Journal of Monetary Economics 160 (2026) 103935

Contents lists available at ScienceDirect

Journal of Monetary Economics

journal homepage: www.elsevier.com/locate/jme

Customer accumulation, returns to scale, and secular trends☆

Andrea Chiavari

Department of Economics, University of Oxford, United Kingdom

ARTICLE INFO

JEL classification:
D21
D24
D83
E20
L11
L16

Keywords:
Customer capital
Firm dynamics
Search and matching
Technological change

ABSTRACT

This paper studies how rising returns to scale contributed to declining business dynamism and increasing markups and expenditures devoted to customer acquisition in the U.S. economy. It introduces a firm dynamics model with heterogeneous markups and customer accumulation based on directed search, in which larger firms gain a competitive edge from higher returns to scale. This makes markets less contestable for new firms and leads to the rise of superstar firms. The model quantitatively accounts for a substantial share of these trends, and the underlying micro-level mechanisms align with empirical evidence.

1. Introduction

Over the past four decades, the U.S. economy has become increasingly dominated by a small number of large superstar firms outperforming competitors and capturing a disproportionate share of customers (Autor et al., 2020). This shift has coincided with several secular trends: rising markups (De Loecker et al., 2020), declining business dynamism (Decker et al., 2014), and growing firm investments in customer acquisition (He et al., 2024). Despite the attention received by these secular trends, no consensus has emerged on the fundamental causes behind them.

The period during which these secular trends emerged has also seen rapid technological change, particularly in the production processes of firms (Brynjolfsson and McAfee, 2014). A growing body of evidence suggests that these changes have led to higher returns to scale (Bloom et al., 2014; De Loecker et al., 2020; Chiavari and Goraya, 2021; Lashkari et al., 2021; Kariel and Savagar, 2022). However, no study has directly examined the role of rising returns to scale in driving these secular trends.

This paper fills this gap and makes two main contributions. First, it introduces a novel model of customer accumulation, based on search-and-matching frictions in the product market, where higher returns to scale give larger firms a cost advantage, turning

☆ This version: September 2024. The first version was circulated in October 2020. I am deeply indebted to my advisors Isaac Baley and Edouard Schaal for their invaluable advice and support. I thank Steve Bond, Andrea Caggese, Ryan A. Decker, Jan De Loecker, Jan Eeckhout, Manuel Garcia-Santana, Sampreet Goraya, François Gourio, Matthias Kehrig, Alexandre N. Kohlhas, Michael McMahon, Virgiliu Midrigan, Morten O. Ravn, Pau Roldan-Blanco, Carolina Villegas-Sanchez, and Thomas Winberry as well as all the participants at the CREi macro lunch, 10th S&M Conference, 3rd QMUL Economics and Finance Workshop, SMYE, ES Asian Meeting, BGSE SF, YES, ES European Meeting, Workshop of the Spanish Macro Network, BoL, Sciences Po, TSE, Oxford, CEMFI, NUS, BdI, Fed Board, BdE, SMU, FGV, UPenn, EIEF, Cornell, Junior S&M Workshop, 11th EIEF-IGIER-UNIBO IO Workshop, Kent Workshop, Warwick-CFM-Vienna Macro Conference, Bristol S&M Workshop, CMA, and London E1 Workshop in Quantitative Macro. I thank the European Economic Association and Unicredit Foundation for the Best Job Market Paper Award. This paper was previously circulated under the title "The Macroeconomics of Rising Returns to Scale: Customer Acquisition, Markups, and Dynamism".

E-mail address: andrea.chiavari@economics.ox.ac.uk.

https://doi.org/10.1016/j.jmoneco.2026.103935
Received 30 September 2024; Received in revised form 30 March 2026; Accepted 30 March 2026
Available online 1 April 2026
0304-3932/© 2026 The Author. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

them into superstar firms. Second, it shows that, when calibrated to realistic increases in returns to scale, the model suggests that this technological shift is a quantitatively important driver of the observed secular trends.

To study how increasing returns to scale drive the rise of superstar firms and shape customer competition, I develop a novel firm dynamics model with endogenous entry and exit à la Hopenhayn (1992) and realistic customer accumulation. Building on Gourio and Rudanko (2014) and Roldan-Blanco and Gilbukh (2020), the model incorporates tools from the labor search literature to capture (i) customer switching between firms—estimated at 10%–25% annually in Gourio and Rudanko (2014), and (ii) the price sensitivity of incumbent customers, as shown in Paciello et al. (2019).

To achieve this, I introduce directed search in the product market and a price-setting environment with firm-side commitment, drawing on Schaal (2017). These features jointly ensure uniquely determined heterogeneous prices in equilibrium and allow for endogenous customer switching. Because search is directed and firms internalize increasing returns, the model yields a constrained-efficient allocation, offering an efficient theory of market power and thus a laboratory to examine how far efficient explanations can account for the secular trends observed in the U.S. economy.

Search frictions in the product market explain why firms invest resources in acquiring customers, in addition to using prices to attract and retain them, and imply that firms grow through customer accumulation—consistent with recent empirical evidence (e.g., Afrouzi et al., 2020; Einav et al., 2020). The customer acquisition motive creates a trade-off between investing in the customer base through low markups and harvesting it through high ones, which is shaped by changes in technology. The model remains computationally tractable and is rich enough to capture key firm-level behaviors, serving as a natural laboratory to assess how far the efficient response of firms to rising returns to scale can go in explaining secular trends—such as higher markups, declining dynamism, and increased customer acquisition efforts.

Analyzing the implications of the model, I show that an increase in returns to scale reduces marginal costs more for larger firms than for smaller ones. Although all firms experience the same technological change, its effects are unequal, giving larger firms a competitive edge. As competition is for the same pool of customers, this advantage allows large firms to manage the investing-harvesting trade-off more effectively, creating a demand-based channel through which they outcompete smaller firms, resulting in winners-and-losers dynamics and the rise of superstar firms.

The demand-based channel arising from marginal cost reductions due to rising returns to scale has several key implications. First, the incomplete pass-through of falling marginal costs results in an increase in the average cost-weighted markup. Second, this uneven cost advantage creates winners-and-losers dynamics: larger firms expand by effectively attracting customers, while smaller firms struggle to compete. As a result, the selection process shifts in favor of larger firms, reducing market contestability and contributing to a decline in business dynamism. Third, with lower production costs, firms have stronger incentives to grow, leading to increased spending on customer acquisition costs relative to production costs. Empirical evidence supports these predictions: higher returns to scale are positively associated with markups, negatively associated with business dynamism, and positively associated with selling costs relative to production costs.

After validating the mechanism of the model, I assess its quantitative implications. I calibrate the model to the 1980s using key moments from firms' life cycles, markups, and business dynamism. To further validate the model, I show that incorporating customer accumulation aligns with findings in Gourio and Rudanko (2014) and Paciello et al. (2019), while also improving the fit on important, often overlooked, life-cycle patterns. Specifically, the model reproduces the observed rise in markups and the decline in selling costs relative to production costs over the firm life cycle. Moreover, I find that quantitatively a 5% rise in returns to scale goes a long way in explaining the increase in average cost-weighted markups, the decline in business dynamism, and the rise in customer acquisition spending relative to production costs.

I show that the model explains these key macro trends by tracing them back to well-established micro-level mechanisms. First, a winners-and-losers dynamic shifts activity toward larger, older firms – consistent with findings by Autor et al. (2020), De Loecker et al. (2020), and Kehrig and Vincent (2021) – helping explain the aging of U.S. firms (Hopenhayn et al., 2018). Second, rising average markups stem partly from a widening right tail in the markup distribution. Third, declining business dynamism reflects firms' weaker responses to productivity shocks (Decker et al., 2020).

However, while rising returns to scale help explain a substantial part of many of the changes in the U.S. economy since 1980, the model aligns more closely with the data after the 2000s, as the transition dynamics reveal a less precise fit for earlier decades. Moreover, since the model captures only a portion – albeit a substantial one – of the broader economic transformation, it suggests that firms' efficient responses to increasing returns alone are insufficient to fully explain the observed trends. Complementary factors, possibly involving less efficient mechanisms, likely played a significant role as well.

Literature Review. This paper contributes to several strands of the literature. It contributes to the literature on directed search in the product market by integrating insights from Gourio and Rudanko (2014) and Roldan-Blanco and Gilbukh (2020), where customers are locked in once matched with a firm, with labor-search tools from Schaal (2017) to uniquely pin down firm prices while also allowing for (i) incumbent customer switching between firms, which ranges from 10% to 25% annually (Gourio and Rudanko, 2014), and (ii) a non-zero response of incumbent customers to firm prices, as documented empirically in Paciello et al. (2019). Thus, the paper relates to Paciello et al. (2019) but enables customers to direct their search for goods and allows firms to invest in demand through selling expenditures, making it possible to study how technological changes affect customer reallocation and firms' cost structures.

Moreover, the paper relates to New Keynesian models that have often adopted alternative microfoundations to search frictions to capture pricing under customer accumulation concerns, such as good-specific habits (e.g., Ravn et al., 2006; Nakamura and Steinsson, 2011) or switching costs (e.g., Kleshchelski and Vincent, 2009; Dupraz, 2024). This paper also relates to tractable models of market

power in the spirit of Atkeson and Burstein (2008). While these approaches account for pricing under customer accumulation or heterogeneous markups, in contrast to this paper, they typically abstract from investment in the customer base – a central element of the theory developed here – and deliver inefficient theories of market power.

Finally, this paper complements the growing literature studying technological factors behind trends such as the rise in markups and the decline in business dynamism. Related works include Akcigit and Ates (2021), Cavenaile et al. (2019), De Ridder (2019), Weiss (2019), and De Loecker et al. (2021). I contribute to this literature by examining a distinct technological change—the rise in returns to scale in production. Combined with a novel model of customer accumulation, this reveals a demand-based channel that triggers winners-and-losers dynamics, helping explain many observed U.S. economic trends.

Outline. Section 2 reviews the secular trends unfolding in the U.S. economy alongside evidence of rising returns to scale. Section 3 introduces the theoretical model. Section 4 discusses the model's implications of rising returns to scale and validates them with empirical data. Section 5 calibrates the model, evaluates its performance using firm-level and cross-sectional data, and quantifies the impact of rising returns to scale on the secular trends in the U.S. economy. Section 6 concludes.

