more rapidly reconfigure itself toward a more scalable structure.

The evolution of the economy's returns to scale has implications for the growth rate of GDP.

Proposition 9. *The growth rate of GDP is given by*

$$\frac{dy}{dt} = \frac{g_\mu}{1-\alpha} \times \left(1 - \frac{1}{\sqrt{1 + \frac{1}{\gamma}\frac{1-\alpha}{\alpha}\left(\frac{g_\mu}{1-\varphi}t + T\right)}}\right) > 0, \tag{35}$$

where

$$T := -\frac{1-\alpha}{\alpha}a'\left(\hat{\eta}_0\right) - 2a\left(\hat{\eta}_0\right) > 0,$$

and where $\hat{\eta}_0$ is the effective returns to scale at $t = 0$.

This proposition provides a closed-form solution for the growth rate of GDP along its entire transition path. The growth rate dy/dt can be understood as the product of two terms: a *potential long-run growth rate*, $g_\mu/\left(1-\alpha\right)$, and a convergence factor (the term in parentheses) that starts below one and asymptotically approaches it as $t \to \infty$. This means that the economy's growth rate is always increasing, accelerating towards its long-run potential, which it reaches asymptotically.

The speed of this convergence is governed by the terms inside the square root. The economy adapts its structure and accelerates more quickly when the forces driving reorganization are stronger. Specifically, the transition is faster when: 1) the pace of innovation (g_μ) is high, providing a strong and persistent incentive to adapt, and 2) the economy is structurally flexible, meaning the net cost of adjusting scale is low (low γ and high φ).

Figure 6 provides an example of the dynamic of the growth rate of GDP. The left panel shows that from its initial condition $\eta\left(0\right) = \eta_0$, growth increases before converging to its long-run level $g_\mu/\left(1-\alpha\right)$. The right-panel depicts the same economy but with a higher γ. In this case, increasing returns to scale is more costly, and the progression to the long-run growth rate is slower.[18]

[17] Several empirical studies document rising returns to scale over time. De Loecker et al. (2020) estimate that firm-level returns to scale have increases over the last few decades in the United States. Chiavari and Goraya (2025) show that these results hold even accounting for intangible capital. Lashkari et al. (2024) provide evidence that the decline in IT prices led to an increase in returns to scale in France.

[18] While growth is always accelerating ($d^2y/dt^2 > 0$), the rate of acceleration is in general modest. One can show

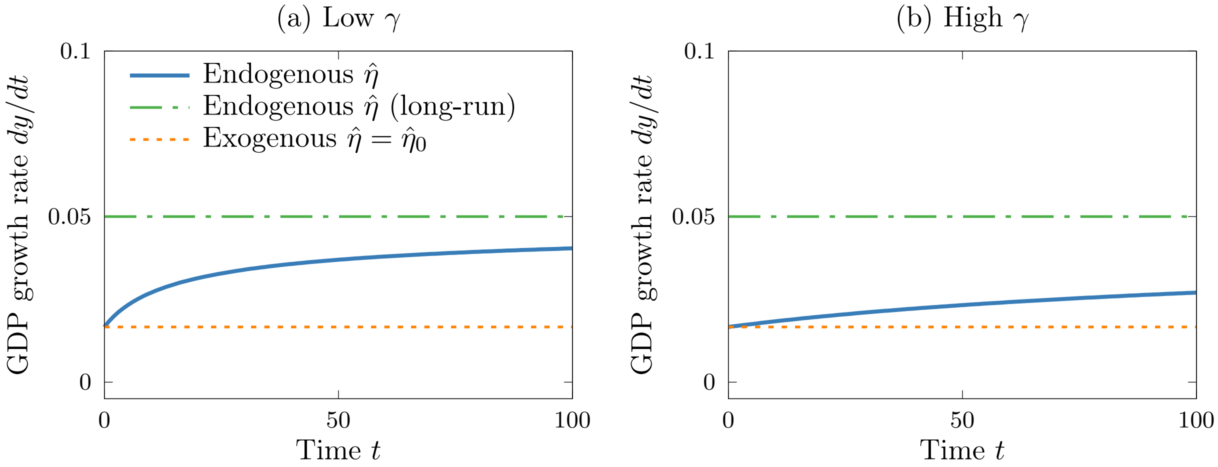


Figure 6: Endogenous returns to scale accelerate growth

Finally, we can compare the growth rate of GDP in the model with endogenous returns to scale, to its counterpart in the fixed returns-to-scale economy.

Corollary 2. *For any $t > 0$, GDP grows faster in the economy with endogenous returns to scale. In the limit as $t \to \infty$, the long-run growth rates satisfy*

$$\lim_{t\to\infty} \frac{dy}{dt} = \frac{1}{1-\alpha} g_\mu > \frac{1}{1-\hat{\eta}_0 \alpha} g_\mu = \lim_{t\to\infty} \frac{d\tilde{y}}{dt},$$

where $\tilde{y}$ is log GDP in the fixed returns-to-scale economy, and where $\hat{\eta}_0$ is the effective returns to scale vector in the baseline economy at $t = 0$.

This result captures the key growth implication of our model: an economy that can adapt its returns to scale grows faster. In the fixed-scale economy, the benefits of technological progress are constrained by a static production structure. In our model, as sectoral productivity increases, the economy adopts more scalable technologies and becomes more interconnected. This increases Domar weights and magnifies the benefit of the higher productivity. This effect grows larger over time, with an increasing gap between the growth rates of the two economies. Our model therefore suggests that the macroeconomic consequences of scaling decisions, like Ford's introduction of the moving assembly line, are not a one-off level effect, but a persistent force that reshapes the economy's long-run growth trajectory.

The quantitative implications of endogenous returns to scale for growth can be substantial. For instance, using parameters calibrated to the Spanish economy (where $\alpha \approx 0.67$ and $\hat{\eta}_0 \approx 0.83$, see next sections), a 1% annual rate of underlying technological progress (g_μ) would translate into a

that the second derivative of log GDP scales with g_μ^2. This implies that for realistic calibrations of annual productivity growth (e.g., $g_\mu \approx 1\%$), the acceleration is a gradual, slow-moving process, that might be hard to notice over short horizons.

2.2% growth rate in the fixed-η economy. With endogenous returns to scale, however, the long-run growth rate converges to 3.0%. This difference of 0.8 percentage points, compounded over decades, can amount to a large welfare gain.

6 Empirical evidence

Our theory describes how returns to scale respond to the economic environment. In this section, we use detailed firm-level data from Spain to provide empirical evidence for these predictions at the firm, sector, and aggregate levels. Consistent with the model, we show a robust positive correlation between productivity and returns to scale in the cross-section of firms. We further use panel data and within-firm variation to demonstrate that returns to scale respond to incentives: firms actively increase scalability as they grow and reduce it when facing higher input costs driven by import tariffs. At the sector level, the theory predicts that industries with stronger endogenous scalability should exhibit fatter firm-size tails. We find strong support for this mechanism in the data. Finally, our theory implies that endogenous scalability is beneficial for GDP. To test the prediction, we extend our analysis to 24 countries, and show that the strength of this mechanism is indeed a predictor of long-run economic development.

6.1 Data

Our primary source of firm-level data is Moody's Orbis Historical database, which covers the near-universe of Spanish firms between 1995 and 2019. This dataset provides detailed information on sales, labor costs, capital stocks, and material costs. After cleaning, our sample comprises 9,754,405 firm-year observations.[19] We use these data to estimate firm-level returns to scale using a production function approach. We briefly describe our estimation procedure below and provide additional details in Appendix A.1.

We complement our analysis with firm-level data from 24 additional countries. We use data from Orbis for 22 European countries with good coverage of the key variables needed for production function estimation. For developing countries, we use China's National Bureau of Statistics (NBS) firm-level database and India's Annual Survey of Industries (ASI), both of which are censuses of above-scale manufacturing firms. The details of data cleaning and variable construction are provided in Appendix A.7.

[19] We deflate all nominal variables using the Spanish GDP deflators and drop any firm-year observation whose average revenue product for any input (fixed assets, wage bills, or material costs) lies above the 99th percentile or below the 1st percentile of that year's distribution.

6.2 Estimating returns to scale and productivity

Our theory predicts that within a sector, firms of similar sizes should have similar returns to scale. We take advantage of this prediction to estimate returns to scale across the firm-size distribution. Specifically, for each sector i and year t, we group firms into 10 deciles based on their 7-year moving average of firm-level log sales (years $t-3$ to $t+3$). This construction smooths out short-run fluctuations and measurement error and thus yields a more reliable measure of a firm's position in the size distribution.

In our baseline approach, we assume that all firms in sector i, year t and decile d_t share the same Cobb-Douglas production technology $Q_{ilt} = \widetilde{A}_{ilt} K_{ilt}^{\beta^K_{i,d_t(l),t}} L_{ilt}^{\beta^L_{i,d_t(l),t}} M_{ilt}^{\beta^M_{i,d_t(l),t}}$, where K_{ilt}, L_{ilt} and M_{ilt} denote capital, labor and materials, respectively. We then estimate the output elasticities for each cell (i, t, d_t) using the Blundell and Bond (2000) IV-GMM estimator on a 7-year rolling-window sample. The estimated returns to scale η_{ilt} for a firm l in sector i and year t is therefore given by the sum of these elasticities:

$$\eta_{ilt} = \hat{\beta}^K_{i,d_t(l),t} + \hat{\beta}^L_{i,d_t(l),t} + \hat{\beta}^M_{i,d_t(l),t}.$$

We then use the estimated returns to scale for the years 1997-2019 in our empirical analysis and in the calibration of Section 7.[20]

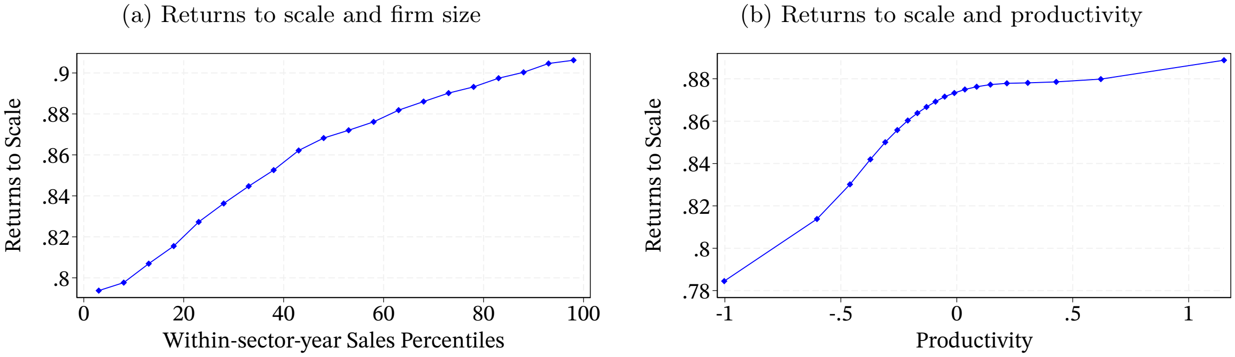
We use the Blundell–Bond estimator as our baseline because it can estimate the gross output production function by leveraging moment conditions that exploits input persistence, without requiring rigid assumptions on input timing. In Appendix A.4, we confirm the robustness of our results using a wide range of alternative strategies. These include: (i) alternative production function estimators, including Olley and Pakes (1996) and Levinsohn and Petrin (2003); (ii) controls for market power to alleviate the measurement errors in output using the Ackerberg et al. (2015) estimator; and (iii) grouping firms by rolling sales percentiles or by contemporaneous rather than moving-average sales. Our main empirical patterns are robust across all of these alternative estimation strategies.

Finally, to compare productivity in the cross-section despite heterogeneous production technologies, we follow best practices in the development accounting literature (Caves et al., 1982a,b; Feenstra et al., 2015) and construct a comparable measure, $\widehat{z}_{ilt}$, using a Törnqvist productivity index. Specifically, for each sector-year, we define a hypothetical "average firm" for each sector-year with mean (log) sales, (log) inputs, and output elasticities. The measure $\hat{z}_{ilt}$ then compares productivity between firm l and the average firm by looking at how much more output one produces relative to the other, adjusting for differences in technology and input use. We formally define this productivity index and provide more detail in Appendix A.3.[21]

[20] We do not estimate the elasticities for 1995 and 1996 as the number of firm-year observations in each sector-year-decile cells is limited in those years. Our 1997-2019 sample covers 9,424,952 firm-year observations.

[21] In our calibrated economy of Section 7, the within-sector correlation between the Törnqvist index and ε_{il} is above 0.99. The same number for $\varepsilon_{il} + a_i(\eta_{il})$ is about 0.92. The Törnqvist index therefore seems to be a good proxy

Figure 7: Returns to scale, productivity, and firm size


(a) Returns to scale and firm size

(b) Returns to scale and productivity

Notes: Panel (a) presents a binned scatter plot of firm-level returns to scale against within–sector–year sales percentiles. Panel (b) presents a binned scatter plot of firm-level returns to scale against productivity, controlling for sector–year fixed effects; the unconditional mean of returns to scale is added back for interpretability. Sectors are defined as the 62 sectors in the Input–Output table of the Annual Spanish National Accounts, approximately corresponding to NACE 2-digit industries. Both panels are constructed using a sample of Spanish firms from Orbis. See Appendix A.1 for details on variable construction and sample selection.

6.3 Firm-level evidence

6.3.1 Cross-sectional patterns: returns to scale, productivity and size

A key prediction of our theory is that larger, more productive firms operate technologies with higher returns to scale. Figure 7 confirms this pattern in the data. Panel (a) plots returns to scale against firms' sales percentiles within a sector-year. It shows that the largest firms (100th percentile) operate with returns to scale of around 0.91, compared to 0.79 for the smallest firms (1st percentile).[22] Furthermore, panel (b) shows that higher productivity is directly associated with higher returns to scale, which is a distinctive feature of our endogenous returns to scale mechanism. Furthermore, the empirical relationship between returns to scale and productivity is concave. This supports our assumption of rising marginal cost of scalability (concave a_i) and qualitatively matches the theoretical prediction in Figure 2.

6.3.2 Panel evidence on endogenous returns to scale

The evidence presented in the last section shows that on *average* larger and more productive firms are more scalable. But the theory also predicts that the returns to scale of *individual firms* should also respond to changes in the economic environment. Indeed, Lemma 3 establishes that

for productivity for reduced-form estimates.

[22] Hubmer et al. (2025) document similar patterns in Canadian and U.S. manufacturing firms. Gao and Kehrig (2025) and McAdam et al. (2024) report that industries with larger average firm size also have higher returns to scale in the United States and in European countries, respectively. Using production data to infer firm- and industry-level returns to scale has a long tradition since Hall (1990), Burnside et al. (1995) and Basu and Fernald (1997). More recent work that uses production-function estimation to recover heterogeneity in returns to scale at the firm or industry level includes De Loecker et al. (2020), Ruzic and Ho (2023), Chiavari (2024), Savagar and Kariel (2024) and Demirer (2025).

beneficial shocks, whether higher productivity or lower input prices, induce firms to adopt more scalable technologies. Since these same shocks also drive firm expansion, the model implies that returns to scale and sales should co-move positively within a firm over time.

Figure 8 tests these predictions by plotting the *within-firm* variation in returns to scale against sales and productivity, controlling for firm and sector-year fixed effects.[23] Both panels show a strong positive relationship: when a firm experiences an increase in sales or productivity, its returns to scale tend to rise. Quantitatively, a 100% increase in sales is associated with a 0.013 increase in returns to scale, while an analogous increase in productivity raises returns to scale by 0.008. Taken together, these within-firm results provide panel evidence for our mechanism: rather than operating fixed production technologies, firms seem to adopt higher returns to scale as they grow.We next examine the response of returns to scale to changes in input costs. In the model, an increase in the cost of the variable input bundle, for example, due to higher tariffs on imported intermediates, reduces returns to scale.[24] To test this prediction, we exploit variation in import tariffs, which differentially affects input costs across sectors and over time. We construct a sector-year measure of exposure to tariff-induced changes in input prices, $\log T_{it}$, combining data from the OECD multi-country input–output tables and the Global Tariff Project by Teti (2024) (see Appendix A.6 for details). We then estimate the dynamic impact of these shocks on returns to scale using panel local projections for horizon years $h = -2, \ldots, 5$:

$$\eta_{il,t+h} - \eta_{il,t-1} = \beta_h \log T_{it} + \gamma_{lh} + \gamma_{th} + \varepsilon_{ilth}, \tag{36}$$

controlling for firm (γ_{lh}) and year (γ_{th}) fixed effects. Figure 9 plots the estimated dynamic response $\hat{\beta}_h$ to a standardized tariff shock. Consistent with the theory, firms that are more exposed to tariff-induced cost increases experience larger declines in returns to scale after the shock. A one-standard-deviation shock is associated with a decline in returns to scale of up to 0.004. This adjustment happens progressively, suggesting that changing returns to scale might take time.

Taken together, this evidence is consistent with the key premise of our model that firms adjust their returns to scale in response to both productivity and input-cost changes. This supports our interpretation of returns to scale as an endogenous choice margin rather than a fixed feature of the production function.

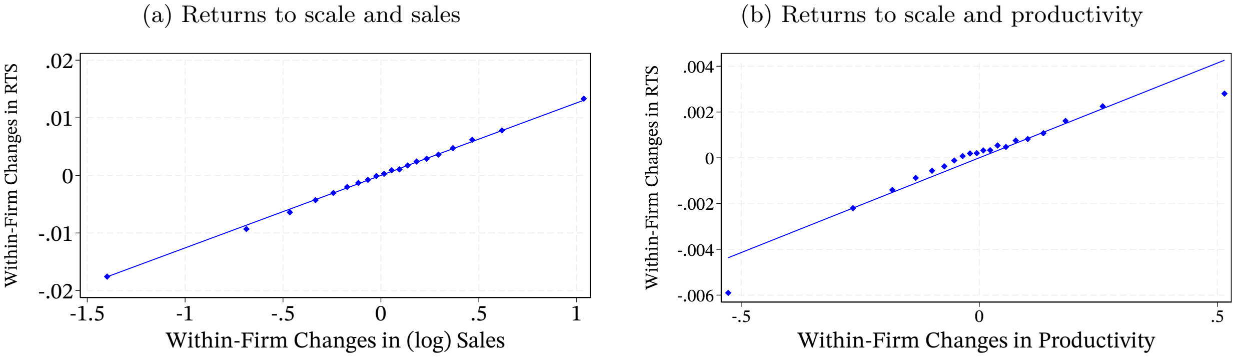
6.4 Sectoral evidence

Our model has implications for cross-industry variation. For instance, because high-productivity firms adopt more scalable technologies to leverage their productivity advantage, sectors with greater

[23] To track within-firm productivity over time, we use a within-firm Törnqvist productivity index that nets out share-weighted input growth from output growth. Full definitions and implementation details are in Appendix A.3.

[24] We formally analyze the model with a tax on intermediate inputs in Appendix D.5.

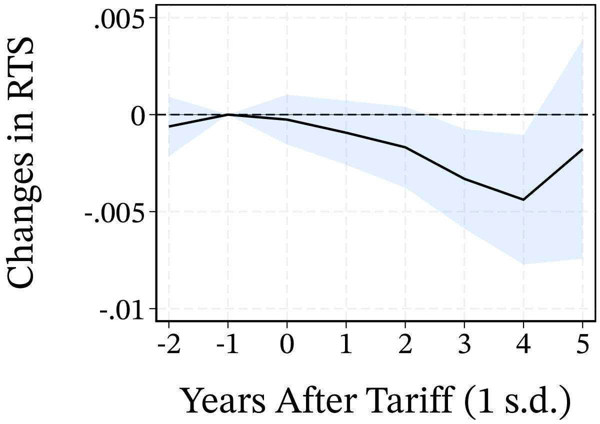
Figure 8: Within-firm variation in returns to scale, productivity, and firm size


(a) Returns to scale and sales

(b) Returns to scale and productivity

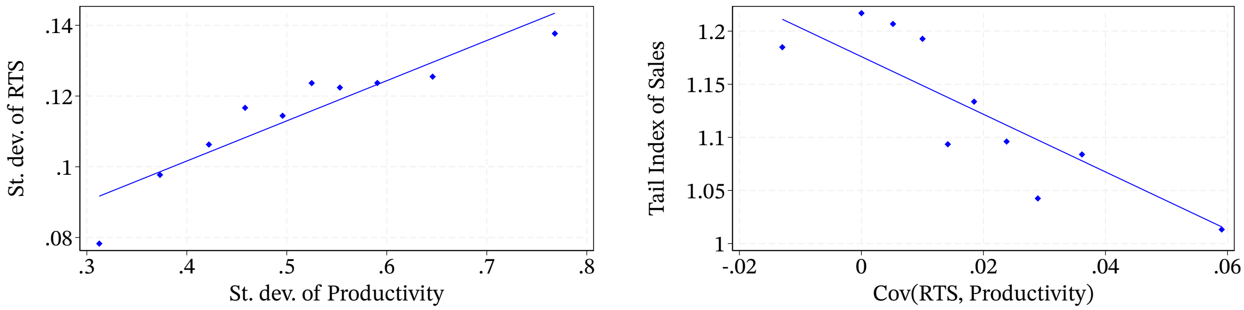
Notes: Panel (a) presents a binned scatter plot of firm-level returns to scale against log sales. Panel (b) presents a binned scatter plot of firm-level returns to scale against productivity. Firm fixed effects and sector-year fixed effects are controlled in both panels. Sectors are defined as the 62 sectors in the Input–Output table of the Annual Spanish National Accounts, approximately corresponding to NACE 2-digit industries. Both panels are constructed using a sample of Spanish firms from Orbis. See the main text for details on variable construction and sample selection.

Figure 9: The impact of a one-standard-deviation tariff shock on firm-level returns to scale



Notes: This figure reports estimation results for equation (36) using a sample of Spanish firms from Orbis. It plots the estimated dynamic response coefficients $\hat{\beta}_h$ for horizons $h = -2, \ldots, 5$ to a standardized tariff shock, normalizing the coefficient at $h = -1$ to zero. 90% confidence intervals are constructed using standard errors two-way clustered at the firm and industry-year levels. Industries are defined according to the OECD Input–Output table. See the main text for details on variable construction and sample selection.

Figure 10: Dispersion in returns to scale, productivity and market concentration


(a) Dispersion in returns to scale and productivity

(b) Tail indices of sales and endogenous scalability

Notes: Panel (a) presents a binned scatter plot of the standard deviation of firm-level returns to scale against the standard deviation of firm-level productivity, for all sector–year observations. Panel (b) presents a binned scatter plot of the tail index of firm sales against the covariance between firm-level returns to scale and productivity, for all sector–year observations with observations more than 50 firms. Year fixed effects are controlled in both panels, the unconditional mean of returns to scale is added back for interpretability. Sectors are defined as the 62 sectors in the Input–Output table of the Annual Spanish National Accounts, approximately corresponding to NACE 2-digit industries. Both panels are constructed using a sample of Spanish firms from Orbis. See Appendix A.1 for details on variable construction and sample selection.

productivity dispersion should also exhibit greater dispersion in returns to scale (Lemma 5). The first panel of Figure 10 confirms this prediction in the Spanish data. We find a strong positive relationship between these moments: sectors in the top decile of productivity dispersion have a standard deviation of returns to scale that is 0.06 higher than those in the bottom decile. This pattern supports the model's "double blessing" mechanism which gives rise to superstar firms through the adoption of high-scalability technologies.

This mechanism also implies that the shape of the firm-size distribution should vary systematically across sectors. Specifically, Proposition 1 shows that sectors where adopting higher scalability is easier (high φ_i) should have a thicker right tail of the sales distribution. Since φ_i is not directly observable, we construct a proxy for the strength of the mechanism: the within-sector covariance between returns to scale and productivity, $\text{Cov}\left(\eta_{il}, \varepsilon_{il} + a\left(\eta_{il}\right)\right)$. Theoretically, this covariance should be strictly positive when returns to scale are endogenous ($\varphi_i > 0$), but zero if returns to scale are fixed.[25] Using this measure, panel (b) of Figure 10 confirms the model prediction: sectors where firms can more easily adjust their returns to scale display significantly thicker firm-size tails.[26]

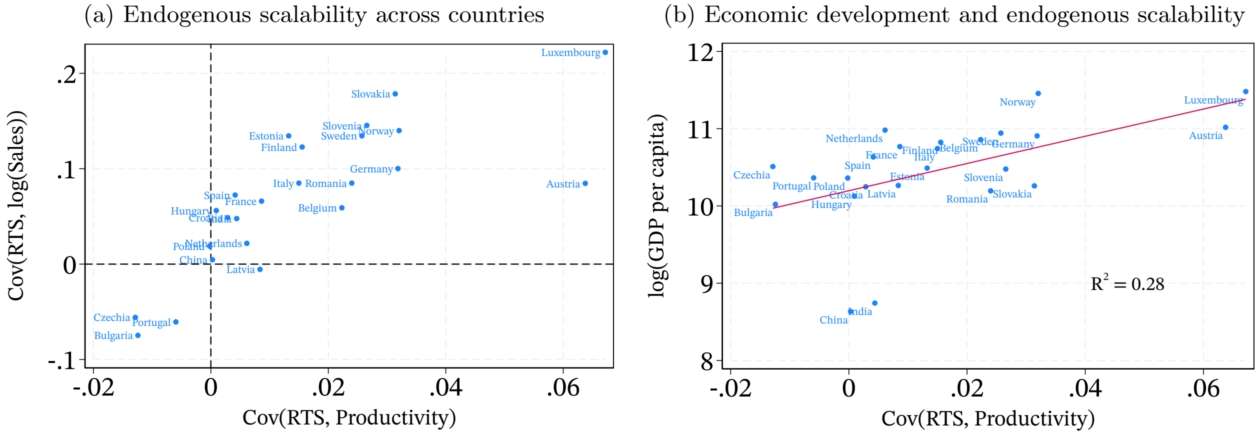
6.5 Cross-country development and endogenous scalability

We conclude this section by providing cross-country evidence for the importance of endogenous returns to scale in our sample of 24 countries. To ensure comparability, our main cross-country analysis focuses on manufacturing firms only. For each country, we select the seven-year window

[25] One can show that in the model this quantity is increasing in φ_i.

