Customer Accumulation, Returns to Scale, and Secular Trends

Andrea Chiavari

Online Appendix

Contents

I Data, Secular Trends, and Returns to Scale

I.I Main Sample, Variables Construction, and Summary Statistics

I.I.I Data cleaning, main variables, and summary statistics.

I use Compustat from 1977 to 2014. I drop all firms whose Foreign Incorporation Code (FIC) is not equal to USA. Then, I linearly interpolate when there is one missing between two available data points SALE, COGS, XSGA, EMP, PPEGT, PPENT, XRD, XLR, XPR, XRENT, RECD, DP for data quality. I exclude utilities (NAICS code 22) due to heavy price regulation, and financial and insurance firms (NAICS code 52) because their balance sheets differ substantially from those of non-financial firms.

To construct the firm-level total stock of capital, I use the perpetual inventory method (PIM). In particular, with PIM, capital is defined as:

$$k_{it} = (1-\delta)k_{it-1} + x_{it}, \tag{1}$$

where $x_{it} - \delta k_{it-1} = \text{PPENT}_{it} - \text{PPENT}_{it-1}$ is the net investment, and the initial capital stock, k_{i0}, is initialized using the first available entry of PPEGT.

For data quality, I interpret as mistakes zero or negative in SALE, k, EMP, or XSGA, and I drop those observations; moreover, if SALE, k, EMP are missing, I drop these observations too; however, if XSGA is missing, I set it to zero. Finally, if XRD, XLR, XPR, XRENT, RECD, or DP are negative or missing, I treat them as zeros. To obtain a real measure of the main variables, I deflate them with the GDP deflator; I deflate investment and capital stock by the investment good deflator.[1] Table I presents a few basic summary statistics for a few leading variables used in the analysis.

I.I.II Selling costs.

I present the two main approaches used to measure firm-level expenditures on customer accumulation. As Compustat lacks a direct measure of this variable, I highlight the strengths and limitations of each approach in turn.

Advertisement. The XAD variable in Compustat captures firms' reported advertising ex-

[1] Deflators are taken from the NIPA tables.

Table I: Summary Statistics (1977-2014)

	Sales	**Cost of Goods Sold**	**Employment**	**Capital Stock (*Book Value*)**	**Capital Stock (*PIM*)**	**Age**
Mean	1,873,553	1,296,868	7,056	1,005,617	728,260	13
25^{th} *Percentile*	22,553	13,896	115	5,756	3,552	5
Median	139,060	84,909	638	36,079	24,323	11
75^{th} *Percentile*	751,619	483,007	3,500	241,352	169,204	19
No. Obs.	168,496	168,496	168,496	167,884	168,496	168,496

Note. Summary statistics of cleaned Compustat dataset between 1977 and 2014. All variables except for Age are in thousands US$. Sales and Costs of Goods Sold are deflated with the GDP deflator using the base year 2012, whereas both types of capital stocks are deflated using the investment deflator with the base year 2012.

penditures, offering insight into the costs incurred through various promotional activities. While it serves as a useful proxy for selling costs, XAD has notable limitations that must be considered when interpreting the data.

Typically, XAD includes spending on advertising through traditional media—such as television, radio, print, and outdoor billboards—as well as promotional expenditures (Landes and Rosenfield, 1994; Belo et al., 2014; Vitorino, 2014). These are direct, explicitly reported expenses tied to marketing campaigns, making XAD a natural starting point for analyzing firm-level selling costs.

However, XAD may not capture the full scope of marketing-related costs. For example, expenditures on in-house advertising teams—such as salaries and benefits—are typically recorded under general personnel costs rather than advertising. Moreover, reporting of XAD varies across firms and industries, with some companies omitting this information altogether. This heterogeneity results in missing or sparse data, which can limit both the precision and representativeness of the measure over time and across sectors.

Adjusted SG&A. As an alternative to the advertising-based measure, an adjusted version of Selling, General, and Administrative Expenses (XSGA) is used. This variable has attracted considerable attention in recent studies, including Gourio and Rudanko (2014), Ptok et al. (2018), Afrouzi et al. (2020), and Morlacco and Zeke (2021). Notably, Ptok et al. (2018) find that XSGA is particularly effective in capturing firm-level sales force expenditures.

However, XSGA includes a range of expenses not directly related to selling activities—such as bad debt expenses, pension and retirement costs, rent, and research and development expenditures. For a breakdown of the components of XSGA, refer to Afrouzi et al. (2020). To partially address these limitations, I construct an adjusted measure as follows:

$$\mathrm{S}_{it} = \mathrm{XSGA}_{it} - \mathrm{XRENT}_{it} - \mathrm{XPR}_{it} - \mathrm{RECD}_{it} - \mathrm{XRD}_{it}, \tag{2}$$

where XSGA denotes total SG&A expenses, XRENT is rent expenditure, XPR captures pension and retirement costs, RECD reflects bad debt expenses, and XRD corresponds to R&D spending. Whenever the adjusted value S_{it} is negative, it is set to zero.

I.I.III User Cost of Capital.

One approach used in this paper to estimate the production function is the cost share method, which requires a measure of the user cost of capital. To construct this measure, I follow the standard procedure in the literature (Hall and Jorgenson, 1967; De Loecker et al., 2020) and use the following expression:

$$r_t = i_t - \mathbb{E}_t \pi_{t+1} + \delta, \tag{3}$$

where i_t is the nominal interest rate, $\mathbb{E}_t \pi_{t+1}$ is the expected inflation rate at time t, and δ is the depreciation rate of capital. Following Barkai (2020), I use the annual Moody's Seasoned Aaa Corporate Bond Yield as a proxy for the nominal interest rate, and calculate expected inflation using the annual growth rate of the Investment Nonresidential Price Deflator. The depreciation rate is calibrated to $\delta = 0.1$, consistent with the rest of the paper.[2,3,4]

I.I.IV Variable input in production.

Recent studies, following De Loecker et al. (2020), have adopted the cost of goods sold (COGS) variable from Compustat as the preferred proxy for variable inputs in production. This choice is motivated by the need for a bundled measure of variable input expenditures when computing firm-level markups. However, despite its usefulness for this purpose, using COGS necessitates estimating a gross output production function, which poses identification challenges (Gandhi et al., 2020). Moreover, this approach requires the strong assumption of perfect substitutability between labor and materials.

[2] Moody's Seasoned Aaa Corporate Bond Yield: https://fred.stlouisfed.org/series/AAA

[3] Investment Price Deflator: https://fred.stlouisfed.org/series/A008RD3Q086SBEA

[4] I estimate an AR(1) process on the annual growth rate of the Investment Nonresidential Price deflator and define the contemporaneous expected inflation as $\mathbb{E}_t \pi_{t+1} = \mu + \rho \pi_t$.

Given that the primary aim of this paper is to estimate returns to scale and output elasticities—rather than markups—I adopt a direct measure of variable input that avoids these complications. Specifically, I use EMP, the number of employees at the firm level, as the benchmark measure of variable input.

To compute cost shares consistent with this measure, I construct firm-level labor costs as $w_{it}\ell_{it}$. For firms that report labor expenditures (XLR), I calculate the wage per worker as $w_{it} = \text{XLR}_{it}/\text{EMP}_{it}$. I then compute the within-sector median wage $\widehat{w}_{st}$ across all firms in a given sector and use it to impute labor costs for firms that do not report XLR, using the formula $w_{it}\ell_{it} = \widehat{w}_{st} \cdot \text{EMP}_{it}$.

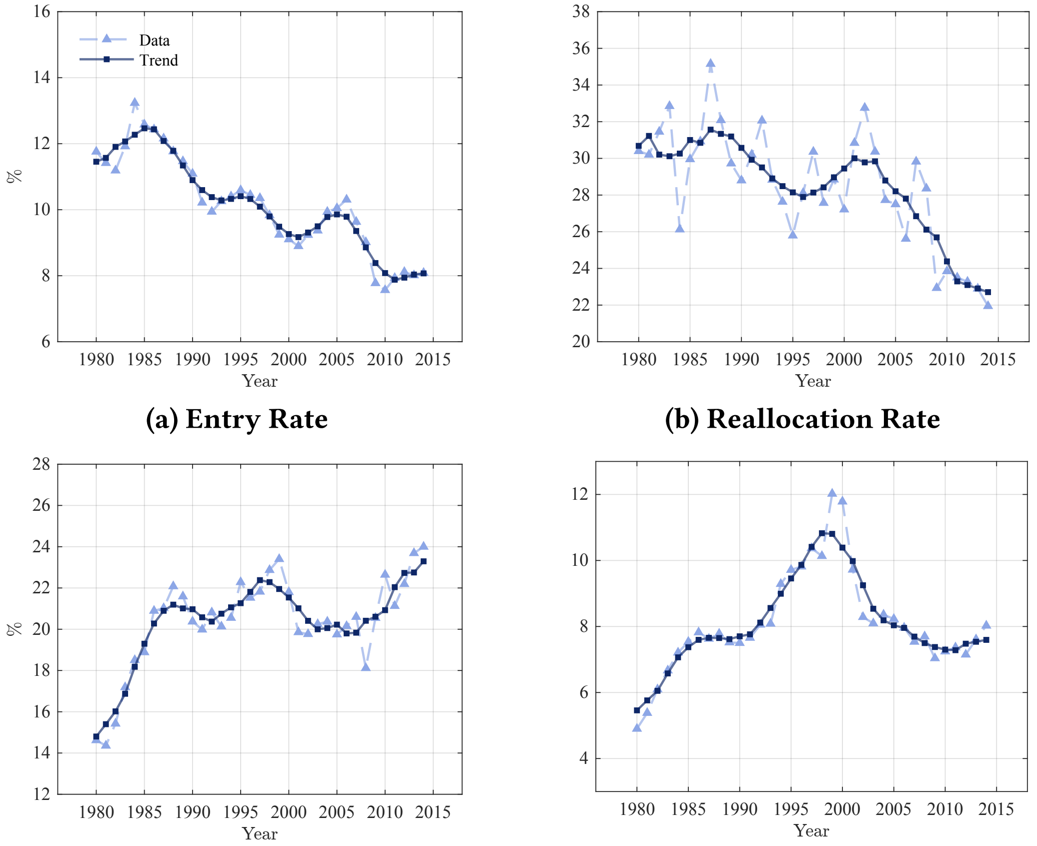
I.II Secular Trends

Figure I shows the evolution of the entry rate, reallocation rate, markups, and selling costs relative to production costs from 1980 to 2014. Dashed light blue lines with triangles represent the raw data, while solid dark blue lines with squares show the smoothed series using a 5-year moving average. All variables are expressed as percentages. Entry and reallocation rates are sourced directly from the BDS. Markups are cost-weighted averages from Compustat, measured following De Loecker et al. (2020). Selling costs relative to production costs are constructed as a simple average of the ratio of advertising expenditures to the cost of goods sold, also using Compustat data.