2. Motivating evidence

This section documents the key empirical facts and motivates the mechanism studied later. It starts by describing the data sources used in the analysis and then reviews three secular trends since 1980: declining business dynamism, rising markups, and rising customer acquisition costs relative to production costs. Then, it presents evidence on the main technological change focus of the paper: that rise in returns to scale in production within all sectors of the U.S. economy.

2.1. Data

In this paper, I use two data sources: Compustat for information on U.S. firms and Business Dynamics Statistics (BDS) data to obtain representative measures for the U.S. economy.

Compustat. The main data source is Compustat, a firm-level database with all the U.S. publicly traded firms between 1977 and 2014. This section discusses this dataset, while Online Appendix I.I.I provides more details on the data cleaning process.

Although publicly traded firms represent a small share of the total number of firms, they are among the largest in the economy, accounting for approximately 30% of U.S. employment (Davis et al., 2006). Compustat provides rich firm-level financial information, including sales, input expenditures, capital stock, and detailed industry classifications. An advantage of the dataset is its inclusion of certain – albeit imperfect – measures of selling costs, most notably firm-level advertisement expenditure. While useful as a benchmark, this measure has two limitations: it captures only a narrow subset of total selling costs and suffers from limited availability. To address these shortcomings, I complement it with an alternative measure from the existing literature based on adjusted selling, general, and administrative (SG&A) costs. Online Appendix I.I.II provides a detailed explanation of both measures.

However, despite its many strengths, the Compustat database has two limitations: (i) it does not allow for the separation of quantities and prices, complicating the production function estimation[1]; and (ii) it includes only publicly traded firms, raising potential selection concerns. To address the first issue, I follow De Loecker et al. (2020) and flexibly proxy for demand using sales shares as sufficient statistics, as explained in detail in Online Appendix I.III.I. To address the second, I compare my empirical results with alternative data sources where possible and, quantitatively, ensure that the model-generated data mimics the selection criteria of Compustat.

BDS data. I use the firm component of the publicly available BDS dataset for 1977–2014 to obtain representative aggregate and sector-level measures of the U.S. firm size distribution and business dynamism, including firm entry and reallocation rates.

2.2. Secular trends

This section reviews several secular trends in the U.S. economy since 1980, including the decline in business dynamism, the rise in markups, and the growing share of resources firms devote to selling costs. Fig. 1 summarizes these trends.

Business dynamism. The decline in business dynamism – measured by the entry and reallocation rates reported in the BDS – has been documented by Decker et al. (2014, 2016), among others. Fig. 1(a) reports their evolution. Since 1980, this trend has included a 33% drop in the entry rate of new firms, measured as the ratio of new to incumbent firms, and a 29% decline in the (excess) reallocation rate of employment across firms, measured as the sum of the job creation and destruction rates, net of the absolute difference between them. One consequence of the declining entry of new firms is the aging of the incumbent firm population (Hopenhayn et al., 2018). Although this paper focuses on the U.S., Biondi et al. (2023) show that similar patterns are evident across most European countries.

Market power. Much of the decline in business dynamism has been accompanied by a reallocation of economic activity toward a small number of large superstar firms, resulting in increased market concentration (Autor et al., 2020; Kehrig and Vincent, 2021). Since 1980, markups have risen by 42% when measured using the production approach in Compustat (De Loecker et al., 2020) and

[1] This challenge is common to most production datasets.



Fig. 1. Secular trends over time.
Note. Fig. 1 presents the evolution of the entry rate, reallocation rate, markups, and selling costs relative to production costs from 1980 to 2014. 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.

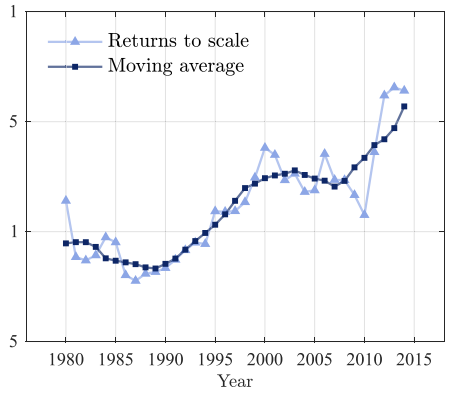
aggregate using cost weights, largely driven by these dominant firms' ability to charge higher prices. Similar conclusions have been reached using the demand approach with product-level data (Döpper et al., 2021). These trends are not unique to the United States and have also been observed globally, particularly in advanced economies (Díez et al., 2021).

Selling costs. Early evidence that firms are devoting increasing resources to customer accumulation–such as advertising expenditures and trademark activities–can be found in De Loecker et al. (2020), Kost et al. (2019), He (2022), He et al. (2024). Additional support comes from Kaplan and Zoch (2020), who show that firms are allocating more labor toward expanding demand rather than production. In Compustat, selling costs relative to production costs, measured by cost of goods sold, have increased by 60%–90% over time, depending on whether the measure is based on advertising expenditures or adjusted SG&A.

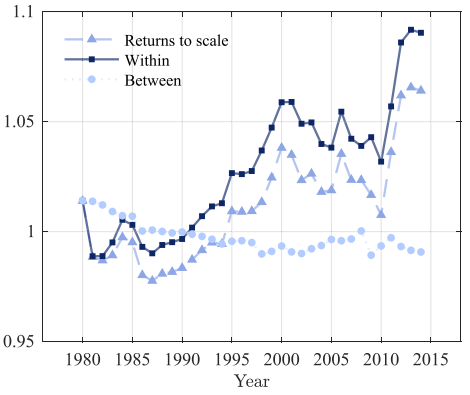
2.3. Rising returns to scale

Here, I review the evidence on rising returns to scale, discuss its possible causes, and discuss the limitations imposed by Compustat in interpreting the results.

Empirical evidence. An early reference that provides evidence of rising returns to scale in production in the U.S. was De Loecker et al. (2020). Subsequently, several contemporaneous studies have highlighted the pervasiveness of this phenomenon: Chiavari and



(a) Average Returns to Scale



(b) Within-Between Sector Decompositions

Fig. 2. Returns to scale over time.
Note. Fig. 2(a) 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. Fig. 2(b) plots the counterfactual evolution implied by the within-between sector decomposition presented in detail in Online Appendix I.III.II. 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Goraya (2021) in the U.S. by augmenting the production function with intangible capital, Lashkari et al. (2021) in France, and Kariel and Savagar (2022) in the UK.

I review here the findings based on Compustat data. I estimate a Cobb–Douglas production function taking as inputs capital and labor with time-varying sector-specific elasticities at the 2-digit NAICS level.[2] The estimation follows the approach of Ackerberg et al. (2015), augmented with sales share controls in the first stage as suggested by De Loecker et al. (2020), to flexibly account for demand heterogeneity across firms. reports the sales-weighted returns to scale. Online Appendix I.III.I details the estimation procedure.

Overall, I confirm that returns to scale have increased over time. In 1980, returns to scale were approximately 1, indicating that firms operated under a *constant* returns to scale production function.[3] By 2014, returns to scale rise to about 1.05, suggesting that firms operate under *increasing* returns to scale. To assess whether this rise reflects reallocation toward sectors with initially higher returns to scale or a broad increase in returns to scale within sectors, Fig. 2(b) plots the counterfactual evolution of each component implied by the within-between sector decomposition described in detail in Online Appendix I.III.II. The figure shows that the latter mechanism drives the observed increase. Online Appendix I.III.III presents a range of robustness checks, confirming the rise in returns to scale.

Potential causes. Anecdotally, the rise in returns to scale coincides with the IT revolution, which began in the 1980s and accelerated in the 1990s. Technologies such as the internet, mobile phones, and software have transformed production and business models by expanding data use, reducing communication costs, and improving internal coordination (Lashkari et al., 2021). Larger firms – operating across more products, markets, and complex hierarchies – face greater coordination challenges, making them especially likely to benefit from technologies that reduce organizational frictions. Bartel et al. (2007) show how IT improves the efficiency of managing multiple production processes, while Bloom et al. (2014) document its impact on firm structure and span of control. As data becomes a byproduct of firm activity Baley and Veldkamp, 2025, larger firms can better leverage these flows with IT, enhancing decision-making and operations. Altogether, these dynamics suggest that IT adoption can reduce coordination frictions and improve information use, particularly for larger firms, thereby increasing returns to scale.

Interpretation of the results and limitations. As already discussed, in Online Appendix I.III.II, I further document that the rise in returns to scale has occurred primarily within sectors. Moreover, using a translog production function that permits firm-specific elasticities, I find limited evidence that larger firms have disproportionately benefited from higher returns to scale (Online Appendix I.III.III). Thus, within the Compustat sample, the increase in returns to scale appears relatively homogeneous across firms.

However, caution is warranted when interpreting these results due to Compustat's limited representativeness, as the database predominantly captures larger firms. If adopting high-returns-to-scale technologies involves fixed costs, smaller firms outside

[2] To obtain time-varying elasticities, I estimate the production function using rolling windows of the five years before and after a given year.
[3] These findings are consistent with Gao and Kehrig (2017), who report nearly constant returns to scale using U.S. Census data from 1982 to 1987.

Compustat might have experienced minimal or no increase. Although this selection issue does not undermine the empirical analysis presented earlier, it remains an essential consideration for interpreting the subsequent quantitative results. If returns to scale have indeed increased disproportionately among the largest firms, their aggregate impact could be magnified—a point I elaborate on when discussing the model mechanism in Section 4.1.

Finally, while the secular trends documented in this section begin in the early 1980s, the estimated rise in returns to scale becomes more pronounced only from the early 1990s onward. This timing difference suggests that complementary forces have likely contributed to the early evolution of these secular trends or that firms' anticipatory behavior played a role prior to 1990, an interpretation explored quantitatively in Section 5.3 through the model's transition dynamics.

3. Model

A growing body of evidence shows that firms grow primarily by acquiring new customers rather than selling more to existing ones (Foster et al., 2008; Sterk et al., 2021; Bernard et al., 2022; Afrouzi et al., 2020; Einav et al., 2020). If competition for these customers unfolds through prices and selling activities, a rise in returns to scale may fundamentally reshape market dynamics. Higher returns to scale would reduce marginal costs disproportionately for larger firms, making it easier for them to attract customers through lower prices while retaining higher margins to fund selling activities. Since firms compete for the same pool of customers, this cost advantage shifts demand toward large firms at the expense of small ones—a winners-and-losers dynamic that makes markets less contestable, potentially accounting for part of the secular trends examined in the previous section.

