UniCredit Foundation

"The Macroeconomics of Rising Returns to Scale"

Andrea Chiavari

8th Econ Job Market Best Paper Award

Working Paper Series

n. 160 ■ December 2021

Statement of Purpose

The Working Paper series of the UniCredit Foundation is designed to disseminate and to provide a platform for discussion of either work of the UniCredit economists and researchers or outside contributors (such as the UniCredit Foundation scholars and fellows) on topics which are of special interest to the UniCredit Group. To ensure the high quality of their content, the contributions are subjected to an international refereeing process conducted by the Scientific Committee members of the Foundation.

The opinions are strictly those of the authors and do in no way commit the Foundation and UniCredit.

These Working Papers often represent preliminary work. Citation and use of such a paper should take account of its provisional character.

The Working Papers are also available on our website (http://www.unicreditfoundation.org)

The Macroeconomics of Rising Returns to Scale:
Customer Acquisition, Markups, and Dynamism

Andrea Chiavari*

First draft: October 25, 2020. This draft: October 29, 2021

Abstract

This paper studies the macroeconomic implications of the rise in firm-level scale economies. My empirical finding is that the average firm-level returns to scale increased within all US sectors, going from 1 to 1.05 between 1980 and 2014. Simultaneously, business dynamism declined, markups rose, and firms devoted increasing resources to customer acquisition, suggesting their active involvement in building and exploiting scales. To jointly account for these facts, I propose a novel theory of firm dynamics grounded in directed search in the product market. Search frictions microfound the customer accumulation process and the presence of heterogeneous markups. The rise in returns to scale explains 62-70% of the decline in business dynamism; 29% of the increase in markups; and 14-45% of the growth in expenditures devoted to customers acquisition. Additionally, the model rationalizes further facts: the aging of US firms, the reallocation of sales toward high markup firms, and firms' declining responsiveness to productivity shocks.

Keywords: Markups, Business Dynamism, Technological Change, Production Function, Search-and-Matching, Customer Acquisition

JEL Codes: D21, D24, L11, L16, E22

*Department of Economics, Universitat Pompeu Fabra. Email: andrea.chiavari@upf.edu. I am grateful to my advisors Isaac Baley and Edouard Schaal for for their invaluable support. I sincerely thank Andrea Caggese and Jan Eeckhout for the long discussions that substantially improved the paper. I also thank Davide Debortoli, Jordi Galí, Manuel Garcia-Santana, Sampreet Goraya, Akhil Ilango, Priit Jeenas, Matthias Kehrig, Marta Morazzoni, Sandro Shelegia, Jaume Ventura as well as all the participants at the CREI macro lunch, 10th Annual Search and Matching Conference, 3rd QMUL Economics and Finance Workshop, Spring Meeting of Young Economists, Asian Summer Meeting of the Econometric Society, BGSE Summer Forum Market Power and the Labor Market, Young Economist Symposium, European Summer Meeting of the Econometric Society, and Workshop of the Spanish Macroeconomics Network. All errors are my own.

1 Introduction

Over the last decades, firm-level production processes have undergone spectacular transformations. The introduction of new technologies, such as information and communication technologies (ICT), and extensive data availability have changed the way firms organize their production. The potential of these technological advancements to expand firms' scale economies—the cost advantages that firms obtain due to their scale of operation—has captured the attention of academic researchers.[1] Meanwhile, US policymakers' concerns about the effect of these changes on firms' pricing strategies and competition for customers have gained momentum.[2] This is because, rising scale economies, manifesting through lower costs for the largest firms, may have enabled these same firms to become highly effective in pricing, attracting customers, and exerting market power. Simultaneously, firms have devoted increasing resources to customers acquisition throughout activities such as advertisement and trademarks, suggesting their active involvement in building and exploiting scales.[3]

These technological transformations in firm-level production processes may explain why the US economy has experienced noteworthy trends over the same period of time. In particular, business dynamism—the entry rate of new firms and the reallocation rate of labor across firms—has declined steadily while markups have risen.[4] This has led some observers to speculate that the engine of US productivity may have slowed down, and that its economy may have moved from a competitive to a rent-based one.[5]

However, to date, few studies have systematically analyzed the evolution of firm-level scale economies. What are the consequences of this technological transformation for the above US trends? This project aims to provide an explanation that links these phenomena and makes two contributions. First, I use firm-level data from Compustat to investigate the evolution of firm-level returns to scale in production in the US between 1980 and 2014. Second, I propose a novel theoretical framework to study the implications of changes in returns to scale through their impact on customer accumulation.

To study the evolution of returns to scale in the US economy, I estimate the firm-level production function. Here, I follow two state-of-the-art techniques. The first is the control function approach, as

[1] Bloom, Garicano, Sadun, and Van Reenen (2014) show the link between better information technologies and a wider firm-level span of control.

[2] Khan (2016), now chair of the Federal Trade Commission, argued extensively about her worries regarding the pricing strategies adopted by Amazon and how this might be the outcome of the firm's scale economies.

[3] Kost, Pearce, and Wu (2019) document the rise in trademark activities in the US and how this is associated with market power, whereas De Loecker, Eeckhout, and Unger (2020) show that firms spending more on selling general and administrative are associated with higher markups.