[26] We use the log-rank tail index estimator of Gabaix and Ibragimov (2011). See Appendix A.5 for details.

Figure 11: Cross-country evidence on productivity, sales, and returns to scale


(a) Endogenous scalability across countries

(b) Economic development and endogenous scalability

Notes: Panel (a) plots, for each country, the covariance between firm-level returns to scale and log sales against the covariance between firm-level returns to scale and productivity; dashed lines denote zero covariances. Panel (b) plots the seven-year-avearge of log GDP per capita against the covariance between firm-level returns to scale and productivity; the solid line reports the fitted linear relationship and the figure reports the associated R^2. Each marker corresponds to a country. See A.7 for details on variable construction and sample selection.

with the largest number of firm-year observations between 2001 and 2018.[27] Within this window, we estimate size-decile–specific production functions using the same Blundell–Bond procedure described above to construct firm-level measures of returns to scale and productivity.

Figure 11 summarizes our results. Panel (a) reports the covariance between returns to scale and log sales, as well as between returns to scale and productivity$\widehat{z}_{clt}$. Mirroring the Spanish evidence in Section 6.3.1, we find that in most countries, larger and more productive firms systematically operate higher returns-to-scale technologies. This suggests that endogenous scalability is at work in a broad set of countries.

Panel (b) relates the strength of the endogenous returns to scale mechanism to economic development. As before, we measure the intensity of endogenous scalability using the firm-level covariance between productivity and returns to scale. Plotting this measure against log GDP per capita reveals a clear relationship: countries with stronger endogenous scalability are systematically richer.[28] This is consistent with the model prediction that endogenous scalability increases the level of GDP. Quantitatively, variation in this measure explains approximately one quarter of the cross-country variation in GDP per capita. This suggests that the capacity of productive firms to adopt more scalable technologies might be an important driver of long-run economic development.

[27] For several countries, the quality of the data is uneven over the sample period. We therefore restrict our analysis to the time window with the most observations to improve the quality of our estimates. We then apply identical data cleaning and estimation procedures to all country samples.

[28] There are two outliers, India and China, but since those countries are at an earlier stage of development, their lower GDP per capita level might be explained by their distance to the technology frontier.

7 Calibration to the Spanish economy

To evaluate the quantitative importance of endogenous returns to scale decisions for the macroeconomy, we provide a basic calibration of the model to the Spanish economy. We rely on the detailed firm-level data introduced in the previous section to discipline the parameters of the model. Our calibration strategy is summarized below, with further details provided in Appendix B.

7.1 Calibration strategy

The preference parameters β can be identified directly from the sectoral consumption shares provided by the 2010 Spanish National Accounts.[29] The input-share matrix α and the firm-level returns to scale η_{il} are taken directly from our estimates of Section 6. Using firm-level returns to scale η_{il}, we compute each sector's effective returns to scale $\hat{\eta}_i$ as the sales-weighted average of these firm-level estimates. We find substantial heterogeneity in $\hat{\eta}_i$ across sectors, ranging from 0.54 to 0.98, with a mean and median of 0.83 and 0.82, respectively. Figure 16 in Appendix B.4 shows $\hat{\eta}_i$ for all sectors.

The remaining parameters (the productivity dispersion σ_i and the cost of scalability γ_i) are jointly identified by targeting within-sector moments that are informative about scalability choices. As we show in Appendix B, the model implies a mapping from the pair (σ_i, γ_i) to the cross-sectional dispersion of firm-level returns to scale η_{il} and profits Π_{il}. Accordingly, we choose σ_i and γ_i for each sector to match the interquartile range of these two variables in the data.[30] Appendix B.4 reports those calibrated values (Figure 15) and shows that the model matches the targeted moments well (Figure 14). The calibrated model also matches the empirical effective returns to scale $\hat{\eta}$ perfectly.[31]

To validate the calibration, we test the model's ability to reproduce untargeted moments related to the cross-section of firms. Recall that our mechanism implies a positive sorting where larger firms operate technologies with higher returns to scale. Panel (a) of Figure 12 compares this relationship in the model against the data. Although this moment was not targeted, the model replicates the empirical patterns well. While the fit is less precise for very small firms, these producers account for a negligible share of aggregate output and thus have little influence on the counterfactual exercises that follow. The model also generates a steeper relationship between sales and returns to scale than what we see in the data. This is likely because model firms adjust their scale instantly in response to productivity shocks, whereas real-world adjustments are subject to frictions and delays.

Panel (b) of Figure 12 shows that the model closely matches the empirical link between pro-

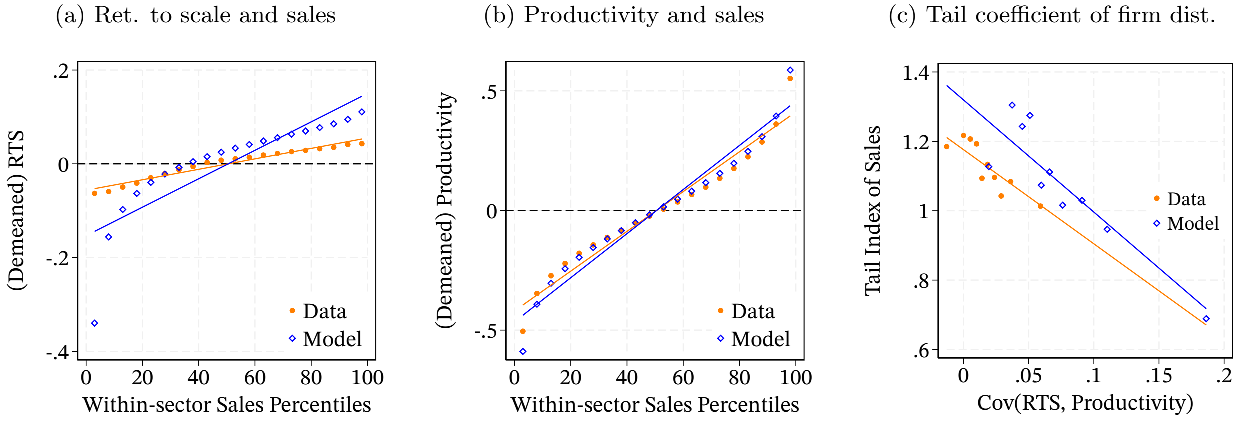
[29] We calibrate the model to the 2010 Spanish economy as it is the most recent year with detailed input-output tables that can easily be matched with the Orbis firm-level data.

[30] To be consistent with the model, we compute the profits of firm l in sector i as $(1 - \eta_{il})\, P_i Q_{il}$. We target interquartile ranges instead of variances to reduce the impact of outliers.

[31] As we study *changes* to the environment, the results presented below are independent of the average productivity μ and the entry cost κ, and so there is no need to specify their values. Given all the other parameters, we can always find a combination of μ and κ to match $\hat{\eta}$ perfectly.

ductivity (measured using the Törnqvist index) and sales. This is reassuring, as the link between these variables in the model is driven by the endogenous returns to scale parameter η_{il}.[32] Finally, panel (c) shows that the tail of the firm-size distribution is thicker is sectors where the endogenous scalability mechanisms is stronger. Once again, the model matches the data reasonably well.

Figure 12: Returns to scale, sales, productivity and tail coefficients in the model and in the data



Notes: *Data* correspond to Spanish firms in Orbis; *Model* corresponds to simulated firm-level outcomes from the calibrated model. Simulated firm observations are reweighted at the sector level so that the sector composition matches the data. Panels (a)–(b) report binscatter plots of returns to scale and productivity against sales percentiles. The y-axis variables are residualized by sector–year fixed effects in the *Data* series and sector fixed effects in the *Model* series. Panel (c) plots the tail index of sales against the within-sector-year covariance of returns to scale and productivity: the *Data* series uses sector–year statistics (computed only for sector–years with at least 50 firms) and includes year fixed effects, whereas the *Model* series uses sector-level statistics computed from simulated data and includes no fixed effects.

7.2 Contribution of endogenous returns to scale to GDP

Using our calibrated model, we first evaluate the importance of endogenous returns to scale for the level of GDP. To do so, we compare our calibrated baseline economy to the "fixed returns-to-scale" counterfactual where each firm's returns to scale is exogenously fixed at its sector's average, $\eta_{il} = \hat{\eta}_i$ (Definition 2). As shown in Proposition 5, the difference in log GDPs between those two economies is given by

$$y - \tilde{y} = \sum_{i=1}^{N} \omega_i \frac{1}{2} \underbrace{(1 - \hat{\eta}_i) \log \left(\frac{1}{1 - \varphi_i} \right)}_{\text{Sectoral flexibility}}. \tag{37}$$

Using our calibrated parameters, we find that this gain is 11.7%, so that allowing high-productivity firms to choose more scalable technologies has a substantial effect on GDP. Figure 17 in Appendix B.4 decomposes this aggregate gain, showing the contribution of each sector.

[32] Productivity and returns to scale are also positively correlated in the calibrated model, as they are in the data.

7.3 Implications for long-run GDP growth

Next, we evaluate the importance of our mechanism for long-run macroeconomic growth. We conduct an experiment in which we increase the mean productivity μ_i of all sectors by one percentage point every year. We then compare the response of GDP in our full model to two counterfactual benchmarks. The purpose of this exercise is to decompose the full effect of endogenous scalability into two components: the gain from the initial *static dispersion* in returns to scale, and the additional gain from allowing these returns to scale to *dynamically adjust* over time. Our first benchmark is the fixed returns-to-scale economy where $\eta_{il}(t) = \hat{\eta}_i(0)$ for all firms, shutting down both channels. Growth in this economy is determined by the sum of the sectoral Domar weights, so that $\tilde{y}(t+1) - \tilde{y}(t) = \sum_{i=1}^{N} \omega_i(0) \times 1\%$. Our second benchmark is a *dispersed return-to-scale* economy, where η_{il} is fixed at each firm's initial level $\eta_{il}(t) = \eta_{il}(0)$, thus featuring the initial dispersion but not the dynamic adjustment.[33]

Panel (a) of Figure 13 shows the evolution of GDP over time in each economy, relative to the fixed returns-to-scale benchmark. The initial gap at $t = 0$ reflects the level effect from (37). The solid blue line captures the full effect of endogenous returns to scale. We see that after one hundred years, GDP has grown an extra 5.5% in the full model. Panel (c) illustrates the mechanism. As μ increases, the price of intermediate inputs falls, which incentivizes firms to adopt more scalable technologies. This increase in returns to scale leads to higher Domar weights, which make the increase in productivity more impactful.[34]

Figure 13 also allows us to decompose this 5.5% gain. It shows that the dispersed returns-to-scale economy grew by an extra 2.8% over a century compared to the fixed benchmark. This implies that the static reallocation channel and the dynamic adjustment channel each contribute roughly half of the total effect. The static channel is powerful because, even when individual η_{il} cannot change, the economy-wide productivity boom disproportionately benefits the firms that can scale more easily—that is, those that started with higher η_{il} (Lemma 4). As these firms expand their output disproportionately, their sales shares increase. This compositional shift has two beneficial effects. First, it endogenously raises the aggregate $\hat{\eta}$ and Domar weights, which amplifies growth even though no firm changes its technology. Second, since these high-η firms are also the most productive, this reallocation of market share directly raises aggregate productivity.[35]

Figure 13 also shows that the evolution of log GDP over time in the full model is convex, with growth constantly accelerating (panel b). This indicates that the effects of endogenous returns to scale become more and more important over time, with the gap between the three curves in panel (a) increasing indefinitely. In the long-run, as $t \to \infty$, the baseline economy converges to a growth

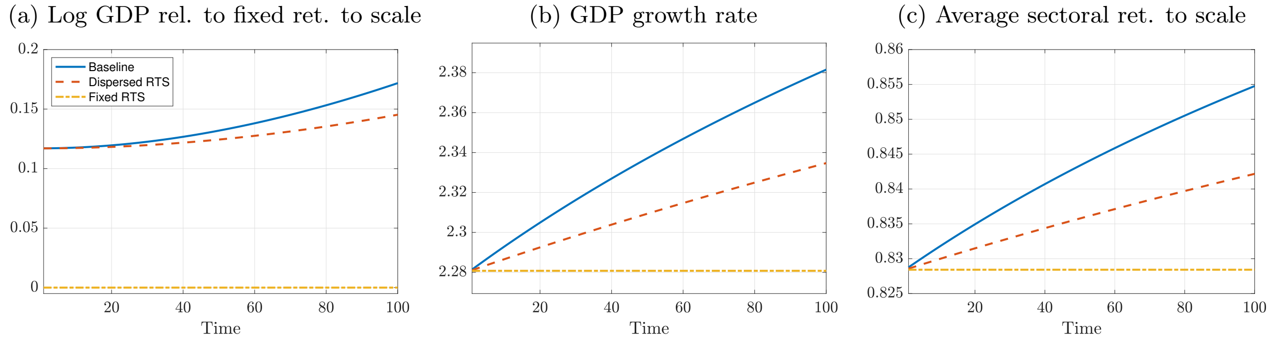
[33] We formally analyze this economy in Appendix D.7.

[34] Over one hundred years, effective returns to scale in the calibrated model increase by about 0.025, which is within the range found by empirical studies. For instance, De Loecker et al. (2020) find that firm-level returns to scale increased between 0.01 and 0.1 over periods of 40 to 60 years, depending on the estimation method.

[35] Endogenous returns to scale are still important in this economy as they create the initial dispersion in η_{il}.

rate of 3.1% under our parametrization. In contrast, in the fixed returns-to-scale economy, where the endogenous returns to scale mechanism is shut down, long-run growth is only 2.3%. The large gap between these two numbers suggests that endogenous scalability might play a significant role in shaping long-run economic outcomes.[36]

Figure 13: Endogenous returns to scale and productivity: Implications for GDP



Notes: Panel (a) shows $y(t) - \tilde{y}(t)$ ("Baseline") and $y^d(t) - \tilde{y}(t)$ ("Dispersed RTS") as productivity of all sectors μ grows at 1% per year, where y, $\tilde{y}$, and y^d are log GDPs in the economies with fully flexible returns to scale, fixed returns to scale, and dispersed returns to scale that do not respond to changes in μ, respectively. Panel (b) shows growth rates of log GDP and panel (c) shows $\frac{1}{N}\sum_{i=1}^{N}\hat{\eta}_i(t)$ as productivity of all sectors grows at 1% per year.

7.4 Distorted economy

One implication of our model is that the ability of high-productivity firms to adopt more scalable technologies is important for the level and growth rate of GDP. Yet, several studies have documented that larger firms often suffer from higher wedges (e.g., Restuccia and Rogerson, 2008). With endogenous returns to scale, those wedges would disproportionately distort the scalability decisions of those superstar firms, with potentially large consequences for welfare. In this section, we provide a simple exercise to quantify this new adverse effect of wedges.

Following Hsieh and Klenow (2009), we compute the sales wedge $\widehat{\tau}_{il}^{S}$ for firm l in sector i as the ratio of the marginal revenue product of labor to the wage:

$$\frac{1}{1-\widehat{\tau}_{il}^{S}} = \frac{\eta_{il}\left(1-\sum_{j=1}^{N}\alpha_{ij}\right)P_iY_{il}}{WL_{il}}.$$

Consistent with the literature, we find these wedges to be large. The average sales-weighted wedge in a sector, $\hat{\tau}_i^S = \int_0^{M_i}\frac{P_iQ_{il}}{P_iQ_i}\widehat{\tau}_{il}^{S}dl$, is 0.40 on average across all sectors. Furthermore, wedges are

[36] As $t \to \infty$, the *growth rates* in the baseline model and the dispersed returns-to-scale economy converge. In that limit, output in both economies is dominated by a vanishingly small fraction of extremely productive firms operating with near-constant returns to scale. In that case, the gains from the *dynamic channel* of endogenous returns to scale are exhausted, and both economies grow at the same pace of 3.1%. This convergence, however, is extremely slow. The half-life of the transition—the time required for the growth rate to close half the gap to its long-run limit—is approximately one thousand years in the baseline model and two and a half thousand years in the dispersed returns-to-scale economy.

positively correlated with firm size: the average within-sector correlation between τ_{il}^S and firm sales is 0.27.

To study the impact of these wedges in our calibrated economy, we assume that each firm is subject to a sales wedge

$$\log\left(1-\tau_{il}^S\right)=\log\left(1-\tau_i^S\right)-b_i\left(\varepsilon_{il}-\mu_i\right), \tag{38}$$

where $\log\left(1-\tau_i^S\right)$ is an average sector-wide term and $-b_i\left(\varepsilon_{il}-\mu_i\right)$ captures size-dependent distortions. With $b_i>0$, more productive firms face larger distortions. As in the case where the wedge corresponds to a tax or to a markup, we assume that net revenues collected from τ_{il} are rebated to the household lump-sum. We calibrate the intercept τ_i^S and the slope b_i to match the average sectoral wedge and the covariance between firm profits and the estimated wedges in the data. After fixing τ_i^S and b_i, we recalibrate the remaining model parameters so that the economy continues to match our empirical targets. Further details on this procedure are provided in Appendix B.3.

To evaluate the importance of these distortions, we conduct an experiment in which we remove all wedges.[37] The first two columns of Table 1 shows that, in the baseline model, this leads to a large increase in both returns to scale and in GDP. To understanding the mechanisms behind this result, the table also reports the outcome of the same experiment in the dispersed and the fixed returns-to-scale economies. In the dispersed return-to-scale economy, where firms choose their returns to scale in the presence of wedges but cannot adjust them once wedges are removed, the impact of eliminating wedges is substantially smaller. In this case, firms at the top of the distribution cannot increase their scalability to fully benefit from the removal of distortions, which limits welfare gains. In the fixed return-to-scale economy, where the endogenous scalability mechanism is shut down entirely, the effect of wedges is even more muted. Because all firms operate with the same returns to scale, the firm-size distribution is more compressed and GDP is produced more evenly across firms. Distorting high-productivity producers is thus less damaging in that economy.

To further show that wedges that disproportionately affect the top firms are particularly harmful when those firms endogenously choose the scalability of their operation, the last two columns of Table 1 repeat the analysis for an economy with flat wedges ($b_i=0$). In this setting, productive firms face no additional penalty relative to smaller firms, so removing distortions yields much smaller GDP gains. Moreover, the results are nearly identical across the three model specifications. This confirms that the interaction between endogenous scalability and size-dependent distortions is the primary driver of our results.

In summary, the quantitative analysis in this section suggests that endogenous returns to scale might play a substantial role in shaping both the level and the growth rate of GDP. When highly

[37] For some sectors, removing wedges would push φ_i above one. For these sectors, we set $\varphi_i=0.99$. We conduct sensitivity analysis to the value of this threshold in Appendix B.5.

productive firms are able to adopt more scalable technologies as the technological frontier advances, they can expand disproportionately, with substantial gains for welfare and GDP growth. Taxes or distortions that fall on those firms, however, can disrupt this process. As our results show, size-dependent distortions discourage the adoption of high-scale technologies, stifle the growth of superstar firms, and generate efficiency losses that can exceed those in standard models. Policy interventions that disproportionately burden high-productivity firms may therefore be particularly harmful for welfare.

Table 1: Returns to scale and GDP when wedges are removed

	Size-dependent wedges		Flat wedges	
	Δ Ret. to scale	Δ GDP	Δ Ret. to scale	Δ GDP
Baseline economy	0.067	167%	0.020	62%
Dispersed ret. to scale	0.046	138%	0.010	60%
Fixed ret. to scale	0	70%	0	58%

Notes: Increases in average effective returns to scale, $\Delta\left[\sum_{i=1}^{N}\hat{\eta}_i/N\right]$, and in log GDP, Δy, due to removal of sales wedges in the baseline economy, and in the economies with fixed and dispersed returns to scale. The "Size-dependent wedges" column reports the results when sales taxes are correlated with firm productivity. The "Flat wedges" column reports the results when sales wedges are identical for all firms within a sector.

8 Conclusion

This paper develops a theory in which returns to scale are endogenous equilibrium objects driven by incentives. At the micro level, this mechanism gives rise to superstar firms and fat-tailed firm-size distributions. At the macro level, it endows the economy with greater resilience by dampening the impact of adverse shocks while amplifying favorable ones. It also provides an engine for long-run growth, as the economy's organizational structure co-evolves with its technological frontier. Input-output connections between sector play a crucial role for these mechanisms.

Several extensions would be worth pursuing. First, introducing capital would change the incentives to increase returns to scale. Since capital can be accumulated, its presence might affect the long-run growth properties of the model. Second, the superstar firms that emerge in our model might, in reality, gain market power, creating a feedback effect that would further increase their incentives to scale. Third, while we have modeled the choice of scalability as a single, abstract decision, future work could explicitly model the distinct margins through which firms achieve scale, such as building distribution networks, creating hierarchical organizations, and many others. Modeling these individual margins would allow for a more precise contact with the data and a deeper understanding of forces shaping returns to scale.

References

Acemoglu, D. and P. D. Azar (2020): "Endogenous Production Networks," *Econometrica*, 88, 33–82.

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012): "The Network Origins of Aggregate Fluctuations," *Econometrica*, 80, 1977–2016.

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification properties of recent production function estimators," *Econometrica*, 83, 2411–2451.

Argente, D., S. Moreira, E. Oberfield, and V. Venkateswaran (2025): "Scalable Expertise: How Standardization Drives Scale and Scope," Tech. rep., National Bureau of Economic Research.

Aw, B. Y., X. Chen, and M. J. Roberts (2001): "Firm-level evidence on productivity differentials and turnover in Taiwanese manufacturing," *Journal of development economics*, 66, 51–86.

Axtell, R. L. (2001): "Zipf distribution of US firm sizes," *science*, 293, 1818–1820.

Baily, M. N., C. Hulten, and D. Campbell (1992): "Productivity dynamics in manufacturing plants," *Brookings papers on economic activity. Microeconomics*, 1992, 187–267.

Baqaee, D. R. and E. Farhi (2019a): "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem," *Econometrica*, 87, 1155–1203.

——— (2019b): "Productivity and Misallocation in General Equilibrium," *Quarterly Journal of Economics*, 135, 105–163.

Basu, S. and J. G. Fernald (1997): "Returns to scale in US production: Estimates and implications," *Journal of political economy*, 105, 249–283.

Bigio, S. and J. La'O (2020): "Distortions in Production Networks," *Quarterly Journal of Economics*, 135, 2187–2253.

Blundell, R. and S. Bond (2000): "GMM estimation with persistent panel data: an application to production functions," *Econometric reviews*, 19, 321–340.

Burnside, C., M. Eichenbaum, and S. Rebelo (1995): "Capital utilization and returns to scale," *NBER macroeconomics annual*, 10, 67–110.

Caves, D. W., L. R. Christensen, and W. E. Diewert (1982a): "Multilateral comparisons of output, input, and productivity using superlative index numbers," *Economic Journal*, 92, 73–86.

——— (1982b): "The economic theory of index numbers and the measurement of input, output, and productivity," *Econometrica*, 1393–1414.

Chandler, A. D. (1977): *The Visible Hand: The Managerial Revolution in American Business*, Harvard University Press.

——— (1990): *Scale and scope: The dynamics of industrial capitalism*, Harvard University Press.

Chiavari, A. (2024): "Customer accumulation, returns to scale, and secular trends," .

Chiavari, A. and S. S. Goraya (2025): "The rise of intangible capital and the macroeconomic implications," .

De Loecker, J., J. Eeckhout, and G. Unger (2020): "The rise of market power and the macroeconomic implications," *Quarterly Journal of Economics*, 135, 561–644.

De Ridder, M., B. Grassi, G. Morzenti, et al. (2022): "The Hitchhiker's Guide to Markup Estimation," .

Demirer, M. (2025): "Production function estimation with factor-augmenting technology: An application to markups," Working paper.

Engbom, N., H. Malmberg, T. Porzio, F. Rossi, and T. Schoellman (2025): "Economic development according to chandler," Tech. rep., Tech. rep., mimeo, New York University.

Feenstra, R. C., R. Inklaar, and M. P. Timmer (2015): "The next generation of the Penn World Table," *American economic review*, 105, 3150–3182.

Gabaix, X. and R. Ibragimov (2011): "Rank- 1/2: a simple way to improve the OLS estimation of tail exponents," *Journal of Business & Economic Statistics*, 29, 24–39.

Gale, D. and H. Nikaido (1965): "The Jacobian matrix and global univalence of mappings," *Mathematische Annalen*, 159, 81–93.

Gao, W. and M. Kehrig (2025): "Returns to scale, productivity and competition: Empirical evidence from US manufacturing and construction establishments," Working paper.

Garicano, L. (2000): "Hierarchies and the Organization of Knowledge in Production," *Journal of political economy*, 108, 874–904.

Garicano, L. and E. Rossi-Hansberg (2006): "Organization and inequality in a knowledge economy," *The Quarterly journal of economics*, 121, 1383–1435.

Gottlieb, C., M. Poschke, and M. Tueting (2025): "Skill Supply, Firm Size, and Economic Development," *Background paper prepared for World Development Report.*

Griliches, Z. and H. Regev (1995): "Firm productivity in Israeli industry 1979–1988," *Journal of econometrics*, 65, 175–203.

Hall, R. E. (1990): "Invariance properties of Solow's productivity residual," *Growth, Productivity, Unemployment: Essays to Celebrate Bob Solow's Birthday*, 71.

Hopenhayn, H. A. (1992): "Entry, exit, and firm dynamics in long run equilibrium," *Econometrica: Journal of the Econometric Society*, 1127–1150.

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and manufacturing TFP in China and India," *Quarterly Journal of Economics*, 124, 1403–1448.