This section presents a model that formalizes this mechanism. The framework builds on directed search in the product market, which provides a microfoundation for costly competition over customers through both prices and selling activities. Moreover, it delivers a constrained-efficient allocation, offering a natural benchmark for assessing how far efficient responses to technological change can go in explaining observed secular trends.

3.1. Population and technology

Time is discrete. The economy features a representative household and an endogenous measure of firms with free entry. The household consists of a mass of customers of measure one and a large mass of workers. It discounts the future at rate β and derives utility from consumption linearly, uC, minus the disutility of labor, $\vartheta L^{1+1/\psi}/(1+1/\psi)$, where ψ is the Frisch elasticity. The household aggregates consumption across customers: $C = \int_{j\in\mathcal{J}} c_j \mathrm{d}j$, where c_j is individual customer consumption, and $\mathcal{J} \subseteq 1$ is the set of customers matched with a firm. Customers can buy at most one unit of a firm's good, so demand operates purely on the extensive margin. Thus, their shopping value equals the household's marginal utility of consumption, u.

Because customers buy one unit, the model focuses on the extensive margin of demand accumulation, a substantial contributor (at least 70%) to firms' growth (Afrouzi et al., 2020; Einav et al., 2020). This emphasis, a key aspect of search-and-matching models, introduces a competitive environment in which firms compete for the *same* customers. Thus, if certain firms succeed in the competition process, others must lose, resulting in the emergence of winners-and-losers dynamics.

Firms also discount the future at rate β and differ in their idiosyncratic productivity z, following an AR(1) process, given by: $z_t = \rho z_{t-1} + \sigma\sqrt{1-\rho^2}\varepsilon_t$, $\varepsilon_t \sim \mathcal{N}(0,1)$, where $\rho \in (0,1)$ is the persistence and σ is the standard deviation. With a workforce of size ℓ, a firm's production technology is $y = e^z \ell^\alpha$, with $\alpha \geq 0$. The production function's curvature, defining its returns to scale, captures the technological change quantified in this paper. As in Hopenhayn (1992), upon entry, firms incur a sunk entry cost κ and then a fixed operating cost f each period, all paid in units of labor. Exogenous exit occurs with probability $\delta \in (0,1)$.

3.2. Frictional product market

The product market is frictional, featuring directed search on customers' and firms' sides. Firms announce contracts to attract customers, and as utility is transferable, a sufficient statistic for a contract is the utility x it delivers. Contracts delivering the same value compete within the same market segment; thus, the product market is segmented into a continuum of submarkets indexed by the promised utility $x \in [\underline{x}, \overline{x}]$. Customers – even those matched with a firm – and firms search and can choose which submarket to go to, but the search process takes time and I restrict firms and customers to visiting one submarket at a time.

A constant returns to scale matching function governs match creation in each market segment. The tightness of submarket x, denoted by $\theta(x) = a/\mu$ (where a is advertisements and μ is the mass of searching customers, including matched and unmatched ones), captures the advertisement-customers ratio. In a submarket with tightness θ, customers find a firm with probability $m(\theta) = \theta(1+\theta)^{-1}$, while firms find potential customers with probability $q(\theta) = m(\theta)/\theta = (1+\theta)^{-1}$. Consistent with the search literature, m is increasing, q is decreasing, and that $m(0) = 0$ and $q(0) = 1$. Customers and firms solve a trade-off between the contract utility and the probability of matching.

These probabilities capture an important idea: while posting more favorable terms increases the likelihood of attracting customers, it does not guarantee it. For instance, a restaurant with limited tables might not be the best choice if it is a popular destination for everyone else. Thus, these probabilities reflect a within-period capacity constraint (Wright et al., 2021).

Firms post a measure of advertisement, a.[4] By the law of large numbers, firms face no uncertainty regarding the mass of customers they attract. In a submarket with tightness θ, each unit of advertising converts into a new customer with probability $q(\theta)$, so a firm

[4] Advertisement is a stand-in for a broader notion of marketing effort and will be interpreted as such later on.

that chooses a attracts $n_i = q(\theta)a$ new customers. Attracting customers requires labor and entails a linear cost ca and a convex cost $\chi_1(q(\theta)a/n)^2 n^{\chi_2}$ (where n is the firm's existing customer mass). The convex component implies that the cost per acquired customer rises with the scale of customer acquisition, generating diminishing returns to advertising expenditures.

This feature plays an important role in the model's dynamics. Capturing the time required to establish stores and reach customers, it constrains firms' customer base expansion, creating a realistic firm life cycle and preventing a degenerate distribution of firms.[5] The separation of convex and linear costs is also necessary for technical reasons: it ensures equilibrium block recursivity even when both matched and unmatched customers engage in search.

The fact that search is directed on both sides of the market and that returns to scale are fully internalized by firms implies that the decentralized equilibrium is constrained-efficient. Further details are provided in Online Appendix II.

3.3. Contractual environment and timing

This section outlines the contractual environment, while Section 3.6 describes the implications for prices. Contracts specify various elements relevant to the firm and its customers. I assume that contracts are state-contingent and that firms fully commit to them. A contract specifies $\{p_{t+j}, \tau_{t+j}, d_{t+j}\}_{j=0}^{\infty}$, where p is the price, τ is a separation probability, and d is an exit dummy.[6] Each element at time $t+j$ is contingent on the entire history of shocks (z^{t+j}). A detailed exposition of the contractual environment and its implications is in Online Appendix II.

The contracts are large objects but can be written recursively. They are rewritten every period after matching occurs when production takes place (stage B in Fig. 3). At this stage, the firm starts with some utility C, promised in the past to its incumbent customers or new ones. A recursive contract $\omega = \{p, \tau, d, C'\}$ specifies the current price p and the next period's quantities $\{\tau(z'), d(z'), C'(z')\}$, contingent on the next period's state, where $C'(z')$ is some future promised utility. Because of the firms's commitment, contract ω delivers at least the promised utility C.

Evidence of explicit long-term contracts is extensive in contexts with long-term customer relationships such as banking, telecommunications, and business-to-business transactions. However, contracts in the model stand in for more broad implicit long-term customer relationships, where purchasing decisions consider present and future prices. Such implicit long-term relationships are prevalent across various contexts; for instance, Bronnenberg et al. (2012) find persistent brand preferences in consumer packaged goods. Previous research, including Dubé et al. (2010), has also documented significant consumer inertia.

The model's timing, shown in Fig. 3, starts with firms deciding to enter at the beginning of period t. Then productivity z and exogenous exit δ shocks are realized. Surviving firms choose whether to exit ($d = 1$). Separation follows with probability τ, then search and matching occur between firms and customers. Production occurs last, and markets clear.

3.4. Customer's problem

As conventional in the search literature, the value functions below are expressed at stage B when production happens. The value of a customer not yet matched to a firm is as follows:

$$\mathcal{U} = \max_{x_u} \beta[m(\theta(x_u))x_u + (1 - m(\theta(x_u)))\mathcal{U}']. \tag{1}$$

If a customer is not matched, she does not receive any per-period utility. In the following period, she chooses a market segment, x_u, where to search by solving a trade-off between the offered utility, x_u, and the likelihood of finding a firm, $m(\theta(x_u))$. When successful, she enjoys the promised utility x_u; otherwise, she remains unmatched.

In the case of a customer matched with a firm with productivity z under the contingent contract $\omega = \{p, \tau(z'), d(z'), C'(z')\}$, the value can be written as:

$$\begin{aligned} \mathcal{C}(z, \omega) = u - p + \beta\mathbb{E}\{&(\delta + (1-\delta)d + (1-\delta)(1-d)\tau)\mathcal{U}' \\ &+ (1-\delta)(1-d)(1-\tau)\max_{x'}[m(\theta(x'))x' + (1 - m(\theta(x')))C'(z')]\}. \end{aligned} \tag{2}$$

A matched customer purchases one unit of output at a price p, valuing it at the household's marginal utility, u. The following period leads to one of three outcomes: (i) exit or relation dissolution, where the customer gets the value $\mathcal{U}'$; (ii) moving to another firm under a contract with value x' with probability $m(\theta(x'))$; or (iii) staying with the current firm and receiving promised utility $C'(z'; \omega)$. Customers entering the unmatched pool cannot search in the same period.

Eq. (2) shows that the model enables a novel margin of customer switching across firms, which in the data is as high as 10%–25% per year (Gourio and Rudanko, 2014). Moreover, this addition imposes further discipline and realism on firms' pricing dynamics by requiring the consideration of the impact of pricing decisions on customers' probability of moving to another firm, a margin highlighted by the empirical work of Paciello et al. (2019).

[5] Absent the convex cost, as the model lacks decreasing returns to scale in production, the distribution of firms would be degenerate—a common feature in this class of models.

[6] While I model separation probabilities endogenously, this is not crucial for the results, as they would hold even with externally calibrated probabilities.