[4] Decker, Haltiwanger, Jarmin, and Miranda (2014) document the decline in business dynamism, that is, the slowdown in the entry rate of new firms and the reallocation rate of labor across firms. De Loecker, Eeckhout, and Unger (2020) show the rise in markups.

[5] Decker, Haltiwanger, Jarmin, and Miranda (2016) explain how declining business dynamism may impair the reallocation process across firms, and hence, lower US productivity. Philippon (2019) and Eeckhout (2021) discuss some of the potential reasons why the US economy has become less competitive.

in Ackerberg, Caves, and Frazer (2015), widely used by the empirical Industrial Organization literature. Second, I use the cost shares approach adopted by Syverson (2004) and Foster, Haltiwanger, and Syverson (2008). Estimating production technologies in two-digit sectors and over time, as in De Loecker, Eeckhout, and Unger (2020), I find a 5% increase in the average returns to scale, going from 1 in 1980 to 1.05 in 2014. Additionally, this rise shows an acceleration around 1990, consistent with the ICT acceleration ongoing in the same period.[6] Estimating production technologies at the sector level makes it possible to go beyond the analysis to the evolution of the average returns to scale—which could neglect distributional changes across sectors—and to study alternative reasons for this rise, exploiting cross-sectional variation. In particular, there are two potential reasons why the average returns to scale may have risen. First, returns to scale may have increased *within* all sectors. Second, there could have been a reallocation of economic activity *between* sectors toward sectors with ex-ante higher returns to scale. To study these two possibilities, I exploit a statistical decomposition at the sector level, which shows that the rise in the average returns to scale is a within-sector phenomenon.

Although other works have noticed the rise in returns to scale, this paper is the first to highlight the within-sector nature of this phenomenon. This novel fact is consistent with the view that US firms have undergone a technological transformation that has enabled them to increase their scale of operations.[7] I interpret the estimated increase in returns to scale as an exogenous technological change, seeking to understand its consequences for the US economy and the recent trends mentioned above.

To understand the consequences of this technological change, I propose a novel model of customer accumulation. The framework builds on Gourio and Rudanko (2014) and Roldan-Blanco and Gilbukh (2020) and brings additional tools from the labor-search literature to model customer switching across firms, which in the data is between 10-25% a year.[8] Accounting for customer switching imposes discipline on market power dynamics, as firms internalize the effect of their pricing decisions on their customer base endogenous attrition. To do so, I introduce directed search in the product market, which implies that firms use prices and markups to compete for customers. Further, search frictions imply that firms devote resources to contact new customers. In the model, the presence of fixed operating costs introduces the endogenous entry and exit of firms as standard in most firm dy-

[6] For instance, the World Wide Web entered everyday life in the first period of the 1990s.

[7] Haskel and Westlake (2018) argue in their book that the rise of intangible capital—which is highly related to the digital revolution—has increased the ability of firms to scale their production. Newman (2014), Agrawal, Gans, and Goldfarb (2018), Begenau, Farboodi, and Veldkamp (2018), Goldfarb and Trefler (2018), Carriere-Swallow and Haksar (2019), and Jones and Tonetti (2020) all emphasize the potential role of data, particularly gathering information from the customer base, for the rise of returns to scale and the presence of increasing returns. Lashkari, Bauer, and Boussard (2021) document, using French data, that the adoption of ICT inputs has allowed firms to improve their organization, helping them to improve their scale economies and giving rise to higher returns to scale.

[8] See, for example, the value surveyed by Gourio and Rudanko (2014) from industry estimates.

namics frameworks à la Hopenhayn (1992). Therefore, while remaining tractable for computational analysis, the framework can manage a rich set of firm-level facts and aggregate trends.

The model is grounded in search frictions in the product market. Search frictions microfound firm-level investments in the customer base and firms' strategic use of prices and markups to attract and retain customers, which are an established feature of the firms' activities.[9] Perhaps most importantly, they align the model with the literature pioneered by Foster, Haltiwanger, and Syverson (2008), which shows that firms mostly grow by accumulating demand. In this vein, recent empirical works by Afrouzi, Dernik, and Kim (2020) and Einav, Klenow, Levin, and Murciano-Goroff (2020) show that customer accumulation accounts for 70% of firms' overall life-cycle growth.

I calibrate the model to the 1980s period using identifying moments of the firms' life-cycle, business dynamism statistics from that period, and moments related to firm-level markups. First, as a validation exercise, I show that the model is consistent with a range of cross-sectional and firm-level facts. Second, I demonstrate that the introduction of customer accumulation through search frictions improves the general fit of the model on a series of important but often neglected firms' life-cycle facts. In particular, the model captures the upward sloping life-cycle path of markups and the downward sloping life-cycle path of selling-expenditures, relative to production costs, as observed in the microdata.