The entry rate declined from 12% to 8%, a drop of 33%. The reallocation rate fell from 31% to 22%, representing a 29% decrease. Markups rose from 15% to 23%, an increase of 53%. Finally, selling costs relative to production costs, measured as costs of goods sold, increased from 5% to 8%, a rise of 60%.

Figure II shows the evolution of an alternative measure of selling costs relative to production costs, based on adjusted SG&A rather than advertising expenditures. This measure also displays a substantial increase, rising from 50% to 95%, which represents a 90% increase.

Figure I: Secular Trends over Time



(a) Entry Rate

(b) Reallocation Rate

(c) Markups

(d) Selling Costs over Production Costs

Note. Figure I presents the evolution of the entry rate, reallocation rate, markups, and selling costs relative to production costs from 1980 to 2014. Dashed light blue lines with triangles represent the raw data, while solid dark blue lines with squares show the smoothed series using a 5-year moving average. All variables are expressed as percentages. Entry and reallocation rates are taken directly from the BDS. Markups and selling costs are constructed from Compustat data, with the former aggregated using cost weights and the latter calculated as a simple average.

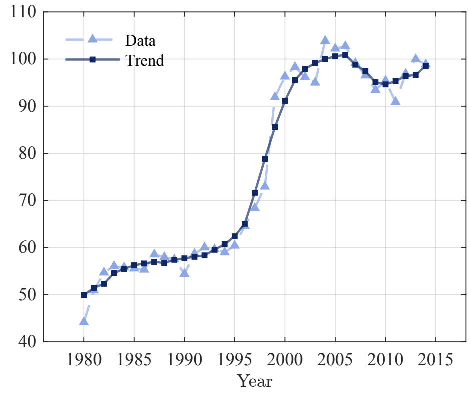
I.III Returns to Scale

I.III.I Production Function Estimation

To estimate returns to scale, I follow De Loecker et al. (2020) and employ the control function approach proposed by Ackerberg et al. (2015). To sidestep the measurement issues associated with materials in Compustat and the identification challenges inherent in gross output production functions (Gandhi et al., 2020), I adopt the following structural value-added production function as the preferred specification:

$$\mathcal{Q}_{it} = \min\left\{K_{it}^{\beta^k} L_{it}^{\beta^\ell} \exp(z_{it} + \varepsilon_{it}),\ \beta^m M_{it}\right\}, \tag{4}$$

Figure II: Alternative Selling Costs to Production Costs over Time



(a) Selling Costs over Production Costs

Note. Figure II displays the alternative selling costs relative to production costs. The light blue line with triangles depicts the data, while the dark blue line with squares illustrates the trend calculated as a 5-year moving average. Selling costs over production costs have been calculated using the alternative measure of adjusted SG&A.

where Q_{it} is output, K_{it} is capital, L_{it} is labor, z_{it} is log-productivity, ε_{it} is the error term, and M_{it} is materials. Despite the advantages of the structural value-added specification, Section I.III.III of the Appendix shows that the results remain robust to alternative specifications of the production function such a gross output and a translog specification. This structural value-added production function yields the following first-order condition:

$$Q_{it} = K_{it}^{\beta^k} L_{it}^{\beta^\ell} \exp(z_{it} + \varepsilon_{it}), \tag{5}$$

This approach justifies regressing output Q_{it}—rather than value-added—on capital and labor, while omitting materials. Robustness checks, reported in Section I.III.III of the Appendix, show that using value-added instead yields similar results. Therefore, under the specification in equation (4), the estimation of the firm-level production function simplifies to:

$$q_{it} = \beta^k k_{it} + \beta^\ell \ell_{it} + z_{it} + \varepsilon_{it}, \tag{6}$$

where $q_{it} = \log(Q_{it})$, $k_{it} = \log(K_{it})$, and $\ell_{it} = \log(L_{it})$. The main challenge in estimating the production function is the *simultaneity bias* arising from the unobserved, time-varying firm-level productivity term, z_{it}. Although equation (6) relates physical output to inputs, in practice the available data only allow the estimation of a relationship between sales and input.

This leads to the following revenue-based production function:

$$\begin{aligned} p_{it} + q_{it} &\equiv y_{it} \\ &= \beta^k k_{it} + \beta^\ell \ell_{it} + z_{it} + p_{it} + \varepsilon_{it}, \end{aligned} \tag{7}$$

where p_{it} denotes the log of the firm's output price, y_{it} is the log of sales, p_t^k is the common log user cost of capital, and p_t^ℓ is the log input price of labor. Consequently, researchers must address not only the simultaneity bias but also the *omitted price bias* (Klette and Griliches, 1996; Bond et al., 2021) to accurately estimate the elasticities in equation (6). As emphasized by De Ridder et al. (2022), omitted price bias can lead to downwardly biased production elasticities, particularly when firms face persistent demand shocks under downward-sloping demand curves or increasing returns to scale. For instance, with increasing returns to scale, positive shocks lower marginal costs and prices, generating a *negative* correlation between prices and inputs. Correcting for this bias is thus crucial to avoid systematically underestimating returns to scale.

To address both biases in estimating equation (7), I follow the control function literature, which relies on the insight that a firm's labor demand can be expressed as $\ell_{it} = \ell(k_{it}, z_{it}, d_{it})$. Under standard regularity conditions, there exists a one-to-one mapping between productivity and labor input, conditional on capital and demand shifters, allowing for control of simultaneity bias, since $z_{it} = \ell^{-1}(k_{it}, \ell_{it}, d_{it})$. The term d_{it} captures factors related to output and input markets that generate variation in labor demand across firms, conditional on productivity and capital. Incorporating d_{it} is crucial for allowing imperfect competition in product markets, ensuring invertibility of the function $\ell(\cdot)$, and addressing the omitted price bias. In particular, observable variables governing cost pass-through—such as market shares—can serve this role (De Loecker et al., 2020), since firms with identical productivity may transmit input cost shocks differently depending on their market position. This strategy enables the separation of price measurement errors from the estimation of production function parameters.

In practice, the production function is estimated in two stages. In the first stage, output is purged of measurement error and unanticipated productivity shocks by regressing it on a second-order polynomial of capital and labor, $\phi(k_{it}, \ell_{it})$, along with market shares:

$$y_{it} = \phi(k_{it}, \ell_{it}) + \gamma d_{it} + \varepsilon_{it}. \tag{8}$$

In the second stage, using purged output $\widehat{\phi}_{it}$, I construct a productivity measure independent of both the measurement error ε_{it} or the price component, as captured by γd_{it}, given by:

$$z_{it}(\beta^k, \beta^\ell) = \widehat{\phi}_{it} - \beta^k k_{it} - \beta^\ell \ell_{it}. \tag{9}$$

Finally, assuming an AR(1) process, shocks to productivity are given by:

$$\xi(\beta^k, \beta^\ell, \rho) = z_{it}(\beta^k, \beta^\ell) - \rho z_{it-1}(\beta^k, \beta^\ell). \tag{10}$$

Therefore, using the productivity shocks, a set of moment conditions can be constructed to estimate the parameters of the production function, given by:

$$\mathbb{E}(\xi(\beta^k, \beta^\ell, \rho) \times \mathbf{z}_{it}) = \mathbf{0}_{Z\times 1}, \tag{11}$$

where Z ≥ 3 and, under the assumption that firms react to unanticipated productivity shocks contemporaneously and that capital is predetermined, the set of admissible instruments is $\mathbf{z}_{it}$ $\in \{k_{it}, \ell_{it-1}, k_{it-1}, \ldots\}$. Returns to scale are recovered as $\alpha = \beta^k + \beta^\ell$. In Compustat, y_{it} is measured by log-sales, k_{it} by log-capital, ℓ_{it} by the number of log-employees, and d_{it} by log-sales shares.

I.III.II Returns to Scale Estimates

Figure IIIa shows the evolution of average sales-weighted returns to scale across two-digit NAICS industries, constructed as follows:

$$\alpha_t = \sum_s \omega_{st} \alpha_{st}, \tag{12}$$

where ω_{st} denotes the sales share of sector s in year t, and α_{st} is the sector-level estimate of returns to scale.

In 1980, the average returns to scale is close to 1, suggesting firms operated under a *constant* returns to scale technology.[5] By 2014, the estimate rises by 5% to approximately 1.05, indicating that firms operate an *increasing* returns to scale production technology.

[5] These results align with findings by Gao and Kehrig (2017), who report nearly constant returns to scale for the 1982–1987 period using U.S. Census data.

The average returns to scale is a useful statistic but it does not fully capture the underlying distributional changes in returns to scale. To study the dynamics of returns to scale across sectors, I decompose the change in returns to scale as follows:

$$\Delta\alpha_t = \underbrace{\sum_s \omega_{st-1}\Delta\alpha_{st}}_{\Delta\text{within}} + \underbrace{\sum_s \Delta\omega_{st}\alpha_{st-1}}_{\Delta\text{between}} + \underbrace{\sum_s \Delta\omega_{st}\Delta\alpha_{st}}_{\Delta\text{cross term}} . \tag{13}$$

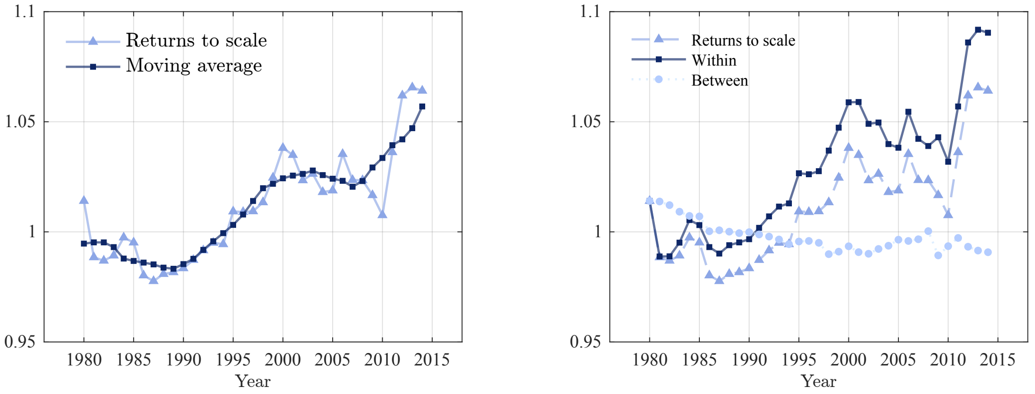
Thus, the change in average returns to scale can be exactly decomposed into three components: (i) a within component, which captures the change in average returns to scale at the sector level; (ii) a between component, which captures the change in average returns to scale due to the reallocation of economic activity toward high returns to scale sectors; and (iii) a cross-term component, which captures the change in average returns to scale due to the joint effect of returns to scale and reallocation.

Figure IIIb illustrates the decomposition of average returns to scale, alongside two counterfactual experiments: the Δwithin and the Δbetween. The Δcross-term experiment is omitted as it holds little economic interest and remains close to zero throughout the period. The analysis begins in 1980, with changes in each component from equation (13) cumulatively added.