Hubmer, J., M. Chan, S. Ozkan, S. Salgado, and G. Hong (2025): "Scalable versus Productive Technologies," Tech. rep., Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Hulten, C. R. (1978): "Growth Accounting with Intermediate Inputs," *Review of Economic Studies*, 45, 511–518.

Jones, C. I. (2011): "Intermediate Goods and Weak Links in the Theory of Economic Development," *American Economic Journal: Macroeconomics*, 3, 1–28.

Kalemli-Özcan, Ş., B. E. Sørensen, C. Villegas-Sanchez, V. Volosovych, and S. Yeşiltaş (2024): "How to Construct Nationally Representative Firm-Level Data from the Orbis Global Database: New Facts on SMEs and Aggregate Implications for Industry Concentration," *American Economic Journal: Macroeconomics*, 16, 353–374.

Kopytov, A., B. Mishra, K. Nimark, and M. Taschereau-Dumouchel (2024a): "Endogenous production networks under supply chain uncertainty," *Econometrica*, 92, 1621–1659.

Kopytov, A., M. Taschereau-Dumouchel, and Z. Xu (2024b): "The Origin of Risk," *Available at SSRN*.

Kuznets, S. (1973): "Modern economic growth: findings and reflections," *The American economic review*, 63, 247–258.

Lashkari, D., A. Bauer, and J. Boussard (2024): "Information technology and returns to scale," *American Economic Review*, 114, 1769–1815.

Levinsohn, J. and A. Petrin (2003): "Estimating production functions using inputs to control for unobservables," *The review of economic studies*, 70, 317–341.

Liu, E. (2019): "Industrial Policies in Production Networks," *Quarterly Journal of Economics*, 134, 1883–1948.

Long, J. B. and C. I. Plosser (1983): "Real Business Cycles," *Journal of Political Economy*, 91, 39–69.

Lucas, R. E. (1978): "On the size distribution of business firms," *The Bell Journal of Economics*, 508–523.

McAdam, P., P. Meinen, C. Papageorgiou, and P. Schulte (2024): "Returns to scale: New evidence from administrative firm-level data," Tech. Rep. 24/2024.

McKenzie, L. W. (1959): "On the existence of general equilibrium for a competitive market," *Econometrica: journal of the Econometric Society*, 54–71.

Oberfield, E. (2018): "A Theory of Input-Output Architecture," *Econometrica*, 86, 559–589.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Restuccia, D. and R. Rogerson (2008): "Policy distortions and aggregate productivity with heterogeneous establishments," *Review of Economic dynamics*, 11, 707–720.

Ruzic, D. and S.-J. Ho (2023): "Returns to scale, productivity, measurement, and trends in US Manufacturing Misallocation," *Review of Economics and Statistics*, 105, 1287–1303.

Savagar, A. and J. Kariel (2024): "Scale Economies and Aggregate Productivity," *Available at SSRN 4776572*.

Smirnyagin, V. (2023): "Returns to scale, firm entry, and the business cycle," *Journal of Monetary Economics*, 134, 118–134.

Söderbom, M. (2009): "Imposing common factor restrictions for AR (1) model" ex post" using a minimum distance procedure," *stata command, Downloaded from: http://www. soderbom. net/comfac_md. zip.*

Star, S. and R. E. Hall (1976): "An approximate Divisia index of total factor productivity," *Econometrica: Journal of the Econometric Society*, 257–263.

Teti, F. A. (2024): "Missing Tariffs," Tech. Rep. 11590, CESifo Working Papers, feodora Teti's Global Tariff Database (v_beta1-2024-12).

Online Appendix

A Appendix for Section 6

This appendix contains details about the reduced-form results of Section 6.

A.1 Details of the Spanish Orbis data

Our Spanish firm-level data are drawn from the Orbis Historical Disk. Orbis is commonly regarded as the most comprehensive cross-country firm database, covering both public and private firms' financial statements and measures of real activity (Kalemli-Özcan et al., 2024). We focus on Spain because firm coverage is close to universal—capturing over 95% of total industry gross output after 2010—making it well suited for economy-wide analysis. Our sample spans 1995–2019.[38]

Sample cleaning Our sample construction closely mirrors the cleaning steps used in our earlier work (Kopytov et al., 2024b). We begin by merging each firm's descriptive information with its financial statements using the unique BVD firm identifier (`BVDID`). We then restrict our analysis to Spanish firms, defined as firms that satisfy two criteria: 1) their latest address is in Spain and 2) their `BVDID` starts with the ISO-2 code `ES`. In the resulting Orbis Spain sample, we implement the following standard cleaning procedure:

1. We harmonize the calendar year of each firm-year observation using the variable `closing_date`: if the closing date is after or on July 1, the current year is assigned as the calendar year. Otherwise, the previous year is assigned.[39]

2. In a given year, a firm in the Orbis database might have multiple financial statements from different sources (local registry, annual report, or others), for consolidated or unconsolidated accounts. When several source-consolidation combinations exist for a firm, we deduplicate by selecting the account that, in order of priority, 1) shows the most consistent reporting frequency (closest to regular annual reporting), 2) offers the longest non-missing time series for key financial variables (fixed assets and/or sales), and 3) is consolidated, if the first two criteria are tied.

3. We only keep firms with non-missing and positive sales (`operating_revenue_turnover`), fixed assets (`fixed_assets`), wage bills (`costs_of_employees`), and material costs

[38] Orbis offers good coverage of the Spanish economy starting from 1995. Moreover, the most recent observations in the version of Orbis Historical Disk Product that we use are from 2021. We therefore use 2019 as the last year of the sample since there is usually a two-year reporting lag for some variables (see Kalemli-Özcan et al. (2024) for details).

[39] This adjustment matters little for the Spanish sample, as 99% of firms close their books on December 31.

(material_costs). We also harmonize the units of all monetary values to be in current euros.

4. To prevent outliers from affecting the production function estimation, we exclude any firm-year observation whose average revenue product for any input (fixed assets, wage bills, or material costs) lies above the 99th percentile or below the 1st percentile of that year's distribution.

A.2 Details of the production function and RTS estimation

This appendix describes in detail how we implement the production-function estimation procedure that delivers the results used in the main text.

We use the Blundell and Bond (2000) IV-GMM estimator to estimate the production functions as our benchmark. This estimator is designed for dynamic panel settings with persistent firm-level variables and, under standard moment conditions, delivers consistent estimates of output elasticities. Our model imposes a competitive output market in a sector. In this setting, the identifying assumptions are most plausible when there is sufficient persistent variations in predetermined inputs and in the cost of flexible inputs, so that observed input choices are not collinear with unobserved productivity. Recent work by De Ridder et al. (2022) further shows in Monte Carlo simulations that this approach performs well when such identifying variation is strong.

Our empirical strategy builds on the model's implication that, within a sector, firms that are similar in size should operate under similar production technologies and therefore exhibit similar returns to scale. We use this prediction to estimate returns to scale across the firm-size distribution. For each sector i and year t, we rank firms by a smoothed measure of size: the 7-year moving average of firm-level log sales computed over the window $t-3$ to $t+3$. We then assign firms to 10 deciles, $d_t = \{1, ..., 10\}$, based on this sector-year ranking.[40] Using a moving average reduces the influence of short-run fluctuations and measurement error in annual sales, and thus provides a more stable proxy for the scale of a firm's production. For each sector-decile-year cell (i, d_t, t) (using the 7-year rolling sample around t), we assume firms share a common Cobb-Douglas technology:

[40] We reassign a few small sectors with few firms to closely related sectors that produce similar goods or services, for the purpose of production function estimation. Specifically, (i) we merge sectors 5, 6, 7, and 9—Manufacture of food products, beverages and tobacco products; Manufacture of textiles, wearing apparel and leather products; Manufacture of wood and of products of wood and cork (except furniture); manufacture of articles of straw and plaiting materials; and Printing and reproduction of recorded media—into sector 8 (Manufacture of paper and paper products). (ii) We merge sector 12—Manufacture of basic pharmaceutical products and pharmaceutical preparations—into sector 11 (Manufacture of chemicals and chemical products). (iii) We merge sector 20—Manufacture of motor vehicles, trailers and semi-trailers—into sector 19 (Manufacture of machinery and equipment n.e.c.).

$$
\begin{aligned}
q_{ilt} &= \beta^{L}_{i,d_t(l),t}\, l_{ilt} + \beta^{K}_{i,d_t(l),t}\, k_{ilt} + \beta^{M}_{i,d_t(l),t}\, m_{ilt} + \gamma_t^{i,d_t(l)} + \kappa_{il}^{i,d_t(l)} + a_{ilt}^{i,d_t(l)},\\
a_{ilt}^{i,d_t(l)} &= \rho^{i,d_t(l)}\, a_{il,t-1}^{i,d_t(l)} + e_{ilt}^{i,d_t(l)}, \qquad |\rho^{i,d_t(l)}| < 1,\\
e_{ilt}^{i,d_t(l)} &\sim MA(0),
\end{aligned}
$$

where q_{ilt}, l_{ilt}, k_{ilt}, m_{ilt} are the log of output, labor, capital, material input for firm l. These are measured as deflated vales of, respectively, sales, wage bills, fixed asset and material costs using the GDP deflator in the Annual Spanish National Accounts. We assume a firm's productivity contains three components: a year-specific component $\gamma_t^{i,d_t(l)}$, a firm-specific effect $\kappa_{il}^{i,d_t(l)}$ and an autoregressive component $a_{ilt}^{i,d_t(l)}$with i.i.d. innovation $e_{ilt}^{i,d_t(l)}$. The model admits the following dynamic representation:

$$
\begin{aligned}
q_{ilt} = \rho^{i,d_t(l)} q_{il,t-1} &+ \beta^{L}_{i,d_t(l),t}\, l_{ilt} - \rho^{i,d_t(l)}\beta^{L}_{i,d_t(l),t}\, l_{il,t-1} + \beta^{K}_{i,d_t(l),t}\, k_{ilt} - \rho^{i,d_t(l)}\beta^{K}_{i,d_t(l),t}\, k_{il,t-1}\\
&+ \beta^{M}_{i,d_t(l),t}\, m_{ilt} - \rho^{i,d_t(l)}\beta^{M}_{i,d_t(l),t}\, m_{il,t-1} + \gamma_t^{*\,i,d_t(l)} + \kappa_{il}^{*\,i,d_t(l)} + e_{ilt}^{i,d_t(l)},
\end{aligned}
\tag{39}
$$

where $\gamma_t^{*\,i,d_t(l)} \equiv (1-\rho^{i,d_t(l)})\gamma_{t-1}^{i,d_t(l)}$ and $\kappa_{il}^{*\,i,d_t(l)} \equiv (1-\rho^{i,d_t(l)})\kappa_{il}^{i,d_t(l)}$. Therefore, we can estimate the following dynamic specification with current and lagged variables:

$$
\begin{aligned}
q_{ilt} = \alpha^{i,d_t(l)}\, q_{il,t-1} &+ \beta^{L0}_{i,d_t(l),t}\, l_{ilt} + \beta^{L1}_{i,d_t(l),t}\, l_{il,t-1} + \beta^{K0}_{i,d_t(l),t}\, k_{ilt} + \beta^{K1}_{i,d_t(l),t}\, k_{il,t-1}\\
&+ \beta^{M0}_{i,d_t(l),t}\, m_{ilt} + \beta^{M1}_{i,d_t(l),t}\, m_{il,t-1} + \gamma_t^{*\,i,d_t(l)} + \kappa_{il}^{*\,i,d_t(l)} + e_{ilt}^{i,d_t(l)},
\end{aligned}
\tag{40}
$$

where, under our assumption, the AR(1) productivity structure implies restrictions across coefficients (e.g., $\alpha^{i,d_t(l)} = \rho^{i,d_t(l)}$ and $\beta^{x1}_{i,d_t(l),t} = -\rho^{i,d_t(l)}\beta^{x0}_{i,d_t(l),t}$ for $x \in \{L,K,M\}$).

Blundell-Bond system-GMM moments We now describe the system-GMM moment conditions we exploited to estimate the model in (40). Our choice of moment conditions follows the exact implementation in Table III column 5 of Blundell and Bond (2000), where we treat $\{q,l,k,m\}$ as potentially endogenous and use two sets of moments:

(i) Difference equation (levels dated $t-2$ and earlier):

$$
\mathbb{E}\Big[x_{il,t-s}\ \Delta e_{ilt}^{i,d_t(l)}\Big] = 0 \quad \text{for } x \in \{q,l,k,m\} \text{ and } s \geq 2. \tag{41}
$$

(ii) Levels equation (first differences dated $t-1$ only):

$$\mathbb{E}\Big[\Delta x_{il,t-1}\left(\kappa_{il}^{*i,d_t(l)}+e_{ilt}^{i,d_t(l)}\right)\Big]=0 \quad \text{for } x\in\{q,l,k,m\}\,. \tag{42}$$

Moreover, year dummies are included as controls and treated as exogenous instruments in the levels equation. We implement this estimation using the `xtabond2` command in Stata.

Obtaining the Firm-level Returns-to-Scale Estimates After estimating (40), we then use the minimum distance estimator by Söderbom (2009) to impose the AR(1)-implied restrictions and get the restricted parameter estimates $\left(\hat{\rho}^{i,d_t(l)},\hat{\beta}^K_{i,d_t(l),t},\hat{\beta}^L_{i,d_t(l),t},\hat{\beta}^M_{i,d_t(l),t}\right)$. The estimated returns to scale η_{ilt} for a firm l in sector i and year t is therefore given by the sum of these elasticities:

$$\eta_{ilt}=\hat{\beta}^K_{i,d_t(l),t}+\hat{\beta}^L_{i,d_t(l),t}+\hat{\beta}^M_{i,d_t(l),t}.$$

Because 1995 and 1996 contain relatively few firm-year observations for production-function estimation, we report results using only the 1997-2019 estimates matched to the firm-level data for our analysis.

A.3 Constructing the Törnqvist productivity index

To compare productivity across firms, we rely on the Törnqvist productivity index. We provide here theoretical results about that index to link our model with our estimation procedure.

Lemma 3 shows that returns to scale are increasing in ε_{il}, but we do not observe ε_{il} directly in the data. In addition, comparing measured productivity $e^{\varepsilon_{il}}A_i\left(\eta_{il}\right)\zeta\left(\eta_{il}\right)$ across firms with different technologies faces well-known issues about the choice of units. When going to the data, we rely instead on a Törnqvist productivity index, which is commonly used to compare productivities across firms or countries with different production functions (Caves et al., 1982a,b) and recently in Penn World Table by Feenstra et al. (2015). Specifically, we use the multilateral Tornqvist productivity index by Caves et al. (1982a) that has been extensively used in the firm dynamics context (Baily et al. 1992, Griliches and Regev, 1995 and Aw et al., 2001).

Definition 3 (Multilateral Törnqvist productivity index)**.** Consider a sector i in year t. Let N_{it} be the number of firms observed in (i,t). Define the sector-year reference firm's moments as $\overline{\log Q_{it}}=\frac{1}{N_{it}}\sum_l \log Q_{ilt}$, $\overline{\log O_{it}}=\frac{1}{N_{it}}\sum_l \log O_{ilt}$, $\overline{\beta_{O,it}}=\frac{1}{N_{it}}\sum_l \beta_{O,ilt}$ where $O\in\{K,L,M\}$ and $\beta_{O,ilt}$ are firm-level output elasticity of input O. The multilateral Törnqvist productivity index of firm l is defined as:

$$z_{ilt}:=\left(\log Q_{ilt}-\overline{\log Q_{it}}\right)-\sum_{O\in\{K,L,M\}}\frac{1}{2}\left(\beta_{O,ilt}+\overline{\beta_{O,it}}\right)\left(\log O_{ilt}-\overline{\log O_{it}}\right).$$

For any two firms k and l in sector i and year t, we say firm k is more productive than firm l if $z_{ikt} > z_{ilt}$.

Intuitively, the measure z_{ilt} compares productivity between firm l and the reference firm by looking at how much more output one produces relative to the other, adjusting for differences in technology and input use. It is "multilateral" because z_{ilt} is defined relative to a common sector-year reference firm constructed from *all firms* in (i, t), so productivity comparisons $z_{ikt} - z_{ilt}$ are base-firm invariant and can be consistently ranked across all firm pairs. In our benchmark case, we set all $\beta_{O,ilt} = \hat{\beta}^{O}_{i,d_t(l),t}$ to obtain the estimated productivity index $\hat{z}_{ilt}$ and use it as our measured productivity in all cross-sectional exercises that involve comparisons between firms within a sector-year. We find the Törnqvist index to be a good proxy for productivity in model-simulated data. In our calibrated economy of Section 7, the within-sector correlation between the Törnqvist index and ε_{il} is above 0.99. The same number for $\varepsilon_{il} + a_i(\eta_{il})$ is about 0.92.

However, when analyzing within-firm productivity changes over time, we use a chained (within-firm) Törnqvist productivity index—i.e., an approximate Divisia index—following the implementation in Star and Hall (1976). Specifically, to account for the fact that firms may simultaneously adjust both their technology (and hence elasticities) and their input mix, we define

$$\Delta \hat{z}^{\text{within}}_{ilt} = \Delta \log Q_{ilt} - \sum_{O \in \{K,L,M\}} \overline{\beta_{O,ilt}} \Delta \log O_{ilt}, \quad \text{where } \overline{\beta_{O,ilt}} \equiv \frac{1}{2}\left(\hat{\beta}^{O}_{i,d_t(l),t} + \hat{\beta}^{O}_{i,d_{t-1}(l),t-1}\right).$$

We then normalize each firm's initial (log) within-firm productivity to zero and construct the level index $\hat{z}^{\text{within}}_{ilt}$ by accumulating changes over time, i.e.,

$$\hat{z}^{\text{within}}_{il,t} = \hat{z}^{\text{within}}_{il,t-1} + \Delta \hat{z}^{\text{within}}_{ilt}, \quad \hat{z}^{\text{within}}_{il,t_0} = 0.$$

This normalization is innocuous because our within-firm analysis in 6.3.2 include firm fixed effects, so only productivity changes (not the level) are identified.

A.4 Robustness of the production function and RTS estimation

This section shows that the documented positive RTS-size and RTS-productivity relationship, both across firms in the cross-section (Section 6.3.1) and within a firm over time (Section 6.3.2), are not driven by a particular estimator, IV-GMM instruments choice, or grouping design. We first vary the Blundell and Bond (2000) system-GMM specification by changing the treatment of year dummies and the internal instrument set, following the implementation in De Ridder et al. (2022) (Appendix A.4.1). We then re-estimate production functions using standard control-function approaches–Olley and Pakes (1996) and Levinsohn and Petrin (2003)–to verify that our results are not specific to IV-GMM (Appendix A.4.2). Moreover, we account for potential market power by adding markup controls (proxied by sales shares) within an Ackerberg et al. (2015) estimator

(Appendix A.4.3). Finally, we show that our conclusions are robust to alternative ways of forming size groups (Appendix A.4.4).

We report coefficients from simple regressions to summarize robustness of our empirical findings both across firms and within firms with these alternative estimates of returns to scale and productivity. To show robustness for 7, which documents the cross-sectional pattern that larger and more productive firms have higher returns to scale within a sector-year, we estimate two simple regressions of returns to scale on log sales and productivity:

$$\eta_{ilt} = \beta_0 \log \left(\text{Sales}_{ilt}\right) + \delta_{it} + \epsilon_{ilt}, \qquad \eta_{ilt} = \beta_1 \hat{z}_{ilt} + \delta_{it} + \epsilon_{ilt}, \tag{43}$$

where δ_{it} denotes sector-year fixed effects. The estimated coefficients of β_0 and β_1 are displayed in 2 across all specifications.

Similarly, to show robustness for 8 and document our within-firm pattern that firms have higher returns to scale when they grow larger or become more productive, we estimate:

$$\eta_{ilt} = \gamma_0 \log \left(\text{Sales}_{ilt}\right) + \kappa_{il} + \delta_{it} + \epsilon_{ilt}, \qquad \eta_{ilt} = \gamma_1 \hat{z}_{ilt}^{\text{within}} + \kappa_{il} + \delta_{it} + \epsilon_{ilt}, \tag{44}$$

where κ_{il} denotes firm fixed effects, so identification comes from within-firm variation over time. In the productivity specification, we use a within-firm Törnqvist productivity index, $\hat{z}_{ilt}^{\text{within}}$, which is appropriate for within-firm comparisons. The estimated coefficients of γ_0 and γ_1 are presented in 3 across all specifications.

Results using our benchmark estimator are reported in column (1) of Tables 2 and 3. We now describe the alternative estimators and grouping designs used in the robustness checks.

A.4.1 With alternative Blundell-Bond specifications

Our baseline specification follows Blundell and Bond (2000) and includes year dummies. Including year effects is recommended in dynamic-panel GMM applications because it absorbs economy-wide shocks and thereby reduces cross-firm correlation in the regression residuals. At the same time, once common year shocks are removed, identification of flexible-input elasticities relies on variation that is not common across firms in a group. In practice, this shifts weight toward persistent within-year differences in flexible input costs or wedges across firms. If such variation is interpreted literally as firm-specific input prices, it can raise concerns about measurement—because input quantities constructed from expenditures may mechanically inherit noise from unobserved firm-level prices. [41]

[41] However. if the relevant heterogeneity operates through non-monetary wedges—e.g., distortions that affect effective input costs without changing the recorded unit prices paid by the firm, in the spirit of Hsieh and Klenow

Table 2: Across-firm variation in returns to scale, productivity, and firm size with alternative production function estimators

	Dependent variable: Firm-level RTS						
	(1) BB baseline	(2) BB alternative	(3) OP	(4) LP	(5) ACF market power	(6) Av.-size percentiles	(7) Cur.-size deciles
$\log(\text{Sales}_{ilt})$	0.023***	0.028***	0.045***	0.037***	0.050***	0.019***	0.007***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
Sector-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952
R^2	0.688	0.688	0.642	0.806	0.564	0.728	0.655
$\hat{z}_{ilt}$	0.050***	0.058***	0.083***	0.080***	0.091***	0.034***	0.021***
	(0.002)	(0.003)	(0.002)	(0.003)	(0.005)	(0.002)	(0.002)
Sector-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952
R^2	0.655	0.648	0.567	0.760	0.507	0.684	0.652

Notes: This table reports coefficients from cross-sectional regressions of firm-level returns to scale (RTS) on (i) log sales and (ii) firm productivity ($\hat{z}_{ilt}$), each including sector-year fixed effects. Column (1) uses our benchmark Blundell-Bond (BB) estimates; column (2) uses an alternative BB specification as implemented in De Ridder et al. (2022). Columns (3) and (4) use the Olley-Pakes (OP) and Levinsohn-Petrin (LP) control-function estimators, respectively. Column (5) reports results from the Ackerberg-Caves-Frazer (ACF) estimator with market power controls (proxied by firms' sales shares). Columns (6) and (7) use alternative grouping methods for estimating elasticities: rolling average-size percentiles and contemporaneous size deciles. The regressions use a sample of Spanish firms from Orbis. See Appendix A.1 for details on variable construction and sample selection. Standard errors (in parentheses) are two-way clustered at the firm and sector-year level. *,** ,*** indicate significance at the 10%, 5%, and 1% levels, respectively.

As a robustness check, we therefore also implement the Blundell–Bond estimator specification used by De Ridder et al. (2022), which omits year dummies and uses a more conservative internal-instrument set. Concretely, our baseline estimates a dynamic sales equation with current and one-lag terms for labor, capital, and materials, includes year fixed effects, and instruments the endogenous variables with lags starting at $t-2$ (and deeper) in the transformed equation, while treating the year dummies as standard instruments in the levels equation. In contrast, the De Ridder et al. (2022) specification removes year dummies and restricts the GMM-style instruments to a single deeper lag (the third lag) for output and inputs. Relative to our baseline, this alternative places less weight on within-year cross-sectional price/wedge variation as the driver of instrument relevance and also reduces instrument proliferation by construction. The results using the alternative Blundell-Bond estimates are reported in column (2) of Tables 2 and 3.

A.4.2 With different production function estimators

We also use other commonly used production function estimators as robustness check. In particular, we consider the control-function approach and implement the Olley and Pakes (1996) and Levinsohn and Petrin (2003) estimators.

(2009)—then this concern is mitigated because observed input quantities are not mechanically distorted by unobserved prices.

Table 3: Within-firm variation in returns to scale, productivity, and firm size with alternative production function estimators

	Dependent variable: Firm-level RTS						
	(1) BB baseline	(2) BB alternative	(3) OP	(4) LP	(5) ACF market power	(6) Av.-size percentiles	(7) Cur.-size deciles
$\log(\text{Sales}_{ilt})$	0.013***	0.018***	0.022***	0.019***	0.027***	0.010***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sector-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,248,461	9,424,952	9,248,461	9,248,461	9,424,952	9,248,461	9,248,461
R^2	0.799	0.813	0.875	0.927	0.693	0.853	0.739
$\hat{z}_{ilt}$	0.008***	0.008***	0.012***	0.014***	0.021***	0.005***	0.004***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sector-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,254,839	5,254,839	5,254,839	5,254,839	5,254,839	5,254,839	5,254,839
R^2	0.829	0.845	0.911	0.947	0.741	0.882	0.768

Notes: This table reports coefficients from within-firm regressions of firm-level returns to scale (RTS) on (i) log sales and (ii) within-firm productivity, each including firm fixed effects and sector-year fixed effects. Column (1) uses our benchmark Blundell-Bond (BB) estimates; column (2) uses an alternative BB specification as implemented in De Ridder et al. (2022). Columns (3) and (4) use the Olley-Pakes (OP) and Levinsohn-Petrin (LP) control-function estimators, respectively. Column (5) reports results from the Ackerberg-Caves-Frazer (ACF) estimator with market power controls (proxied by firms' sales shares). Columns (6) and (7) use alternative grouping methods for estimating elasticities: rolling average-size percentiles and contemporaneous size deciles. The regressions use a sample of Spanish firms from Orbis. See Appendix A.1 for details on variable construction and sample selection. Standard errors (in parentheses) are two-way clustered at the firm and sector-year level. *,** ,*** indicate significance at the 10%, 5%, and 1% levels, respectively.