In the model, a rise in returns to scale reduces the marginal cost of production and, due to the properties of increasing returns to scale in production, reduces it by more for the biggest firms. This implies that the biggest firms in the economy become very effective in pricing, attracting customers, and charging markups. Therefore, although all firms are subject to the same change, its outcome is highly unequal, as it favors the biggest firms in the economy. This decline in marginal costs has three direct implications: (i) it increases the willingness of firms to scale up, and hence, their expenditures devoted to customer acquisition; (ii) it raises the firm-level markups due to the presence of incomplete pass-through; and (iii) it weakens the selection process in the model, implying a lower entry and reallocation rate. It is noteworthy that the first prediction—that is, the *endogenous* rise in selling-related expenditures relative to production costs after a rise in returns to scale—is a unique feature of this model, where firms invest in their demand through selling-related expenditures.[10] I test and confirm all the predictions in the cross-section of sectors of the Compustat data: I find that higher returns to scale in a sector are positively associated with higher average markups and higher average

[9] Dubé, Hitsch, and Rossi (2010) and Bronnenberg, Dubé, and Gentzkow (2012) document the prevalence of long-term customer relations. Ruhl and Willis (2008) and Eaton, Eslava, Kugler, and Tybout (2009) show that the buildup of market shares is a slow process. Paciello, Pozzi, and Trachter (2019) show that customers are sensitive to prices and that firms consider this while setting them.

[10] Models in which market power comes from horizontal differentiation (Dixit and Stiglitz (1977), Kimball (1995), and Atkeson and Burstein (2008)) or search frictions, with only strategic pricing (Paciello, Pozzi, and Trachter (2019) and Roldan-Blanco and Gilbukh (2020)), would not be able to produce the aforementioned facts as, normally, the only non-production costs they feature are fixed costs.

selling-related expenditures, relative to production costs, and negatively associated with entry and reallocation rates.

I use the calibrated model to study the macroeconomic consequences of the observed rise in returns to scale. This technological change explains 62-70% of the decline in business dynamism; 29% of the increase in markups; and 14-45% of the growth in expenditures devoted to customer acquisition. Additionally, I show that this technological change is consistent with the phenomenon of the aging of firms, as documented in the data by Hopenhayn, Neira, and Singhania (2018). It reproduces the reallocation of economic activity toward high markup firms, which gives rise to the fattening of the right tail of the markup distribution, as documented by Autor, Dorn, Katz, Patterson, and Van Reenen (2020), De Loecker, Eeckhout, and Unger (2020), and Kehrig and Vincent (2021). It explains the decline in firm-level responsiveness to productivity shocks, which Decker, Haltiwanger, Jarmin, and Miranda (2020) document as a central component of the decline in business dynamism. Although the rise in returns to scale does not fully account for the markup increase, my investigation suggests that they are an important factor.

Literature Review. This paper contributes to several strands of the literature. It first relates to the search and matching literature on both the labor and the product market. Labor market papers that first introduced some of the techniques used in this paper are Moen (1997), Menzio and Shi (2010), and Menzio and Shi (2011). I build on the methodology developed by Schaal (2017), which, however, focuses on the labor market. Closer to my focus are Gourio and Rudanko (2014), Paciello, Pozzi, and Trachter (2019), and Roldan-Blanco and Gilbukh (2020), which all develop heterogeneous firms models with search frictions in the product market.[11] Relative to Gourio and Rudanko (2014) and Roldan-Blanco and Gilbukh (2020), I allow incumbent customers to search, which is a feature of reality and an important factor for firms' pricing decisions. Moreover, compared to Gourio and Rudanko (2014), I allow for commitment on the firm side, which enables firms to charge different prices, even to their incumbent customers. In the absence of commitment, all firms would ask the same price to the incumbent customers, equal to their marginal evaluation, which would make the model quantitatively unsuited to study dispersion in markups coming from different pricing strategies. Differently from Paciello, Pozzi, and Trachter (2019) and Roldan-Blanco and Gilbukh (2020), I allow for increasing returns production technology and firm-level expenditures for customer accumulation, which are all fundamental features for the objective of this paper.

This paper also contributes to the empirical literature that has analyzed technological changes in

[11] Burdett and Coles (1997) study the role of firm size for pricing when firms use the price to attract new customers. Dinlersoz and Yorukoglu (2012) provide a theoretical model of industry dynamics in the presence of information frictions. Burdett and Judd (1983), Menzio and Trachter (2015), Burdett and Menzio (2018), and Menzio and Trachter (2018) study equilibrium price dispersion without relying on firm heterogeneity.

the firm-level production process. Chiavari and Goraya (2021) show that firms' production technology has become more intangible intensive, at the expense of labor, and that this has had significant implications for the changes in the US factor shares. More closely related to this paper is the work by Lashkari, Bauer, and Boussard (2021), using French data to show that firms employed ICT investment to increase their firm-level returns to scale; however, they do not analyze its implications for markups. Relative to them, I focus on the US, documenting the within-sector increase in firm-level returns to scale, showing that this has had sizeable consequences for the rise in markups. Despite the focus on the evolution of markups, De Loecker, Eeckhout, and Unger (2020) also document a rise in returns to scale. Yet, they do not focus on sector-level patterns, which I claim are essential in understanding the source of this increase.

Furthermore, this paper complements the growing literature that studies the potential explanations behind the rise in markups and the decline in business dynamism. A strand of this literature emphasizes demographic changes as a relevant factor behind these trends. Papers of this kind are Karahan, Pugsley, and Şahin (2019), Hopenhayn, Neira, and Singhania (2018), Peters and Walsh (2019), and Bornstein (2018). Alternatively, Liu, Mian, and Sufi (2020) hypothesize that lower interest rates can explain certain recent trends. Relative to this strand of the literature, this project emphasizes technological factors as a potential force driving these trends.