The first experiment (solid dark blue line with squares) isolates the Δwithin component, showing that it exceeds the average returns to scale. The second experiment (dotted light blue line with circles) isolates the Δreallocation component, revealing a decreasing profile. These experiments indicate that the rise in average returns to scale is primarily driven by changes *within* sectors, while any cross-sector reallocation of economic activity has had a minor dampening effect on the increase.

Additionally, Section I.III.III uses firm-level estimates of returns to scale derived from a translog production function to show that there appears to have been little reallocation within sectors. This observation together with the finding that returns to scale have been rising within all sectors supports treating the rise in returns to scale as a homogeneous technological change affecting all firms—at least within the Compustat sample.

Figure III: Returns to Scale over Time



(a) Average Returns to Scale

(b) Within-Between Sector Decompositions

Note. Figure IIIa displays the evolution of the estimated returns to scale. The light blue line with triangles shows the point estimates, while the dark blue line with squares presents a 7-year smoothed moving average. Figure IIIb plots the counterfactual evolution implied by the decomposition in equation (13). The dashed light blue line with triangles represents the average returns to scale. The solid dark blue line with squares shows the evolution of average returns to scale driven solely by the Δwithin component, while the dotted light blue line with circles reflects the evolution driven only by the Δbetween component. Sector-level estimates are winsorized at the 1% level. Output elasticities are time-varying and calculated using rolling windows from 1980 to 2014.

I.III.III Robustness

Robustness 1 — Cost shares. The cost shares approach relies on the firm's first-order conditions, assuming constant returns to scale and all inputs being variable, to calculate output elasticities from cost shares as:

$$\theta^{\ell} = \text{median}\left\{\frac{w_{it}\ell_{it}}{w_{it}\ell_{it} + r_t k_{it}}\right\}, \quad \text{and} \quad \theta^{k} = 1 - \theta^{\ell}; \tag{14}$$

where $w_{it}\ell_{it}$ is the wage bill, and $r_t k_{it}$ is the rental cost of capital. Following Syverson (2004), returns to scale can be then measured through the following regression:

$$q_{it} = \alpha\left[\theta^{k} k_{it} + \theta^{\ell}\ell_{it}\right] + \boldsymbol{\delta}\boldsymbol{X'}_{it} + z_{it} \tag{15}$$

with all variables in logs, θ^{k} and θ^{ℓ} are given by (14), and $\boldsymbol{X}_{it}$ is a vector of controls, such as sector-level fixed effects. Thus, while each cost share determines the output elasticities, the returns to scale are captured by α, recovered via simple OLS.

Robustness 2 — Intangible capital. To assess the robustness of the rise in returns to scale to the presence of intangible capital in production I estimate a new production function that incorporates intangible capital instead of equation (6):

$$y_{it} = \beta^k k_{it} + \beta^x x_{it} + \beta^\ell \ell_{it} + z_{it} + \varepsilon_{it}, \tag{16}$$

where x_{it} represents the intangible capital in production. The returns to scale implied by the augmented production technology in equation (16) are given by $\alpha = \beta^k + \beta^x + \beta^\ell$, which can be estimated as in Section I.III.I.

To measure intangible capital I follow Chiavari and Goraya (2021) and assume that intangible capital is made of balance sheet and knowledge intangible capital. The balance sheet intangible capital is given by:

$$\chi_{it}^{balance\ sheet} = \texttt{INTAN}_{it} + \text{AM}_{it} - \text{GDWL}_{it}, \tag{17}$$

where $\texttt{INTAN}$ represents the net balance sheet intangible capital, AM is the amortization of the balance sheet intangible capital, and GDWL is goodwill. Knowledge capital is defined as:

$$\chi_{it}^{knowledge} = (1 - 0.30)\chi_{it-1}^{knowledge} + \text{XRD}_{it}, \tag{18}$$

where the depreciation rate is set to 30%, similar to the estimates by Ewens et al. (2019). Here, XRD represents the firm-level expenditure on research and development, and $\chi_{i0}^{knowledge}$ is set to zero. Finally, the total firm-level intangible capital is given by $\chi_{it} = \chi_{it}^{balance\ sheet} + \chi_{it}^{knowledge}$. To approximate log-intangible capital $x_{it} = \log(\chi_{it})$, I use the inverse hyperbolic sine transformation: $\log\left(\chi_{it} + \sqrt{\chi_{it} + 1}\right)$. This transformation retains observations with $\chi_{it} = 0$.

Robustness 3 — Translog production function. Here, I examine the robustness of the rise in returns to scale by considering an alternative specification to equation (6) using a translog production function:

$$q_{it} = \theta_1^k k_{it} + \theta_1^\ell \ell_{it} + \theta_2^k k_{it}^2 + \theta_2^\ell \ell_{it}^2 + \theta_3^{k\ell} k_{it}\ell_{it} + z_{it} + \varepsilon_{it}. \tag{19}$$

To estimate the translog production function, I follow the methodology outlined in Section I.III.I. Since the translog production function involves estimating more parameters and provides firm- and time-specific estimates, I estimate it only once in the data to maximize statistical power.[6] I then compute the median estimate within each industry and year, ensuring that the estimates are not overly influenced by outliers, given by:

$$\beta^k = \text{median}\Big\{\theta_1^k + 2\theta_2^k k_{it} + \theta_3^{k\ell}\ell_{it}\Big\} \quad \text{and} \quad \beta^\ell = \text{median}\Big\{\theta_1^\ell + 2\theta_2^\ell \ell_{it} + \theta_3^{k\ell} k_{it}\Big\}. \tag{20}$$

The returns to scale implied by the production technology in equation (19) for each sector and time are given by $\alpha = \beta^k + \beta^\ell$.

Robustness 4, 5, 6 — Alternative Variable Inputs. Recent studies have carefully examined which Compustat expenditure items should be included as measures of variable input in production, recognizing that accounting classifications do not necessarily align with economic variability. It is therefore important to assess how different input definitions affect estimates of returns to scale over time. In this study, I show that returns to scale rise consistently over time, regardless of the specific input measure used. To demonstrate this, I consider three specifications: (i) treating cost of goods sold as the sole variable input, following De Loecker et al. (2020); (ii) treating the sum of cost of goods sold and selling, general, and administrative expenses as variable input, following Traina (2018); and (iii) treating cost of goods sold as the primary input while allowing selling, general, and administrative expenses as an additional input.

In the first two cases, as opposed to equation (6), the production function takes the following form:

$$q_{it} = \beta^k k_{it} + \beta^v v_{it} + z_{it} + \varepsilon_{it}, \tag{21}$$

where v_{it} is the variable input in production. The variable input in production in these two cases is measured as:

$$\text{Case I:} \quad v_{it} = \text{COGS}_{it} \quad \text{and} \quad \text{Case II:} \quad v_{it} = \text{COGS}_{it} + \text{XSGA}_{it}; \tag{22}$$

hence, the first case assumes that the variable input in production is the cost of goods sold, as in De Loecker et al. (2020), while the second case assumes that the variable input in production is the cost of goods sold plus selling, general and administrative expenditures, as in Traina (2018). On top of these two specifications, I consider a third case where both expen-

[6] In this specification I allow for a higher-order polynomial relative to the main text to accommodate the higher degree of nonlinearities implied by this production function.

ditures enter production, but as different inputs. In this case, as opposed to equation (6), the production function is given by:

$$\text{Case III:} \quad q_{it} = \beta^k k_{it} + \beta^{v^1} v^1_{it} + \beta^{v^2} v^2_{it} + z_{it} + \varepsilon_{it}, \tag{23}$$

where v^1_{it} is COGS_{it} and v^2_{it} is XSGA_{it}.[7] The returns to scale implied by the production technology in equation (23) is given by $\alpha = \beta^k + \beta^{v^1} + \beta^{v^2}$, which can be estimated following the methodology outlined in Section I.III.I.[8]

Robustness 7 — Value-added as output. A long-standing tradition in the literature on production function estimation has focused on estimating value-added production functions, where value-added (sales net of materials) is regressed on capital and labor. Therefore, in this robustness exercise, I estimate the following specification as an alternative to equation (6) in Section I.III.I:

$$\log\left(Q_{it} - M_{it}\right) = \beta^k k_{it} + \beta^\ell \ell_{it} + z_{it} + \varepsilon_{it}, \tag{24}$$

where $\log\left(Q_{it} - M_{it}\right)$ represents the logarithm of value-added. Since materials are not directly reported in Compustat but are included within the cost of goods sold alongside labor costs, an imputation procedure is required. Following common practice in the literature using Compustat data, I impute material expenditures by first calculating the firm-level wage bill and then subtracting it from the cost of goods sold.

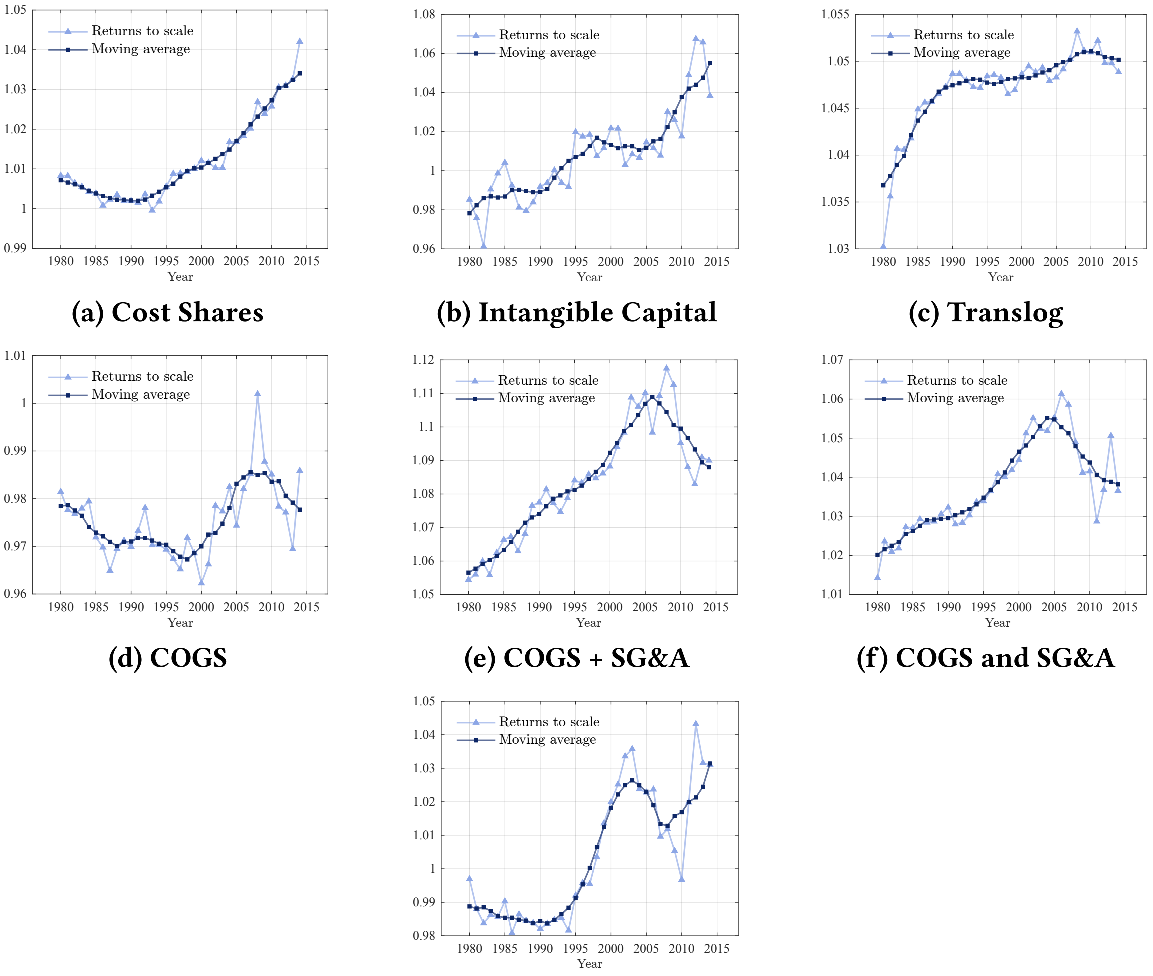
Robustness results I. Here, we present the implied rise in returns to scale across the seven robustness exercises outlined above, as shown in Figure IV. Overall, all specifications exhibit an upward trend in returns to scale, though they differ somewhat in baseline levels depending on the specific exercise. These results confirm the robustness of the rising trend in returns to scale observed in the baseline specification.