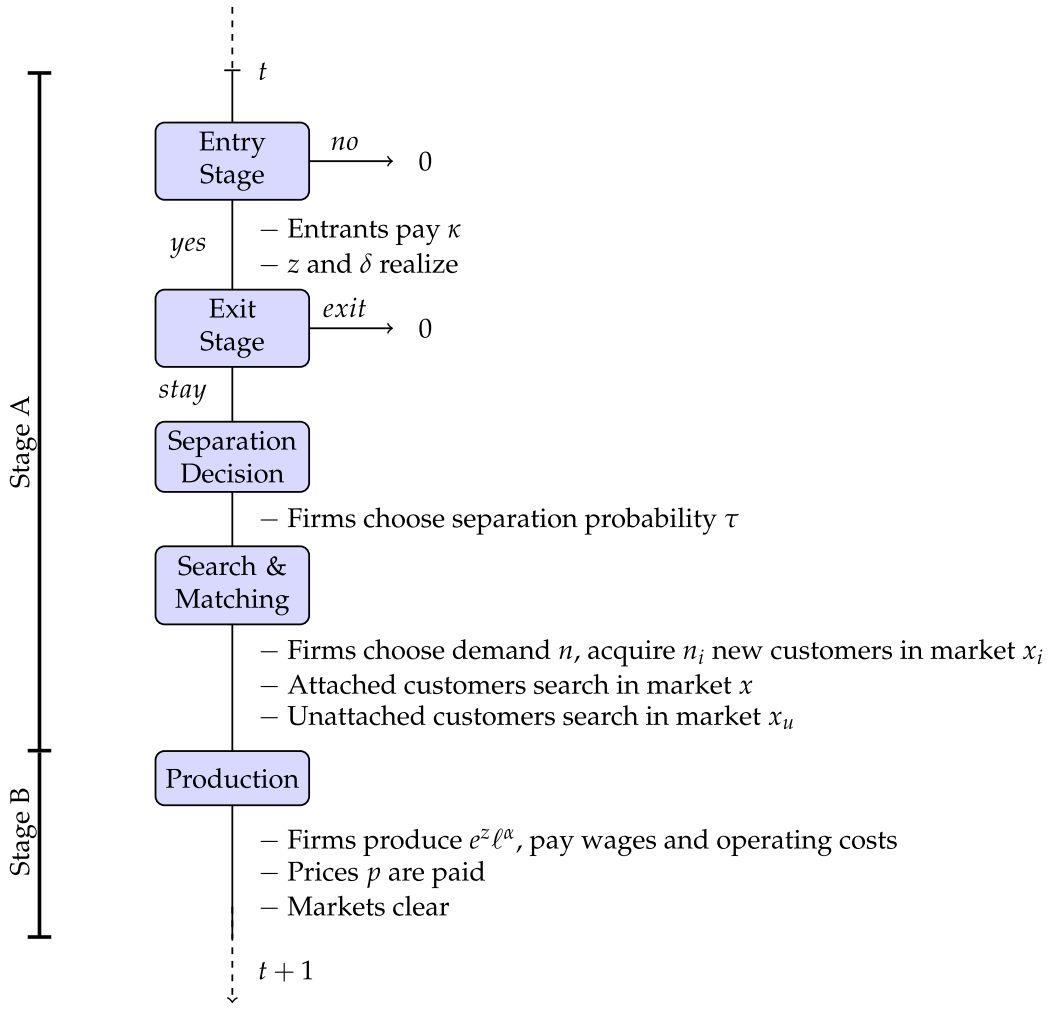


Fig. 3. Timing of the model.

3.5. Firm's problem

Consider the problem of a firm at the production stage with a measure n of customers differing in their level of promised utility. Each customer is indexed by $j \in [0, n]$ and a level of promised utility $C(j)$. The firm chooses the contracts for its customers: $\omega(j) = \{p(j), \tau(z'; j), d(z'), C'(z'; j)\}, \ \forall j \in [0, n]$. In addition, the firm decides on a submarket $x_i(z')$ where to search for new customers and chooses the number of new customers to acquire $n_i(z')$. Thus, the firm problem is as follows:

$$\begin{aligned}\mathcal{V}(z, n, \{C(j)\}_{j\in[0,n]}) &= \max_{n_i'(z'), x_i'(z'), \{\omega(j)\}_{j\in[0,n]}} \int_0^n p(j)\mathrm{d}j - W\ell - Wf \\ &\quad + (1-\delta)\beta\mathbb{E}\left\{-n_i'\frac{Wc}{q(\theta(x_i'))} - W\chi_1(n_i'/n)^2 n^{\chi_2} + \mathcal{V}(z', n', \{\widehat{C}(z'; j')\}_{j'\in[0,n']})\right\}^+,\end{aligned} \tag{3}$$

subject to:

$$y = e^z \ell^\alpha, \tag{4}$$

$$y = n, \tag{5}$$

$$n'(z') = \int_0^n (1 - \tau(z'; j))(1 - m(\theta(x'(z'; j))))\mathrm{d}j + n_i'(z'), \tag{6}$$

$$\widehat{C}(z'; j') = \begin{cases} C(z'; j) & \text{for} \quad j' \in [0, n'(z') - n_i'(z')] \quad \text{and} \quad j' = \Phi(z'; j), \\ x_i(z') & \text{for} \quad j' \in [n'(z') - n_i'(z'), n'(z')], \end{cases} \tag{7}$$

where $\Phi(z';j) = \int_0^j (1-\tau(z';k))(1-m(\theta(x'(z';k))))\mathrm{d}k$.

In the current period, the firm earns revenue, $\int_0^n p(j)\mathrm{d}j$, minus the labor cost $W\ell$ and the fixed cost, Wf. It reaches the following period if survives with probability $(1-\delta)$ and chooses to stay, as captures by the notation $\{\cdot\}^+$, standing for $\max(\cdot,0)$, which is summarize by the dummy $d(z') \in \{0,1\}$ ($d=1$ for exit). In the next period, the firm chooses the number of new customers to get $n_i'(z')$ and the submarket $x_i'(z')$ where to direct its advertisement. Because each unit of advertisement succeed with probability $q(\theta(x_i'))$, the firm pays the linear cost $n_i' Wc/q(\theta(x_i'))$ and the convex cost $W\chi_1(n_i'/n)^2 n^{\chi_2}$ of advertisement. In the rest of the paper, I define the model-implied firm-level *selling costs* as the sum of linear and convex advertisement costs, $n_i' Wc/q(\theta(x_i')) + W\chi_1(n_i'/n)^2 n^{\chi_2}$, and define the model-implied firm-level *production costs* as the wage bill, $W\ell$.

Eqs. (4) and (5) state the production and the demand constraints. Eq. (6) is the evolution of total customers, where next period customers n' are the sum of the new customers $n_i'(z')$, minus those separating with probability $\tau(z';j)$ and those leaving with probability $m(\theta(x'(z';j)))$. Thus, search frictions lead firms to expand through demand accumulation, aligning with empirical findings (Foster et al., 2008; Sterk et al., 2021; Bernard et al., 2022). Additionally, in such a setting, demand accumulation comes from new customers rather than higher sales per customer, in line with empirical evidence showing that 70% of firms' life cycle growth comes from this extensive margin (Afrouzi et al., 2020; Einav et al., 2020).

Eq. (7) keeps track of the promised utilities across customers. Because the measure of customers evolves, I use the mapping Φ to re-index the customers that stay and make sure that a customer with an original index $j \in [0, n'(z') - n_i'(z')]$ is assigned a new index $\Phi(z';j) \in [0, n'(z') - n_i'(z')]$ in the next period. Newly acquired customers with promised utility, $x_i'(z')$, are assigned an index in $[n'(z') - n_i'(z'), n'(z')]$. In addition to (4)–(7), and because of commitment on the firm side, the firm is subject to the following *promise-keeping* constraint:

$$\forall j \in [0,n], \quad C(j) \leq \boldsymbol{C}(z,\omega(j)). \tag{8}$$

Eq. (8) ensures that the contract $\omega(j)$ delivers at least the promised lifetime utility $C(j)$. Online Appendix II discusses the problem and its solution in further detail.

The firm problem involves solving a trade-off between *investing* in the customer base searching in high-value submarkets, leading to lower prices (as explained in the next section), and increasing profits *harvesting* the customer base through higher prices. More productive firms solve this investing-harvesting trade-off better, attracting customers through lower prices while maintaining higher profits through lower marginal costs of production. Thus, a rise in returns to scale by changing the curvature of the production function (4) gives a cost advantage to larger firms, explained in detail in Section 4.1, allowing them to solve the trade-off more effectively. Since competition is for the same customers, as larger firms expand, the others contract, generating a demand-based channel leading to winners-and-losers dynamics. Section 4.1 discusses the macroeconomic implications of this competition change due to higher returns to scale.

3.6. Firm's pricing

Because firms have commitment but customers do not, when a firm designs a contract, it must consider two constraints. First, the contract must take into account a *participation constraint*, given by:

$$m(\theta(x'))x' + (1-m(\theta(x')))C(z') \geq \boldsymbol{U}', \tag{9}$$

which states that the customer continuation value, conditional on remaining matched, must be higher than the value of being unmatched. Second, the contract must take into account the following *incentive constraint*:

$$x' = \arg\max_{\tilde{x}} m(\theta(\tilde{x}))\tilde{x} + (1-m(\theta(\tilde{x})))C'(z'), \tag{10}$$

which states that the submarket in which the customer is incentivized to search is the one that maximizes its continuation value conditional on remaining matched. A contract satisfying constraints (9) and (10) is said to be incentive-compatible. Given these two constraints, prices can be derived from the promise-keeping constraint (8). Hence, the price for a customer j is given by:

$$p(j) = \boldsymbol{C}(z,\{0,\tau,d,C'\}) - \varkappa(j), \tag{11}$$

where $\varkappa(j) \in \{C(j), x(j), x_u\}$, depending on the customer's past history.

Prices are the difference between the present value of being matched evaluated at today's price equal to zero, i.e., $\boldsymbol{C}(z,\{0,\tau,d,C'\})$, minus the history-dependent promised utility $\varkappa(j)$. Since incumbent customers attached to the same firm are identical, they all prefer searching in the same submarket. The firm, therefore, has an incentive to guide them toward this optimal submarket. By the incentive constraint (10), it follows that $C(j) = C$ for all incumbents, resulting in a unique price charged to incumbent customers, specifically $p^{inc} = \boldsymbol{C}(z,\{0,\tau,d,C'\}) - C$. In contrast, new customers initially differ based on the promised utility of the submarket in which their firm searches, leading to a price $p^{new} = \boldsymbol{C}(z,\{0,\tau,d,C'\}) - x$.[7] Consequently, initial prices differ according to these submarket distinctions. However, once new customers join the firm, they become identical to incumbent customers, share the same incentives, and thus search in the same submarket. From that point onward, they face the same price as incumbents. Hence, in each period, the

[7] This submarket could be either where unmatched customers search, x_u, or where customers attached to other firms search, $x \in [\underline{x}, \overline{x}]$.

firm effectively charges two distinct prices: one to incumbents and one to new customers, the latter conditional on the submarket in which the firm searches.

Eq. (11) illustrates that firms giving high value to customers can charge higher prices. Conversely, committing to provide a high promised utility requires charging lower prices. Thus, pricing reflects the investment–harvest trade-off discussed in Section 3.5. Conditional on productivity, firms expanding their customer base offer greater promised utility and, therefore, lower prices. In contrast, firms focused on extracting value from their customer base offer less utility and charge more. The prevailing incentive in equilibrium depends on firm size: smaller firms want to grow, while larger firms do not. Since firms enter the market small (as detailed in the next section), this dynamic generates a price life cycle—young firms invest via low prices, while mature firms harvest through higher prices. As this mechanism operates conditional on productivity – and thus marginal costs – the model predicts a corresponding life cycle of markups, which I validate empirically in Section 5.2.