Another strand of the literature, closer to this project, emphasizes the technological factors behind the rise in markups and the decline in business dynamism. Papers in this vein are Akcigit and Ates (2021), De Ridder (2019), Weiss (2019), and De Loecker, Eeckhout, and Mongey (2021).[12] Akcigit and Ates (2021) argue that a decline in productivity spillovers from leaders to laggards is a driver of some recent trends. De Ridder (2019) emphasizes that the rise of firms that are better at using intangibles (as intangibles make other factors more productive) is important for the rise in markups, the decline in business dynamism, and productivity growth. Weiss (2019) shows how intangibles can explain the rise in markups and concentration. De Loecker, Eeckhout, and Mongey (2021) document that the rise in fixed costs and the decline in the number of potential entrants can jointly explain the rise in markups and the decline in business dynamism. I contribute to this literature by studying a different technological change—the rise of returns to scale in production—grounded outside the model in a detailed micro-level analysis. Leveraging Industrial Organization techniques to estimate the firm-level production function allows me to infer the strength of the technological change occurring in the US, bringing extra discipline outside the model to the quantitative analysis. The analysis

[12] Korinek, Ng, and Hopkins (2018) and Martinez (2018) relate automation to the rise in concentration and to the labor share decline. Crouzet and Eberly (2019) and Zhang (2019) relate the rise in intangibles with the rise in concentration. Hsieh and Rossi-Hansberg (2019) suggest that the shift toward more productive technologies with higher fixed costs can explain the divergence behind local and aggregate concentration. Aghion, Bergeaud, Boppart, Klenow, and Li (2019) and Olmstead-Rumsey (2019) link the rise in concentration to the decline in productivity growth.

of an alternative technological transformation also provides a new perspective to the ongoing debate regarding the causes of these US trends. Moreover, using a novel quantitative framework, I study additional implications compared to the previous literature. In particular, the model explains the rise in firm-level expenditures devoted to customer accumulation as firms desire to increase their scale of operation to take full advantage of the rise in scale economies.

Outline. Section 2 presents the empirical methodology and empirical findings of the paper. Section 3 introduces the theoretical model. Section 4 calibrates the model and evaluates the performance of the model using firm-level and cross-sectional facts. Section 5 analyzes and discusses the impact of rising returns to scale before quantifying implications for the aggregate trends objective of this paper. Section 6 concludes.

2 Empirical Evidence

In this section, I present the empirical analysis of this project: (i) I introduce the main dataset used throughout the analysis; (ii) then, I introduce the main empirical methodology used to estimate firm-level returns to scale; (iii) finally, I document a rise in returns to scale in production within the last three decades.

2.1 Data

In this paper, I use two main data sources: Compustat and BDS data. The former is used to obtain information on US firms, while the latter is used to obtain representative measures for the US economy.

Compustat. The main data source is Compustat, a firm-level database with all US publicly traded firms between 1977 to 2014.[13] In this section, I discuss the strengths and limitations of this dataset. I provide more details on the data-cleaning process in Appendix A.1.

The choice of data is driven solely by the ability of these data to cover the period of interest and the largest number of sectors. These characteristics make these data an excellent source of firm-level information to study technological changes in production undertaken by US firms.

Even though publicly traded firms are few relative to the total number of firms ()as they tend to be the largest firms in the economy) they account for roughly 30% of US employment (see, Davis, Haltiwanger, Jarmin, Miranda, Foote, and Nagypal (2006)). The Compustat data contain information on firm-level financial statements, including measures of sales, input expenditures, capital stock information, and a detailed industry activity classification.

[13] This is also the frame for which the BDS data are available.

However, despite its many virtues, these data present two main limitations: (i) the fact that it is impossible to distinguish quantity and prices, which makes measurement of the production function elasticities significantly more challenging as extensively explained in the next section;[14] and (ii) the possible selection issues arising from using only publicly traded firms. To address the first concern, I follow the methodologies explained in Section 2.2.1. Moreover, whenever possible, I compare my results with additional data sources to isolate the potential bias of using only publicly traded firms.

BDS data. To obtain representative aggregate US measures of the firms' size distribution and business dynamism, I use the publicly available dataset from the Business Dynamics Statistics (BDS) program of the Census Bureau.[15]

2.2 Production Function Estimation

To estimate firm-level returns to scale, I follow De Loecker, Eeckhout, and Unger (2020) and use two main approaches: (i) the control function approach and (ii) an "augmented" cost shares approach. Both of these approaches are popular methods used to estimate firm-level production functions. I review here the two methodologies, emphasizing their virtues and their limitations.

2.2.1 Control Function Approach

The control function approach was pioneered by Olley and Pakes (1996), and developed further by Levinsohn and Petrin (2003) and Ackerberg, Caves, and Frazer (2015). The main insight from this literature is that firm-level unobservable productivity can be proxied by some variable expenditure.