Robustness results II. Here, we examine whether the rise in returns to scale has been heterogeneous across firms, potentially favoring larger ones. To investigate this, we use estimates from the translog production function, which yield firm-level returns to scale, and regress them on time, relative firm size, and their interaction. The interaction term allows us to test whether the increase in returns to scale disproportionately benefits larger firms.

[7] In this third specification, I treat only cost of goods sold as variable and use it as an instrument.

[8] In these specifications I allow for a higher-order polynomial relative to the main text to accommodate the fact that the variable input used as an instrument is the sum of several expenditures and may require a more flexible mapping with productivity.

Figure IV: Returns to Scale over Time



(a) Cost Shares

(b) Intangible Capital

(c) Translog

(d) COGS

(e) COGS + SG&A

(f) COGS and SG&A

(g) Value-Added

Note. Figure IV displays the estimated returns to scale from the seven robustness exercises discussed above. Figure IVa shows estimates using the cost shares approach. Figure IVb reports results from the baseline production function augmented with intangible capital. Figure IVc presents estimates from a translog production function. Figure IVd uses the baseline production function with COGS as the variable input. Figure IVe extends this by including COGS plus SG&A as variable input. Figure IVf considers COGS and SG&A as separate inputs. Finally, Figure IVg shows estimates using value-added as the measure of output. In all figures, light blue lines with triangles indicate point estimates, and dark blue lines with squares show the 7-year moving averages. Sector-level estimates are winsorized at the 1% level. Output elasticities are time-varying and calculated with rolling windows from 1980 to 2014.

Table II presents the results. Overall, we find no clear pattern within the Compustat sample: most estimates for the interaction between time and relative firm size are small and statistically insignificant. This suggests that the rise in returns to scale has not systematically favored larger firms.

Table II: Translog: Returns to Scale, Trends, and Heterogeneity

	Returns to scale			
	(1)	(2)	(3)	(4)
Time	0.001***	0.001***		
	(0.000)	(0.000)		
Relative size × *Time*	-0.001***	0.000	-0.001***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Relative size	-0.048***	-0.041***	-0.045***	0.041
	(0.002)	(0.001)	(0.002)	(0.098)
Fixed effects				
Firm		✓		✓
Year			✓	✓
Observations	153,665	153,173	153,665	153,173

Note: Table II presents the results regressing returns to scale on time, relative size, and their interaction. Returns to scale at the firm level are the output of the estimation of the translog production function. Relative size is measured as the relative size of the firm within a 4-digit NAICS industry by year. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

II Model

In this section, I explain additional details of the model, emphasizing important steps related to its solution method. Most of the discussion follows the logic developed in Schaal (2017). In particular, first, I present a complementary and less general contractual environment than that in the main that allows us to find the allocation implied by the firm problem in the main text without taking care of the entire distribution of promised utilities within the firm. Then, I explain how having the allocation at hand, we can recover the prices implied by the main text contractual environment. Finally, I discuss the efficiency properties of the equilibrium.

II.I Alternative Contractual Environment

The alternative contractual environment discussed in this appendix assumes that contracts are complete and state-contingent and that there is full commitment on both the customer and firm side. Relative to the main text, contracts are complete, and customers also have commitment; this is a very convenient formulation of the contractual environment despite its lack of realism. Therefore, in this case, the contract specifies $\{p_{t+j}, \tau_{t+j}, x_{t+j}, d_{t+j}\}_{j=0}^{\infty}$, where p is the price, x is the submarket where the customer searches while being matched, τ is a separation probability, and d is an exit dummy. Each element at time $t + j$ is contingent on the entire history of shocks (z^{t+j}). The fact that the contract specifies x, i.e., the submarket in which a firm's customer must search, is a feature of completeness.

II.II Joint Surplus

The additional assumptions embedded in this alternative contractual environment allow for the simplification of the problem of the firm presented in the main text. The completeness of contracts, the commitment assumption, and the transferability of utility guarantee that the optimal policies always maximize the joint surplus of a firm and its customers. The model can thus be solved in two stages: a first stage in which I maximize the surplus and a second stage in which I design the contracts that implement the allocation. The following Bellman equation gives the joint surplus maximization problem for a firm and its current customers:

$$
\begin{aligned}
\boldsymbol{\mathcal{S}}(z,n) = \max_{\ell,d,n_i',x_i',\tau,x'} \; & nu - W\ell - Wf \\
& + \beta\mathbb{E}\Bigg\{(\delta + (1-\delta)d)n\boldsymbol{\mathcal{U}}' + (1-\delta)(1-d)\Bigg[\tau n\boldsymbol{\mathcal{U}}' \\
& + (1-\tau)m(\theta(x'))nx' - \left(\frac{Wc}{q(\theta(x_i'))} + x_i'\right)n_i' - W\chi_1(n_i'/n)^2 n^{\chi_2} + \boldsymbol{\mathcal{S}}(z',n')\Bigg]\Bigg\},
\end{aligned} \tag{25}
$$

subject to:

$$n' = (1-\tau)(1-m(\theta(x')))n + n_i', \tag{26}$$

$$y = e^z \ell^\alpha, \tag{27}$$

$$y = n. \tag{28}$$

The first element in the surplus maximization problem is the total utility of the customers nu followed by the wages $W\ell$ and operating costs Wf paid by the firm. In the next period, conditional on surviving the exit shock δ, the firm chooses whether or not to exit, a decision captured by the exit dummy d. If a firm chooses to exit, all the customers become unmatched while the firm's value is set to zero, yielding a total utility of $n\boldsymbol{\mathcal{U}}'$. If it chooses not to exit, the firm may then proceed with its separations. The total mass of separations is τn, which provides a total expected utility of $\tau n\boldsymbol{\mathcal{U}}'$ to the customer-firm group. After searching, some customers move to other firms with the value x' and contribute the amount $(1-\tau)m(\theta(x'))nx'$ to the total surplus. Simultaneously, the firm proceeds with its customer acquisitions. For each new customer acquisition in the product market segment x_i', the firm incurs a cost of $Wc/q(\theta(x_i'))$ and must offer on average a lifetime utility-price x_i' to its new customer, which appears as a cost to the current customer-firm group, and pays, to adjust its customer base, the convex cost $W\chi_1(n_i'/n)^2 n^{\chi_2}$.

The surplus maximization problem characterizes the optimal allocation of all physical resources within a firm: the optimal amount of separations, firm-to-firm transitions, the number of new customers, and the decision of whether or not to exit. Because the utility is transferable, transfers between the firm and its customers leave the surplus unchanged. Elements of the contracts describing the way profits are split, such as prices and continuation utilities, disappear in the surplus maximization problem. In particular, the distribution of promised utilities, $\{\mathcal{C}(j)\}_{j\in[0,n]}$, is not part of the state space, and only the size of the customer base at the production stage n matters. Hence, equations (25)-(28) allow for the characterization of all physical resources within a firm with standard recursive methods.

II.III Free Entry

Under this different contractual environment, the entering value for a firm stated in equation (13) in the main text can be restated in terms of the joint surplus maximization problem. I redefine the problem faced by an entering firm of type z as follows:

$$\boldsymbol{\mathcal{V}}_e(z) = (1-\delta)\max_{x_e}\left[\boldsymbol{\mathcal{S}}(z,n_e) - n_e\left(x_e + \frac{Wc}{q(\theta(x_e))}\right)\right]^+. \tag{29}$$

Having drawn the idiosyncratic productivity z, the potential entrant first decides whether to exit, a decision captured by the notation $\{\cdot\}^+$ and summarized in the dummy d_e. If it stays, the firm acquires a measure of customers, $n_e \in \mathbf{R}^+$, and chooses a market x_e in which to search, to maximize the joint surplus minus the linear advertisement cost $n_eWc/q(\theta(x_e))$ and the total utility n_ex_e that the firm must deliver to its new customers.

An important feature of this economy is that the submarket in which customers are acquired, x_e, solely appears through the term $Wc/q(\theta(x_e)) + x_e$, which is an acquisition cost per customer common to both entering and incumbent firms. The fact that the per customer acquisition cost is the same across firms is what leads to equilibrium block recursivity and is possible because of the separation of total acquiring costs into a linear and a convex component, as explained in the main text. The first term of this common per customer costs, $Wc/q(\theta(x_e))$, captures the linear advertisement cost of acquiring exactly one customer.[9] The second term, x_e, is the utility price that firms offer to their new customers. Firms choose submarkets that minimize the acquisition cost per customer defined as:

$$\xi = \min_{x}\left[x + \frac{Wc}{q(\theta(x))}\right]. \tag{30}$$

[9] If the cost were not linear, this first term would depend on the number of new customers the firm wishes to acquire, resulting in different firms facing varying per customer acquisition costs, thereby breaking block recursivity and computational tractability.

The optimal entry further requires that only the submarkets that minimize this acquisition cost per customer are open in equilibrium, which I summarize in the following complementarity slackness condition:

$$\forall x, \quad \theta(x)\left[x + \frac{Wc}{q(\theta(x))} - \xi\right] = 0. \tag{31}$$

This condition means that submarkets either minimize the acquisition cost, $\xi = x + c/q(\theta(x))$, or remain unvisited, $\theta(x) = 0$. In equilibrium, active submarkets will have the same acquisition cost, and firms will be indifferent between them. Therefore, the equilibrium market tightness in every active market is:

$$\theta(x) = q^{-1}\left(\frac{Wc}{\xi - x}\right). \tag{32}$$

Notice that because q is a decreasing function, the equilibrium market tightness decreases with the level of utility promised to the customers, as these offers succeed in attracting more customers, while firms refrain from posting such expensive contracts. The probability of finding a firm for customers thus declines with the attractiveness of the offer.