Comparing firms of different productivity levels, more productive ones are better at solving this trade-off, offering lower prices while enjoying higher profits due to lower marginal costs of production. This edge is amplified by the cost advantage to large firms generated by a rise in returns to scale, allowing them to reduce prices further while still retaining high profits, thus outcompeting smaller firms. Crucially, since firms compete for the same demand, as larger firms attract customers, others lose them, generating a demand-based channel resulting in winners-and-losers dynamics affecting macroeconomic outcomes, as discussed in Section 4.1.

3.7. Free entry

Every period, before the idiosyncratic shock z is realized, the entrants decide whether or not to enter. Upon entry, firms pay an entry cost κ, then draw their z from a distribution $g_z(z)$. Depending on the outcome, firms may decide to exit or stay, in which case they can start searching for customers. Thus, the entering value for a firm of type z is as follows:

$$\mathcal{V}^e(z) = (1-\delta)\max_{x_e}\left\{-n_e\frac{Wc}{q(\theta(x_e))} + \mathcal{V}(z, n_e, \{C(j)\}_{j\in[0,n_e]})\right\}^+. \tag{12}$$

After drawing z and surviving the exit shock $\delta \in (0,1)$, the entrant first decides whether or not to exit, a decision captured by $\{\cdot\}^+$ and summarized in the dummy $d_e(z)$. If it stays, the firm searches for new customers $n_e \in \mathbb{R}^+$, which is a parameter of the model, and chooses a submarket, x_e, to maximize its expected value of operating, minus the linear advertisement cost $n_e Wc/q(\theta(x_e))$. Calibrating n_e to the average entrant size, with convex advertisement costs slowing down firm growth, allows a realistic life cycle in the model. Section 5.3 reports a robustness exercise showing that the results are similar when firms are allowed to choose their initial size, n_e.

The presence of free entry means that firms enter until expected profits equal entry cost κ, paid in labor units, implying the following equilibrium condition:

$$W\kappa = \int \mathcal{V}^e(z) g_z(dz). \tag{13}$$

3.8. Firm distribution and equilibrium definition

Let $g(z,n)$ be the distribution of customers across firms in stage B of the period. The dynamics of the distribution of customers across firms can be described by:

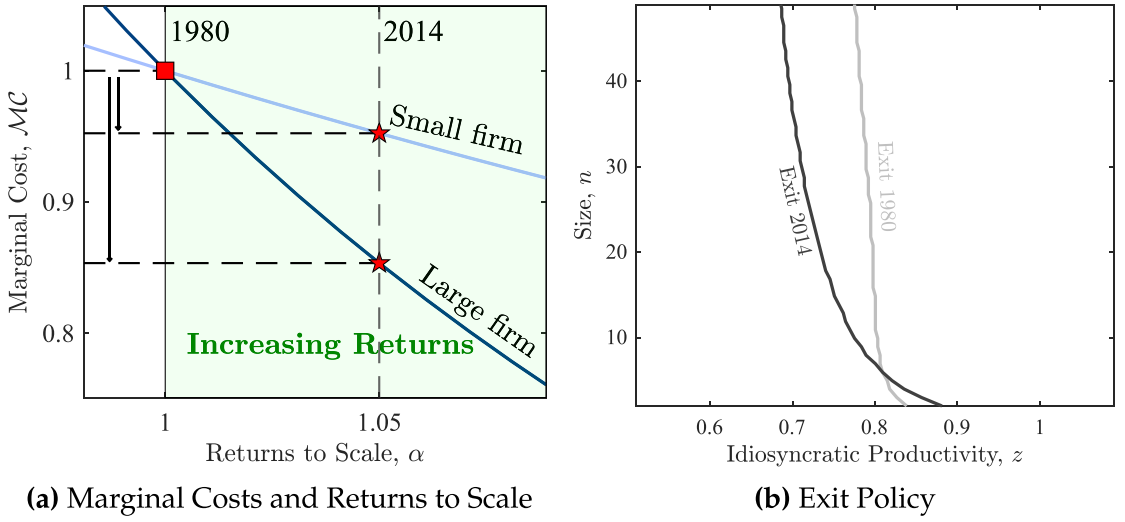
$$\begin{aligned} g(z',n') = \sum_{z,n} \mathbb{1}\{n'(z';n) = n'\}(1-d(z';n))(1-\delta)\pi(z'|z)g(z,n) \\ + \; m_e \mathbb{1}\{n_e(z') = n'\}(1-d_e(z'))(1-\delta)g_z(z'), \end{aligned} \tag{14}$$

where $\mathbb{1}\{\cdot\}$ denotes an indicator function. Eq. (14) defines the mass of firms with state (z',n') in the next period as the sum of surviving incumbent and entering firms that end up in this state. The term m_e is the endogenous measure of new entrants, defined as the number of entering firms required to reach the equilibrium market tightness in every market segment.

Finally, a *stationary recursive competitive equilibrium* consists of value functions $\{\mathcal{U}, \mathcal{C}, \mathcal{V}, \mathcal{V}^e\}$, policy functions $\{x_u, x, p, \tau, d, C', n_i, x_i, d_e, x_e\}$, a wage $\{W\}$, an invariant measure of incumbents g, and a measure of entrant firms m_e, such that: (i) $\mathcal{U}$ and x_u solve the unmatched customers' problem (1); (ii) $\mathcal{C}$ and x solve the matched customers' problem (2); (iii) $\mathcal{V}$, τ, d, n_i, and x_i solve the firms' problem (3)–(8); (iv) $\mathcal{V}^e$, d_e, and x_e solve the entrants' problem (12); (v) p and C' solve (10) and (11); (vi) the labor market clears; and (vii) the incumbent measure g satisfies (14) and the entrants measure m_e satisfies the free-entry condition (13).

4. Mechanism exploration and validation

This section uses the model to examine how rising returns to scale affect customer competition. It first shows through which mechanism higher returns to scale favor larger firms, leading to the unfolding of the secular trends of interest. The mechanism is then tested by relating changes in returns to scale to entry, reallocation, markups, and selling costs using sector- and firm-level data. Finally, it discusses how heterogeneous changes in returns to scale might affect the quantitative results.



(a) Marginal Costs and Returns to Scale **(b)** Exit Policy

Fig. 4. Returns to scale, marginal costs, and selection.
Note. Fig. 4(a) shows the relation between the firm-level marginal cost of production and the returns to scale for different firm size levels. The dark blue line represents the marginal cost of production for a large firm, and the light blue line represents the marginal cost of production for a small firm. Fig. 4(b) shows the exit threshold in the 1980 (light gray line) and 2014 (dark gray line) over the firms' state space. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Inspecting the mechanism

This section explores through which mechanism a rise in returns to scale operates. It links this technological change to changes in the marginal cost of production and how this affects firms' competition for customers and, through this, markups, business dynamism, and expenditures devoted to the accumulation of customers.

Given the production structure, as in Gourio and Rudanko (2014), the firm-level marginal cost of production is given by:

$$\mathcal{MC}(z,n) = \ell(n,z)^{1-\alpha}\frac{1}{\alpha}\frac{W}{e^z}, \tag{15}$$

where α is the firm-level returns to scale, ℓ is the number of employees, e^z is the idiosyncratic productivity, and W is the wage. Notice that, in the presence of constant returns to scale, i.e., when α equals 1, the marginal cost of production reduces to the familiar W/e^z. However, in the presence of increasing returns to scale, i.e., when α is greater than 1, the marginal cost of production varies depending on the firm's size $\ell(n,z)$.

illustrates how the marginal cost of production varies with firm size as returns to scale increase. Higher returns to scale induce a clockwise rotation of the marginal cost schedule by firm size. When calibrated to match a realistic firm size distribution, this implies that the marginal cost for large firms (dark blue line) declines more significantly than for small firms (light blue line). Although the technological change is uniform across firms, its effects are *unequal*, disproportionately benefiting larger firms and enhancing their competitive advantage.

As large firms disproportionately benefit from this technological change, gaining a competitive advantage, they become more effective at navigating the investment–harvest trade-off described in Sections 3.5 and 3.6. In particular, lower marginal costs allow them to attract customers through lower prices while sustaining higher markups. Since firms compete for the same pool of customers, the expansion of larger firms comes at the expense of smaller ones, generating a winners-and-losers dynamic. The model thus highlights a demand-driven channel through which changes in returns to scale shape firm outcomes in general equilibrium.

This demand-based channel triggered by changes in returns to scale has three main macroeconomic implications: (i) it raises the firm-level markups, (ii) it lowers business dynamism, and (iii) it increases expenditures devoted to the acquisition of customers.

First, the decrease in firms' marginal cost of production increases the value generated by the customer-firm relationship. However, because of the incomplete pass-through of costs, only part of this increase in value is passed on to customers in the form of lower prices. Firms retain the remaining part in the form of higher markups, thus leading to their aggregate increase in the new steady state. Moreover, as quantitatively the gain in marginal costs is larger for large firms, the model predicts a larger markup increase for these firms.

Second, the unequal reduction in marginal costs between large and small firms intensifies competition for customers, raising the cost of acquiring them. This shifts the selection process in favor of larger firms. Fig. 4(b) plots exit thresholds over firms' productivity and customer base in the 1980 and 2014 steady states. The shift in the 2014 threshold reflects two opposing effects of higher returns to scale: (i) lower marginal costs enable large firms to remain viable at lower productivity levels despite higher customer acquisition costs; (ii) heightened competition forces only the most productive small firms to survive, as operating in the

Table 1
Returns to scale and secular trends.

Dependent variable	Entry rate	Reallocation rate	Markups	Selling costs over production costs	
				Advertisement	Adjusted SG&A
	(1)	(2)	(3)	(4)	(5)
Returns to scale	−2.89***	−1.16***	3.15***	1.85^{+}	8.52***
	(0.34)	(0.25)	(0.63)	(1.24)	(1.30)
Sector FE	✓	✓	✓	✓	✓
Observations	592	592	586	591	592

Note: Table 1 reports regression results where entry rates, reallocation rates, 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. All variables are in logs. Entry and reallocation data are sourced from the BDS, while markups and selling costs are from Compustat. Returns to scale are the author's own sector-level estimates, derived using the ACF approach. Firm-level variables are averaged at the sector level. Sectors correspond to 2-digit NAICS classifications, consistent with the BDS definition. The time period is 1978–2014. Observations are weighted by relative sector size to reflect aggregate relevance. Robust standard errors are shown in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, + $p < 0.15$.

market becomes increasingly costly. Higher customer acquisition costs thus act as a barrier to entry, making markets less contestable, firm entry, and the reallocation of resources. As customer reallocation slows, business dynamism declines.