To overcome some of the criticism emphasized in Gandhi, Navarro, and Rivers (2020), I work with a structural value-added specification, as in Ackerberg, Caves, and Frazer (2015) and De Loecker and Scott (2016), given by:

$$\mathcal{Q}_{it} = \min\left\{K_{it}^{\beta^k} L_{it}^{\beta^\ell} \exp(\omega_{it} + \varepsilon_{it}),\ \beta^m M_{it}\right\}, \tag{1}$$

where $\mathcal{Q}_{it}$ is output, K_{it} is capital, L_{it} is labor, ω_{it} is log-productivity, ε_{it} is the error term, and M_{it} is the materials. This structural value-added production function yields the following first-order condition:

$$\mathcal{Q}_{it} = K_{it}^{\beta^k} L_{it}^{\beta^\ell} \exp(\omega_{it} + \varepsilon_{it}), \tag{2}$$

justifying the regression of $\mathcal{Q}_{it}$ on capital and labor while ignoring materials. One caveat is that,

[14] This challenge is present in most of the production data.

[15] https://www.census.gov/programs-surveys/bds/data/data-tables.html.

in theory, equation (2) may not be satisfied in certain situations. If capital and labor are quasi-fixed, and the materials are a flexible input, then when output prices are sufficiently low relative to the price of materials, it will be better to set $M_{it} = 0$ and not produce at all. However, given that my data only include actively producing firms, I assume that equation (2) always holds.[16] Therefore, under the specification in equation (1), the estimation of the firm-level production function reduces to:

$$q_{it} = \beta^k k_{it} + \beta^\ell \ell_{it} + \omega_{it} + \varepsilon_{it}, \tag{3}$$

where $q_{it} = \log(\mathcal{Q}_{it})$, $k_{it} = \log(K_{it})$, and $\ell_{it} = \log(L_{it})$. As usual, the main identification challenge to the production function estimation is the simultaneity bias induced by the unobserved time-varying firm-level productivity, ω_{it}. I follow the control function literature, and in particular Ackerberg, Caves, and Frazer (2015) and De Loecker, Eeckhout, and Unger (2020), to estimate the production function in (3) using a two-step approach based on the use of a control function for the productivity process. The identification relies on the observation that the firm's labor demand is given by a policy function of the form $\ell_{it} = \ell(k_{it}, \omega_{it})$. Then, providing that the policy function is invertible, the productivity process can be proxied by a control function given by $\omega_{it} = \omega(k_{it}, \ell_{it})$, where $\omega(\cdot) = \ell^{-1}(\cdot)$.[17]

Therefore, in the first stage of this estimation procedure, I clean the firm-level output value from the measurement errors and unanticipated productivity shocks, regressing output on a polinomial of capital, labor, and potential demand shifters, given by:

$$q_{it} = \mathcal{P}(k_{it}, \ell_{it}, \mathbf{d}_{it}) + \varepsilon_{it}. \tag{4}$$

Then, in the second stage, using the estimate $\widehat{\mathcal{P}}$ from the previous stage, I can construct a measure of productivity that does not depend on the measurement error ε_{it}, given by:

$$\omega_{it}(\beta^k, \beta^\ell) = \widehat{\mathcal{P}}(k_{it}, \ell_{it}, \mathbf{d}_{it}) - \beta^k k_{it} - \beta^\ell \ell_{it}. \tag{5}$$

Finally, taking advantage of the assumption that productivity follows an AR(1) process, it is possible to construct a measure of productivity innovations given by:

$$\xi(\beta^k, \beta^\ell, \rho) = \omega_{it}(\beta^k, \beta^\ell) - \rho\omega_{it-1}(\beta^k, \beta^\ell). \tag{6}$$

Therefore, using the productivity innovations, I construct a set of moment conditions to estimated

[16] For a more detailed discussion on this issue, see Ackerberg, Caves, and Frazer (2015).

[17] The assumptions needed to ensure the invertibility of the policy functions associated with a wide class of production functions have been discussed extensively by Pakes (1994), Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg, Caves, and Frazer (2015).

the parameters of the production function, given by:

$$\mathbb{E}(\tilde{\xi}(\beta^k, \beta^\ell, \rho) \times \mathbf{z}_{it}) = \mathbf{0}_{Z\times 1}, \tag{7}$$

where $Z \geq 3$ and, under the assumption that firms react to unanticipated productivity shocks contemporaneously and that capital is predetermined, the set of admissible instruments is $\mathbf{z}_{it} \in \{k_{it}, \ell_{it-1}, k_{it-1}, \dots\}$. Once the output elasticities are obtained, it is straightforward to recover the returns to scale as:

$$\alpha = \beta^k + \beta^\ell. \tag{8}$$

Units. It is well known that most of the time, standard production data, such as Compustat, record revenues and expenditures rather than the physical production and input used. In the presence of product differentiation (be it through physical attributes or location), an additional source of endogeneity presents itself through unobserved output and input prices.[18] This implies that, when bringing the model to the data, the structural value-added production function takes the following form:

$$q_{it} + p_{it} = \beta^k(k_{it} + p_t^k) + \beta^\ell(\ell_{it} + p_{it}^\ell) + \omega_{it} + \varepsilon_{it}, \tag{9}$$

where p_{it} is the output price, p_t^k is the common user cost of capital, and p_{it}^ℓ is the price of labor. This empirical specification produces the following structural error term:

$$\omega_{it} + p_{it} - \beta^k p_t^k - \beta^\ell p_{it}^\ell. \tag{10}$$

I follow De Loecker, Goldberg, Khandelwal, and Pavcnik (2016) and let the wedge between the output and input price (scaled by the output elasticity) be a function of the demand shifters and productivity difference.[19] Including demand shifters $\mathbf{d}_{it}$ in the control function, constructed using the measures of market shares, as in De Loecker, Eeckhout, and Unger (2020), should therefore capture the relevant output and input market forces that generate differences in the output and input price.[20] As discussed in De Loecker, Goldberg, Khandelwal, and Pavcnik (2016), this is an exact control when output prices, conditional on productivity, reflect input price variation, and when the demand is of

[18] See De Loecker, Goldberg, Khandelwal, and Pavcnik (2016) for a recent treatment of these issues.