II.IV Prices and the Main Model

Building on the results in Schaal (2017), one can first recover the optimal policies of the firm problem in the main text by solving the joint surplus maximization problem from equations (25)-(28) together with equations (29)-(32). Then one can construct the prices presented in equation (12) in the main text that implement the exact same allocation of physical resources as the one retrieved from the joint surplus maximization problem using equations (9) and (11) in the main text.

In conclusion, the contractual environment in the main text and the one in Section II.I produce the same allocation of physical resources within the firm, making the two allocations isomorphic. However, the contractual environment in the main text, on top of being more realistic, constrains the set of available contracts pinning down prices uniquely (for an in-depth discussion of this issue, refer to Schaal, 2017).[10] This implies that, even under the contractual environment specified in the main text, one can solve the model restated in this appendix

[10] The contractual environment specified in Section II.I leaves prices undetermined.

through the joint surplus maximization problem with standard recursive methods instead of solving the firm's problem stated in the main text, which depends on an infinitely dimensional object such as the contracts distribution, and only later solve for the optimal pricing strategy of the firms that sustain the allocation.

II.V Discussion About Equilibrium Efficiency

One of the main implications of the model is that the presence of directed search ensures efficiency in resource allocation within the market (Schaal, 2017). This means that market power, as depicted in the framework, operates in an efficient manner. This approach is by itself valuable because it enables us to gauge the extent to which inefficient explanations can account for the observed increase in market power. By demonstrating that market power can arise through efficient mechanisms, the model provides a benchmark against which alternative explanations can be evaluated.

However, it is important to acknowledge that the model deliberately abstracts from explicit efficiency considerations. The focus is primarily on search frictions, and as a result, normative considerations are not explicitly incorporated. Consequently, the analysis in the paper remains primarily descriptive and positive, providing insights into the workings of market power within the specific framework.

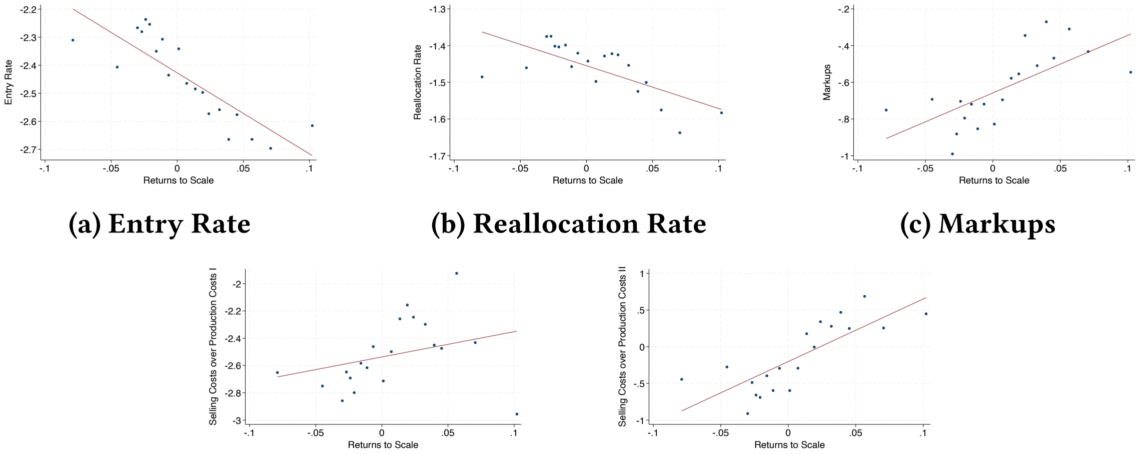
Finally, the fact that the model can only explain a fraction of the observed rise in market power suggests that inefficient explanations are likely to play a significant quantitative role. This indicates that there are additional factors and mechanisms at play in the dynamics of market power that are not fully captured by the model. The substantial explanatory power left for inefficient explanations implies that there may be ample opportunities for policy interventions that extend beyond the scope of this study.

III Mechanism Validation

III.I Sector-Level Validations

Figure V presents binned scatter plots corresponding to the sector-level regressions of secular trends on returns to scale, controlling for sector fixed effects. These plots correspond to columns (2) and (6) in Table III, and columns (2), (6), and (10) in Table IV. Overall, the linear fit appears to capture the relationship between the various secular trends and returns to scale reasonably well.

Figure V: Secular Trend Regressions: Binned Scatter Fit



(a) Entry Rate

(b) Reallocation Rate

(c) Markups

(d) Selling Costs over Production Costs I

(e) Selling Costs over Production Costs II

Note. Figure V presents binned scatter plots corresponding to the sector-level regressions of secular trends on returns to scale, controlling for sector fixed effects.

Tables III and IV present the results of regressing various secular trends on returns to scale at the sector level, allowing for all combinations of sector and time fixed effects. Overall, the estimated coefficients align with the theoretical predictions in all but one case and are statistically significant in most specifications.

III.II Firm-Level Validations

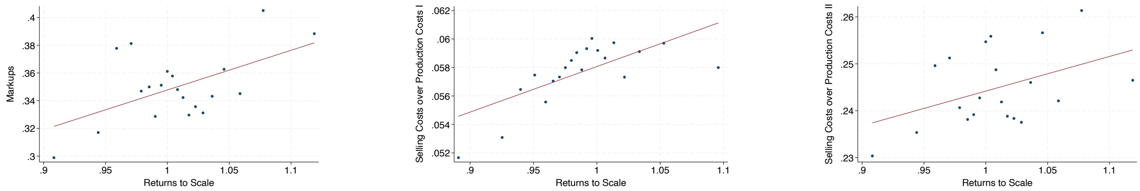
Here, we present firm-level validation exercises, where we regress markups and various measures of selling costs relative to production costs on returns to scale, allowing for all combinations of firm and time fixed effects. Table V reports the full-sample results, while Table VI presents results for the subsample of firms with positive advertising expenditures.

Table III: Returns to Scale and Secular Trends I

	Entry Rate				Reallocation Rate			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Returns to scale	-2.02***	-2.89***	-1.39***	-0.21	-1.12***	-1.17***	-0.95***	-0.08
	(0.29)	(0.34)	(0.20)	(0.15)	(0.23)	(0.24)	(0.18)	(0.13)
Fixed effects								
Sector		✓		✓		✓		✓
Time			✓	✓			✓	✓
Observations	592	592	592	592	592	592	592	592

Note: Table III presents regression results where entry and reallocation rates are regressed on returns to scale. The data source is the BDS and the author's own estimates. Observations are at the 2-digit NAICS industry by year level. All variables are expressed in logs, so coefficients can be interpreted as elasticities. Observations are weighted by relative sector size to reflect aggregate effects. The time period is 1978-2014. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1, + p<0.15.

Figure VI: Markups and Selling Costs over Production Costs Regressions: Binned Scatter Fit



(a) Markups

(b) Selling Costs over Production Costs I

(c) Selling Costs over Production Costs II

Note. Figure V presents binned scatter plots corresponding to the firm-level regressions of markups and the different measures of selling costs over production costs on returns to scale, controlling for firm fixed effects.

Overall, we find that all but one coefficient align with the theoretical prediction: higher returns to scale are associated with higher markups and higher selling costs relative to production costs. In particular, when exploiting all time variation—i.e., when including only firm fixed effects, which constitutes the most natural test of the theory—the results are all in the right direction and statistically significant. Figure VI presents the binned scatter fit of the baseline specification with firm fixed effects only, showing that the linear fit captures the data reasonably well.

Finally, we test for heterogeneous effects of returns to scale on markups as a function of firm size. Table VII presents the results. Overall, we consistently find that higher returns to scale are associated with higher markups—particularly for firms that are larger to begin with—in line with the prediction of the theory.

Table IV: Returns to Scale and Secular Trends II

	Markups				**Selling Costs over Production Costs**							
					Advertisement				Adjusted SG&A			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Returns to scale	4.45***	3.15***	4.28***	0.81	2.81***	1.85^{+}	2.74***	-0.11	8.73***	8.53***	7.46***	1.81**
	(0.36)	(0.63)	(0.38)	(0.59)	(0.79)	(1.25)	(0.60)	(0.80)	(0.62)	(1.30)	(0.50)	(0.92)
Fixed effects												
Sector		✓		✓		✓		✓		✓		✓
Time			✓	✓			✓	✓			✓	✓
Observations	586	586	586	586	591	591	591	591	592	592	592	592

Note: Table IV presents regression results where markups and two alternative measures of selling costs relative to production costs—based on advertising expenditures and an adjusted measure of SG&A—are regressed on returns to scale. The data source is Compustat and the author's own estimates, and observations are averages at the 2-digit NAICS industry by year level. Observations are weighted by relative sector size to reflect aggregate effects. The time period is 1978-2014. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1, + p<0.15.

Table V: Returns to Scale, Markups, and Selling Costs over Production Costs I

	Markups				**Selling Costs over Production Costs**							
					Advertisement				Adjusted SG&A			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Returns to scale	0.96***	0.29***	0.84***	0.18***	0.18***	0.03*	0.18***	0.00	0.20***	0.07***	0.11***	-0.01
	(0.09)	(0.06)	(0.09)	(0.07)	(0.02)	(0.02)	(0.02)	(0.01)	(0.04)	(0.03)	(0.04)	(0.03)
Fixed effects												
Firm		✓		✓		✓		✓		✓		✓
Time			✓	✓			✓	✓			✓	✓
Observations	158,850	158,384	158,850	158,384	69,299	68,829	69,299	68,829	158,850	158,384	158,850	158,384

Note: Table V presents regression results where markups and two alternative measures of selling costs relative to production costs—based on advertising expenditures and an adjusted measure of SG&A—are regressed on returns to scale. The data source is Compustat and the author's own estimates. Observations are at the firm-by-year level and are weighted by relative sector size to reflect aggregate effects. The time period is 1977-2014. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1, + p<0.15.

Table VI: Returns to Scale, Markups, and Selling Costs over Production Costs II

	Markups				**Selling Costs over Production Costs**							
					Advertisement				Adjusted SG&A			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Returns to scale	2.29***	0.32***	2.17***	0.04	0.18***	0.03*	0.18***	0.00	0.85***	0.07^{+}	0.78***	-0.07^{+}
	(0.17)	(0.09)	(0.17)	(0.10)	(0.02)	(0.02)	(0.02)	(0.01)	(0.08)	(0.04)	(0.07)	(0.05)
Fixed effects												
Firm		✓		✓		✓		✓		✓		✓
Time			✓	✓			✓	✓			✓	✓
Observations	69,299	68,829	69,299	68,829	69,299	68,829	69,299	68,829	69,299	68,829	69,299	68,829

Note: Table VI presents regression results where markups and two alternative measures of selling costs relative to production costs—based on advertising expenditures and an adjusted measure of SG&A—are regressed on returns to scale. The data source is Compustat and the author's own estimates, with the sample restricted to firm-year observations with non-missing advertising data. Observations are at the firm-by-year level and are weighted by relative sector size to reflect aggregate effects. The time period is 1977-2014. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1, + p<0.15.