The Olley-Pakes Estimator We first implement the Olley and Pakes (1996) (OP) estimator to estimate the production function for each sector–decile–year cell. The Olley and Pakes (1996) estimator is a semiparametric control-function method that addresses simultaneity bias, since unobserved productivity affects firms' input choices. It assumes that investment is a function of capital and productivity and, under a monotonicity condition, can be inverted to express unobserved productivity in terms of observed investment and capital. Substituting this inverted control function into the production function, the method first estimates the elasticities of freely adjustable inputs (labor and materials in our case) while controlling for productivity, and then uses a Markov assumption on productivity to recover the coefficient on the quasi-fixed input, capital. To implement this approach, we measure real investment as the change in the capital stock net of depreciation, and we recognize that this can generate zero or negative investment values, which reduces the usable sample. The results using the OP estimator are reported in column (3) of Tables 2 and 3.

The Levinsohn-Petrin Estimator Because investment can be lumpy in practice and the Olley and Pakes (1996) procedure may force us to drop observations with zero or negative investment, we also apply the Levinsohn and Petrin (2003) (LP) estimator as an additional robustness check. Instead of using investment, this method uses intermediate inputs (materials in our case) as the control variable. It assumes that materials are flexibly chosen after observing productivity, while capital is still treated as quasi-fixed. Under the assumption that materials demand is a function of capital and productivity and is monotone in productivity (conditional on capital), the materials demand function can be inverted to recover unobserved productivity. This control function allows consistent estimation of the labor elasticity, and additional moment conditions then recover the elasticities of capital and materials. The results using the LP estimator are reported in column (4) of Tables 2 and 3.

A.4.3 With controls for market power

Our model abstracts from market power and markups, but these forces could hinder the identification of output elasticities and obscure the positive RTS–size relationship we identify in the cross section. When firms have market power, they may charge different output prices, so elasticities estimated using deflated sales can be closer to revenue elasticities rather than physical output elasticities.

That said, we expect this channel to be unlikely to explain our results. Under monopolistic competition, larger firms typically have higher markups. Higher markups mechanically dampen the sensitivity of revenue to input expansion, implying lower revenue elasticities for larger firms relative to smaller firms. If anything, this would bias against finding a positive RTS–size relationship. Therefore, the presence of markups would tend to weaken our estimated positive RTS–size relationship, suggesting that the underlying relationship could be even stronger.

Nonetheless, we follow common practice and re-estimate the production function with explicit controls for market power, treating price variation as an additional measurement component to be partialled out. Specifically, following Baqaee and Farhi (2019b) and De Loecker et al. (2020), we control for markups using firms' sales shares (measured at the NACE 3-digit and 4-digit levels) and estimate production functions using the Ackerberg et al. (2015) (ACF) estimator. The results using the ACF estimator are reported in column (5) of Tables 2 and 3.

A.4.4 With different size-based grouping methods

Grouping firms by 7-year average sales percentiles Our benchmark approach groups firms in sector i and year t into deciles based on their 7-year average log sales. While straightforward, this discretization can generate non-smooth variation across firm sizes. As a robustness check, we therefore implement a rolling-percentile approach based on firms' 7-year average sales. For each sector-year, we rank firms into 100 percentiles using their 7-year average (log) sales. In each sector i, for each percentile p_t, we construct a local sample consisting of firms whose percentile rank lies between $p_t - 15$ and $p_t + 15$ in year t. We then estimate output elasticities for each cell (i, t, p_t) using the Blundell–Bond estimator on the corresponding 7-year rolling-window sample. The results using the rolling-percentile grouping approach are reported in column (6) of Tables 2 and 3.

Grouping firms by contemporaneous sales deciles Alternatively, we group firms into deciles based on contemporaneous firm-level (log) sales in year t, rather than the 7-year average. We then estimate output elasticities for each cell (i, t, d_t) using the same Blundell–Bond estimator on a 7-year rolling-window sample. The results using the contemporaneous sales-decile grouping approach are reported in column (7) of Tables 2 and 3.

A.4.5 Summary

Overall, the main empirical patterns are robust. Across all alternative production-function estimators (alternative Blundell-Bond specifications, Olley-Pakes, Levinsohn-Petrin, and ACF with market-power controls) and alternative grouping methods (rolling percentiles and contemporaneous deciles), we continue to find a positive relationship between firm-level returns to scale and firm size, as well as between returns to scale and productivity, both in the cross section within sector-years and within firms over time. While magnitudes vary across specifications, the sign and statistical significance of these relationships are stable (see Tables 2 and 3).

A.5 Estimation of the tail index

This appendix describes how we estimate the tail index of the firm-size distribution in each sector-year using the log-rank estimator of Gabaix and Ibragimov (2011). For each sector i and year t, let S_{ilt} denote firm l's sales, and let N_{it} be the number of firms observed in (i, t). We assign ranks $r = 1, \dots, N_{it}$ according to their sales, where $r = 1$ corresponds to the firm with the largest sales. Let $S_{i(1)t} \geq S_{i(2)t} \geq \cdots \geq S_{i(N_{it})t}$ denote sales sorted in descending order within sector-year (i, t).

We focus on the right tail of the sales distribution and select the tail sample as follows: If $N_{it} > 5000$, we use firms in the top 1% of the sales distribution in (i, t). If $N_{it} \leq 5000$, we use the 50 firms with the largest sales in (i, t).[42] For each sector-year (i, t), we estimate the Pareto tail index ζ_{it} within the tail sample using the Gabaix and Ibragimov (2011) bias-corrected log-rank regression:

$$\log\left(r - \frac{1}{2}\right) = a_{it} - \zeta_{it} \log S_{i(r)t} + u_{irt}. \tag{45}$$

This regression relates the log bias-corrected rank $\log\left(r - \frac{1}{2}\right)$ to log sales. We recover $\widehat{\zeta}_{it}$ as the negative of the OLS slope coefficient on $\log S_{i(r)t}$ and use it as the tail index of sales in 10.

[42] If fewer than 50 firms are observed, we use all available firms.

A.6 Details of the imported input tariff shock exercise

This appendix provides additional details on the imported-input tariff shock used in Section 6.3.2. Our goal is to measure changes in input costs driven by changes in import tariffs. To isolate variation that differs across downstream sectors and over time, we construct a shift-share exposure measure that combines (i) predetermined import input shares from the OECD multi-country input–output tables and (ii) tariff changes from the Global Tariff Project (Teti, 2024).

Let downstream sectors in Spain be indexed by i. Index a foreign exporter-sector pair by $n = (c, s)$, where c denotes the exporting country and s the exporting sector. For each Spanish downstream sector i and year t, we define the tariff-based input cost shifter as

$$\log T_{it} = \sum_{c,s} \Big(\text{ImportShare}^{\text{Intermediate}}_{(\text{Spain},i)\leftarrow(c,s),t-1} \cdot \log\big(1 + \text{TariffRate}^{\text{Spain}}_{(c,s),t}\big) \Big), \tag{46}$$

where $\text{ImportShare}^{\text{Intermediate}}_{(\text{Spain},i)\leftarrow(c,s),t-1}$ is the share of sector i's total intermediate inputs imported from exporter-sector $n = (c, s)$, measured in year $t-1$ using the OECD multi-country input-output tables. $\text{TariffRate}^{\text{Spain}}_{(c,s),t}$ is the ad valorem tariff rate applied by Spain to imports from exporter-sector (c, s) in year t, taken from the Global Tariff Project. Sector i and foreign sectors s are defined according to the OECD input–output classification, which is slightly more aggregated than the NACE 2-digit level. When tariff data are available at a more disaggregated level in Teti (2024), we aggregate to (c, s) using a simple (unweighted) mean across subsectors. Note that $\log T_{it}$ is essentially a weighted average of log tariff factors across upstream foreign inputs, with weights given by the downstream sector's lagged import input structure. It rises when tariffs increase on inputs that the sector i relies on more intensively. The shift-share structure uses lagged import shares to reduce concerns that contemporaneous changes in sourcing respond mechanically to tariff changes.

We then estimate the dynamic impact of these shocks on returns to scale using panel local projections for horizon years $h = -2, \ldots, 5$:

$$\eta_{il,t+h} - \eta_{il,t-1} = \beta_h \log T_{it} + \gamma_{lh} + \gamma_{th} + \varepsilon_{ilth},$$

controlling for firm (γ_{lh}) and year (γ_{th}) fixed effects. Under the assumption that tariff changes for a given exporter-sector pair (c, s) are not systematically correlated with unobserved, time-varying shocks to Spanish downstream sector i (conditional on these fixed effects), variation in $\log T_{it}$ provides plausibly exogenous movements in input costs across sectors and over time.

A.7 Details of the cross-country firm-level data

This appendix describes the firm-level data sources and sample construction for our cross-country analysis. We augment the analysis with firm-level data from a total of 24 countries (including Spain). For 22 European countries, we use Orbis and restrict attention to countries with good coverage of the variables required for production-function estimation. For developing countries, we use China's National Bureau of Statistics (NBS) manufacturing firm database and India's Annual Survey of Industries (ASI). Both the NBS and ASI datasets are censuses of above-scale manufacturing firms. To make cross-country comparisons comparable, we restrict all datasets to manufacturing firms. For each country, we select a seven-year window that maximizes the number of firm-year observations. We briefly discuss the data cleaning below.

Orbis For Orbis, we start from the raw firm-year panel for each country and apply the same four-step cleaning procedure used in Section A.1 for Spain. We then (i) restrict the sample to manufacturing firms (corresponding to USSIC codes 2000-3999) and (ii) deflate all nominal financial variables using the country-specific GDP deflator from the World Bank. After cleaning and deflation, we implement the seven-year window selection described above and keep the window with the largest number of firm-year observations for each country.

India ASI Our India data come from the Annual Survey of Industries (ASI) for 1998-2018. We harmonize industry codes to NIC-2004 and then map them to the USSIC division level, retaining only manufacturing divisions. We measure sales using the gross sale value of all products. We measure capital using the average of the opening and closing gross book value of total capital. We measure labor using total wage bills. All variables are deflated using India's GDP deflator from the World Bank. We then select the seven-year window with the largest number of firm-year observations (2012-2018).

China NBS The China data are annual firm-level surveys collected by the National Bureau of Statistics (NBS). We use the 1998-2007 sample period. We measure sales using product sales revenue, capital using total fixed assets, and labor using total annual wages payable. Firms are classified by a four-digit Chinese Industry Classification (CIC) code, which we harmonize to the USSIC division level. We retain manufacturing divisions only. All nominal variables are deflated using China's GDP deflator from the World Bank. We then select the seven-year window with the largest number of firm-year observations within the available sample period (2001-2007).

Production function and RTS estimation We estimate production functions using the Blundell-Bond approach, following our baseline estimation strategy. We treat manufacturing as a single sector within each country. For each country c, let $[t_m(c)-3, t_m(c)+3]$ denote the selected seven-year window and $t_m(c)$ is the median year of that window. We only estimate production functions for firms existing in the median year $t_m(c)$. We group firms into deciles for year $t_m(c)$ based on their seven-year average log sales. We then estimate a decile-specific Cobb-Douglas production function using the full seven-year panel.

Let $\hat{\beta}^{O}_{c,d(l),t_m(c)}$ denote the estimated output elasticity of input $O \in \{K, L, M\}$ for country c and sales decile d. The returns to scale assigned to firm l in country c at year $t_m(c)$ is computed as the sum of the estimated input elasticities:

$$\eta_{clt_m(c)} = \hat{\beta}^{K}_{c,d(l),t_m(c)} + \hat{\beta}^{L}_{c,d(l),t_m(c)} + \hat{\beta}^{M}_{c,d(l),t_m(c)}.$$

We then construct the Törnqvist productivity index $\widehat{z}_{clt_m(c)}$ using these estimates and compute the covariance between returns to scale and log sales, as well as between returns to scale and productivity $\widehat{z}_{clt_m(c)}$ used in 11 panel (a). In panel (b), we plot the seven-year average ($t_m(c) - 3$ to $t_m(c) + 3$) of log GDP per capita obtained from Penn World Table version 11.0 against the covariance between returns to scale and productivity $\widehat{z}_{clt_m(c)}$.

B Appendix for Section 7

This appendix contains details about the calibration of Section 7.

B.1 Calibration data

This appendix describes the datasets used in the calibration and how the associated sectoral moments are computed.

1. We calibrate the sectoral parameters using the 2010 input-output table from the Annual Spanish National Accounts. This table partitions the Spanish economy into 62 sectors which are usually defined at the 2-digit NACE industry level.[43] Conforming to the accounting conventions in the data, we calibrate the input elasticities of good s' in the production of sector s as

$$\begin{aligned}\hat{\alpha}_{ss'} =& \frac{\text{Input from } s' \text{ at basic prices}_s}{\text{Total input at basic prices}_s} \times \\ & \frac{\text{Intermediate consumption at purchaser's prices}_s}{\text{Intermediate consumption at purchaser's prices}_s + \text{total labor expenditure}_s}\end{aligned}$$

and the labor elasticity as

$$1 - \sum_{s'} \hat{\alpha}_{ss'} = \frac{\text{total labor expenditure}_s}{\text{Intermediate consumption at purchaser's prices}_s + \text{total labor expenditure}_s},$$

[43] Sector 63 (household-related production activities) and sector 64 (services by extraterritorial organizations and bodies) are also present in the 2010 input-output table, but their input-output data is missing.

corresponds to the labor share of total cost in the data.[44,45] We calibrate the consumption share β_s to be the share of final consumption expenditure of good s in the sum of consumption expenditure spent on the 62 sectors.

2. We compute cross-sectional moments from the Orbis sample. After steps 1-4 in Appendix A.1 and the production function estimation in Appendix A.4.1, we perform a few additional steps:

   (a) We winsorize the estimated returns to scale η_{ilt} at the top or bottom 0.5% of the firm-year distribution. In addition, we cap values above 0.99 at 0.99. Using firm-level returns to scale η_{ilt}, we compute each sector's effective returns to scale $\hat{\eta}_{it}$ as the sales-weighted average of these firm-level estimates.

   (b) We compute profits as $\Pi_{ilt} = (1 - \eta_{ilt}) P_{it} Q_{ilt}$ and winsorize it at the top or bottom 0.5% within each sector–year.

   (c) We then compute the interquartile range of Π_{ilt} and η_{ilt} at the sector-year level.

   (d) Finally, we average these sector–year moments over time to obtain sector-level moments used in our static model.

B.2 Interquartile ranges for returns to scale and profits

From (12), we have

$$\eta_{il} = 1 - \frac{1}{\frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\varepsilon_{il}-\mu_i}{2\gamma_i}},$$

which implies[46]

$$\mathrm{IQR}\left(\eta_{il}\right) = \frac{1}{\frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\sigma_i}{2\gamma_i}\Phi^{-1}\left(0.25\right)} - \frac{1}{\frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\sigma_i}{2\gamma_i}\Phi^{-1}\left(0.75\right)}, \tag{47}$$

where $\Phi\left(\cdot\right)$ is the cumulative distribution function of the standard normal random variable.

Profit of firm l in sector i is given by (55). Plugging (10) and (12) in this expression, we get

$$\log \Pi_{il} = \frac{1}{4\gamma_i}\left(2\gamma_i \frac{1-\varphi_i}{1-\hat{\eta}_i} - \mu_i + \varepsilon_{il}\right)^2 + \log H_i, \tag{48}$$

[44] Because sector-to-sector data at purchasers' prices (i.e., adjusted for taxes less subsidies on products) are unavailable, we calibrate the intermediate-input expenditure share of inputs from sector s' used by sector s with the share computed in basic prices, $\frac{\text{Input from } s' \text{ at basic prices}_s}{\text{Total input at basic prices}_s}$.

[45] We deliberately omit capital in this calibration because the Spanish input–output table does not distinguish the user cost of capital from profits.

[46] For or all firms with $\eta_{il} \in (0,1)$, η_{il} is strictly increasing in ε_{il}. In the calibrated economy, the fraction of firms with $\eta_{il} \notin (0,1)$ is very small.

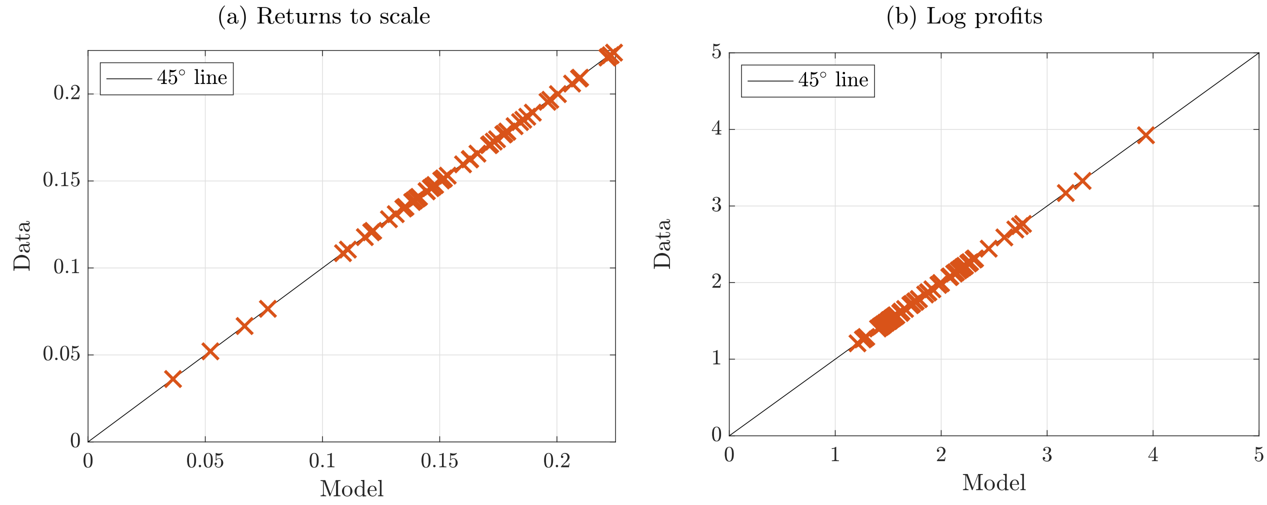
which implies[47]

$$\mathrm{IQR}\left(\log\Pi_{il}\right) = \frac{\sigma_i^2}{4\gamma_i}\left(F^{-1}_{\chi_1^2\left(\frac{2\gamma_i(1-\varphi_i)}{\sigma_i(1-\hat{\eta}_i)}\right)}(0.75) - F^{-1}_{\chi_1^2\left(\frac{2\gamma_i(1-\varphi_i)}{\sigma_i(1-\hat{\eta}_i)}\right)}(0.25)\right), \tag{49}$$

where $F_{\chi_1^2(x)}(\cdot)$ is the cumulative distribution function of noncentral χ^2 distribution with one degree of freedom and the non-centrality parameter x, and $\varphi_i = \frac{\sigma_i^2}{2\gamma_i}$.

Equations (47) and (49) make clear that IQRs of returns to scale and log profits are functions of σ_i, γ_i, and $\hat{\eta}_i$. We can, therefore, use them to identify σ_i and γ_i. We choose σ_i and γ_i to minimize the distance between model-implied and empirical IQRs, with a constraint $\varphi_i \in [0,1]\ \forall i$. Figure 14 shows that the calibrated model matches the targeted IQRs well.

Figure 14: Interquartile ranges in returns to scale and profits


(a) Returns to scale

(b) Log profits

Notes: Panels (a) and (b) report sectoral interquartile ranges in returns to scale and log profits in the calibrated model and in the data.

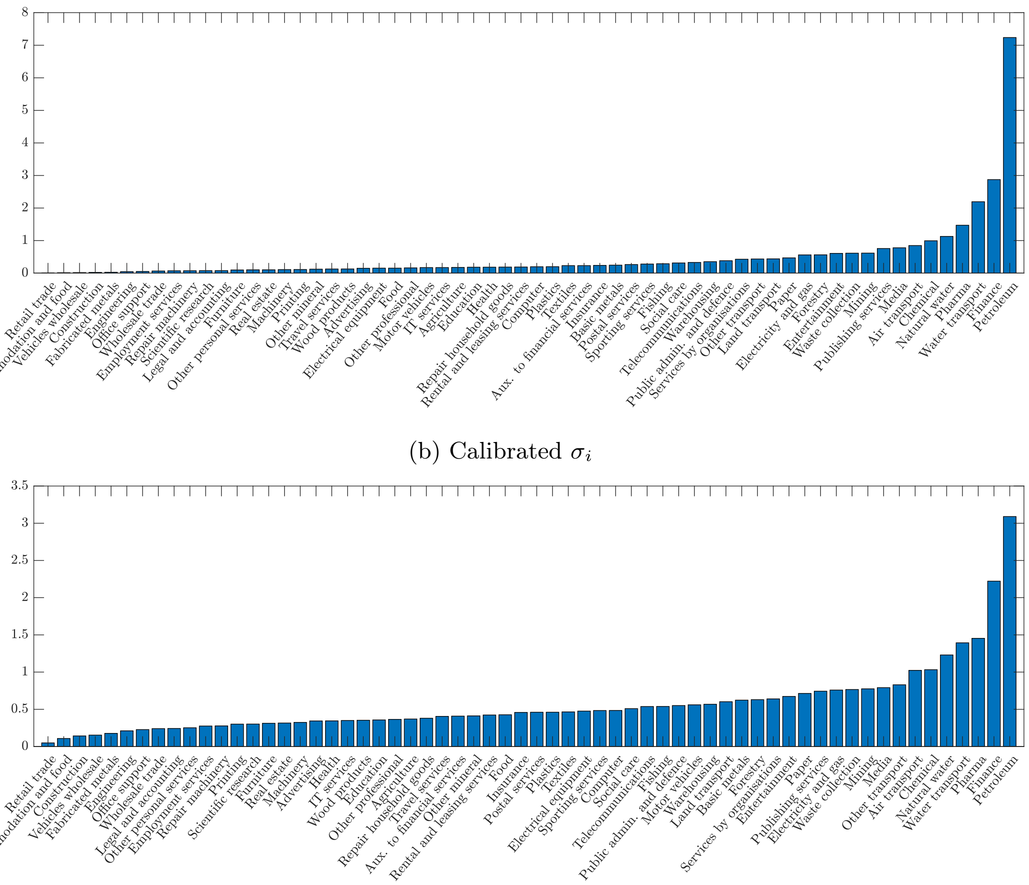
Figure 15 shows calibrated values of σ_i and γ_i for all sectors. The sector with most volatile productivity is "Petroleum", with $\sigma_i = 3.09$. At the same time, this sector has a high cost of adjusting returns to scale, $\gamma_i = 7.24$, meaning that its effective productivity dispersion is not too large, $\varphi_i = 0.66$.

B.3 Calibration details for Section 7.4

We analyze the model with sales tax in Appendix D.6. In that appendix, we show that the model with sales taxes can be analyzed analogously to the main model if we properly redefine the

[47] From (10) and (12) , $2\gamma_i\frac{1-\varphi_i}{1-\hat{\eta}_i} - \mu_i + \varepsilon_{il} > 0$ for all firms with $\eta_{il} \in (0,1)$. For these firms, $\log\Pi_{il}$ is strictly increasing in ε_{il}. In the calibrated economy, the fraction of firms with $\eta_{il} \notin (0,1)$ is very small.

Figure 15: Calibrated γ_i and σ_i



(a) Calibrated γ_i

(b) Calibrated σ_i

mean and the variance of sectoral shocks:

$$\tilde{\mu}_i = \mu_i + \log\left(1 - \tau_i^S\right) \quad \text{and} \quad \tilde{\sigma}_i = (1 - b_i)\,\sigma_i.$$

We can identify $\tilde{\sigma}_i$ and γ_i in the same way as described in Appendix B.2. The only difference is that we need to use after-tax profits in (49).

To pin down the parameters of the tax process (38), we proceed as follows. In the data, we compute the covariance of pre-tax profits with $\log\left(1 - \tau_{il}^S\right)$ for each sector. Using (48), we can compute the model analogue of this quantity as

$$\operatorname{Cov}\left(\log \Pi_{il}, \log\left(1 - \tau_{il}^S\right)\right) = \operatorname{Cov}\left(\frac{1}{4\gamma_i}\left(2\gamma_i \frac{1 - \tilde{\varphi}_i}{1 - \hat{\eta}_i} - \tilde{\mu}_i + \tilde{\varepsilon}_{il}\right)^2 - \log\left(1 - \tau_{il}^S\right), \log\left(1 - \tau_{il}^S\right)\right) =$$

$$= -\frac{b_i}{1 - b_i}\tilde{\sigma}_i^2\left(\frac{1 - \tilde{\varphi}_i}{1 - \hat{\eta}_i} + \frac{b_i}{1 - b_i}\right).$$

We can identify b_i from this equation.

To compute τ_i^S,we rely on equation (88), derived in Appendix D.6, which we reproduce below:

$$\frac{1}{1-\hat{\tau}_i^S} = \frac{1}{1-\tau_i^S}\left(1+\frac{\tilde{\varphi}_i}{\frac{1-\tilde{\varphi}_i}{1-\hat{\eta}_i}}\frac{b_i}{1-b_i}\right)\exp\left(-\frac{b_i}{1-b_i}\frac{4\tilde{\varphi}_i\gamma_i\frac{1-\tilde{\varphi}_i}{1-\hat{\eta}_i}+\tilde{\sigma}_i^2\frac{b_i}{1-b_i}}{2}\frac{1}{1-\tilde{\varphi}_i}\right). \tag{89}$$

In the data, we can observe $\hat{\tau}_i^S$ as the sales-weighted average tax rate in sector i. Then, (88) can be used to identify τ_i^S.

Finally, equation (87) makes clear that the proper measure of sectoral returns to scale $\hat{\eta}_i$ uses after-tax sales as weights.