[19] De Loecker, Eeckhout, and Unger (2020) note that not observing output prices perhaps has the unexpected benefit that output price variation absorbs input price variation, thus eliminating part of the variation in the error term.

[20] I also use industry dummies to capture persistent variation in the demand across sectors.

the (nested) logit form.

This is clearly a second-best solution to address the above challenge in estimating the production function; however, it is impossible to go beyond this second-best solution to the problem without more detailed data on the output quantities.

2.2.2 Cost Shares

The cost shares approach has been prominently adopted in Foster, Haltiwanger, and Syverson (2008), and it exploits the first-order conditions of the firm. To make fruitful use of the firm's first-order conditions, two assumptions are needed: (i) there are constant returns to scale in production and (ii) all inputs are variable. With these assumptions, we can calculate output elasticities from the cost shares. The cost shares of both inputs are defined as:

$$\theta^{\ell} = \text{median}\left\{ \frac{w_{it}\ell_{it}}{w_{it}\ell_{it} + r_t k_{it}} \right\} \quad \text{and} \quad \theta^{k} = 1 - \theta^{\ell}, \tag{11}$$

where $w_{it}\ell_{it}$ is the wage bill, and $r_t k_{it}$ is the rental cost of capital. Therefore, an extra requirement in this method involves the possibility of calculating the return on the physical capital, r_t.

The assumptions required to apply this methodology seem to be incompatible with the objective of this project, that is, the estimation of returns to scale in production. However, I explain how these assumptions have been relaxed by the literature, rendering this methodology flexible for a wide scope of applications.

First, following Foster, Haltiwanger, and Syverson (2008), one can use moving averages of the cost shares to accommodate for slow adjustments of the inputs due, for example, to adjustment costs. Second, following Syverson (2004), returns to scale can be calculated, even when using a cost shares approach. In particular, he assumes the following functional form for the technology based on cost shares but without constant returns:

$$q_{it} = \alpha\left[\theta^{k} k_{it} + \theta^{\ell}\ell_{it}\right] + \boldsymbol{X}_{it}'\delta + \omega_{it} \tag{12}$$

with all variables in logs, θ^{k} and θ^{ℓ} are given by (11), and $\boldsymbol{X}_{it}$ is a vector of potential controls. Therefore, while each cost share determines the output elasticity, the technology does not need to be constant returns, and the curvature is captured by α, which can be estimated with a simple OLS.

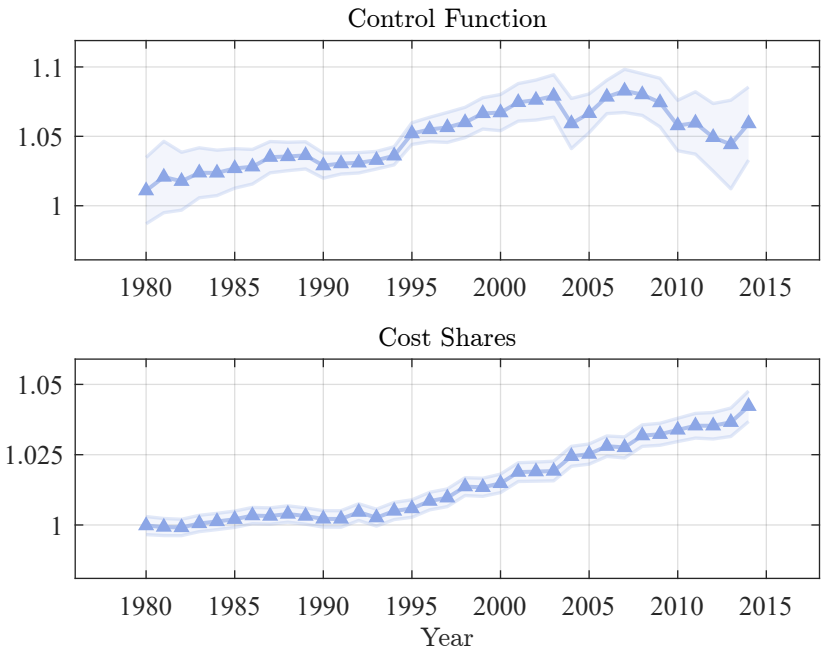
2.3 The Rise in Returns to Scale

Here, I document the rise in returns to scale under both specifications. Then, I look into the sectoral distribution of returns to scale, finding that this rise is due to an increase across all sectors.