Third, lower marginal costs of production incentivize firms to scale up. At the same time, higher search costs mean that acquiring customers requires greater resource expenditure. Together, these forces lead firms to spend more to expand their customer base. As a result, in the second steady state, firms allocate a larger share of resources to selling costs relative to production costs.

4.2. Mechanism validation

This section tests the predictions outlined in Section 4.1, examining the relationship between returns to scale and the relevant secular trends at both the sector level and, where possible, the firm level.

Sector-level validation. A central prediction of the model is that rising returns to scale over time should lead to a decline in business dynamism – specifically, lower entry and reallocation rates – alongside higher markups and a greater share of selling costs relative to production costs. A natural test of this prediction is to examine the time-series association between sector-level returns to scale and corresponding sector-level outcomes. Since the model's predictions pertain to changes over time rather than cross-sectional differences, Table 1 presents estimates with sector fixed effects, which absorb time-invariant heterogeneity and identify the coefficients using within-sector time variation.

Overall, the model's predictions are supported by the data: time variation in sector-level returns to scale is associated with lower entry and reallocation rates – reflecting lower business dynamism – as well as higher markups and higher selling costs relative to production costs, regardless of the specific measure used. Online Appendix III.I presents binscatter plots of the regressions and robustness checks using alternative fixed effects specifications, confirming that the results are broadly consistent across most cases.

Firm-level validation. For markups and selling costs relative to production costs, firm-level measures are available. This allows us to test the association between time variation in returns to scale and corresponding variation in these firm-level outcomes. Results are presented in Online Appendix III.II and show that increases in returns to scale over time are associated with higher markups and higher selling costs relative to production costs, consistent with the model's predictions. Moreover, for markups, the theory predicts a stronger positive effect for larger firms. I test this by examining whether the association between returns to scale and markups is heterogeneous by firm size. The results, also reported in Online Appendix III.II, confirm this prediction as well.

4.3. Discussion of heterogeneous rise in returns to scale

This section briefly discusses the implications of a heterogeneous rise in returns to scale. While Section 2.3 finds no strong evidence of such heterogeneity within Compustat, the database only moderately represents the full firm population. Therefore, the possibility that Compustat firms have experienced a greater increase in returns to scale cannot be ruled out. As discussed above, the core mechanism in the model hinges on the idea that higher returns to scale confer a competitive advantage to larger firms. If Compustat firms, which tend to be larger, are also those experiencing steeper increases, this would reinforce the mechanism. Although direct evidence comparing trends between Compustat and non-Compustat firms is unavailable – since such a comparison would require observing the entire population of U.S. firms, which is not available – the quantitative exercises in Section 5 likely provide a conservative lower bound on the implications of rising returns to scale.

5. Rising returns to scale and secular trends

This section quantifies how much of the post-1980 secular trends the model explains via rising returns to scale. It first calibrates the model to match key moments in the 1980 steady state and assess its fit on untargeted dimensions. Then, it presents the quantitative results of changing returns to scale to match the estimated rise in the data. Finally, it shows that the same mechanism generates micro-level patterns consistent with the aggregate trends.

Table 2
Parameters and moments.

Fixed	Value	Description
β	0.97	Annual discount rate
u	1	Customer valuation of goods
α	1	Returns to scale
ρ	0.8	Autocorrelation idiosyncratic productivity
σ	0.2	Standard deviation idiosyncratic productivity
ψ	2.84	Frisch elasticity
ϑ	1.05	Labor supply shifter

Fitted	Value	Description	Moments	Model	Data
c	0.45 ·1e−3	Linear adv. cost	Avg. cost-weighted markup	0.20	0.17
n_e	6.79	New firms' customers	Avg. size at age zero	5.98	5.97
χ_1	0.46	Convex adv. cost 1	Avg. size at age five	12.32	10.16
χ_2	1.91	Convex adv. cost 2	Share of old (11+) firms	0.32	0.32
f	0.78	Fixed operating cost	Avg. firm size	20.24	20.25
κ	6.92	Entry cost	Entry rate	0.14	0.13
δ	0.02	Exit shock probability	Reallocation rate	0.29	0.31

Note. The table presents parameter values and target moments from both the model and the data. Data moments are calculated for the period 1977–1985. The average cost-weighted markup is from Compustat, while other target moments are from BDS. Firm size in the model is measured by the total labor employed, consistent with BDS measures.

5.1. Parametrization

The model is yearly and calibrated to the 1980 period, the onset of the secular trend of interest, in two steps. First, a set of parameters is fixed to match standard targets in the steady state. Second, given these parameter values, the remaining parameters are chosen to match identifying moments from the data.

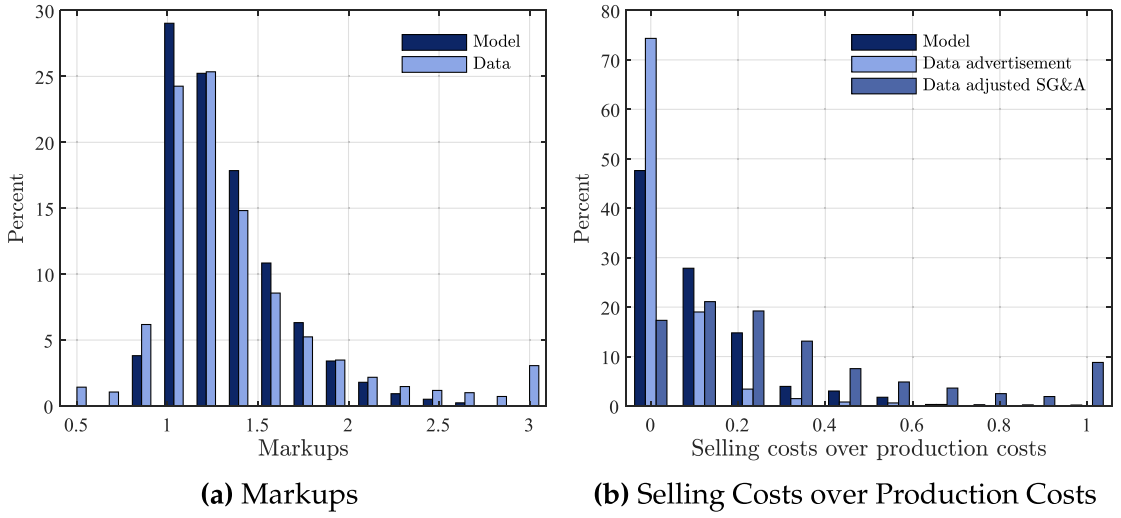
Fixed parameters. The discount rate β is fixed at 0.97, corresponding to an annual interest rate of approximately 3%. The customer valuation of goods u is set to 1, implying a marginal utility of the same value. The firm-level returns to scale α is set to 1, representing constant returns to scale, as found in Section 2.3. The persistence of the productivity shock ρ is 0.8, and the standard deviation σ is 0.2, in line with the estimates of Foster et al. (2008). The Frisch elasticity, ψ, is set to 2.84, in line with the average value commonly used in macroeconomics as reported by Chetty et al. (2011). However, since they argue for lower elasticities more consistent with micro evidence, Section 5.3 reports robustness to smaller values and shows that the main results do not depend on this choice. The labor supply shifter ϑ is set equal to 1.05, ensuring a normalized wage equal to one.

Fitted parameters. The remaining parameters, $\{n_e, \chi_1, \chi_2, f, \kappa, \delta, c\}$, are calibrated internally using cross-sectional and life-cycle moments. The parameter n_e, governing the customer base for entering firms, is informed by the number of employees of entrant firms. Given the entrant size, convex cost parameter χ_1, which governs the pace of customer accumulation, is informed by number of employees in firms of age five. The other convex cost parameter χ_2, affecting firms' likelihood to exit young by imposing a disproportionate cost of growing larger, is disciplined based on the share of firms aged 11 years or older. The operating cost f, affects the selection of firms and is chosen to match the average firm size. The entry cost κ is identified using the entry rate. The exit shock probability δ is informed based on the aggregate excess reallocation rate. Finally, the linear cost parameter c is set based on the average cost-weighted markup, reflecting the need for firms to recover sunk costs through higher markups. This is the only moment not based on statistics designed to represent the full population of firms, as it is derived from Compustat. To ensure consistency between the model and the Compustat sample, I construct the model counterpart by conditioning on firms that have survived at least five years – approximately the median time to IPO before the 2000s, according to Ewens and Farre-Mensa (2020) – and by imposing a minimum employment threshold of ten workers, below which firms are rarely observed in Compustat. This sample selection criterion is applied whenever model-implied data are compared with Compustat data.

Calibration results. Table 2 presents the model parameters and implied moments. Despite its nonlinearity, the model fits the targeted moments well. Although the calibration targets only the average cost-weighted markup, the model implies a sales-weighted markup of 0.28—close to the 0.25 reported by De Loecker et al. (2020). It also endogenously generates a customer turnover rate of approximately 9%, consistent with empirical estimates such as the 15% reported by Gourio and Rudanko (2014). In addition, the model yields an elasticity of the shrinkage of the customer base to price of 0.08, within the 0.01–0.16 range estimated by Paciello et al. (2019). However, it accounts for only 28%–53% of the price dispersion documented by Kaplan and Menzio (2015), suggesting that the difficulty search-and-matching models face in generating wage dispersion, as noted by Hornstein et al. (2011), also applies to explaining price dispersion. Additional validations of the calibration strategy and model performance are discussed below.

5.2. Validation

This section validates the model by examining its firm-level predictions for markups and selling costs. Since all validations are based on Compustat data, I construct a Compustat-like sample from the model using the same selection criteria applied in the calibration strategy in Section 5.1.