B.4 Additional quantitative results

Figure (16) shows effective sectoral returns to scale $\hat{\eta}_i$ for all sectors. In our data, the sector with lowest returns to scale is "Water transport" with $\hat{\eta}_i = 0.54$, and the sector with the highest returns to scale is "Retail trade" with $\hat{\eta}_i = 0.98$. The mean and median returns to scale are both 0.83 and 0.82, respectively.

Figure 16: Effective returns to scale $\hat{\eta}_i$ across sectors

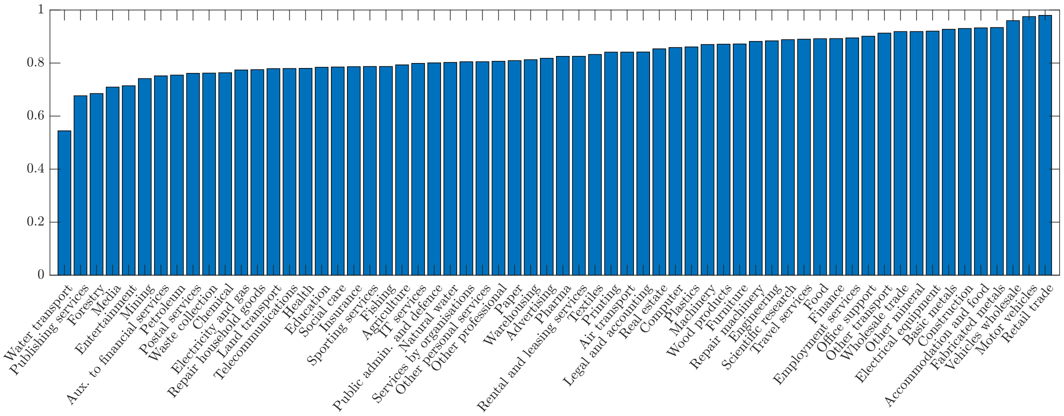
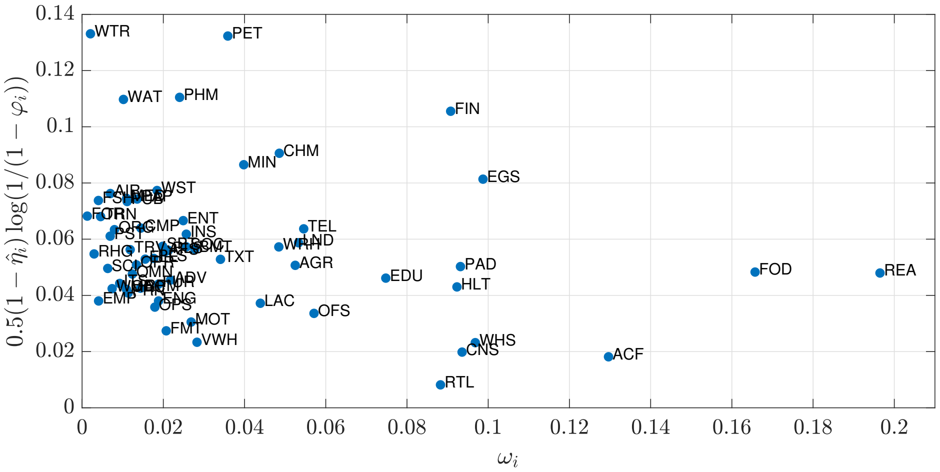


Figure 17 decomposes the gap in GDP between our baseline model and the fixed returns-to-scale economy in its sectoral components. It reports the two terms in (37) that captures a sector's importance: 1) its Domar weight ω_i and 2) the flexibility of its sectoral productivity $\frac{1}{2}\left(1-\hat{\eta}_i\right)\log\frac{1}{1-\varphi_i}$. We see that the "Water transport" sector is the most flexible one. However, since its Domar weight is only 0.0021, its importance for the economy is small. High-Domar-weights sectors like "Finance", "Real estate", and "Electricity and gas" that are also flexible are where the endogenous returns to scale mechanism has the most impact on GDP.

B.5 Sensitivity analysis for Section 7.4

In Section 7.4, we experiment with removing wedges that are correlated with firm productivity. As we discuss there, removing these wedges leads to higher productivity dispersion. For some

Figure 17: Domar weights, ω_i, and productivity gain due to endogenous returns to scale, $\frac{1}{2}(1-\hat{\eta}_i)\log\frac{1}{1-\varphi_i}$, across sectors



sectors, removing wedges would imply that $\varphi_i = \frac{\sigma_i^2}{2\varphi_i} > 1$, which is not allowed by our model. For these sectors, we set $\varphi_i = 0.99$. In this appendix, we explore how sensitive our results are to this threshold. Table 4 shows log GDP gains due to removal of sales wedges if we set $\varphi_{max} = 0.985$, 0.99 (main text), and 0.995. We see that the GDP gains become larger as φ_{max} increases. In the model, having sectors with $\varphi_i \to 1$ is particularly valuable because they feature a larger mass of firms with very high productivity draws operating at nearly constant returns to scale, which makes these sectors especially productive.

Table 4: Log GDP change after removal of sales wedges: Sensitivity analysis

	$\varphi_{max} = 0.985$	$\varphi_{max} = 0.99$	$\varphi_{max} = 0.995$
Baseline economy	160%	167%	177%
Dispersed RTS	134%	138%	142%
Fixed RTS	69%	70%	70%

Notes: Increases in log GDP due to removal of sales wedges in the baseline economy, and in the economies with fixed and dispersed returns to scale, for three values of maximum effective productivity dispersion φ.

C Proofs

C.1 Sectoral Domar weights

Multiplying the resource constraint for good i, given by (9), by P_i we get

$$P_i Q_i = P_i C_i + \sum_j P_i \int_0^{M_i} X_{ji,l}\, dl.$$

From the problem of the household we know that $P_i C_i = \beta_i \bar{P} Y$. It follows that

$$\frac{P_i Q_i}{\bar{P} Y} = \beta_i + \sum_j \frac{P_i}{\bar{P} Y} \int_0^{M_i} X_{ji,l} dl,$$

where we have divided by nominal GDP $\bar{P} Y$. Next, from the problem of firm l in sector j we know that

$$P_i X_{ji,l} = \alpha_{ji} \eta_{jl} P_j Q_{jl}.$$

Combining with the previous expression yields

$$\frac{P_i Q_i}{\bar{P} Y} = \beta_i + \sum_j \int_0^{M_i} \alpha_{ji} \eta_{jl} \frac{P_j Q_{jl}}{\bar{P} Y} dl,$$

or

$$\omega_i = \beta_i + \sum_j \alpha_{ji} \omega_j \hat{\eta}_j.$$

Solving this linear system leads to (18).

C.2 Proof of Lemma 1

Lemma 1. *The firm's marginal cost of production λ_{il} is given by*

$$\lambda_{il} = \frac{1}{e^{\varepsilon_{il}} A_i (\eta_{il})} H_i^{\eta_{il}} \Pi_{il}^{1-\eta_{il}}, \tag{3}$$

where $H_i := W^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N P_j^{\alpha_{ij}}$ is the price of the variable input bundle used by firms in sector i, and

$$\Pi_{il} = (1 - \eta_{il}) \lambda_i Q_{il} \tag{4}$$

is profits.

Proof. We tackle problem (2) through its cost minimization dual:

$$\min_{\eta_{il}, L_{il;}, X_{ij,l}} W L_{il} + \sum_{j=1}^N P_j X_{ij,l}, \text{ subject to } F_i (L_{il}, X_{il}, \eta_{il}) \geq Q_{il}. \tag{50}$$

The Lagrangian is

$$\mathcal{L} = W L_{il} + \sum_{j=1}^N P_j X_{ij,l} - \lambda_{il} \left(e^{\varepsilon_{il}} A_i (\eta_{il}) \zeta (\eta_{il}) \left(L_{il}^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N X_{ij,l}^{\alpha_{ij}} \right)^{\eta_{il}} - Q_{il} \right),$$

and the first-order conditions with respect to L_{il} and $X_{ij,l}$ are

$$\eta_{il}\left(1-\sum_{j=1}^{N}\alpha_{ij}\right)\lambda_{il}Q_{il}=WL_{il}, \tag{51}$$

$$\eta_{il}\alpha_{ij}\lambda_{il}Q_{il}=P_jX_{ij,l}. \tag{52}$$

Plugging back into the constraint, we find

$$\lambda_{il}=\frac{1}{\left(e^{\varepsilon_{il}}A_i\left(\eta_{il}\right)\right)^{\frac{1}{\eta_{il}}}}H_i\left(\left(1-\eta_{il}\right)Q_{il}\right)^{\frac{1-\eta_{il}}{\eta_{il}}}. \tag{53}$$

Using the definition of Π_{il} from (4) yields the result.

Note also that the envelope theorem implies that λ_{il} is the marginal production cost of the firm. Notice that λ_{il} is increasing in Q_{il} for $\eta_{il}<1$. As usual, we can the write the profit maximization problem of the firm as

$$\max_{Q_{il}} P_iQ_{il}-\int_0^{Q_{il}}\lambda_{il}\left(x\right)dx,$$

where the notation makes clear the dependence of $\lambda_{il}\left(Q_{il}\right)$ on the size of the firm. This problem's first-order condition implies that $P_i=\lambda_{il}\left(Q_{il}\right)$, so that the firm sets Q_{il} to equalize its marginal cost to the price of its good. □

C.3 Proof of Lemma 2

Lemma 2. *At an interior solution, the firm chooses its returns to scale $\eta_{il}\in(0,1)$ according to*

$$\frac{da_i\left(\eta_{il}\right)}{d\eta_{il}}=\log H_i-\log\Pi_{il}, \tag{5}$$

where $a_i\left(\eta_{il}\right):=\log A_i\left(\eta_{il}\right)$.

.

Proof. The first-order condition for η_{il} in the cost-minimization problem (50) is

$$\frac{dA_i\left(\eta_{il}\right)}{d\eta_{il}}\zeta\left(\eta_{il}\right)\left(L_{il}^{1-\sum_{j=1}^{N}\alpha_{ij}}\prod_{j=1}^{N}X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}}+A_i\left(\eta_{il}\right)\frac{d\zeta\left(\eta_{il}\right)}{d\eta_{il}}\left(L_{il}^{1-\sum_{j=1}^{N}\alpha_{ij}}\prod_{j=1}^{N}X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}} \tag{54}$$
$$+A_i\left(\eta_{il}\right)\zeta\left(\eta_{il}\right)\frac{d}{d\eta_{il}}\left(L_{il}^{1-\sum_{j=1}^{N}\alpha_{ij}}\prod_{j=1}^{N}X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}}=0.$$

Note that we do not include Lagrange multipliers for the constraints $0\leq\eta_{il}\leq1$ since we focus on interior solutions. Dividing by Q_{il} yields

$$\frac{d\log A_i\left(\eta_{il}\right)}{d\eta_{il}}+\frac{d\log\zeta\left(\eta_{il}\right)}{d\eta_{il}}+\frac{d}{d\eta_{il}}\log\left(L_{il}^{1-\sum_{j=1}^{N}\alpha_{ij}}\prod_{j=1}^{N}X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}}=0.$$

Combining this with (51) and (52) yields (5). □

C.4 Proof of Lemma 3

Lemma 3. *At an interior solution, the returns to scale parameter η_{il} satisfies*[48]

$$\frac{d\eta_{il}}{d\varepsilon_{il}}=\frac{d\eta_{il}}{d\log P_i}=-\left[\left(1-\eta_{il}\right)\frac{d^2a_i}{d\eta_{il}^2}\right]^{-1}>0,\qquad \textit{and}\qquad \frac{d\eta_{il}}{d\log H_i}=\left[\left(1-\eta_{il}\right)\frac{d^2a_i}{d\eta_{il}^2}\right]^{-1}<0.$$

Proof. We can combine (3) with the firm's optimality condition $\lambda_i=P_i$ to write

$$\log\Pi_{il}=\frac{1}{1-\eta_{il}}\left(\log P_i+\varepsilon_{il}+a_i\left(\eta_{il}\right)-\eta_{il}\log H_i\right). \tag{55}$$

Together with (5), we can write the first-order condition with respect to η_{il} as

$$\underbrace{\log H_i-\log P_i-\varepsilon_{il}}_{K}=\left(1-\eta_{il}\right)\frac{d\log A_i\left(\eta_{il}\right)}{d\eta_{il}}+\log A_i\left(\eta_{il}\right), \tag{56}$$

where we use K as a temporary variable to denote the left-hand side of (56). Full differentiation yields

$$1=-\frac{d\eta_{il}}{dK}\frac{d\log A_i\left(\eta_{il}\right)}{d\eta_{il}}+\left(1-\eta_{il}\right)\frac{d^2\log A_i\left(\eta_{il}\right)}{d\eta_{il}^2}\frac{d\eta_{il}}{dK}+\frac{d\log A_i\left(\eta_{il}\right)}{d\eta_{il}}\frac{d\eta_{il}}{dK}.$$

Simplifying we find

$$\frac{d\eta_{il}}{dK}=\frac{1}{\left(1-\eta_{il}\right)\frac{d^2\log A_i\left(\eta_{il}\right)}{d\eta_{il}^2}},$$

and the result follows. □

C.5 Proof of Lemma 4

Lemma 4. *At an interior solution, the elasticity of output Q_{il} with respect to productivity ε_{il} is given by*

$$\frac{d\log Q_{il}}{d\varepsilon_{il}}=\underbrace{\frac{1}{1-\eta_{il}}}_{\textit{Fixed }\eta\textit{ effect}}+\frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\varepsilon_{il}}>0.$$

[48] When increasing P_i, we keep the price of the variable input bundle constant to distinguish the two channels that affect η_{il}.

In addition, the elasticities of output Q_i with respect to prices are given by

$$\frac{d\log Q_{il}}{d\log P_i} = \underbrace{\frac{\eta_{il}}{1-\eta_{il}}}_{\textit{Fixed } \eta \textit{ effect}} + \frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\log P_i} > 0, \;\; \textit{and} \;\; \frac{d\log Q_{il}}{d\log H_i} = \underbrace{-\frac{\eta_{il}}{1-\eta_{il}}}_{\textit{Fixed } \eta \textit{ effect}} + \frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\log H_i} < 0.$$

Furthermore, the impact of a change in ε_{il}, $\log P_i$ or $\log H_i$ on $\log Q_i$ is amplified because of the endogenous response of η_{il}.

Proof. Profit maximization implies that the firm's marginal cost of production λ_i is equal to the price P_i, and so we can invert (3) and (4) to write

$$Q_{il} = \frac{1}{1-\eta_{il}}\left(e^{\varepsilon_{il}}A_i\left(\eta_{il}\right)\right)^{\frac{1}{1-\eta_{il}}}\left(\frac{P_i}{H_i}\right)^{\frac{\eta_{il}}{1-\eta_{il}}},$$

or, in log form, as

$$\log Q_{il} = -\log\left(1-\eta_{il}\right) + \frac{1}{1-\eta_{il}}\varepsilon_{il} + \frac{1}{1-\eta_{il}}a_i\left(\eta_{il}\right) + \frac{\eta_{il}}{1-\eta_{il}}\left(\log P_i - \log H_i\right). \quad (57)$$

Without endogenous returns to scale, it is immediate that

$$\frac{\partial \log Q_{il}}{\partial \varepsilon_{il}} = \frac{1}{1-\eta_{il}} \quad \text{and} \quad \frac{\partial \log Q_{il}}{\partial \log P_i} = -\frac{\partial \log Q_{il}}{\partial \log H_i} = \frac{\eta_{il}}{1-\eta_{il}}.$$

With endogenous returns to scale, we can combine (5) and (55) to find

$$-\left(1-\eta_{il}\right)\frac{da_i\left(\eta_{il}\right)}{d\eta_{il}} = \log P_i + \varepsilon_{il} + a_i\left(\eta_{il}\right) - \log H_i. \quad (58)$$

Combining (57) and (58), we get

$$\log Q_{il} = -\log\left(1-\eta_{il}\right) - \frac{da_i\left(\eta_{il}\right)}{d\eta_{il}} - \left(\log P_i - \log H_i\right). \quad (59)$$

Differentiating with respect to ε_{il}, we find

$$\frac{d\log Q_{il}}{d\varepsilon_{il}} = \frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\varepsilon_{il}} - \frac{d^2 a_i\left(\eta_{il}\right)}{d\eta_{il}^2}\frac{d\eta_{il}}{d\varepsilon_{il}}.$$

Combining with Lemma 3 yields the result. The derivatives with respect to $\log P_i$ and $\log H_i$ can be computed in a similar way. The last part of the result follows from the signs of the derivatives in Lemma 3. □

C.6 Proof of Proposition 1

Proposition 1. *Suppose that Assumption 1 holds. Without endogenous returns to scale, the distribution of Q_{il} in sector i is log-normal. With endogenous returns to scale, the right tail of the distribution of Q_{il} behaves like a Pareto distribution with tail index $1/\varphi_i$, in the sense that*

$$\log\left(\mathbb{P}\left(Q_{il} > q\right)\right) \sim -\frac{1}{\varphi_i}\log q, \text{ as } q \to \infty.$$

Proof. Without endogenous returns to scale, the log of Q_{il} is given by (57). The only random term is ε_{il} and so Q_{il} is log-normal. We now turn to the case with endogenous returns to scale. Under Assumption 1, we can write (58) as

$$\frac{1}{1-\eta_{il}} = \frac{\varepsilon_{il}+B_i}{2\gamma_i},$$

where we define $B_i := \log P_i - \log H_i$ as a temporary variable to simplify the notation. Combining with (59), we can write

$$\log Q_{il} = \log\left(\frac{\varepsilon_{il}+B_i}{2\gamma_i}\right) + \gamma_i\left(\frac{\varepsilon_{il}+B_i}{2\gamma_i}\right)^2 - B_i.$$

We want to characterize the right tail of Q_{il}. Because of the logarithm, we need to be careful about eventual bounds on ε_{il}. We impose here that $\varepsilon_{il} \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ is normally distributed with a truncation such that $\varepsilon_{il} > -B_i$. We provide a full treatment of the model with truncated normal distribution in Appendix D.1. To simplify the notation, we drop the subscripts i and l from now on.

Step 1. We want to characterize the Complementary CDF (CCDF) $S_Q(q) = \mathbb{P}(Q > q)$ as $q \to \infty$. Let us define $g : (-B, \infty) \to \mathbb{R}$ as the function that maps ε to $\log Q$:

$$g(x) = \log\left(\frac{x+B}{2\gamma}\right) + \gamma\left(\frac{x+B}{2\gamma}\right)^2 - B.$$

One can show that g is a strictly increasing function. It is therefore invertible, and we can write

$$S_Q(q) = \mathbb{P}(\log Q > \log q) = \mathbb{P}(g(\varepsilon) > \log q) = \mathbb{P}\left(\varepsilon > g^{-1}(\log q)\right).$$

Given the properties of g, the right tail of Q corresponds to the right tail of ε.

Step 2. Let $y = g(x)$. We need to characterize the asymptotic behavior of $x = g^{-1}(y)$ as $y \to \infty$. Letting $X = \frac{x+B}{2\gamma}$, the equation $y = g(x)$ can be rewritten as

$$y + B = \log X + \gamma X^2.$$

As $y \to \infty$, it must be that $X \to \infty$. In this limit, the quadratic term γX^2 dominates $\log X$ and we can write[49]

$$y + B \sim \gamma X^2, \text{ as } y \to \infty.$$

This implies that $X \sim \sqrt{y/\gamma}$ for large y.

Now, we relate this to $x = g^{-1}(y)$. Since $x = 2\gamma X - B$, we have

$$g^{-1}(y) = x = 2\gamma X - B \sim 2\sqrt{\gamma y}.$$

since the constant B is negligible as $y \to \infty$. We will come back to this expression momentarily.

Step 3. The CCDF of the truncated normal ε is given by

$$S_{\varepsilon}(x) = \frac{1}{K_1} S_{\bar{\varepsilon}}(x)$$

where $\bar{\varepsilon}$ is the untruncated normal with the same mean and variance, and where K_1 is a constant. It is well-known that approximating the Mills ratio implies that

$$S_{\bar{\varepsilon}}(x) \sim \frac{\sigma}{(x-\mu)\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ as } x \to \infty.$$

We can therefore write

$$\log S_{\varepsilon}(x) \sim \log\left(\frac{1}{K_1}\right) - \frac{(x-\mu)^2}{2\sigma^2} + \log\left(\frac{\sigma}{(x-\mu)\sqrt{2\pi}}\right).$$

As $x \to \infty$, the quadratic term dominates the others and thus

$$\log S_{\varepsilon}(x) \sim -\frac{x^2}{2\sigma^2}, \text{ as } x \to \infty.$$

Step 4. We now combine the results. Let $x_q = g^{-1}(\log q)$. From Step 1, $S_Q(q) = S_{\varepsilon}(x_q)$. From Step 3, for large q, and consequently large x_q,

$$\log S_Q(q) \sim -\frac{x_q^2}{2\sigma^2}.$$

We can now substitute the asymptotic form for x_q from Step 2. Let $y = \log q$. As $q \to \infty$, $y \to \infty$ and

$$x_q = g^{-1}(\log q) \sim 2\sqrt{\gamma \log q}.$$

[49] As usual, we write "$f(x) \sim g(x)$ as $x \to \infty$" if and only if $\lim_{x\to\infty} f(x)/g(x) = 1$.

It is well-known that if $f \sim g$ then $f^r \sim g^r$ for r real. Therefore,

$$x_q^2 \sim 4\gamma \log q.$$

Since $\sim$ is transitive, we can substitute in the expression for $\log S_Q$ to find

$$\log S_Q\left(q\right) \sim -\frac{2\gamma}{\sigma^2}\log q, \text{ as } q \to \infty,$$

which is the result. □

C.7 Proof of Lemma 5

Lemma 5. *The returns to scale η_{il} of firm l in sector i is given by*

$$\frac{1}{1-\eta_{il}} = \frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\varepsilon_{il}-\mu_i}{2\gamma_i}. \tag{12}$$

Furthermore, the moments of the firm-level returns to scale distribution in sector i are given by

$$\mathrm{E}_i\left[\frac{1}{1-\eta_{il}}\right] = \frac{1-\varphi_i}{1-\hat{\eta}_i}, \quad \mathrm{V}_i\left[\frac{1}{1-\eta_{il}}\right] = \frac{\varphi_i}{2\gamma_i}, \quad \textit{and} \quad \mathrm{Cov}_i\left[\frac{1}{1-\eta_{il}}, \varepsilon_{il}\right] = \varphi_i > 0. \tag{13}$$

Proof. Given Assumption 1, we can write the returns to scale first-order condition (56) as

$$\log P_i - \log H_i + \varepsilon_{il} = \frac{2\gamma_i}{1-\eta_{il}},$$

Combining that expression with itself when $\varepsilon_{il} = \mu_i$ yields

$$\frac{1}{1-\eta_{il}} = \frac{1}{1-\eta_i\left(\mu_i\right)} + \frac{\varepsilon_{il}-\mu_i}{2\gamma_i},$$

and the result follows from combining with (65), derived below, and taking the moments. □

C.8 Proof of Proposition 2

Proposition 2. *The marginal cost of sector i is given by*

$$\lambda_i = \frac{1}{Z_i\left(\hat{\eta}_i\right)} W^{1-\hat{\eta}_i\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N P_j^{\hat{\eta}_i \alpha_{ij}}, \tag{15}$$

where sectoral total factor productivity $Z_i\left(\hat{\eta}_i\right)$ is defined as

$$\log Z_i\left(\hat{\eta}_i\right) := \underbrace{\mu_i + a_i\left(\hat{\eta}_i\right) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i}}_{\textit{Exogenous returns to scale}} + \underbrace{\frac{1}{2}\left(1-\hat{\eta}_i\right)\log\left(\frac{1}{1-\varphi_i}\right)}_{\textit{Superstar effect}} - \underbrace{\left(1-\hat{\eta}_i\right)\log\kappa_i}_{\textit{Entry cost}}. \tag{16}$$

Furthermore, the effective returns to scale $\hat{\eta}_i$ is given by

$$\frac{1}{1-\hat{\eta}_i} = \frac{1}{2\gamma_i \left(1-\varphi_i\right)} \left(\mu_i + \log P_i - \log H_i\right). \tag{17}$$

Proof. Since firms in a sector all face the same sales price, they have the same marginal cost through profit maximization. We therefore define the marginal cost λ_i of a sector i as the marginal cost of any firm in that sector, such that $\lambda_i := \lambda_{il}$ for any l.