Table VII: Returns to Scale and Markups: Heterogeneous Effects

	Markups			
	(1)	(2)	(3)	(4)
Returns to scale	0.62^{***}	0.30^{***}	0.29^{***}	0.11^{*}
	(0.04)	(0.05)	(0.04)	(0.06)
Lagged employment	-0.32^{***}	-0.09^{***}	-0.35^{***}	-0.09^{***}
	(0.02)	(0.02)	(0.02)	(0.02)
Returns to scale × *Lagged employment*	0.27^{***}	0.06^{**}	0.29^{***}	0.05^{*}
	(0.02)	(0.02)	(0.02)	(0.02)
Fixed effects				
Firm		✓		✓
Time			✓	✓
Observations	147,022	146,582	147,022	146,582

Note: Table VII presents regression results where markups, past employment, and their interaction are regressed on returns to scale. The data source is Compustat and the author's own estimates. Observations are at the firm-by-year level and are weighted by relative sector size to reflect aggregate effects. The time period is 1977-2014. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1, + p<0.15.

IV Quantitative Implications

IV.I Additional Quantitative Validations

Here, we present additional validation exercises. We begin by examining the joint conditional correlations of markups and selling costs relative to production costs with respect to firm age and size. We then turn to validations beyond markups and selling costs, focusing on the behavior of employment levels and employment growth across firms and over the life cycle.

Table VIII present the regression results from the data and the model of the joint conditional correlations of markups and selling costs relative to production costs with respect to firm age and size. Although none of these correlations were directly targeted, the model captures both the qualitative and quantitative patterns of the unconditional correlations between markups and selling costs relative to production costs with firm age and size, as shown in the main text. However, its performance is more limited when examining joint conditional correlations—that is, when jointly regressing markups and selling costs over production costs on age and size. For markups (comparing column (3) with column (6)), the model replicates the qualitative pattern but underestimates the magnitude. For selling costs (comparing columns (9) and (12) with column (15)), the model performs is limited both qualitatively and quantitatively, missing the observed empirical relationship along the size dimension. This shortcoming likely arises because, in the model, age conditional on size primarily reflects productivity, whereas in the data it may capture richer dynamics that lie outside the model's scope.

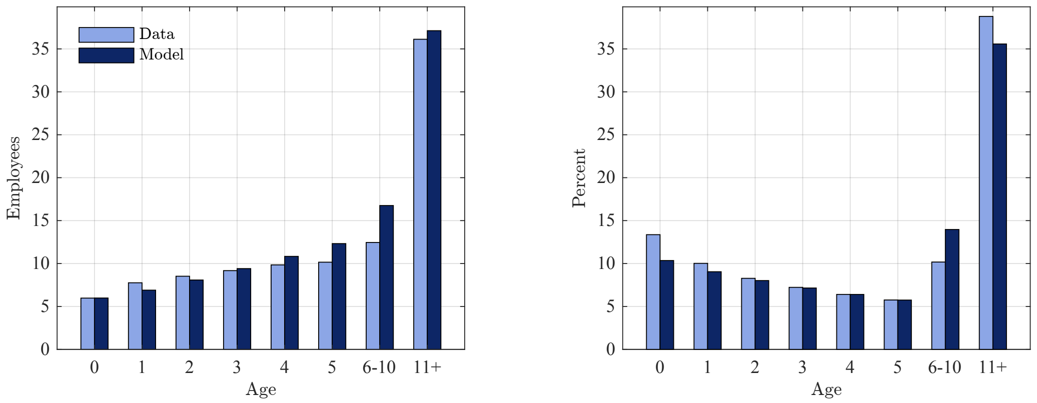
Table VIII: Conditional Correlations of Markups and Selling Costs over Production Costs with Age and Sales

	Markups						**Selling Costs over Production Costs**								
	Data			Model			Data						Model		
							Advertisement			Adjusted SG&A					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Age	0.00		-0.02**	0.01***		-0.22***	-0.19***		-0.12***	-0.13***		-0.03***	-0.62***		-0.98***
	(0.00)		(0.00)	(0.00)		(0.00)	(0.02)		(0.02)	(0.01)		(0.01)	(0.01)		(0.01)
Sale		0.07***	0.06***		0.8***	0.30***		-0.22***	-0.20***		-0.31***	-0.30***		-0.46***	0.48***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.01)	(0.01)		(0.00)	(0.00)		(0.01)	(0.02)
Fixed effects															
Firm	✓	✓	✓				✓	✓	✓	✓	✓	✓			
Sector× Time	✓	✓	✓				✓	✓	✓	✓	✓	✓			
Observations	47,146	49,293	47,146	50,000	50,000	50,000	22,663	23,555	22,663	41,930	43,728	41,930	50,000	50,000	50,000

Note: Table VIII presents the unconditional and conditional elasticities of markups and selling costs over production costs to age and sales. The data are from Compustat (1977–1990), and the model corresponds to the Compustat-like subsample of the 1980 initial steady state. All variables are in logs. Robust standard errors are reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

The model aims to capture key cross-sectional patterns in the micro-data beyond markups and selling costs, particularly those related to the firm life cycle. Firms enter the market small and gradually grow by acquiring customers, resulting in cohort-specific differences in firm size—young firms tend to have fewer employees, consistent with patterns observed in the BDS data. In addition, firm survival rates are low, leading to a declining share of older cohorts in the overall firm population.

Figure VII: Model Cross Section



(a) Cohort Size

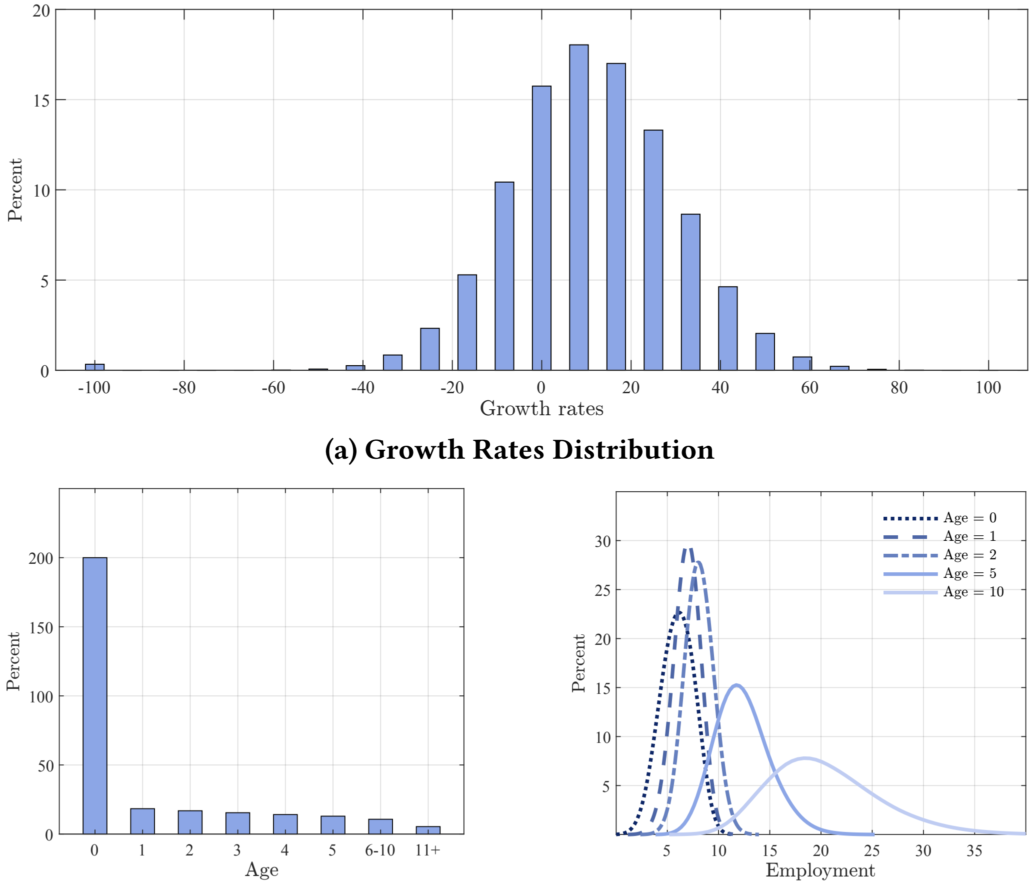
(b) Age Distribution

Note. Figure VIIa shows the size of each cohort, measured as the number of employees within firms. Figure VIIb shows the distribution of firms across cohorts. The light blue bars represent BDS data; the dark blue bars show the model predictions. Data reported are between 1977 and 1985.

Figure VII visually compares key cross-sectional patterns in the model and the data. Panel VIIa shows average firm size by cohort, measured by the number of employees. The model closely matches the data, accurately capturing life-cycle dynamics. Panel VIIb displays the distribution of firms across cohorts, which the model replicates well. Overall, the model captures the core features of the selection dynamics observed in the data.

Empirical studies based on firm-level data have identified several robust patterns in the life cycle of firms. One key finding is that firm-level growth rates are, on average, positive but highly dispersed, and that growth tends to decline with firm age—a relationship first established by Dunne et al. (1989). This negative correlation between growth and age has been documented across a wide range of sectors and countries, as shown by Coad (2009). Furthermore, Cabral and Mata (2003), using data on Portuguese manufacturing firms, find that as cohorts age, the employment distribution shifts to the right and becomes less skewed, reflecting a gradual convergence toward larger firm sizes.

Figure VIII: Additional Validations



(a) Growth Rates Distribution

(b) Growth Rates by Age

(c) Distribution by Cohort

Note. VIIIa shows the distribution of employment growth rates calculated as $g^{\ell}_{it} \equiv (\ell_{it} - \ell_{it-1})/\frac{1}{2}(\ell_{it} + \ell_{it-1})$. Figure VIIIb shows the employment growth rate by age. Figure VIIIc shows the employment distribution across cohorts.

Figure VIII presents the model-implied behavior corresponding to the empirical patterns described above. Panel VIIIa shows the distribution of employment growth rates, which are, on average, positive but highly dispersed—consistent with the data. Panel VIIIb illustrates that the model accurately replicates life-cycle dynamics, driven by the assumption that firms enter with a small customer base and grow gradually by acquiring new customers over time. This mechanism also explains the patterns shown in Panel VIIIc, where firm size increases with age, leading to a rightward shift in the cohort size distribution. Overall, the model successfully reproduces several non-targeted features of the firm life cycle.

IV.II Rising Returns to Scale and the Macroeconomy

IV.II.I Transitional Dynamics

Here, we examine the model's predictions along the transition dynamics. This exercise requires additional assumptions about firms' knowledge of the path of returns to scale over time. To this end, we follow the standard—though admittedly strong—assumption commonly used in the literature: firms are assumed to know the entire future trajectory of returns to scale from the outset, which in our case begins in the 1980s.