(a) Markups **(b)** Selling Costs over Production Costs

Fig. 5. Distributions of markups and selling costs over production costs.
Note: Fig. 5 presents the distributions of markups and selling costs over production costs. The data are from Compustat (1977–1990), and the model corresponds to Compustat-like subsample of the 1980 initial steady state. Dark blue bars represent the model, while light blue bars represent the data. Panel 5(a) shows the distribution of markups, and Panel 5(b) displays the distribution of selling costs over production costs, using both the baseline measure based on advertisement expenditures and an alternative measure based on adjusted SG&A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We begin by analyzing the unconditional distributions of markups and selling costs relative to production costs. Markups in the model are defined as the average price charged by the firm relative to marginal cost, as specified in Eq. (15), while the ratio of selling to production costs is measured as described in Section 3.5. Both definitions align exactly with their counterparts used in the empirical measurement. Fig. 5 presents the quantitative results. Overall, the model closely matches the distribution of markups (panel 5(a)) and performs reasonably well in capturing the distribution of selling costs (panel 5(b)), falling within the range defined by the two alternative empirical measures—one based on advertising expenditures and the other on adjusted SG&A.

Next, I examine the conditional correlations of markups and selling costs relative to production costs with firm age and size (measured by sales). Fig. 6 presents the results. Panel 6(a) shows the evolution of markups over the life cycle, while Panel 6(b) presents markups across size deciles. Panel 6(c) displays the evolution of selling costs relative to production costs over the life cycle, and panel 6(d) shows this relationship across size deciles. Overall, the model captures markup dynamics well, both over the life cycle and across firm sizes. It also performs well in replicating the life-cycle pattern of selling costs over production. However, it falls short in capturing their relationship with firm size: while the data show a steady decline across the entire size distribution, the model replicates this pattern only from the third decile onward.

Online Appendix IV.I presents regression specifications in which markups and selling costs relative to production costs are jointly regressed on firm age and size, both in the model and the data. The results confirm the patterns highlighted in Fig. 6. However, controlling for size weakens the model's quantitative performance, as it effectively leaves productivity as the sole source of firm-level heterogeneity. This suggests that while our parsimonious two-state-variable model captures many of the novel cross-sectional correlations, additional sources of heterogeneity may be needed to fully account for the conditional patterns observed in the data. Moreover, Online Appendix IV.I includes also a range of additional validation exercises. It shows that the distribution of employment and firms over the life cycle closely matches the patterns observed in BDS data. It also demonstrates that the distribution of employment growth, as well as its correlation with age and the distribution of employment across cohorts, aligns well with established empirical regularities.

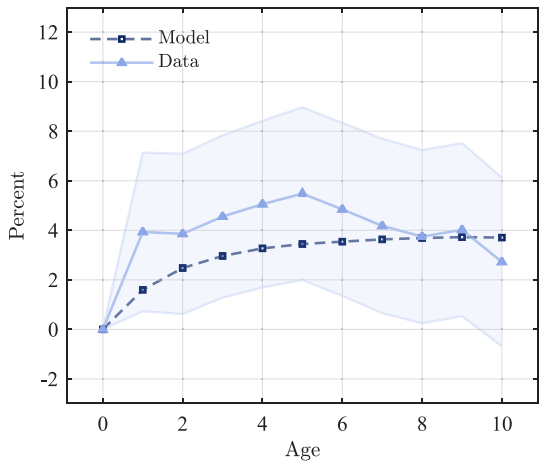
5.3. Quantitative implications

5.3.1. Quantitative implications for the secular trends

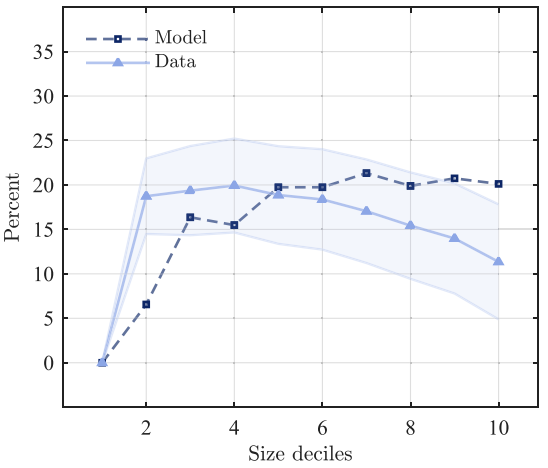
This section examines the quantitative implications of the 5% rise in returns to scale documented in Section 2.3. To do so, I compare two steady states: the 1980 steady state, whose calibration is discussed in Section 5.1, and the 2014 steady state, where I increase the returns to scale parameter α to 1.05 while keeping all other parameters fixed.

Table 3 presents the model's quantitative predictions in response to the rise in returns to scale. Overall, the model explains a substantial share of the increase in the average cost-weighted markup, the decline in business dynamism, and the rise in selling costs relative to production costs observed in the data.

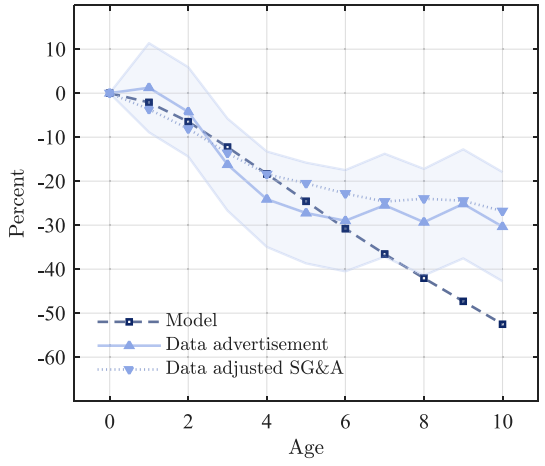
Transitional dynamics. Online Appendix IV.II.I presents results under the assumption that, by 1980, firms have full knowledge of the entire future path of returns to scale, and discusses the implications if firms instead gradually acquired knowledge of the trend



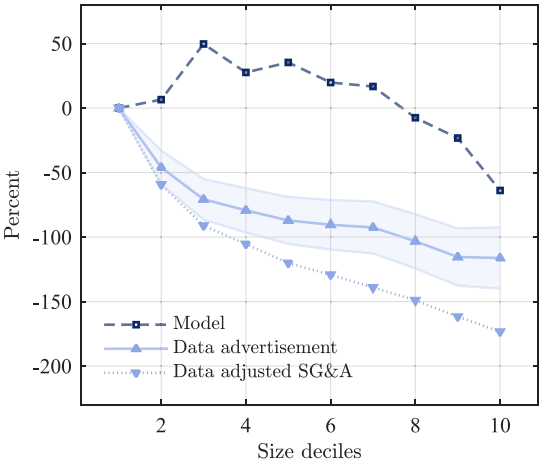
(a) Markups over the Life-Cycle



(b) Markups across the Size Distribution



(c) Selling Costs over Production Costs over the Life-Cycle



(d) Selling Costs over Production Costs across the Size Distribution

Fig. 6. Markups and selling costs over production costs over the life cycle and the size distribution.
Note: Fig. 6 compares markups and selling costs over production costs across the life cycle and the size distribution in both the data and the model. The data come from Compustat (1977–1990), while the model corresponds to the Compustat-like subsample of the 1980 initial steady state. In the data, results are obtained by estimating the following regression:

$$\log(y_{it}) = \alpha + \sum_{j=1}^{J} \beta_j \mathbf{1}\{x_{it} \in \mathcal{I}^j(x_{it})\} + \gamma_i + \gamma_{st} + \epsilon_{it}$$

Here, y_{it} denotes either markups or selling costs over production costs, and x_{it} is either firm age or sales. γ_i and γ_{st} denote firm and sector-time fixed effects, respectively. Panels 6(a) and 6(b) show the evolution of markups over the life cycle and across the sales distribution, respectively. Panels 6(c) and 6(d) present selling costs over production costs over the life cycle and across the sale distribution, respectively. The baseline measure of selling costs is constructed from advertisement. In each panel, the dashed dark blue line with squares represents the model and the solid light blue line with triangles shows the data. In Panels 6(c) and 6(d), an additional dotted light blue line with inverted triangles represents the alternative data measure based on adjusted SG&A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Quantitative implications of the rise in returns to scale.

	1980 S.S.	2014 S.S.	Model	Data
Markups				
Average				
cost-weighted markup	0.202	0.232	+15%	+42%
Business Dynamism				
Entry rate	0.139	0.093	−33%	−33%
Reallocation rate	0.295	0.232	−21%	−29%
Others				
Average selling costs				
over production costs	0.119	0.147	+23%	+60/90%

Note. Columns 1 and 2 report steady-state variables from the model. Columns 2, 3, and 4 report changes in the model and the data (BDS and Compustat). The last column indicates the fraction of empirical variation explained by the model. The average markup is calculated using cost weights, and the average selling ratio is calculated using a simple average. Empirical moments follow from Section 2. All variables in the model align with their data definitions.

in returns to scale over time. Quantitatively, the model performs well in replicating the trends in reallocation rates and selling costs over production costs, but fits entry rates and markups less accurately, especially before the 2000s. I conclude that while rising returns to scale can account for a substantial portion of the observed secular trends, particularly in magnitude, they fall short of fully explaining their timing and evolution. This suggests a significant role for additional mechanisms to complement the effects of returns to scale in explaining the full set of empirical patterns.

Robustness exercises. Here, I consider two distinct robustness exercises: (i) allowing for a lower Frisch elasticity, and (ii) permitting firms to choose their initial mass of customers. Details on the implementation of these exercises are provided in Online Appendix IV.II.II. Overall, I find that incorporating these extensions does not substantially affect the results presented in Table 3.

5.3.2. Firm-level patterns linked to the secular trends

This section shows that the model also explains a broad range of micro-level phenomena established in the literature. All facts are summarized briefly and presented in more detail in Online Appendix IV.II.III.

Firm Aging. The model accounts for most of the aging of U.S. firms emphasized by Hopenhayn et al. (2018) through a winners-and-losers mechanism that favors larger firms, which are older on average. Quantitatively, it generates a 53% increase in the share of firms aged 11 years or older, compared with 50% in the data, and a 58% decline in the employment share of firms aged 5 years or younger, compared with 56% in the data.

Evolution of Markups Distribution and Reallocation. The model explains the rise in markups through a fattening of the right tail of the firm distribution, consistent with the evidence in De Loecker et al. (2020) and the broader notion of superstar firms. In line with Autor et al. (2020), the model generates substantial reallocation toward larger firms: in the second steady state, there are more large firms, and they are larger. In particular, both the cumulative share of firms by size and the cumulative share of sales by size become flatter in the second steady state, indicating the rising importance of superstar firms in the upper tail of the distribution. Since the model features persistent but temporary productivity differences, these superstar firms are best interpreted as "shooting stars", consistent with Kehrig and Vincent (2021).