Together with (55), the free-entry condition (8) imposes that

$$\int_{-\infty}^{\infty} \underbrace{\left(\lambda_i \frac{e^{\varepsilon_{il}} A_i\left(\eta_{il}\right)}{H_i^{\eta_{il}}}\right)^{\frac{1}{1-\eta_{il}}}}_{\Pi_{il}} f_i\left(\varepsilon_{il}\right) d\varepsilon_{il} = \kappa_i W, \tag{60}$$

where f_i is the probability density function of a normal distribution $\mathcal{N}\left(\mu_i, \sigma_i^2\right)$. Multiplying the term inside the parentheses by one, we find

$$\int_{-\infty}^{\infty} \left(\frac{\lambda_i}{H_i^{\eta_{il}}} \frac{H_i^{\left(1-\eta_{il}\right)\frac{\hat{\eta}_i}{1-\hat{\eta}_i}}}{H_i^{\left(1-\eta_{il}\right)\frac{\hat{\eta}_i}{1-\hat{\eta}_i}}} \frac{\lambda_i^{\frac{1-\eta_{il}}{1-\hat{\eta}_i}}}{\lambda_i^{\frac{1-\eta_{il}}{1-\hat{\eta}_i}}} e^{\varepsilon_{il}} A_i\left(\eta_{il}\right)\right)^{\frac{1}{1-\eta_{il}}} f_i\left(\varepsilon_{il}\right) d\varepsilon_{il} = \kappa_i W,$$

which can be reorganized as

$$\lambda_i = \frac{1}{\tilde{Z}_i} \left(\kappa_i W\right)^{1-\hat{\eta}_i} \left(W^{1-\sum_j \alpha_{ij}} \prod_{j=1}^{N} P_j^{\alpha_{ij}}\right)^{\hat{\eta}_i}, \tag{61}$$

where $\tilde{Z}_i$ is defined as

$$\tilde{Z}_i := \left[\int_{-\infty}^{\infty} \left(\left(\frac{\lambda_i}{H_i}\right)^{\frac{\eta_{il}-\hat{\eta}_i}{1-\hat{\eta}_i}} e^{\varepsilon_{il}} A_i\left(\eta_{il}\right)\right)^{\frac{1}{1-\eta_{il}}} f_i\left(\varepsilon_{il}\right) d\varepsilon_{il}\right]^{1-\hat{\eta}_i}.$$

To simplify the notation, define $s_i := \log \lambda_i - \log H_i$. Using the definition of s_i and A_i, we can write

$$\tilde{Z}_i = \left[\int_{-\infty}^{\infty} \left(e^{s_i \frac{\eta_{il}-\hat{\eta}_i}{1-\hat{\eta}_i} + \varepsilon_{il} - \frac{\gamma_i}{1-\eta_{il}}}\right)^{\frac{1}{1-\eta_{il}}} f_i\left(\varepsilon_{il}\right) d\varepsilon_{il}\right]^{1-\hat{\eta}_i}. \tag{62}$$

For an arbitrary set of firm-level returns to scale $\{\eta_{il}\}$ this integral cannot be computed analytically, but we can do so here, given the relationship between ε_{il} and η_{il} implied by the model. Using

Assumption 1, we can write the returns to scale first-order condition (56) as[50]

$$\underbrace{\log\lambda_i-\log H_i}_{:=s_i}+\varepsilon_{il}=\frac{2\gamma_i}{1-\eta_{il}}=-2a_i\left(\eta_{il}\right), \tag{63}$$

which implies that

$$\frac{1}{1-\eta_{il}}=\frac{s_i+\varepsilon_{il}}{2\gamma_i},\ \text{and}\ \frac{\eta_{il}}{1-\eta_{il}}=\frac{s_i+\varepsilon_{il}-2\gamma_i}{2\gamma_i}.$$

Combining with $\tilde{Z}_i$, we find

$$\tilde{Z}_i=\left[\int_{-\infty}^{\infty}e^{\frac{(s_i+\varepsilon_{il})^2}{4\gamma_i}-\frac{s_i}{1-\hat{\eta}_i}}f_i\left(\varepsilon_{il}\right)d\varepsilon_{il}\right]^{1-\hat{\eta}_i}.$$

Given the structure of the normal distribution f_i, this integral can be computed when $2\gamma_i>\sigma_i^2$ and yields

$$\tilde{Z}_i=\left[\sqrt{\frac{2\gamma_i}{2\gamma_i-\sigma_i^2}}\exp\left(\frac{\left(s_i+\mu_i\right)^2}{2\left(2\gamma_i-\sigma_i^2\right)}-\frac{s_i}{1-\hat{\eta}_i}\right)\right]^{1-\hat{\eta}_i}.$$

We will rewrite this expression using $\hat{\eta}_i$. To do so, notice that we can write

$$\hat{\eta}_i=\frac{\int_l\eta_{il}P_iQ_{il}dl}{\int_l P_iQ_{il}dl}=1-\frac{\int_l\left(1-\eta_{il}\right)P_iQ_{il}dl}{\int_l P_iQ_{il}dl}=1-\frac{\int_l\Pi_{il}dl}{\int_l\frac{1}{1-\eta_{il}}\Pi_{il}dl}. \tag{64}$$

Using the profit expression (55), we can compute these integrals and find

$$1-\hat{\eta}_i=2\gamma_i\frac{1-\varphi_i}{s_i+\mu_i}=\left(1-\varphi_i\right)\left(1-\eta_i\left(\mu_i\right)\right), \tag{65}$$

where $\eta_i\left(\mu_i\right)$ is the returns to scale chosen by the firm with $\varepsilon_{il}=\mu_i$ (computed from (63)). Notice that (65) implies (17) because s_i is given by (63).

Combining (65) with our expression for $\tilde{Z}_i$, we find

$$\tilde{Z}_i=\left[\sqrt{\frac{1}{1-\varphi_i}}\exp\left(\frac{1-\varphi_i}{1-\hat{\eta}_i}a_i\left(\hat{\eta}_i\right)+\frac{\mu_i}{1-\hat{\eta}_i}\right)\right]^{1-\hat{\eta}_i}.$$

Taking the log yields

$$\log\tilde{Z}_i\left(\hat{\eta}_i\right):=\mu_i+a_i\left(\hat{\eta}_i\right)+\frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i}-\left(1-\hat{\eta}_i\right)\log\left(\sqrt{1-\varphi_i}\right),$$

where we have used the definition of φ_i and Assumption 1. The quantity $\tilde{Z}_i$ corresponds to the total factor productivity of sector i if we treat the mass of firms in that sector as an independent factor. But it will be often convenient to lump that input together with labor. In that case, we can

[50] In equilibrium, the price charged by firms in sector i must be equal to their marginal costs, so that $\lambda_i=P_i$.

rewrite (61) as

$$\lambda_i = \frac{1}{Z_i\left(\hat{\eta}_i\right)} W^{1-\hat{\eta}_i \sum_{j=1}^N \alpha_{ij}} \left(\prod_{j=1}^N P_j^{\alpha_{ij}} \right)^{\hat{\eta}_i},$$

where $Z_i := \tilde{Z}_i\left(\hat{\eta}_i\right) / \kappa_i^{1-\hat{\eta}_i}$, which completes the proof. □

C.9 Proof of Proposition 3

Proposition 3. *Equilibrium log GDP $y := \log Y$ is given by*

$$y\left(\hat{\eta}\right) = \underbrace{\left[\omega\left(\hat{\eta}\right)\right]^\top z\left(\hat{\eta}\right)}_{\textit{Aggregate productivity}} + \underbrace{\log \bar{L}}_{\textit{Labor endowment}}. \tag{20}$$

Proof. The equilibrium price vector $P = \left(P_1, \ldots, P_N\right)$ satisfies

$$\log \frac{P}{W} = -\mathcal{L}\left(\hat{\eta}\right) z\left(\hat{\eta}\right), \tag{19}$$

where $z\left(\hat{\eta}\right) = \left(\log Z_1\left(\hat{\eta}_1\right), \ldots, \log Z_N\left(\hat{\eta}_N\right)\right)$ is the vector of log sectoral productivities (16). Since in equilibrium prices must be equal to marginal costs, we can use (15) to write

$$\frac{P_i}{W} = \frac{1}{Z_i\left(\hat{\eta}_i\right)} \prod_{j=1}^N \left(\frac{P_j}{W}\right)^{\hat{\eta}_i \alpha_{ij}}.$$

Taking the log of this equation leads to

$$\log \frac{P_i}{W} = -\log Z_i\left(\hat{\eta}_i\right) + \hat{\eta}_i \sum_{j=1}^N \alpha_{ij} \log \frac{P_j}{W}.$$

In vector notation, this becomes $\log\left(P/W\right) = -z\left(\hat{\eta}\right) + \operatorname{diag}\left(\hat{\eta}\right) \alpha \log\left(P/W\right)$. Solving it for $\log\left(P/W\right)$ yields (19).

We now turn to the GDP equation. The budget constraint of the household is $\bar{P}Y = W\bar{L}$. Together with the definition of the price index, $\bar{P} = \prod_{i=1}^N P_i^{\beta_i} = 1$, we can therefore write

$$y = -\sum_{i=1}^N \beta_i \log \frac{P_i}{W} + \log \bar{L},$$

and the result follows from combining this expression with (18) and (19). □

C.10 Proof of Proposition 4

Proposition 3. *There exists a unique equilibrium, and it is efficient. Furthermore, the equilibrium vector of effective returns to scale $\hat{\eta}$ maximizes GDP $y\left(\hat{\eta}\right)$, as given by (20).*

Proof. This proof proceeds in two steps. First, we write down the maximization problem of the social planner and show that its first-order conditions coincide with the equilibrium conditions. Since there exists at least one maximizer to the planner's problem, there is at least one solution to the planner's first-order conditions and so at least one efficient equilibrium exists. Second, we show that the equilibrium conditions imply that there can be at most one equilibrium.

Step 1. The planner maximizes

$$\max_{C,X,L,M,\eta} \sum_{i=1}^{N} \beta_i \log\left(C_i\right)$$

subject to the goods resource constraint

$$C_i + \sum_{j=1}^{N} M_j \int X_{ji}(\varepsilon) f_j(\varepsilon) d\varepsilon \leq M_i \int Q_i(\varepsilon) f_i(\varepsilon) d\varepsilon \quad \forall i \in \{1, ..., N\} \quad \text{(multiplier } \lambda_i\text{)},$$

and the labor resource constraint

$$\sum_{i=1}^{N} M_i \int L_i(\varepsilon) f_i(\varepsilon) d\varepsilon + \sum_{i=1}^{N} M_i \kappa_i \leq \bar{L} \quad \text{(multiplier } \mu\text{)}.$$

The first-order necessary conditions are as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial C_i} &: \quad \lambda_i = \frac{\beta_i}{C_i}, \\
\frac{\partial \mathcal{L}}{\partial L_i(\varepsilon)} &: \quad \lambda_i \frac{\partial Q_i(\varepsilon)}{\partial L_i(\varepsilon)} - \mu = 0, \\
\frac{\partial \mathcal{L}}{\partial X_{ij}(\varepsilon)} &: \quad \lambda_i \frac{\partial Q_i(\varepsilon)}{\partial X_{ij}(\varepsilon)} - \lambda_j = 0, \\
\frac{\partial \mathcal{L}}{\partial \eta_i(\varepsilon)} &: \quad \frac{\partial Q_i(\varepsilon)}{\partial \eta_i(\varepsilon)} = 0, \\
\frac{\partial \mathcal{L}}{\partial M_i} &: \quad \int \left[\lambda_i Q_i(\varepsilon) - \sum_{j=1}^{N} \lambda_j X_{ij}(\varepsilon) - \mu L_i(\varepsilon) \right] f_i(\varepsilon) d\varepsilon = \mu \kappa_i.
\end{aligned}$$

Now, we demonstrate that the competitive equilibrium allocation satisfies the planner's optimality conditions. To do this, we identify the planner's shadow prices with the equilibrium market prices. Set $\mu = W$. Consequently, the planner's shadow price for good i, λ_i, corresponds to the market price P_i. The first condition corresponds to the household's optimality condition (Section 2.4). The second and third optimality conditions correspond to the standard firm equilibrium optimality conditions (51) and (52). The planner's fourth optimality condition coincides with the firm equilibrium condition (54). Finally, the last optimality condition of the planner coincide with the free entry condition (8). Since the resource constraints are the same in the planner's problem and the equilibrium definition, we have shown that the planner's first-order conditions coincides with the equilibrium conditions. Since the planner's constraint set is closed and bounded, and that the objective function is continuous, the Extreme Value Theorem implies that there exists a maximizer to the planner's problem. This maximizer must satisfy the first-order necessary conditions. Therefore, there exists an equilibrium and that equilibrium is efficient.

Step 2. We now show that there can be at most one equilibrium. The equilibrium of the model boils down to equations (17) and (19). Indeed, if we let $p := \log\left(P/W\right)$ we can write these equations as

$$p = -\mathcal{L}\left(\hat{\eta}\right) z\left(\hat{\eta}\right), \tag{66}$$

where

$$z_i\left(\hat{\eta}\right) := \mu_i + a_i\left(\hat{\eta}_i\right) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i} + \left(1-\hat{\eta}_i\right)\log\left(\frac{1}{\sqrt{1-\varphi_i}}\right) - \left(1-\hat{\eta}_i\right)\log\kappa_i, \tag{67}$$

and

$$\frac{1}{1-\hat{\eta}_i} = \frac{1}{2\gamma_i\left(1-\varphi_i\right)}\left(\mu_i + p_i - \sum_{j=1}^{N}\alpha_{ij}p_j\right). \tag{68}$$

There is a unique equilibrium if there are unique vectors $\hat{\eta}$ and p that solve these equations. We can combine these equations into a single one. Let us introduce the variable $v := (I-\alpha)\,p$ and a constant $C_i = 2\gamma_i\left(1-\varphi_i\right) > 0$. We can then rewrite (68) as

$$1-\hat{\eta}_i = \frac{C_i}{\mu_i + v_i} \Leftrightarrow \hat{\eta}_i = 1 - \frac{C_i}{\mu_i + v_i}.$$

We are interested in equilibrium of the firm $0 < \hat{\eta}_i < 1$ for all i.[51] This implies that we can restrict the relevant domain of v to be

$$\mu_i + v_i > C_i.$$

Using that notation, we can simplify the equation (67) for z_i as

$$z_i = \frac{\mu_i - v_i}{2} + \frac{C_i}{\mu_i + v_i}\log\left(K_i/\kappa_i\right),$$

where $K_i := 1/\sqrt{1-\varphi_i}$. Next, we can premultiply (66) by $\mathcal{L}\left(\hat{\eta}\right)^{-1} = \left(I - \text{diag}\left(\hat{\eta}\right)\alpha\right)$ to find

$$\left(I - \text{diag}\left(\hat{\eta}\right)\alpha\right)\left(I-\alpha\right)^{-1} v = -z$$

or

$$\left(I + \text{diag}\left(1-\hat{\eta}\right)\alpha\left(I-\alpha\right)^{-1}\right) v = -z.$$

[51] It is straightforward to write sufficient conditions on the parameters so that the equilibrium is of that form. In particular, large μ lead to higher equilibrium $\hat{\eta}$.

Substituting the expression for z and $1-\hat{\eta}$, we find

$$F_i\left(v\right) := \frac{1}{2}\left(\mu_i+v_i\right)^2 + C_i\left(\left(\alpha\left(I-\alpha\right)^{-1}v\right)_i + \log\left(K_i/\kappa_i\right)\right) = 0.$$

There is a unique equilibrium if there is a unique solution v to the equation $F\left(v\right)=0$. Recall that $p=\left(I-\alpha\right)^{-1}v$. Then

$$\hat{F}_i\left(p\right) := F_i\left(v\left(p\right)\right) = \frac{1}{2}\left(\mu_i+\left(\left(I-\alpha\right)p\right)_i\right)^2 + C_i\left(\left(\alpha p\right)_i + \log\left(K_i/\kappa_i\right)\right).$$

The Jacobian of $\hat{F}$ is

$$M_{ik}\left(p\right) = \left(\mu_i+\left(\left(I-\alpha\right)p\right)_i\right)\left(I-\alpha\right)_{ik} + C_i\alpha_{ik}.$$

In matrix form,

$$M\left(p\right) = \operatorname{diag}\left(\mu+v\left(p\right)\right)\left(I-\alpha\right) + \operatorname{diag}\left(C\right)\alpha.$$

The diagonal elements of M are

$$M_{ii} = \left(\mu_i+v_i\right)\left(1-\alpha_{ii}\right) + C_i\alpha_{ii} > 0,$$

which is positive given our domain restriction that $\mu_i+v_i>C_i$. For off-diagonal terms $i\neq k$,

$$M_{ik} = -\left(\mu_i+v_i\right)\alpha_{ik} + C_i\alpha_{ik} = \alpha_{ik}\left(C_i-\left(\mu_i+v_i\right)\right) < 0,$$

such that M is a Z-matrix. Further notice that

$$\sum_{k\neq i}\left|M_{ik}\right| = \left(\left(\mu_i+v_i\right)-C_i\right)\sum_{k\neq i}\alpha_{ik}.$$

For M to be strictly diagonally dominant, it must be that

$$\left(\mu_i+v_i\right)\left(1-\alpha_{ii}\right) + C_i\alpha_{ii} > \left(\left(\mu_i+v_i\right)-C_i\right)\sum_{k\neq i}\alpha_{ik},$$

which we can reorganize as

$$\mu_i+v_i > -C_i\frac{\sum_k\alpha_{ik}}{1-\sum_k\alpha_{ik}}.$$

This condition is true since $C_i>0$ and $\mu_i+v_i>C_i$. Therefore, M is diagonally dominant. It follows that $M\left(p\right)$ is a non-singular M-matrix for every p. Since nonsingular M-matrices are a subset of P-matrices, $M\left(p\right)$ is also a P-matrix for every p. By the Gale and Nikaido (1965) theorem, $\hat{F}\left(p\right)$ is therefore injective and can have at most one solution $\hat{F}\left(p\right)=0$. There is therefore a unique p that solves our original system of equations. From the vector p, it is straightforward to recover all other equilibrium quantities in a unique fashion. There is therefore a unique equilibrium and it is efficient. □

C.11 Proof of Lemma 6

Corollary ??. *An increase in average productivity μ_j increases returns to scale in all other sectors, such that*

$$\frac{d\hat{\eta}_i}{d\mu_j} = \Psi_i^{-1}\mathcal{K}_{ij} \geq 0. \tag{22}$$

Furthermore, the impact of productivity dispersion σ_j^2 on $\hat{\eta}_i$ is given by

$$\frac{d\hat{\eta}_i}{d\sigma_j^2} = \Psi_i^{-1}\left(\mathcal{K}_{ij}\frac{\partial z_j}{\partial \sigma_j^2} - \mathbb{1}\left(i=j\right)\frac{\partial^2 z_i}{\partial \sigma_i^2 \partial \hat{\eta}_i}\right), \tag{23}$$

where

$$\frac{\partial z_j}{\partial \sigma_j^2} = \frac{1}{2\left(1-\hat{\eta}_j\right)} + \frac{1-\hat{\eta}_j}{4\gamma_j\left(1-\varphi_j\right)} > 0, \text{ and } \frac{\partial}{\partial \sigma_i^2}\left(\frac{\partial z_i}{\partial \hat{\eta}_i}\right) = \frac{1}{2\left(1-\hat{\eta}_i\right)^2} - \frac{1}{4\gamma_i\left(1-\varphi_i\right)}.$$

In particular, $d\hat{\eta}_i/d\sigma_j^2 \geq 0$ for $i \neq j$.

Proof. This proof proceeds as follows. First we derive the first-order conditions of the social planner. Second, we write down the derivative of the first-order conditions with respect to $\hat{\eta}_i$. Third, we use this expression together with the implicit function theorem to derive the impact of μ_j and σ_j^2 on $\hat{\eta}_i$.

First step. Let us first compute the first-order conditions of the planner's problem. Differentiating (20) with respect to $\hat{\eta}_i$ and setting that expression to zero implies that

$$\frac{dy}{d\hat{\eta}_i} = \beta^\top \frac{d\mathcal{L}}{d\hat{\eta}_i} z\left(\hat{\eta}\right) + \left[\omega\left(\hat{\eta}\right)\right]^\top \frac{dz\left(\hat{\eta}\right)}{d\hat{\eta}_i} = 0.$$

Computing the derivative of $z\left(\hat{\eta}\right)$, we find

$$\left(\frac{dz\left(\hat{\eta}\right)}{d\hat{\eta}_i}\right)_j = \begin{cases} \frac{da_i(\hat{\eta}_i)}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2}\frac{1}{(1-\hat{\eta}_i)^2} + \frac{1}{2}\log\left(1-\varphi_i\right) + \log \kappa_i & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}$$

Next, the derivative of the Leontief inverse yields

$$\begin{aligned}\frac{d\mathcal{L}}{d\hat{\eta}_i} &= \frac{d\left(1-\operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}}{d\hat{\eta}_i} = -\left(1-\operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}\left[\frac{d\left(1-\operatorname{diag}\left(\hat{\eta}\right)\alpha\right)}{d\hat{\eta}_i}\right]\left(1-\operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}. \\ &= \mathcal{L}\left[1_i 1_i^\top \alpha\right]\mathcal{L} = \mathcal{L}_{\cdot i}\alpha_i^\top \mathcal{L}.\end{aligned}$$

Putting the pieces together, we have

$$\omega_i(\hat{\eta})\,\alpha_i^\top \mathcal{L}(\hat{\eta})\,z(\hat{\eta}) + \omega_i(\hat{\eta})\left[\frac{da_i(\hat{\eta}_i)}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2}\frac{1}{(1-\hat{\eta}_i)^2} + \frac{1}{2}\log(1-\varphi_i) + \log\kappa_i\right] = 0.$$

Since Domar weights are positive, we can write that condition as

$$\mathcal{F}_i := \alpha_i^\top \mathcal{L}(\hat{\eta})\,z(\hat{\eta}) + \frac{da_i(\hat{\eta}_i)}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2}\frac{1}{(1-\hat{\eta}_i)^2} + \frac{1}{2}\log(1-\varphi_i) + \log\kappa_i = 0, \tag{69}$$

where we have defined $\mathcal{F}_i$.

Second step. The implicit function theorem states that

$$\frac{d\hat{\eta}}{d\mu} = -\left[\frac{\partial \mathcal{F}}{\partial \hat{\eta}}\right]^{-1}\left[\frac{\partial \mathcal{F}}{\partial \mu}\right].$$

First, let us compute the Jacobian matrix $\partial\mathcal{F}/\partial\hat{\eta}$. Consider an off-diagonal element $k \neq i$

$$\begin{aligned}\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_k} &= \frac{\partial}{\partial \hat{\eta}_k}\left(\alpha_i^\top \mathcal{L}(\hat{\eta})\,z(\hat{\eta})\right) = \alpha_i^\top\left(\frac{\partial \mathcal{L}}{\partial \hat{\eta}_k}\right)z + \alpha_i^\top \mathcal{L}\left(\frac{\partial z}{\partial \hat{\eta}_k}\right)\\ &= \alpha_i^\top \mathcal{L}\mathbf{1}_k\mathbf{1}_k^\top \alpha \mathcal{L} z + \alpha_i^\top \mathcal{L}\mathbf{1}_k\frac{\partial z_k}{\partial \hat{\eta}_k}.\end{aligned}$$

Factoring this expression gives

$$\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_k} = \left(\alpha_i^\top \mathcal{L}_{\cdot k}\right)\left[\alpha_k^\top \mathcal{L} z + \frac{\partial z_k}{\partial \hat{\eta}_k}\right] = 0,$$

where the last equality follows since the term in bracket is the first-order condition of the planner.

For a diagonal element,

$$\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_i} = \frac{\partial}{\partial \hat{\eta}_i}\left(\alpha_i^\top \mathcal{L} z + \frac{\partial z_i}{\partial \hat{\eta}_i}\right).$$

Through the logic above, the first term is 0, so we need only focus on the second part

$$\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_i} = \frac{\partial}{\partial \hat{\eta}_i}\frac{\partial z_i}{\partial \hat{\eta}_i} = \frac{d^2 a_i}{d\hat{\eta}_i^2} + \frac{\sigma_i^2}{(1-\hat{\eta}_i)^3} = (1-\varphi_i)\frac{d^2 a_i}{d\hat{\eta}_i^2}. \tag{70}$$

Third step. Next, we can compute the element (i,j) of the matrix $\frac{\partial \mathcal{F}}{\partial \mu}$, which is $\frac{\partial \mathcal{F}_i}{\partial \mu_j}$. The FOC for sector i is $\mathcal{F}_i = \alpha_i^\top \mathcal{L} z + \frac{\partial z_i}{\partial \hat{\eta}_i}$. The parameter μ_j only enters through the vector z, specifically through its j-th element. Therefore,

$$\frac{\partial z}{\partial \mu_j} = \mathbf{1}_j.$$

Thus,

$$\frac{\partial \mathcal{F}_i}{\partial \mu_j} = \frac{\partial}{\partial \mu_j}\left(\alpha_i^\top \mathcal{L} z\right) = \alpha_i^\top \mathcal{L}\left(\frac{\partial z}{\partial \mu_j}\right) = \alpha_i^\top \mathcal{L}\mathbf{1}_j,$$

which is simply the (i,j)-th element of the matrix $\alpha\mathcal{L}$. Putting the pieces together,

$$\frac{d\hat{\eta}}{d\mu}=-\left(\frac{\partial\mathcal{F}}{\partial\hat{\eta}}\right)^{-1}(\alpha\mathcal{L})=-\frac{(\alpha\mathcal{L})_{ij}}{\frac{d^2a_i}{d\hat{\eta}_i^2}+\frac{\sigma_i^2}{(1-\hat{\eta}_i)^3}}=-\left((1-\varphi_i)\frac{d^2a_i}{d\hat{\eta}_i^2}\right)^{-1}(\alpha\mathcal{L})_{ij}\,.$$

Fourth step. We now turn to the impact of σ_j^2. We use the implicit function theorem once more. Note that

$$\frac{\partial\mathcal{F}_i}{\partial\sigma_j^2}=\alpha_i^\top\mathcal{L}\frac{\partial z}{\partial\sigma_j^2}+\frac{\partial}{\partial\sigma_j^2}\left(\frac{\partial z_i}{\partial\hat{\eta}_i}\right).$$

The vector $\partial z/\partial\sigma_j^2$ is zero everywhere except for it j-th element

$$\frac{\partial z_j}{\partial\sigma_j^2}=\frac{\partial}{\partial\sigma_j^2}\left(\frac{\sigma_j^2}{2\,(1-\hat{\eta}_j)}-\frac{1-\hat{\eta}_j}{2}\log(1-\varphi_j)\right)=\frac{1}{2\,(1-\hat{\eta}_j)}+\frac{1-\hat{\eta}_j}{4\gamma_j\,(1-\varphi_j)}>0.$$

Similarly, $\frac{\partial}{\partial\sigma_j^2}\left(\frac{\partial z_i}{\partial\hat{\eta}_i}\right)$ is zero whenever $i\neq j$. We can compute

$$\frac{\partial}{\partial\sigma_j^2}\left(\frac{\partial z_j}{\partial\hat{\eta}_j}\right)=\frac{\partial}{\partial\sigma_j^2}\left(\frac{\sigma_j^2}{2\,(1-\hat{\eta}_j)^2}+\frac{1}{2}\log(1-\varphi_j)\right)=\frac{1}{2\,(1-\hat{\eta}_j)^2}-\frac{1}{4\gamma_j\,(1-\varphi_j)}.$$

Putting the pieces together, we find the result. □

C.12 Proof of Proposition 5

Proposition 5. *The difference in log GDP between the baseline model and the fixed returns-to-scale economy is given by*

$$y-\tilde{y}=\sum_{i=1}^{N}\frac{1}{2}\omega_i\,(1-\hat{\eta}_i)\log\left(\frac{1}{1-\varphi_i}\right)>0.26$$

Proof. We first compute GDP in the fixed returns-to-scale economy (denoted by $\tilde{\cdot}$), in which all firms in sector i have the same returns to scale $\eta_{il}=\hat{\eta}_i$. The free-entry condition is $\mathrm{E}\left[\tilde{\Pi}_{il}\right]=\kappa_i\tilde{W}$. Using the expression for profit (55), this condition becomes

$$\int_{-\infty}^{\infty}\exp\left(\frac{1}{1-\hat{\eta}_i}\left(\log\tilde{P}_i+\varepsilon_{il}+a_i\,(\hat{\eta}_i)-\hat{\eta}_i\log H_i\right)\right)f\,(\varepsilon_{il})\,d\varepsilon_{il}=\kappa_i\tilde{W}.$$

Solving the integral and following the same aggregation steps as in the baseline model (Proposition 2), but without the endogenous choice of returns to scale, this condition yields a sectoral productivity of

$$\tilde{z}_i=\mu_i+a_i\,(\hat{\eta}_i)+\frac{\sigma_i^2}{2\,(1-\hat{\eta}_i)}-(1-\hat{\eta}_i)\log\kappa_i.$$

Because $\hat{\eta}_{il}$ is fixed, the term related to the choice of scale and the resulting amplified selection (i.e., the fourthterm on the right-hand side of (16)) is absent. Since the sectoral production function and cost shares are still governed by $\hat{\eta}_i$, the pricing equation is analogous to the baseline model: $\log\left(\tilde{P}/\tilde{W}\right) = -\mathcal{L}(\hat{\eta})\,\tilde{z}$. Log GDP is therefore given by:

$$\tilde{y} = [\omega(\hat{\eta})]^\top \tilde{z}(\hat{\eta}) + \log \bar{L}.$$

Note that the Domar weights $\omega(\hat{\eta})$ are identical to the baseline model because the sectoral input shares are the same in both economies.