Figure IX presents the transition dynamics implied by the model. Panels IXa to IXd show the evolution of the model's predicted secular trends, while Panel IXe displays the trajectory of returns to scale used to simulate the transition. The entry rate in the model initially rises in anticipation of the perfectly foreseen increase in returns to scale, and only subsequently declines, as observed in the data. The reallocation rate tracks the data reasonably well, showing a steady decline across the whole period. Markups initially fall, as firms—anticipating higher future returns to scale—lower prices to attract and build customer bases; only after this initial decline do markups begin to rise, in line with the data. Finally, the ratio of selling costs to production costs follows a similar inverted U-shape as in the data, despite starting from a different initial level (as these moments are untargeted). Selling costs rise disproportionately at first, reflecting front-loaded investments in customer acquisition, and then plateau at a higher, though lower-than-peak, level in the new steady state.

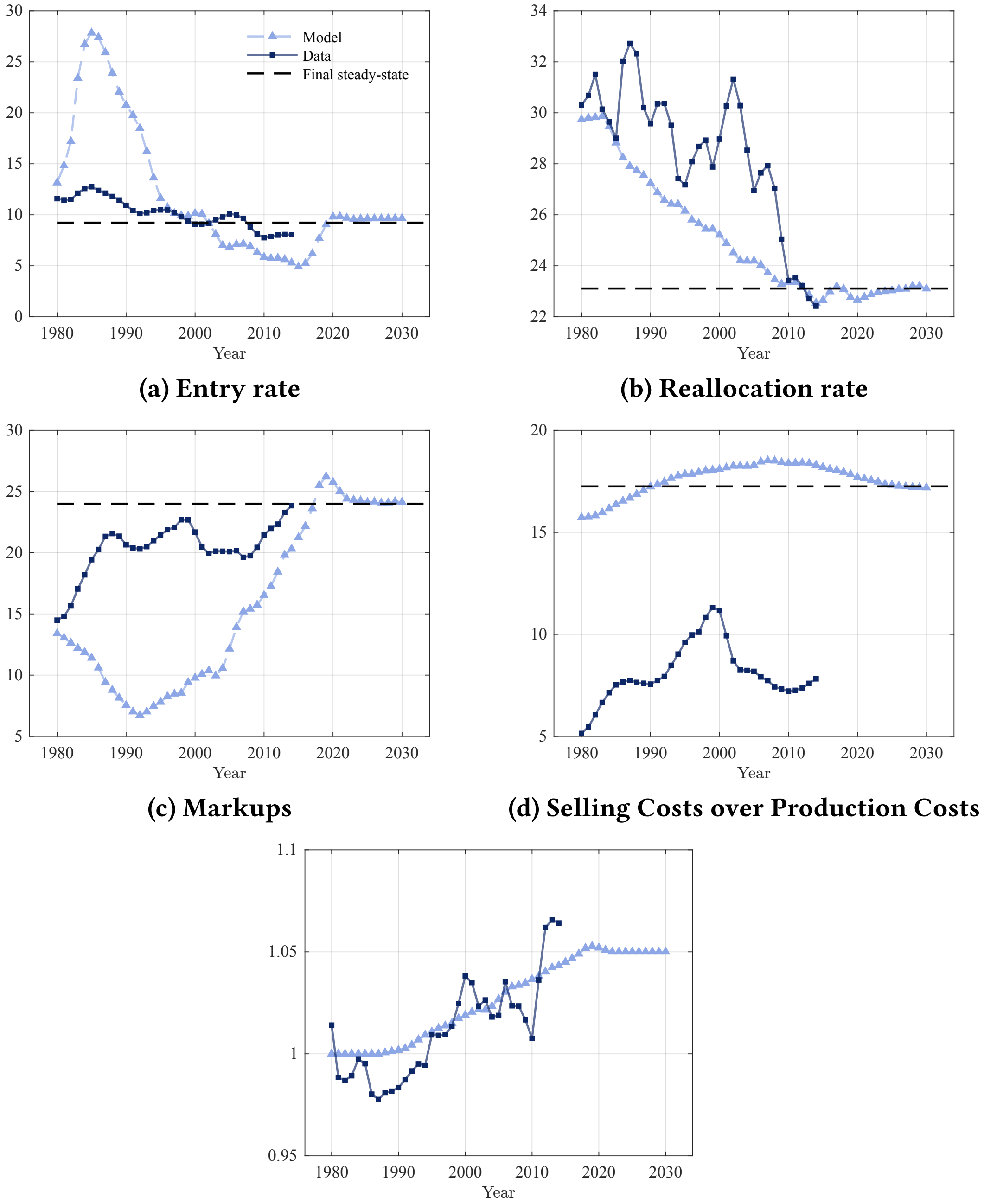
Overall, we find that while most of the secular trends implied by the model align with the data after 2000, strong anticipation effects generate patterns in entry rates and markups that deviate from the empirical evidence. This suggests that the rise in returns to scale may not have been fully anticipated in 1980, but rather gradually discovered by firms over time. Alternatively, it indicates that other complementary forces—beyond returns to scale—likely played a role in shaping the secular trends of interest.

IV.II.II Robustness

Here, we present two additional robustness exercises. First, we solve for the counterfactual steady state using a lower Frisch elasticity. Second, we compute the counterfactual steady state allowing firms to optimally choose their initial size.

Table IX presents the robustness results. Column 3 reports the outcomes using a Frisch

Figure IX: Transitional Dynamics



(a) Entry rate

(b) Reallocation rate

(c) Markups

(d) Selling Costs over Production Costs

(e) Returns to Scale

Note. Figure IX presents the transition dynamics from 1980 onward for the entry rate (Panel IXa), reallocation rate (Panel IXb), markups (Panel IXc), and selling costs relative to production costs (Panel IXd). Light dashed blue lines with triangles represent the 3-year moving average from the model, solid blue lines with squares represent the 3-year moving average from the data, and the dashed black line indicates the model's final steady state. Panel IXe displays the evolution of returns to scale in the data, along with the 15-year moving average used as input for the model.

elasticity of 0.284 instead of the baseline value of 2.84. Overall, we find that the results are quantitatively similar.

Table IX: Quantitative Implications of the Rise in Returns to Scale

	Baseline	**Lower Frisch Elasticity**	**Endogenous Initial Size Choice**
Markups			
Average cost-weighted markup	+15%	+14%	+15%
Business Dynamism			
Entry rate	-33%	-36%	-42%
Reallocation rate	-21%	-22%	-19%
Others			
Average selling costs over production costs	+23%	+24%	+23%

Column 4 shows the results when firms are allowed to choose their initial size optimally. To implement this, we impose the following free entry condition:

$$W\kappa = \max_{n_e} -\gamma W n_e^2 + \int \boldsymbol{\mathcal{V}}^e(z, n_e) g_z(dz), \tag{33}$$

where k is recalibrated to match the initial entry rate, and γ is calibrated to match the initial value of n_e. Again, we find that the results remain quantitatively similar.

IV.II.III Additional Results on Firm-Level Patterns Linked to the Secular Trends

Firm Aging. The model explains the aging of U.S. firms, as emphasized by Hopenhayn et al. (2018), as a consequence of the winners-and-losers mechanism that favors larger—and thus, on average, older—firms. Specifically, Table X shows that the model predicts an increase in the share of firms aged 11 years or older by approximately 53%, compared to 50% in the data. Additionally, the model implies a decline in the employment share of firms aged 5 years or younger by about 58%, closely matching the 56% decline observed in the data.

Table X: Firm Aging

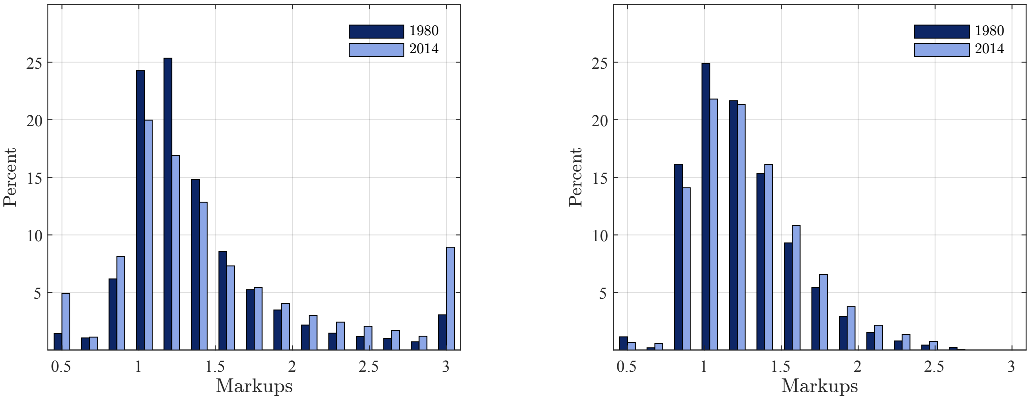
	1980 S.S.	**2014 S.S.**	**Model**	**BDS**
Firm Aging				
Share of old firms	0.322	0.495	+53%	+50%
Employment share of young firms	0.204	0.086	-58%	-56%

Note. Columns 1 and 2 report steady-state variables from the model. Columns 2 and 3 report changes in the model and the data (BDS). Empirical variables come from Hopenhayn et al. (2018). All variables in the model align with their data definitions.

Evolution of Markups Distribution and Reallocation. According to De Loecker et al.

(2020), the main change in the markup distribution since 1980 is the widening of its right tail, explaining most of the rise in the aggregate markup. Here, I examine the model's ability to replicate this empirical observation.

Figure X: Distributions of Markups over Time: Model vs. Data



(a) Data **(b) Model**

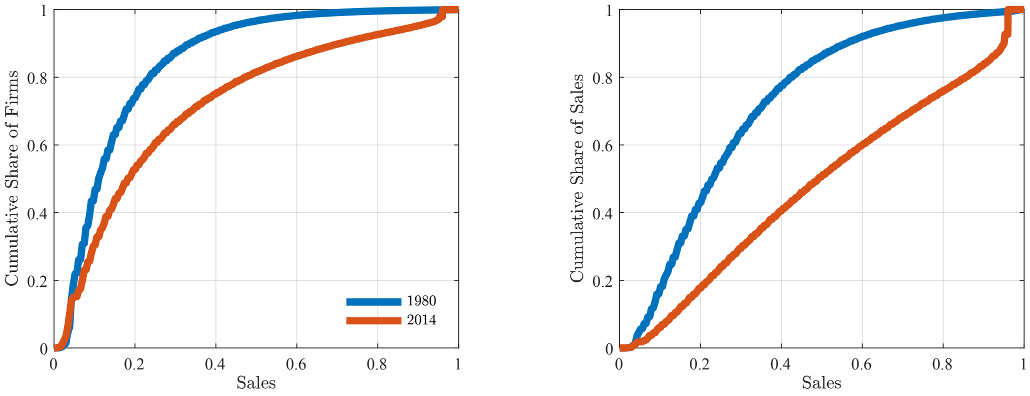
Note. Figure Xa displays the empirical markup distribution for 1977-1990 (light blue) and 2010-2014 (dark blue). Figure Xb shows the model-implied markup distribution for 1980 (light blue) and 2014 (dark blue).