Declining Firm-Level Responsiveness. Finally, the model replicates the decline in firm-level responsiveness documented by Decker et al. (2020), namely the weakening of firms' employment responses to productivity shocks over time. Quantitatively, responsiveness declines by 0.08 in the model, compared with a decline of about 0.01 in Compustat and 0.09 in Decker et al. (2020). In the model, this attenuation arises because rising market power leads firms to pass through a smaller fraction of productivity-driven cost shocks into quantity adjustments.

6. Conclusion

This paper proposes an efficient mechanism based on firms' dynamic competition for customers, linking rising returns to scale to several U.S. secular trends through a winners-and-losers dynamic that gives rise to superstar firms. While the mechanism goes far in quantitatively explaining these trends in a manner consistent with many micro-level regularities observed during the period, analysis of the transition dynamics reveals a less satisfactory fit in earlier decades. This suggests that additional mechanisms – likely involving inefficiencies – also played a role in the evolution of U.S. secular trends since the 1980s.

I conclude by highlighting some promising avenues for future research. For example, it would be natural to examine how the rise in returns to scale, which increases the private gains from mergers and acquisitions, relates to the recent surge in such activity. Furthermore, introducing classical sources of market power, such as horizontal product differentiation, would provide a natural setting to study how much firm-level market power is efficient – arising from search-and-matching frictions – and how much is inefficient—stemming from output suppression. These extensions are left for future work.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmoneco.2026.103935.

Data availability

The authors do not have permission to share data.

References

Ackerberg, Daniel A., Caves, Kevin, Frazer, Garth, 2015. Identification properties of recent production function estimators. Econometrica 83 (6), 2411–2451.
Afrouzi, Hassan, Dernik, Andres, Kim, Ryan, 2020. Growing by the masses. Revisiting the link between firm size and market power. Working Paper.
Akcigit, Ufuk, Ates, Sina T., 2021. Ten facts on declining business dynamism and lessons from endogenous growth theory. Am. Econ. J.: Macroecon. 13 (1), 257–298.
Atkeson, Andrew, Burstein, Ariel, 2008. Pricing-to-market, trade costs, and international relative prices. Am. Econ. Rev. 98 (5), 1998–2031.
Autor, David, Dorn, David, Katz, Lawrence F, Patterson, Christina, Van Reenen, John, 2020. The fall of the labor share and the rise of superstar firms. Q. J. Econ. 135 (2), 645–709.
Baley, Isaac, Veldkamp, Laura L., 2025. The Data Economy: Tools and Applications. Princeton University Press.
Bartel, Ann, Ichniowski, Casey, Shaw, Kathryn, 2007. How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. Q. J. Econ. 122 (4), 1721–1758.
Bernard, Andrew B, Dhyne, Emmanuel, Magerman, Glenn, Manova, Kalina, Moxnes, Andreas, 2022. The origins of firm heterogeneity: A production network approach. J. Politi. Econ. 130 (7), 1765–1804.
Biondi, Filippo, Inferrera, Sergio, Mertens, Matthias, Miranda, Javier, 2023. Declining business dynamism in Europe: The role of shocks, market power, and technology. Technical Report,, Jena Economic Research Papers.
Bloom, Nicholas, Garicano, Luis, Sadun, Raffaella, Van Reenen, John, 2014. The distinct effects of information technology and communication technology on firm organization. Manag. Sci. 60 (12), 2859–2885.
Bronnenberg, Bart J., Dubé, Jean-Pierre H., Gentzkow, Matthew, 2012. The evolution of brand preferences: evidence from consumer migration. Am. Econ. Rev. 102 (6), 2472–2508.
Brynjolfsson, Erik, McAfee, Andrew, 2014. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. WW Norton & Company.
Cavenaile, Laurent, Celik, Murat Alp, Tian, Xu, 2019. Are markups too high? competition, strategic innovation, and industry dynamics. Compét. Strat. Innov. Ind. Dyn. (August 1, 2019).
Chetty, Raj, Guren, Adam, Manoli, Day, Weber, Andrea, 2011. Are micro and macro labor supply elasticities consistent? A review of evidence on the intensive and extensive margins. Am. Econ. Rev. 101 (3), 471–475.
Chiavari, Andrea, Goraya, Sampreet, 2021. The rise of intangible capital and the macroeconomic implications. Work. Pap..
Davis, Steven J, Haltiwanger, John, Jarmin, Ron, Miranda, Javier, Foote, Christopher, Nagypal, Eva, 2006. Volatility and dispersion in business growth rates: publicly traded versus privately held firms. NBER Macroecon. Annu. 21, 107–179.
De Loecker, Jan, Eeckhout, Jan, Mongey, Simon, 2021. Quantifying market power and business dynamism in the macroeconomy. NBER Working Paper.
De Loecker, Jan, Eeckhout, Jan, Unger, Gabriel, 2020. The rise of market power and the macroeconomic implications. Q. J. Econ. 135 (2), 561–644.
De Ridder, Maarten, 2019. Market power and innovation in the intangible economy. Working Paper.
Decker, Ryan, Haltiwanger, John, Jarmin, Ron, Miranda, Javier, 2014. The role of entrepreneurship in US job creation and economic dynamism. J. Econ. Perspect. 28 (3), 3–24.
Decker, Ryan A, Haltiwanger, John, Jarmin, Ron S, Miranda, Javier, 2016. Declining business dynamism: what we know and the way forward. Am. Econ. Rev. 106 (5), 203–207.
Decker, Ryan A., Haltiwanger, John, Jarmin, Ron S., Miranda, Javier, 2020. Changing business dynamism and productivity: shocks versus responsiveness. Am. Econ. Rev. 110 (12), 3952–3990.
Díez, Federico J, Fan, Jiayue, Villegas-Sánchez, Carolina, 2021. Global declining competition? J. Int. Econ. 132, 103492.
Döpper, Hendrik, MacKay, Alexander, Miller, Nathan, Stiebale, Joel, 2021. Rising markups and the role of consumer preferences. Available At SSRN 3939126.
Dubé, Jean-Pierre, Hitsch, Günter J., Rossi, Peter E., 2010. State dependence and alternative explanations for consumer inertia. Rand J. Econ. 41 (3), 417–445.
Dupraz, Stéphane, 2024. A kinked-demand theory of price rigidity. J. Money, Credit. Bank. 56 (2–3), 325–363.
Einav, Liran, Klenow, Peter J., Levin, Jonathan D, Murciano-Goroff, Raviv, 2020. Customers and retail growth. Working Paper.
Ewens, Michael, Farre-Mensa, Joan, 2020. The deregulation of the private equity markets and the decline in IPOs. Rev. Financ. Stud. 33 (12), 5463–5509.
Foster, Lucia, Haltiwanger, John, Syverson, Chad, 2008. Reallocation, firm turnover, and efficiency: selection on productivity or profitability? Am. Econ. Rev. 98 (1), 394–425.
Gao, Wei, Kehrig, Matthias, 2017. Returns to scale, productivity and competition: empirical evidence from US manufacturing and construction establishments. Working Paper.
Gourio, Francois, Rudanko, Leena, 2014. Customer capital. Rev. Econ. Stud. 81 (3), 1102–1136.
He, Bianca, 2022. Unveiling intangibles. Available At SSRN 4321762.
He, Bianca, Mostrom, Lauren I., Sufi, Amir, 2024. Investing in Customer Capital. Technical Report, National Bureau of Economic Research.
Hopenhayn, Hugo A., 1992. Entry, exit, and firm dynamics in long run equilibrium. Econ.: J. Econ. Soc. 1127–1150.
Hopenhayn, Hugo, Neira, Julian, Singhania, Rish, 2018. The rise and fall of labor force growth: implications for firm demographics and aggregate trends. NBER Working Paper.
Hornstein, Andreas, Krusell, Per, Violante, Giovanni L., 2011. Frictional wage dispersion in search models: a quantitative assessment. Am. Econ. Rev. 101 (7), 2873–2898.
Kaplan, Greg, Menzio, Guido, 2015. The morphology of price dispersion. Internat. Econom. Rev. 56 (4), 1165–1206.
Kaplan, Greg, Zoch, Piotr, 2020. Markups, Labor Market Inequality and the Nature of Work. Technical Report,, National Bureau of Economic Research.
Kariel, Joel, Savagar, Anthony, 2022. Returns to scale and productivity in the macroeconomy. Working Paper.
Kehrig, Matthias, Vincent, Nicolas, 2021. The micro-level anatomy of the labor share decline. Q. J. Econ. 136 (2), 1031–1087.
Kleshchelski, Isaac, Vincent, Nicolas, 2009. Market share and price rigidity. J. Monet. Econ. 56 (3), 344–352.
Kost, Kyle, Pearce, Jeremy, Wu, Liangjie, 2019. Market power through the lens of trademarks. Working Paper.
Lashkari, Danial, Bauer, Arthur, Boussard, Jocelyn, 2021. Information technology and returns to scale. Working Paper.
Nakamura, Emi, Steinsson, Jón, 2011. Price setting in forward-looking customer markets. J. Monet. Econ. 58 (3), 220–233.
Paciello, Luigi, Pozzi, Andrea, Trachter, Nicholas, 2019. Price dynamics with customer markets. Internat. Econom. Rev. 60 (1), 413–446.

Ravn, Morten, Schmitt-Grohé, Stephanie, Uribe, Martin, 2006. Deep habits. Rev. Econ. Stud. 73 (1), 195–218.
Roldan-Blanco, Pau, Gilbukh, Sonia, 2020. Firm dynamics and pricing under customer capital accumulation. J. Monet. Econ..
Schaal, Edouard, 2017. Uncertainty and unemployment. Econometrica 85 (6), 1675–1721.
Sterk, Vincent, Sedláček, Petr, Pugsley, Benjamin, 2021. The nature of firm growth. Am. Econ. Rev. 111 (2), 547–579.
Weiss, Joshua, 2019. Intangible investment and market concentration. Working Paper.
Wright, Randall, Kircher, Philipp, Julien, Benoît, Guerrieri, Veronica, 2021. Directed search and competitive search equilibrium: A guided tour. J. Econ. Lit. 59 (1), 90–148.