Recall from 16 and 20 that in the baseline model

$$y = [\omega(\hat{\eta})]^\top z(\hat{\eta}) + \log \bar{L},$$

where

$$z_i = \mu_i + a_i(\hat{\eta}_i) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i} + \frac{1}{2}(1-\hat{\eta}_i)\log\left(\frac{1}{1-\varphi_i}\right) - (1-\hat{\eta}_i)\log\kappa_i.$$

As a result,

$$y - \tilde{y} = [\omega(\hat{\eta})]^\top (z(\hat{\eta}) - \tilde{z}(\hat{\eta})). \tag{71}$$

The difference in the sectoral productivity vectors, $z - \tilde{z}$, is a vector where the i-th element is

$$z_i(\hat{\eta}_i) - \tilde{z}_i(\hat{\eta}_i) = \frac{1}{2}(1-\hat{\eta}_i)\log\left(\frac{1}{1-\varphi_i}\right).$$

Substituting in (71) yields (26). The inequality $y - \tilde{y} > 0$ holds since $0 < \varphi_i < 1$ for all i. □

C.13 Proof of Proposition 7

Proposition 7. *The response of log GDP y to a shock $\Delta\mu_i$ is given by*

$$\Delta y = \omega_i \Delta\mu_i + \frac{1}{2}\frac{d\omega_i}{d\mu_i}(\Delta\mu_i)^2 + o\left((\Delta\mu_i)^2\right). \tag{30}$$

Furthermore, the second-order term is non-negative,

$$\frac{d\omega_i}{d\mu_i} = \left(-\sum_{k=1}^N \mathcal{K}_{ki}\omega_k \frac{d\hat{\eta}_k}{d\mu_i}\right) \geq 0.$$

Proof. The second-order expansion of y with respect to productivity shocks is

$$\Delta y = \sum_{i=1}^N \frac{dy}{d\mu_i}\Delta\mu_i + \frac{1}{2}\sum_{i,j}^N \frac{d^2 y}{d\mu_i d\mu_j}\Delta\mu_i\Delta\mu_j + o\left(\Delta^2\mu\right).$$

By Proposition 6, $dy/d\mu_i = \omega_i$ which yields (30) when $\Delta\mu_j = 0$ for all $j \neq i$. Next, Corollary 7 implies that

$$\frac{d\omega_i}{d\mu_i} = -\sum_{k=1}^{N} \mathcal{K}_{ki}\omega_k \frac{d\hat{\eta}_k}{d\mu_i} \geq 0,$$

where the inequality follows since $d\hat{\eta}_k d\mu_i \geq 0$ from Corollary 6. □

C.14 Proof of Lemma 8

Lemma 8. *An increase in the wedge τ_i^S decreases the returns to scale in all downstream sectors, such that*

$$\frac{d\hat{\eta}_i}{d\tau_j^S} = -\frac{1}{1-\tau_j^S}\Psi_i^{-1}\mathcal{K}_{ij} \leq 0. \tag{32}$$

Proof. See proof of Proposition 10 in Appendix D.5. □

C.15 Proof of Proposition 8

Proposition 8. *In the presence of sales wedges, the impact of a parameter $\chi \in \{\mu_j, \sigma_j, \kappa_j, \gamma_j\}$ on GDP is given by*

$$\frac{dy}{d\chi} = \frac{\partial y}{\partial \chi} + \sum_{i=1}^{N} \frac{\partial y}{\partial \hat{\eta}_i}\frac{d\hat{\eta}_i}{d\chi},$$

6910where $\partial y/\partial \chi$ is given by Proposition 6, $d\hat{\eta}_i/\partial \chi$ is given by Corollaries 6 to 10, and $\partial y/\partial \hat{\eta}_i \geq 0$.

Proof. See proof of Proposition 12 in Appendix D.5. □

C.16 Proof of Corollary 1

Corollary 1. *The growth of effective returns to scale $\hat{\eta}$ is given by*

$$\frac{d\hat{\eta}}{dt} = \Psi^{-1}\mathcal{K}g_\mu > 0. \tag{34}$$

Furthermore, as $t \to \infty$, effective returns to scale $\hat{\eta}$ converges to 1.

Proof. The first equation follows directly from (22). Note that the right-hand side is strictly positive for $0 < \hat{\eta} < 1$ and converges to 0 as $\hat{\eta} \to 1$. The second result follows. □

C.17 Proof of Proposition 9

Proposition 9. *The growth rate of GDP is given by*

$$\frac{dy}{dt} = \frac{g_\mu}{1-\alpha} \times \left(1 - \frac{1}{\sqrt{1 + \frac{1}{\gamma}\frac{1-\alpha}{\alpha}\left(\frac{g_\mu}{1-\varphi}t + T\right)}}\right) > 0, \tag{35}$$

where

$$T := -\frac{1-\alpha}{\alpha} a'\left(\hat{\eta}_0\right) - 2a\left(\hat{\eta}_0\right) > 0,$$

and where $\hat{\eta}_0$ is the effective returns to scale at $t = 0$.

Proof. The envelope theorem implies that

$$\frac{dy}{dt} = \left(1 - \hat{\eta}\alpha\right)^{-1} g_\mu. \tag{72}$$

Therefore, to characterize $\frac{dy}{dt}$, we need to solve for $\hat{\eta}\left(t\right)$. Equation (34) can be written as

$$\frac{d\hat{\eta}}{dt} = \frac{1}{2\gamma - \sigma^2} \frac{\alpha\left(1-\hat{\eta}\right)^3}{1-\hat{\eta}\alpha} g_\mu$$

and reorganized as

$$\frac{\alpha g_\mu}{\left(\gamma - \frac{\sigma^2}{2}\right)2} dt = \left(\frac{1-\alpha}{\left(1-\hat{\eta}\right)^3} + \frac{\alpha}{\left(1-\hat{\eta}\right)^2}\right) d\hat{\eta}.$$

Integrating on both sides yields

$$\frac{\alpha g_\mu}{\left(\gamma - \frac{\sigma^2}{2}\right)2} t + K = \frac{1-\alpha}{2\left(1-\hat{\eta}\right)^2} + \frac{\alpha}{1-\hat{\eta}}, \tag{73}$$

where K is a constant that can be pinned down using an initial condition. Suppose that at $t = 0$, the equilibrium is such that $\hat{\eta} = \hat{\eta}_0$. Then,

$$K = \frac{1-\alpha}{2\left(1-\hat{\eta}_0\right)^2} + \frac{\alpha}{\left(1-\hat{\eta}_0\right)} > 0.$$

Equation provides the evolution of $\hat{\eta}$ over time. Since $\gamma > \sigma^2/2$ by assumption, it shows that $\hat{\eta} \to 1$ as $t \to \infty$.

Combining with (73) with 72 yields (35). □

C.18 Proof of Corollary 2

Corollary 2. *For any $t > 0$, GDP grows faster in the economy with endogenous returns to scale. In the limit as $t \to \infty$, the long-run growth rates satisfy*

$$\lim_{t\to\infty} \frac{dy}{dt} = \frac{1}{1-\alpha} g_\mu > \frac{1}{1-\hat{\eta}_0\alpha} g_\mu = \lim_{t\to\infty} \frac{d\tilde{y}}{dt},$$

where $\tilde{y}$ is log GDP in the fixed returns-to-scale economy, and where $\hat{\eta}_0$ is the effective returns to scale vector in the baseline economy at $t = 0$.

Proof. In the economy with exogenous returns to scale, (20) implies that $\frac{dy}{dt} = \frac{1}{1-\alpha\eta_0} g_\mu$. In the economy with endogenous returns to scale, the envelope theorem implies at, at any point in time we have $\frac{dy}{dt} = \frac{1}{1-\alpha\eta(t)} g_\mu$. This implies that the two economies have the same growth rate at $t = 0$ since $\eta(t) = \eta_0$ by definition. But since $d\eta/dt > 0$ by Corollary 1, the growth rate of the economy with endogenous returns to scale is larger for any $t > 0$. The second part of the result follows from taking the limit $t \to \infty$ in (35). □

D Robustness, extensions, and additional analysis

In this appendix, we provide additional analysis of the benchmark model presented in the main text. We also show that that model can be extended in different ways.

D.1 Truncated normal shocks

In the baseline model, we assume that productivity shocks ε_{il} follow a normal distribution. While this allows for a tractable analytical solution, it theoretically permits firms to draw arbitrarily low productivity shocks, which could imply returns to scale $\eta_{il} \notin (0,1)$. In this appendix, we solve the model assuming that productivity follows a *Truncated Normal* distribution. We show that the equilibrium conditions converge to those of the baseline model as the truncation point goes to negative infinity.

Specifically, productivity shocks of firms in industry i follow truncated normal distribution with support $[\underline{\varepsilon_i}, \infty)$. We assume that $\underline{\varepsilon_i}$ is sufficiently high such that

$$\underline{\varepsilon_i} > 2\gamma_i - s_i,$$

where s_i is given in (63). Under this restriction, all firms choose $\eta_{il} \in (0,1)$, as is evident from (63). The analogue of the free-entry condition, given by (8) in the main text, is

$$\exp(-s_i) \int_{\underline{\varepsilon_i}}^{\infty} \exp\left\{ \frac{\varepsilon_{il}^2 - s_i^2}{4\gamma_i} + \frac{s_i^2 + \varepsilon_{il} s_i}{2\gamma_i} \right\} \exp\left(-\frac{(\varepsilon_{il} - \mu_i)^2}{2\sigma_i^2} \right) \frac{1}{1 - \Phi\left(\frac{\underline{\varepsilon_i} - \mu_i}{\sigma_i} \right)} d\varepsilon_{il} = \kappa_i \frac{W}{\lambda_i}.$$

Simplifying this expression, we get

$$\frac{W}{\lambda_i}\kappa_i = \frac{W^{1-\sum_j \alpha_{ij}} \prod_{j=1}^{N} P_j^{\alpha_{ij}}}{\lambda_i} \sqrt{\frac{1}{1-\varphi_i}} \exp\left(\frac{1}{2} \frac{(s_i + \mu_i)^2}{2\gamma_i (1-\varphi_i)} \right) T_{1i}, \tag{74}$$

where, as above, $\varphi_i = \frac{\sigma_i^2}{2\gamma_i}$, and

$$T_{1i} = \frac{1 - \Phi\left(\frac{\sqrt{1-\varphi_i}}{\sigma_i}\left(\underline{\varepsilon_i} - \mu_i - \frac{\varphi_i}{1-\varphi_i}\left(\mu_i + s_i\right)\right)\right)}{1 - \Phi\left(\frac{1}{\sigma_i}\left(\underline{\varepsilon_i} - \mu_i\right)\right)}.$$

In the baseline model, $\underline{\varepsilon_i} = -\infty$, and $T_{1i} = 1$.

Next, following the same steps as in Appendix C.8, we can compute $\hat{\eta}_i$:

$$\hat{\eta}_i = \frac{\int_l \eta_{il} P_i Q_{il} dl}{\int_l P_i Q_{il} dl} = 1 - \frac{1 - \varphi_i}{\frac{s_i + \mu_i}{2\gamma_i} + \left(1 - \varphi_i\right) T_2} = 1 - \frac{\left(1 - \varphi_i\right)\left(1 - \eta_i\left(\mu_i\right)\right)}{1 + \left(1 - \varphi_i\right)\left(1 - \eta_i\left(\mu_i\right)\right) T_{2i}},$$

where

$$T_{2i} = \frac{\frac{1}{2\gamma_i}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\sqrt{1-\varphi_i}}{\sigma_i}\left(\underline{\varepsilon_i} - \mu_i - \frac{\varphi_i}{1-\varphi_i}\left(\mu_i + s_i\right)\right)\right)^2\right)}{1 - \Phi\left(\frac{\sqrt{1-\varphi_i}}{\sigma_i}\left(\underline{\varepsilon_i} - \mu_i - \frac{\varphi_i}{1-\varphi_i}\left(\mu_i + s_i\right)\right)\right)}.$$

Clearly, if $\underline{\varepsilon_i} = -\infty$, then $T_{2i} = 0$, and we are back to the baseline model (see Equation (65)).

Finally, we can derive the analogue of (19). Following the same steps as in Appendix C.8, we get

$$\begin{aligned}\log \frac{P}{W} = &-\left(I - \left[\operatorname{diag}\left(\varphi\right) - \left(I - \operatorname{diag}\left(\varphi\right)\right)\operatorname{diag}\left(\eta\left(\mu\right)\right)\right]\alpha\right)^{-1} \times \\ &\left[\mu - a\left(\eta\left(\mu\right)\right) - \left(I - \operatorname{diag}\left(\varphi\right)\right)\left(I - \operatorname{diag}\left(\eta\left(\mu\right)\right)\right)\left(\frac{1}{2}\log\left(1 - \varphi\right) + \log\kappa - \log T_1\right)\right].\end{aligned}$$

Again, if $\underline{\varepsilon_i} = -\infty$, we are back to the baseline model.

Clearly, if firms with $\varepsilon_{il} = \mu_i$ choose $\eta_{il} \in (0,1)$, $\underline{\varepsilon_i}$ can be chosen such that $\underline{\varepsilon_i} < \mu_i$. Furthermore, if σ_i is sufficiently small, T_{1i} is arbitrarily close to one and T_{2i} is arbitrarily close to zero. In that case, the mass of firms choosing $\eta_{il} \notin (0,1)$ in our baseline model is negligible, and the baseline economy is almost equivalent to the model with truncated normal shocks.

D.2 The impact of γ and κ on returns to scale

In this appendix, we characterize how entry costs κ and the cost of scalability γ affect returns to scale.

D.3 Entry cost

We examine the impact of entry costs on returns to scale decisions.

Lemma 9. *The impact of the entry cost κ_j on the effective returns to scale $\hat{\eta}_j$ is given by*

$$\frac{d\hat{\eta}_i}{d\log\kappa_j} = \Psi_i^{-1}\left[-\mathcal{K}_{ij}\left(1 - \hat{\eta}_j\right) - \mathbb{I}_{\{i=j\}}\right]. \tag{75}$$

In particular, $d\hat{\eta}_i/d\log\kappa_j \leq 0$ for $i \neq j$.

Proof. Applying the implicit function theorem to (69), we get

$$\frac{d\hat{\eta}}{d\log\kappa_j} = -\left[\frac{\partial\mathcal{F}}{\partial\hat{\eta}}\right]^{-1}\frac{\partial\mathcal{F}}{\partial\log\kappa_j}.$$

We have already computed the first term in the proof of Lemma 6, so consider the second one. We have

$$\frac{\partial\mathcal{F}_i}{\partial\log\kappa_j} = \alpha_i^\top\mathcal{L}\left(\hat{\eta}\right)\frac{\partial z\left(\hat{\eta}\right)}{\partial\log\kappa_j} + \frac{\partial\log\kappa_i}{\partial\log\kappa_j} = -\alpha_i^\top\mathcal{L}\left(\hat{\eta}\right)1_j\left(1-\hat{\eta}_j\right) + \mathbb{1}\left(i=j\right).$$

Putting the pieces together we find the result. □

An increase in the entry cost in sector j always reduces the effective returns to scale of any other sector $i \neq j$. The mechanism is similar to that of a shock to μ_j. Increasing κ_j decreases j's productivity z_j, which increases the price of the input bundle of any sector that relies on j. Firms in those sectors then reduce their returns to scale to rely less on expensive intermediate inputs. At the same time, the effective returns to scale $\hat{\eta}_j$ of sector j itself typically increases with κ_j. This is because, when entry costs are large, there is more pressure to have fewer but larger firms, which requires large $\hat{\eta}_j$.

D.4 Cost of adjusting returns to scale

The productivity cost γ_i of adjusting returns to scale also affects firms' scalability decisions.

Lemma 10. *The impact of the productivity cost of higher returns to scale γ_j on the effective returns to scale $\hat{\eta}_i$ is given by*

$$\frac{d\hat{\eta}_i}{d\gamma_j} = \Psi_i^{-1}\left(\mathcal{K}_{ij}\frac{\partial z_j}{\partial\gamma_j} - \mathbb{I}_{\{i=j\}}\frac{\partial^2 z_i}{\partial\gamma_i\partial\hat{\eta}_i}\right) \tag{76}$$

where $\frac{\partial z_j}{\partial\gamma_j} = -\frac{1}{1-\hat{\eta}_j} - \frac{1}{2\gamma_j}\frac{\varphi_j}{1-\varphi_j}\left(1-\hat{\eta}_j\right) < 0$ and $\frac{\partial^2 z_i}{\partial\gamma_i\partial\hat{\eta}_i} = -\frac{1}{\left(1-\hat{\eta}_i\right)^2} + \frac{1}{2\gamma_i}\frac{\varphi_i}{1-\varphi_i}$. In particular, $d\hat{\eta}_i/d\gamma_j \leq 0$ for $i \neq j$.

Proof. Applying the implicit function theorem to (69), we get

$$\frac{d\hat{\eta}}{d\gamma_j} = -\left[\frac{\partial\mathcal{F}}{\partial\hat{\eta}}\right]^{-1}\left[\frac{\partial\mathcal{F}}{\partial\gamma_j}\right].$$

We have already computed the first term in the proof of Lemma 6, so consider the second one. We have

$$\frac{\partial\mathcal{F}_i}{\partial\gamma_j} := \alpha_i^\top\mathcal{L}\left(\hat{\eta}\right)\frac{\partial z\left(\hat{\eta}\right)}{\partial\gamma_j} + \frac{\partial^2 a_i\left(\hat{\eta}_i\right)}{\partial\gamma_j d\hat{\eta}_i} + \frac{1}{2}\frac{\partial}{\partial\gamma_j}\log\left(1-\varphi_i\right).$$

If $i \neq j$,

$$\frac{\partial\mathcal{F}_i}{\partial\gamma_j} := -\mathcal{K}_{ij}\frac{\partial z_j}{\partial\gamma_j},$$

where

$$\frac{\partial z_j}{\partial \gamma_j} = -\frac{1}{1-\hat{\eta}_j} - \frac{1}{2\gamma_j}\frac{\varphi_j}{1-\varphi_j}\left(1-\hat{\eta}_j\right) < 0.$$

For $i = j$, we have an extra term,

$$\frac{\partial \mathcal{F}_i}{\partial \gamma_i} := -\mathcal{K}_{ii}\frac{\partial z_i\left(\hat{\eta}\right)}{\partial \gamma_i} - \frac{1}{\left(1-\hat{\eta}_i\right)^2} + \frac{1}{2\gamma_i}\frac{\varphi_i}{1-\varphi_i}.$$

□

Consider first the impact of a higher γ_j on the effective returns to scale of another sector $i \neq j$. Unsurprisingly, a higher productivity cost of adjusting returns to scale leads to a lower productivity in sector j. Through input-output linkages, that lower productivity increases the price of the intermediate input bundles of firms that rely, directly or indirectly, on j as an input ($\mathcal{L}_{ij} > 0$). Those firms, to limit the negative impact of higher inputs, lower their returns to scale. A similar impact is at work when considering the impact of a higher γ_j on j itself, but in addition, j is also affected more directly by the increase in γ_j. Indeed, a larger γ_j mechanically makes a high $\hat{\eta}_j$ more expensive, which amplifies the negative movement in $\hat{\eta}_j$. In general, these forces combine to create a stronger negative impact of γ_j on $\hat{\eta}_j$.

D.5 Wedges

In this appendix, we consider an economy with wedges. In the presence of wedges, the firm's problem (2) becomes

$$\Pi_{il} := \max_{\eta_{il}, L_{il}, X_{il}} \left(1-\tau_i^S\right) P_i F_i\left(L_{il}, X_{il}, \eta_{il}\right) - \left(1+\tau_i^L\right) W L_{il} - \sum_{j=1}^{N}\left(1+\tau_{ij}^X\right) P_j X_{ij,l}. \quad (77)$$

Firms in sector i have to pay $\left(1+\tau_{ij}^X\right) P_j$ for each unit of good j, $\tau_{ij}^X > -1$, and $\left(1+\tau_i^L\right) W$ for each unit of labor, $\tau_i^L > -1$. Firms face an effective sales tax $\tau_i^S < 1$. Finally, we introduce a corporate tax rate τ_i^Π. This tax does not directly affect the profit-maximization problem (77). However, it affects the free-entry condition (8):

$$\mathrm{E}_i\left[\left(1-\tau_i^\Pi\right)\Pi_i\left(\varepsilon_{il}, P^*, W^*\right)\right] = \kappa_i W^*.$$

As we can see, the profit tax effectively increases the entry cost.

Wedges $\left\{\tau^X, \tau^L, \tau^S, \tau^\Pi\right\}$ can capture a variety of economic factors, such as tariffs, transportation costs, taxes, markups, etc. Some of those wedges can be associated with loss of resources, while others only lead to resource redistribution. To capture this, we assume that a fraction of wedge income is rebated to the household, such that its budget constraint (7) becomes

$$\sum_{i=1}^{N} P_i C_i \leq W\bar{L} + \mathcal{T},$$

where

$$\mathcal{T} = \sum_{i=1}^{N} \theta_i^S \tau_i^S P_i Q_i + \sum_{i=1}^{N} \theta_i^L \tau_i^L W L_i + \sum_{i=1}^{N}\sum_{j=1}^{N} \theta_{ij}^X \tau_{ij}^X P_j X_{ij} + \sum_{i=1}^{N} \theta_i^\Pi \tau_i^\Pi \Pi_i.$$

Here $\theta_i^S, \theta_i^L, \theta_{ij}^X, \theta_i^\Pi \in [0,1]$. Note that wedges $\left\{\tau^X, \tau^L, \tau^S, \tau^\Pi\right\}$ can be both positive or negative. For example, τ_{ij}^X is positive in case of transportation costs. If those are iceberg costs, nothing is rebated to the household, and $\theta_{ij}^X = 0$. Tariffs would also correspond to a positive τ_{ij}^X. Different from transportation costs, tariff income is likely partially rebated to the household, in which case θ_{ij}^X is positive. On the other hand, τ_{ij}^X would be negative in case of government subsidies. Such subsidies are financed by lump-sum taxation of the household, such that $\theta_{ij}^X = 1$.[52]

The model can be analyzed analogously to our baseline model. In particular, we can derive that the equilibrium price vector is given by

$$\log \frac{P}{W} = -\mathcal{L}\left(\hat{\eta}\right) z\left(\hat{\eta}\right), \tag{78}$$

where, as in the baseline model, $\hat{\eta}$ is a vector of sales-weighted average returns to scale, $\mathcal{L}\left(\hat{\eta}\right) = \left(I - \operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}$, and

$$z_i\left(\hat{\eta}_i\right) = \mu_i - T_i + a_i\left(\hat{\eta}_i\right) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i} + \frac{1}{2}\left(1-\hat{\eta}_i\right)\log\left(\frac{1}{1-\varphi_i}\right) - \left(1-\hat{\eta}_i\right)\log\frac{\kappa_i}{1-\tau_i^\Pi}. \tag{79}$$

The productivity shifter T_i is

$$T_i = T_i\left(\tau_i^L, \tau_i^S, \tau_i^X, \hat{\eta}_i\right) = \log\left[\frac{\left(1+\tau_i^L\right)^{\hat{\eta}_i\left(1-\sum_j \alpha_{ij}\right)} \prod_{j=1}^{N}\left(1+\tau_{ij}^X\right)^{\hat{\eta}_i \alpha_{ij}}}{1-\tau_i^S}\right]. \tag{80}$$

Introducing wedges $\left\{\tau^X, \tau^L, \tau^S\right\}$ is, therefore, equivalent to a change in sectoral total factor productivities. An increase in wedges τ_i^L, τ_i^S or τ_{ij}^X reduces the effective productivity of sector i, resulting in a reduction in the returns to scale in all sectors. This result is analogous to the effect of a reduction in μ_i, described in Corollary 6. At the same time, an increase in the corporate tax τ_i^Π effectively increases the entry cost, and so its impact on returns to scale is analogous to that of $\log \kappa_i$, described in Corollary 9.