2.3.1 Average Returns to Scale in Production

To estimate the returns to scale for the US economy over a period spanning three decades, I need to assume the particular level at which the production technology is shared across firms. I begin by estimating the returns to scale under the assumption that all firms in the economy share the same production technology. I relax this seemingly unrealistic assumption later on in the analysis. Moreover, to allow for time variation in the elasticities, I estimate equation (3) using a ten-year rolling window around the year of interest.[21] Finally, for the choice of variable input in the production, I refer the interested reader to Appendix A.1.3.

Figure 1: Returns to Scale with Common Technology



Note. The figure on the top shows the evolution of the returns to scale computed with the control function approach. The figure on the bottom shows the evolution of the returns to scale computed with the cost shares approach. The dashed dark blue line shows the point estimates, whereas the solid light blue line shows the 90% confidence interval. Output elasticities are time-varying and calculated from 1980 to 2014.

Figure 1 shows the evolution of returns to scale for both the control function approach and cost shares approach. The dashed dark blue lines show the point estimates of the returns to scale, whereas the solid light blue lines show the 90% confidence interval. Despite some qualitative differences between the two approaches, the overall quantitative message is similar. In 1980, returns to scale were 1, that is, there were constant returns to scale that rose approximately by 5% by 2014. Therefore, both estimation techniques suggest that, in recent years, US firms' production technology exhibits increasing returns to scale.

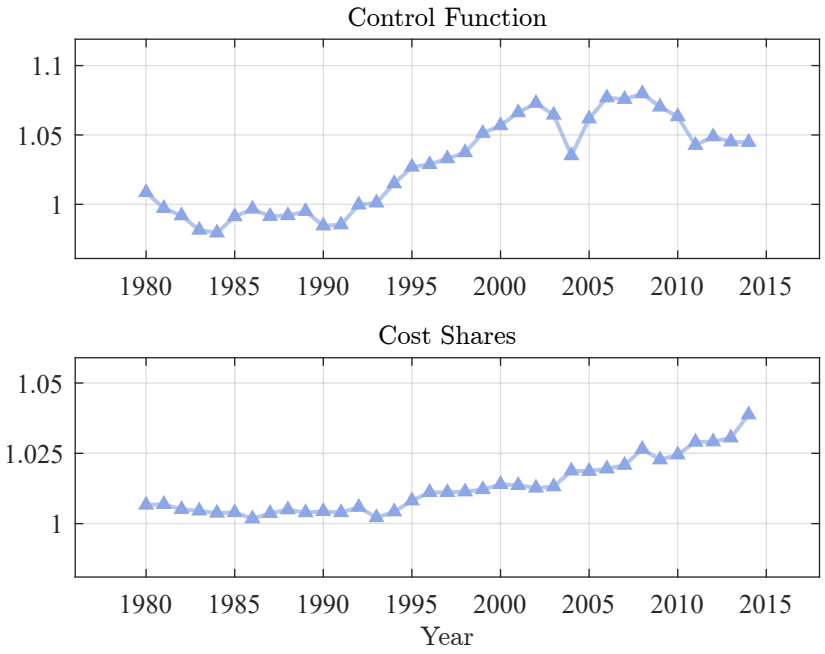
[21] Because of data scarcity, I choose a relatively long rolling window. However, the results do not depend on this assumption and are robust to different rolling window schemes.

Now I relax the previous assumption of common technology across sectors. To do so, I re-estimate the production technology from equation (3) for each two-digit NAICS industry, again using a ten-year rolling window around the year in which I estimate the technology.[22] Therefore, as I estimate a different production technology for each two-digit NAICS industry and year, I define the average returns to scale in the US economy as:

$$\alpha_t = \sum_s m_{st} \cdot \alpha_{st}, \tag{13}$$

where m_{st} is the weight of each sector, and α_{st} is the sectoral returns to scale. In the main specification, I use sales shares as weights.

Figure 2: Returns to Scale with Sector-Level Technology



Note. The figure on the top shows the evolution of the returns to scale computed with the control function approach. The figure on the bottom shows the evolution of the returns to scale computed with the cost shares approach. Output elasticities are time varying and sector specific (two-digit). The average is sales-weighted. The figure illustrates the evolution of the average returns to scale in production from 1980 to 2014.

The graph on the top in Figure 2 reports the evolution of the baseline measure—obtained with the control function approach—of average returns to scale across the economy over time. At the beginning of the sample, returns to scale are equal to 1 and remain constant until the end of the 1980*s*; then, they start to rise steeply and by the end of the sample, are around 1.05.[23] In 2014, the average returns to scale is 5% higher compared to the one in 1980.

[22] The assumption that firms within a two-digit NAICS industry share the same technology makes the results comparable with those in De Loecker, Eeckhout, and Unger (2020).

[23] My estimates are consistent with those reported by Gao and Kehrig (2017) using census data; they find that production technology in the US between 1982 and 1987 had constant returns to scale.

To validate the robustness of the result from the benchmark measure, the graph on the bottom in Figure 2 shows the evolution of the average returns to scale calculated with the cost shares approach. The salient characteristics of this measure closely resemble the patterns of the benchmark measure. From the beginning of the sample to the end of the 1980*s*, returns to scale are flat and close to 1; then from the 1990*s* onward, they start to rise, reaching approximately 1.04 in 2014. Therefore, under the cost shares approach, the average returns to scale is roughly 4% higher relative to 1980.