Figure X compares the markup distributions between the model and the data over time. In Figure Xa, the empirical markup distribution is shown for the 1980s (light blue) and 2014 (dark blue), while Figure Xb presents the model's distributions for the same periods. The model broadly captures the observed change in markup distribution, notably displaying a wider right tail in the 2014 steady state, consistent with De Loecker et al. (2020).

In line with the literature on superstar firms (e.g., Autor et al., 2020), the model generates substantial reallocation toward larger firms. This pattern can be seen inspecting the change over time in the distributions of firms by sales and the cumulative share of total sales accounted for by firms below a given sales level. Figure XI presents these distributions. In both panels, the cumulative distribution for 2014 lies below that for 1980, indicating fatter right tails in 2014: there are more large firms, and those large firms account for a larger share of total sales.

Moreover, panel XIb shows a steeper decline in the cumulative share of sales over time than panel XIa does for the firm count. This implies not only that there are more large firms over time, but also that these firms have grown disproportionately larger. This is consistent with the rise of superstar firms documented by Autor et al. (2020) and Kehrig and Vincent (2021). Importantly, the model generates this pattern through temporary but persistent pro-

Figure XI: Cumulative Share of Firms and Sales by Sales: 1980 vs. 2014



(a) Cumulative Share of Firms **(b) Cumulative Share of Sales**

Note: Figure XI presents the cumulative distribution of firms and sales by firm-level sales in the 1980 (blue line) and 2014 (red line) steady states. Panel XIa shows the cumulative share of firms by sales level—that is, the percentage of firms with sales below a given threshold. Panel XIb shows the cumulative share of total sales accounted for by firms with sales below a given level.

ductivity advantages combined with customer accumulation dynamics. Since firms are ex-ante homogeneous, the resulting large firms are what Kehrig and Vincent (2021) refer to as ”shooting stars”—temporarily dominant firms that gradually revert to the average over time.

Declining Firm-Level Responsiveness. Part of the decline in business dynamism has been shown by Decker et al. (2020) to result from lower responsiveness of firms, i.e., they contract or expand less after productivity shocks. This section tests the model's ability to capture this micro-level observation.

To investigate this phenomenon in the Compustat data, I use the following regression:

$$g_{it+1}^{\ell} = \alpha_1 + \beta z_{it} f(t) + \theta z_{it} + \boldsymbol{\delta X}_{it}' + \phi_{st} + \varepsilon_{it}. \tag{34}$$

Here, g_{it+1}^{ℓ} denotes the employment growth rate, calculated as $2 \times (\ell_{it} - \ell_{it-1})/(\ell_{it} + \ell_{it-1})$. z_{it} represents the empirical total factor productivity residual from the production function estimation used in the empirical section. $f(t)$ is a function of time, $\boldsymbol{X}_{it}$ is a vector of control variables, and ϕ_{st} denotes sector-time fixed effects. The coefficient of interest, β, measures the changing effect of productivity on employment growth over time.

In the model, the following modified regression, mirroring equation (34), is estimated using a simulated panel for each of the two distinct steady states:

$$g^{\ell}_{it+1} = \alpha^{c} + \theta^{c}\, z_{it} + \boldsymbol{\delta^{c}} \boldsymbol{X}'_{it} + \varepsilon_{it}, \quad c \in \{1980, 2014\}. \tag{35}$$

This regression uses the same variables as before but excludes the time-dependent function.[11] Similar to equation (34), θ captures the effect of a marginal change in productivity on employment growth. To assess the decline in responsiveness, I compare the estimated coefficients between the two steady states, given by $\beta \equiv \theta^{2014} - \theta^{1980}$.

Table XI: Declining Firm-Level Responsiveness

	Model	**Compustat**		
Diff. b/w steady states	-0.08***			
	(0.00)			
$z_{it} \times time_t$		-0.05***		
		(0.00)		
$z_{it} \times \mathcal{I}_{t\in[2000,2015)}$			-0.01*	
			(0.00)	
$z_{it} \times \mathcal{I}_{t\in[1990,2000)}$				−0.01***
				(0.00)
$z_{it} \times \mathcal{I}_{t\in[2000,2010)}$				-0.01***
				(0.00)
$z_{it} \times \mathcal{I}_{t\in[2010,2015)}$				-0.01
				(0.01)
Fixed effects				
Firm		✓	✓	✓
Sector-Time		✓	✓	✓
Observations	136,150	136,150	136,150	136,150

Note. The table presents changes in firm-level responsiveness to productivity shocks between 1977 and 2014. Column (1) reports the change in responsiveness in the model, calculated using the Compustat-like subsample as the difference in steady-state estimates from equation (35). Columns (2)–(4) show corresponding estimates from Compustat, based on equation (34), capturing the decline in responsiveness over time. Column (2) includes a linear trend, while columns (3) and (4) use flexible time dummies. The indicator variable $\mathcal{I}_{t\in T}$ equals 1 during the time interval T. For the model, standard errors are computed using a two-sample t-test. Robust standard errors are reported in parentheses for the data regressions. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table XI summarizes the declining firm-level responsiveness to productivity shocks. The first column shows the decrease in responsiveness observed between the two steady states of the model. The last three columns present results from Compustat using different specifications of the time-dependent function. The first column uses a linear trend capturing responsiveness changes over the entire period, the second adds a dummy variable for responsiveness differences after 2000, and the third includes decade-specific dummy variables, using the first decade as the reference. The data and the model show a statistically significant decline in

[11] The time-dependent function is excluded as the regression is conducted separately on the two steady states of the model, where its contribution to labor growth would be zero by construction.

firm-level responsiveness across all specifications.

To better understand the underlying forces driving the decline observed in the model, the concept of firm-level responsiveness is reformulated as:

$$\frac{\partial \log \ell_{it}}{\partial z_{it}} = \frac{1}{\alpha} \cdot \left[\frac{\partial \log y_{it}}{\partial z_{it}} - 1 \right] > 0, \tag{36}$$

where α represents the returns to scale, and $\partial \log y_{it} / \partial z_{it}$ denotes the output growth associated with productivity growth. Thus, differentiating with respect to α, we get

$$\frac{\partial}{\partial \alpha} \frac{\partial \log \ell_{it}}{\partial z_{it}} = -\frac{1}{\alpha^2} \cdot \left[\frac{\partial \log y_{it}}{\partial z_{it}} - 1 \right] + \frac{1}{\alpha} \cdot \frac{\partial}{\partial \alpha} \frac{\partial \log y_{it}}{\partial z_{it}} \tag{37}$$

$$= -\frac{1}{\alpha^2} \cdot \left[\frac{\partial \log y_{it}}{\partial z_{it}} - 1 \right] + \frac{1}{\alpha} \cdot \frac{\partial}{\partial \alpha} \left[\frac{\partial \log y_{it}}{\partial \log p_{it}} \frac{\partial \log p_{it}}{\partial z_{it}} \right], \tag{38}$$

where $\frac{\partial \log y_{it}}{\partial \log p_{it}}$ is the endogenous demand elasticity and $\frac{\partial \log p_{it}}{\partial z_{it}}$ is the endogenous pass-through. Therefore, it can be seen that if changes in demand elasticity and pass-thorough to changes in returns to scale are small enough, as it appears from the quantitative analysis, responsiveness declines with a rise in returns to scale.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–2451.

Afrouzi, H., A. Dernik, and R. Kim (2020). Growing by the masses. revisiting the link between firm size and market power. *Working Paper*.

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics 135*(2), 645–709.

Barkai, S. (2020). Declining labor and capital shares. *The Journal of Finance 75*(5), 2421–2463.

Belo, F., X. Lin, and M. A. Vitorino (2014). Brand capital and firm value. *Review of Economic Dynamics 17*(1), 150–169.

Bond, S., A. Hashemi, G. Kaplan, and P. Zoch (2021). Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data. *Journal of Monetary Economics 121*, 1–14.

Cabral, L. and J. Mata (2003). On the evolution of the firm size distribution: facts and theory. *American Economic Review 93*(4), 1075–1090.

Chiavari, A. and S. Goraya (2021). The rise of intangible capital and the macroeconomic implications. *Working Paper*.

Coad, A. (2009). *The growth of firms: a survey of theories and empirical evidence*. Edward Elgar Publishing.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics 135*(2), 561–644.

De Ridder, M., B. Grassi, and G. Morzenti (2022). *The Hitchhiker's Guide to Markup Estimation*. Centre for Economic Policy Research.

Decker, R. A., J. Haltiwanger, R. S. Jarmin, and J. Miranda (2020, December). Changing business dynamism and productivity: shocks versus responsiveness. *American Economic Review 110*(12), 3952–90.

Dunne, T., M. J. Roberts, and L. Samuelson (1989). The growth and failure of us manufacturing plants. *The Quarterly Journal of Economics 104*(4), 671–698.

Ewens, M., R. H. Peters, and S. Wang (2019). Acquisition prices and the measurement of intangible capital. *NBER Working Paper* (w25960).

Gandhi, A., S. Navarro, and D. A. Rivers (2020). On the identification of gross output production functions. *Journal of Political Economy 128*(8), 2973–3016.

Gao, W. and M. Kehrig (2017). Returns to scale, productivity and competition: empirical evidence from us manufacturing and construction establishments. *Working Paper*.

Gourio, F. and L. Rudanko (2014). Customer capital. *Review of Economic Studies 81*(3), 1102–1136.

Hall, R. E. and D. W. Jorgenson (1967). Tax policy and investment behavior. *The American economic review 57*(3), 391–414.

Hopenhayn, H., J. Neira, and R. Singhania (2018). The rise and fall of labor force growth: implications for firm demographics and aggregate trends. *NBER Working Paper*.

Kehrig, M. and N. Vincent (2021). The micro-level anatomy of the labor share decline. *The Quarterly Journal of Economics 136*(2), 1031–1087.

Klette, T. J. and Z. Griliches (1996). The inconsistency of common scale estimators when output prices are unobserved and endogenous. *Journal of applied econometrics 11*(4), 343–361.

Landes, E. M. and A. M. Rosenfield (1994). The durability of advertising revisited. *The Journal of Industrial Economics*, 263–276.

Morlacco, M. and D. Zeke (2021). Monetary policy, customer capital, and market power. *Journal of Monetary Economics.*

Ptok, A., R. P. Jindal, and W. J. Reinartz (2018). Selling, general, and administrative expense (sga)-based metrics in marketing: conceptual and measurement challenges. *Journal of the Academy of Marketing Science 46*(6), 987–1011.

Schaal, E. (2017). Uncertainty and unemployment. *Econometrica 85*(6), 1675–1721.

Syverson, C. (2004). Market structure and productivity: a concrete example. *Journal of Political Economy 112*(6), 1181–1222.

Traina, J. (2018). Is aggregate market power increasing? production trends using financial statements. *Production Trends Using Financial Statements (February 8, 2018).*

Vitorino, M. A. (2014). Understanding the effect of advertising on stock returns and firm value: Theory and evidence from a structural model. *Management Science 60*(1), 227–245.