Proposition 10. *An increase in wedges* $\left\{\tau^X, \tau^L, \tau^S\right\}$ *reduces returns to scale in all sectors. An*

[52] Deadweight losses of subsidies can be captured by setting $\theta_{ij}^X > 1$.

increase in the profit tax τ_i^{Π} reduces returns to scale in other sectors but can increase returns to scale in sector i.[53]

The market-clearing conditions (9) also change. Specifically, for good i, the resource constraint becomes

$$\left(1-\left(1-\theta_i^S\right)\tau_i^S\right)Q_i = C_i + \sum_{j=1}^{N}\left(1+\left(1-\theta_{ji}^X\right)\tau_{ji}^X\right)X_{ji}.$$

Then the Domar weight of sector i is

$$\tilde{\omega}_i = \frac{P_iQ_i}{\bar{P}Y} = \mathbf{1}_i\left(I - \operatorname{diag}\left[\left(1-\theta^S\right)\circ\tau^S\right] - \tilde{\alpha}^{\top}\operatorname{diag}\left(\hat{\eta}\right)\right)^{-1}\beta,$$

where $\circ$ denotes element-wise product of two vectors, and

$$\tilde{\alpha}_{ji} = \alpha_{ji}\left(1-\tau_j^S\right)\left(\frac{1+\left(1-\theta_{ji}^X\right)\tau_{ji}^X}{1+\tau_{ji}^X}\right) \leq \alpha_{ji}.$$

Using these results, we can derive how wedges affect the expression for the aggregate output.

Proposition 11. *Equilibrium log GDP $y := \log Y$ is given by*

$$y\left(\hat{\eta}\right) = \underbrace{\beta^{\top}\mathcal{L}\left(\hat{\eta}\right)z\left(\hat{\eta}\right)}_{\textit{Contribution of productivity}} + \underbrace{\log\bar{L}}_{\textit{Labor endowment}} - \underbrace{\log\Gamma_{\tau}}_{\textit{Wedges income}}, \tag{81}$$

where

$$\Gamma_{\tau} =$$
$$1-\sum_{i=1}^{N}\tilde{\omega}_i\left(\left(1-\tau_i^S\right)\hat{\eta}_i\left(\sum_{j=1}^{N}\alpha_{ij}\frac{\theta_{ij}^X\tau_{ij}^X}{1+\tau_{ij}^X} + \left(1-\sum_{j=1}^{N}\alpha_{ij}\right)\frac{\theta_i^L\tau_i^L}{1+\tau_i^L}\right) + \theta_i^S\tau_i^S + \theta_i^{\Pi}\tau_i^{\Pi}\left(1-\tau_i^S\right)\left(1-\hat{\eta}_i\right)\right).$$

As discussed above, some wedges can lead to a destruction of resources while others may lead to redistribution of resources. In the latter case, aggregate output needs to be adjusted for wedges income. This is the last term in expression (81). Naturally, if $\theta_{ij}^X = \theta_i^L = \theta_i^S = \theta_i^{\Pi} = 0$ for all i and j, then nothing is rebated to the household, and $\log\Gamma_{\tau} = 0$. If all the wedges are nonnegative, and some of the wedge income is rebated back to the household, then $\log\Gamma_{\tau} < 0$, which leads to a higher y.[54]

The presence of wedges distorts the economy. Intuitively, firms do not internalize that part of the wedge income is rebated to the household, and their decisions are inefficient as a result. If none of the wedge income is rebated to the household, then $\log\Gamma_{\tau} = 0$, and the economy is efficient. In

[53] We provide expressions for derivatives of returns to scale with respect to wedges in the proof of this proposition.

[54] Of course, in that case, sectoral productivities (79) are also lower than in the no-wedges economy.

that case, firms correctly perceive wedges as resource-destructive. In the inefficient economy, the equilibrium returns to scale do not maximize GDP, and any marginal change in returns to scale can have a nontrivial impact on GDP. Specifically, a change in the underlying parameter χ leads to the following response of GDP:

$$\frac{dy}{d\chi} = \frac{\partial y}{\partial \chi} + \sum_{j=1}^{N} \frac{\partial y}{\partial \hat{\eta}_j} \frac{d\hat{\eta}_j}{d\chi}.$$

In general, the sign of the response of GDP to a marginal change in returns to scale, $\frac{\partial y}{\partial \hat{\eta}_j}$, depends on the sign of wedges. However, we can provide a sharp characterization in a few important special cases.

Proposition 12. *Suppose that there are no profit taxes, $\tau_i^{\Pi} = 0$, and all other wedges are positive, $\tau_{ij}^X > 0$, $\tau_i^L > 0$, and $\tau_i^S > 0$ for all i, j, and suppose that some of the wedge income is rebated to the household, $\log \Gamma_\tau < 0$. Then any marginal increase in the returns to scale leads to an increase in GDP, $\frac{\partial y}{\partial \hat{\eta}_j} > 0$.*

Consider first the case with no profit taxes. If other wedges are positive, the equilibrium returns to scale are too low (Proposition 10) as the firms do not internalize that part of the wedge income is rebated to the household. Then, any change in the parameter that leads to an increase in returns to scale is beneficial for GDP. For example, if the economy becomes more productive, as captured by a higher μ_j, the equilibrium returns to scale increase (Corollary 6).[55] Such a change has a positive impact on GDP because the equilibrium returns to scale were inefficiently low before the change.

Profit taxes affect the equilibrium returns to scale differently. As Proposition 10 suggests, an increase in the profit tax τ_i^{Π} is equivalent, from the firms' perspective, to an in increase in the entry cost κ_i. Such an increase typically leads to a higher $\hat{\eta}_i$ (see our discussion following Corollary 9). Therefore, if profit taxes are rebated to the household, equilibrium returns to scale tend to be inefficiently high as firms incorrectly perceive entry costs as being too high. In that case, any change in the parameter that leads to a further increase in returns to scale is harmful for GDP. Expression (86) in the proof of Proposition 12 provides an exact expression for $\frac{\partial y}{\partial \hat{\eta}_j}$ in that case.

D.5.1 Proof of Proposition 10

Proposition 10. *An increase in wedges $\left\{\tau^X, \tau^L, \tau^S\right\}$ reduces returns to scale in all sectors. An increase in the profit tax τ_i^{Π} reduces returns to scale in other sectors but can increase returns to scale in sector i.*

Proof. Taking first-order conditions of (77) with respect to L_{il} and X_{il}, we can derive the following

[55] If $\chi \in \left\{\mu_j, \sigma_j^2, \kappa_j, \gamma_j\right\}$, it is straightforward to show that $\frac{\partial y}{\partial \chi}$ and $\frac{d\hat{\eta}_j}{d\chi}$ are given by the same expressions as in the baseline model.

expression for log profit of firm l in sector i:

$$\log \Pi_{il} = \log P_i + \log\left(1-\tau_i^S\right) + \frac{a_i\left(\eta_{il}\right) + \eta_{il}\left(\log \frac{P_i}{W} - \sum_{j=1}^N \alpha_{ij} \log \frac{P_j}{W}\right)}{1-\eta_{il}}$$
$$+ \frac{\varepsilon_{il} - \eta_{il}\left(\log \frac{\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}}{\left(1-\tau_i^S\right)}\right)}{1-\eta_{il}}.$$

Then the first-order condition with respect to η_{il} yields

$$\frac{d \log \Pi_{il}}{d\eta_{il}} = 0 \Leftrightarrow \tag{82}$$
$$\varepsilon_{il} - \log\left[\frac{\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}}{\left(1-\tau_i^S\right)}\right] + \log\frac{P_i}{W} - \sum_{j=1}^N \alpha_{ij} \log \frac{P_j}{W} + a_i\left(\eta_{il}\right) + \left(1-\eta_{il}\right)\frac{da_i}{d\eta_{il}} = 0.$$

Following the same steps as in the proof of Proposition 2, we can derive that the the equilibrium price vector is given by (78), and the sales-weighted average of firm-level returns to scale $\hat{\eta}_i$ satisfies

$$\frac{1}{1-\eta_{il}} = \frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\varepsilon_{il} - \tilde{\mu}_i}{2\gamma_i}, \tag{83}$$

where $\tilde{\mu_i} = \mu_i - \log\left[\frac{\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}}{\left(1-\tau_i^S\right)}\right]$. Plugging (78) and (83) into (82), we get the following equation for $\hat{\eta}$:

$$\mathcal{F}_i = \frac{da_i\left(\hat{\eta}_i\right)}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2}\frac{1}{\left(1-\hat{\eta}_i\right)^2} + \alpha_i^\top \mathcal{L}\left(\hat{\eta}\right) z\left(\hat{\eta}\right) + \frac{1}{2}\log\left(1-\varphi_i\right) + \log\frac{\kappa_i}{1-\tau_i^\Pi} \tag{84}$$
$$- \log\left[\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}\right] = 0.$$

Denote by χ_i any of τ_{ij}^X, τ_i^L or τ_i^S. Then, by the implicit function theorem,

$$\frac{d\hat{\eta}}{d\chi_i} = -\left[\frac{\partial \mathcal{F}}{\partial \hat{\eta}}\right]^{-1}\left[\frac{\partial \mathcal{F}}{\partial \chi_i}\right].$$

As in the baseline model, we have

$$\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_i} = \left(1-\varphi_i\right)\frac{d^2 a_i}{d\hat{\eta}_i^2}$$

and $\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_j} = 0$ if $i \neq j$. Furthermore,

$$\frac{\partial \mathcal{F}_i}{\partial \chi_k} = \alpha_i^\top \mathcal{L}(\hat{\eta}) \frac{\partial z(\hat{\eta})}{\partial \chi_k} - \frac{\partial \log \left[\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}} \right]}{\partial \chi_k}.$$

From (79), it is clear that $\frac{\partial z_k}{\partial \chi_k} < 0$. Therefore, $\frac{d\hat{\eta}}{d\chi_i} \leq 0$. In particular, we have

$$\frac{d\hat{\eta}_i}{d\tau_j^S} = -\frac{1}{1-\tau_j^S} \frac{d\hat{\eta}_i}{d\log\left(1-\tau_j^S\right)} = -\frac{1}{1-\tau_j^S} \left[(1-\varphi_i) \frac{d^2 a_i}{d\hat{\eta}_i^2} \right]^{-1} \mathcal{K}_{ij}.$$

□

D.5.2 Proof of Proposition 11

Proposition 11. *Equilibrium log GDP $y := \log Y$ is given by*

$$y(\hat{\eta}) = \underbrace{\beta^\top \mathcal{L}(\hat{\eta}) z(\hat{\eta})}_{\textit{Contribution of productivity}} + \underbrace{\log \bar{L}}_{\textit{Labor endowment}} - \underbrace{\log \Gamma_\tau}_{\textit{Wedges income}},$$

where

$\Gamma_\tau =$

$$1 - \sum_{i=1}^N \tilde{\omega}_i \left(\left(1-\tau_i^S\right) \hat{\eta}_i \left(\sum_{j=1}^N \alpha_{ij} \frac{\theta_{ij}^X \tau_{ij}^X}{1+\tau_{ij}^X} + \left(1 - \sum_{j=1}^N \alpha_{ij}\right) \frac{\theta_i^L \tau_i^L}{1+\tau_i^L} \right) + \theta_i^S \tau_i^S + \theta_i^\Pi \tau_i^\Pi \left(1-\tau_i^S\right)(1-\hat{\eta}_i) \right).$$

Proof. From the household's budget constraint, we have

$$Y = W\bar{L} + \mathcal{T}, \tag{85}$$

where

$$
\begin{aligned}
\mathcal{T} &= \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \theta_{ij}^{X} \tau_{ij}^{X} P_j X_{ij} + \theta_i^L \tau_i^L W L_i + \theta_i^S \tau_i^S P_i Q_i + \theta_i^{\Pi} \tau_i^{\Pi} \Pi_i \right) \\
&= \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \hat{\eta}_i \alpha_{ij} \frac{\theta_{ij}^{X} \tau_{ij}^{X}}{1+\tau_{ij}^{X}} \left(1-\tau_i^S\right) P_i Q_i + \hat{\eta}_i \left(1 - \sum_{j=1}^{N} \alpha_{ij}\right) \frac{\theta_i^L \tau_i^L}{1+\tau_i^L} \left(1-\tau_i^S\right) P_i Q_i \right. \\
&\left. + \theta_i^S \tau_i^S P_i Q_i + \theta_i^{\Pi} \tau_i^{\Pi} \left(1-\tau_i^S\right) \left(1-\hat{\eta}_i\right) P_i Q_i \right) \\
&= Y \sum_{i=1}^{N} \tilde{\omega}_i \left(\left(1-\tau_i^S\right) \hat{\eta}_i \left(\sum_{j=1}^{N} \frac{\alpha_{ij} \theta_{ij}^{X} \tau_{ij}^{X}}{1+\tau_{ij}^{X}} + \frac{\left(1 - \sum_{j=1}^{N} \alpha_{ij}\right) \theta_i^L \tau_i^L}{1+\tau_i^L} \right) + \theta_i^S \tau_i^S + \theta_i^{\Pi} \tau_i^{\Pi} \left(1-\tau_i^S\right) \left(1-\hat{\eta}_i\right) \right).
\end{aligned}
$$

Plugging this into (85) gives the result. □

D.5.3 Proof of Proposition

Proposition 10. *Suppose that there are no corporate taxes, $\tau_i^{\Pi} = 0$, and all other wedges are positive, $\tau_{ij}^X > 0$, $\tau_i^L > 0$, and $\tau_i^S > 0$ for all i, j, and suppose that some of the wedge income is rebated to the household, $\log \Gamma_\tau < 0$. Then any marginal increase in the returns to scale leads to an increase in GDP, $\frac{\partial y}{\partial \hat{\eta}_j} > 0$.*

Proof. Differentiating y, given by (81), with respect to $\hat{\eta}_i$ (noting that the Envelope Theorem eliminates the productivity terms at the firm's optimum), we get

$$
\frac{\partial y}{\partial \hat{\eta}_i} = -\frac{\partial \log \Gamma_\tau}{\partial \hat{\eta}_i} = \frac{\mathrm{Num}_i}{\Gamma_\tau},
$$

where the numerator is

$$
\begin{aligned}
\mathrm{Num}_i &= \tilde{\omega}_i \left(1-\tau_i^S\right) \left(\sum_{j=1}^{N} \alpha_{ij} \frac{\theta_{ij}^{X} \tau_{ij}^{X}}{1+\tau_{ij}^{X}} + \left(1 - \sum_{j=1}^{N} \alpha_{ij}\right) \frac{\theta_i^L \tau_i^L}{1+\tau_i^L} - \theta_i^{\Pi} \tau_i^{\Pi} \right) \\
&+ \sum_{k=1}^{N} \frac{d\tilde{\omega}_k}{d\hat{\eta}_i} \left(\left(1-\tau_k^S\right) \hat{\eta}_k \left(\sum_{j=1}^{N} \alpha_{kj} \frac{\theta_{kj}^{X} \tau_{kj}^{X}}{1+\tau_{kj}^{X}} + \left(1 - \sum_{j=1}^{N} \alpha_{kj}\right) \frac{\theta_k^L \tau_k^L}{1+\tau_k^L} \right) + \theta_k^S \tau_k^S + \theta_k^{\Pi} \tau_k^{\Pi} \left(1-\tau_k^S\right) \left(1-\hat{\eta}_k\right) \right).
\end{aligned}
$$

The derivative of the Domar weights is given by

$$
\frac{d\tilde{\omega}_k}{d\hat{\eta}_i} = \sum_{j=1}^{N} \tilde{\alpha}_{ij} \left(I - \mathrm{diag}\left[\left(1-\theta^S\right) \circ \tau^S\right] - \mathrm{diag}\left(\hat{\eta}\right) \tilde{\alpha}\right)_{jk}^{-1} \tilde{\omega}_i > 0.
$$

Therefore, if taxes are positive ($\tau^X, \tau^L, \tau^S > 0$) but there is no profit tax ($\tau^{\Pi} = 0$), all terms in the numerator are positive (assuming some rebates $\theta > 0$), implying $\frac{\partial y}{\partial \hat{\eta}_i} > 0$. □

In contrast, if $\tau^X = \tau^L = \tau^S = 0$ and $\tau^\Pi > 0$ with $\theta^\Pi = 1$, then $\tilde{\omega} = \omega$, and we get

$$\frac{\partial y}{\partial \hat{\eta}_i} = \omega_i \frac{-\tau_i^\Pi + \sum_{k=1}^N \left(\sum_{j=1}^N \alpha_{ij} \mathcal{L}_{jk} \right) \tau_k^\Pi \left(1 - \hat{\eta}_k\right)}{1 - \sum_{i=1}^N \omega_i \tau_i^\Pi \left(1 - \hat{\eta}_i\right)}. \tag{86}$$

As $\hat{\eta}_k \to 1$ for all k, the term $(1 - \hat{\eta}_k)$ vanishes, leaving only the negative term $-\omega_i \tau_i^\Pi$. Thus, $\frac{\partial y}{\partial \hat{\eta}_i}$ becomes negative.

D.6 Sales wedge correlated with productivity

In this appendix, we consider an economy in which firms face sales tax (38). The firm's problem (2) becomes

$$\Pi_{il} := \max_{\eta_{il}, L_{il}, X_{il}} \left(1 - \tau_{il}^S\right) P_i F_i \left(L_{il}, X_{il}, \eta_{il}\right) - W L_{il} - \sum_{j=1}^N P_j X_{ij,l},$$

where $F_i \left(L_{il}, X_{il}, \eta_{il}\right)$ is given by (1). Clearly, this problem is equivalent to the one in the main text if we redefine the productivity as

$$\tilde{\varepsilon}_{il} = \varepsilon_{il} + \log\left(1 - \tau_{il}^S\right) = \left(1 - b_i\right)\left(\varepsilon_{il} - \mu_i\right) + \mu_i + \log\left(1 - \tau_i^S\right),$$

such that $\tilde{\varepsilon}_{il} \sim \text{iid } \mathcal{N}\left(\tilde{\mu}_i, \tilde{\sigma}_i^2\right)$, where

$$\tilde{\mu}_i = \mu_i + \log\left(1 - \tau_i^S\right) \quad \text{and} \quad \tilde{\sigma}_i = \left(1 - b_i\right) \sigma_i.$$

Similar to the baseline model, define

$$\hat{\eta}_i := \int_0^{M_i} \frac{\tilde{\omega}_{il}}{\tilde{\omega}_i} \eta_{il} dl, \tag{87}$$

with $\tilde{\omega}_{il} := \left(1 - \tau_{il}\right) \omega_{il}$ and $\tilde{\omega}_i := \left(1 - \hat{\tau}_i^S\right) \omega_i$, where

$$\hat{\tau}_i^S = \int_0^{M_i} \frac{\omega_{il}}{\omega_i} \tau_{il}^S dl. \tag{88}$$

As in the baseline model, we get

$$\frac{1}{1 - \hat{\eta}_i} = \frac{\tilde{\mu}_i + s_i}{2\gamma_i \left(1 - \tilde{\varphi}_i\right)},$$

where $s_i = \log P_i - \log H_i$. Integrating (88), we get

$$\frac{1}{1 - \hat{\tau}_i^S} = \frac{1}{1 - \tau_i^S} \left(1 + \frac{\tilde{\varphi}_i}{\frac{1 - \tilde{\varphi}_i}{1 - \hat{\eta}_i}} \frac{b_i}{1 - b_i}\right) \exp\left(-\frac{b_i}{1 - b_i} \frac{4\tilde{\varphi}_i \gamma_i \frac{1 - \tilde{\varphi}_i}{1 - \hat{\eta}_i} + \tilde{\sigma}_i^2 \frac{b_i}{1 - b_i}}{2} \frac{1}{\left(1 - \tilde{\varphi}_i\right)}\right). \tag{89}$$

Then, following the same steps as in the main model, we can derive

$$\log W = \beta^{\top} \left(I - \operatorname{diag}(\hat{\eta})\, \alpha\right)^{-1} \tilde{z}(\hat{\eta}),$$

where

$$\tilde{z}_i(\hat{\eta}_i) = \tilde{\mu}_i + a_i(\hat{\eta}_i) + \frac{\tilde{\sigma}_i^2}{2} \frac{1}{1-\hat{\eta}_i} + \frac{1}{2}(1-\hat{\eta}_i) \log\left(\frac{1}{1-\tilde{\varphi}_i}\right) - (1-\hat{\eta}_i) \log \kappa_i$$

and $\tilde{\varphi}_i = \frac{\tilde{\sigma}_i^2}{2\gamma_i}$.

We assume that all tax proceeds are rebated to the household. Therefore, using the market clearing condition (9), we get

$$\omega_i := \frac{P_i Q_i}{\bar{P} Y} = \beta^{\top} \left(I - \operatorname{diag}\left(1 - \hat{\tau}^S\right) \operatorname{diag}(\hat{\eta})\, \alpha\right) \mathbf{1}_i.$$

The rebate amount is

$$\mathcal{T} = \sum_{i=1}^{N} \int_0^{M_i} \tau_{il}^S P_i Q_{il} dl = \sum_{i=1}^{N} \hat{\tau}_i^S \omega_i \bar{P} Y.$$

GDP is then

$$C = W\bar{L} + \mathcal{T} \Leftrightarrow \log C = \log W + \log \bar{L} - \log\left(1 - \sum_{i=1}^{N} \hat{\tau}_i^S \omega_i\right).$$

D.7 Dispersed returns-to-scale economy

In this appendix, we consider the dispersed returns-to-scale economy. Specifically, consider the initial economy (we will use subscripts b to mark any quantities in that economy). From (10), firm l in sector i chooses the following returns to scale:

$$\frac{1}{1-\eta_{il}^b} = \frac{1}{2\gamma_i}\left(\varepsilon_{il}^b + s_i^b\right), \tag{90}$$

where $s_i^b = \log P_i^b - \log H_i^b$ in the initial economy. Furthermore, from (17), we know that

$$\frac{1}{1-\hat{\eta}_i^b} = \frac{1}{2\gamma_i\left(1-\varphi_i^b\right)}\left(\mu_i^b + s_i^b\right), \tag{91}$$

where $\varphi_i^b = \frac{\left(\sigma_i^b\right)^2}{2\gamma_i}$.

Suppose now that there is a change in the distribution of ε_{il}^b, such that the mean changes from μ_i^b to μ_i, and the standard deviation changes from σ_i^b to σ_i. Such a change can reflect an increase in μ for all sectors (Section 7.3) or removal of sales tax (Section 7.4). Then, productivity ε_{il}^b shifts to ε_{il}, where

$$\frac{\varepsilon^b - \mu_i^b}{\sigma_i^b} = \frac{\varepsilon_{il} - \mu_i}{\sigma_i}.$$

In the dispersed economy, firms can adjust all their choices except returns to scale. The free-entry condition (8)

$$\int_{-\infty}^{\infty} \underbrace{\exp\left(\log P_i + \frac{\varepsilon_{il} - a_i\left(\eta_{il}^b\right) + \eta_{il}^b s_i}{1-\eta_{il}^b}\right)}_{=\Pi_{il}\left(\varepsilon_{il},\eta_{il}^b\right)} f_i\left(\varepsilon_{il}\right) d\varepsilon_{il} = \kappa_i W,$$

where $s_i = \log P_i - \log H_i$, and η_{il}^b is given by (90). Taking this integral, we get

$$\exp\left(\frac{\left[\left(\mu_i^b + s_i^b\right)\left(1-\varphi_i^b\frac{\sigma_i}{\sigma_i^b}\right) + \varphi_i^b\left(\mu_i + s_i\right)\right]^2 \frac{1}{1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)} - \left(\mu_i^b + s_i^b\right)^2}{2\left(\sigma_i^b\right)^2}\right) \times \tag{92}$$

$$\exp\left(\sum_{j=1}^{N} \alpha_{ij}\log\frac{P_j}{W}\right)\frac{1}{\sqrt{1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)}} = \kappa_i.$$

Next, we can define $\hat{\eta}_i$ in the same way as usual, $\hat{\eta}_i = \int_0^{M_i}\frac{\omega_{il}}{\omega_i}\eta_{il}^b dl$, where again η_{il}^b is given by (90). Omitting tedious yet straightforward calculations, we get

$$\frac{1}{1-\hat{\eta}_i} = \frac{1}{2\gamma_i\left(1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)\right)}\left[\left(\mu_i^b + s_i^b\right) + \varphi_i^b\left(\mu_i + s_i - \left(\mu_i^b + s_i^b\right)\frac{\sigma_i}{\sigma_i^b}\right)\right]. \tag{93}$$

Combining (91), (92), and (93), we get

$$\log W = \beta^{\top}\left(I - \operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1} z,$$

where

$$\begin{aligned} z_i =& \mu_i - \left[\frac{1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)}{1-\hat{\eta}_i} - 2\frac{1-\varphi_i^b}{1-\hat{\eta}_i^b}\left(1-\varphi_i^b\frac{\sigma_i}{\sigma_i^b}\right) + \left(1-\hat{\eta}_i\right)\left(\frac{1-\varphi_i^b}{1-\hat{\eta}_i^b}\right)^2\right]\frac{\gamma_i}{\varphi_i^b} \\ & - \left(1-\hat{\eta}_i\right)\left[\frac{1}{2}\log\left(1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)\right) + \log\kappa\right]. \end{aligned}$$

GDP is then $Y = W\bar{L}$.