Overall, the rise in returns to scale does not seem to be driven by the specific methodology applied and follows very close patterns across the different specifications. Appendix A.2 reports further robustness exercises using an additional form of capital (such as intangible capital) and an alternative specification of the functional form of the production function (for example, the translog production function). The bottom line is that the finding for the benchmark measure of average returns to scale is robust.

2.3.2 Sectoral Analysis of Rising Returns to Scale

Although the average returns to scale is a useful statistics, it does not fully capture the underlying distributional changes in returns to scale. The advantage of estimating sector-specific production functions is that I obtain a distribution of returns to scale. This allows me to study whether the documented rise in returns to scale is due to a reallocation of economic activity across sectors or whether it is due to a rise in all sectors.

To do so, I decompose the rise in the average returns to scale into the component that is attributable to the rise in returns to scale at the sector level and the component that is attributable to the reallocation of economic activity toward high-returns to scale sectors. Formally, the rise in the average returns to scale can be decomposed as:

$$\Delta\alpha_t = \underbrace{\sum_s m_{st-1}\Delta\alpha_{st}}_{\Delta\text{within}} + \underbrace{\sum_s \Delta m_{st}\alpha_{st-1}}_{\Delta\text{between}} + \underbrace{\sum_s \Delta m_{st}\Delta\alpha_{st}}_{\Delta\text{cross term}}. \tag{14}$$

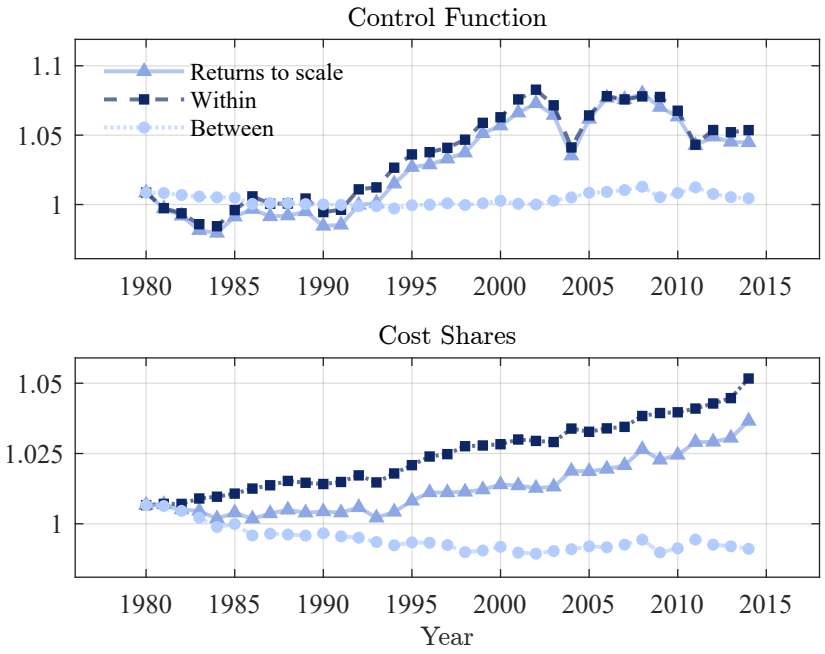
Therefore, the change in average returns to scale can be exactly decomposed into three components: (i) a *within* component, which captures the portion of the change in the average returns to scale at the industry level; (ii) a *between* component, which captures the portion of the change in the average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

I perform this decomposition across sectors in the entire economy. To best present this decomposition, Figure 3 plots the average returns to scale, calculated with both methodologies, as well as

two counterfactual experiments, the within and between experiments, based on the decomposition starting in 1980. I do not plot the cross-term experiment, as it is of little economic interest and substantially zero across the entire period. Finally, I set the initial level to 1980 and then cumulatively add the changes of each component from equation (14).

The first experiment (dashed dark blue line with squares) shows the counterfactual evolution of the average returns to scale as if there were only the Δwithin component, and all the other components were zero. This experiment shows that the within component tightly follows the average returns to scale in the case of the control function approach and exceeds the average returns to scale in the cost shares approach.[24] The second experiment (dotted light blue line with circles) shows the path of the counterfactual returns to scale if the only change had been due to Δreallocation. This shows a flat profile over the period for the control function approach and a decreasing profile for the cost shares approach. From these two experiments, it is apparent that the rise in the average returns to scale is indeed a within-sector phenomenon and, if anything, the cross-sectoral reallocation of economic activity has slightly dampened its rise.

Figure 3: Decomposition of Returns to Scale Growth at Sector Level



Note. The figure plots the counterfactual evolution implied by the decomposition from equation (14) for the control function approach (upper figure) and the cost shares approach (lower figure). The solid blue line with triangles shows the (benchmark) average returns to scale. The dashed dark blue line with squares shows the evolution of the average returns to scale only if the Δwithin component is at play. The dotted light blue line with circles shows the evolution of the average returns to scale only if the Δbetween component is at play.

[24] With the cost shares approach, the within component exceeds the average returns to scale. Thus, in the absence of reallocation of economic activity across sectors, the rise in returns to scale with this methodology would have been even higher.

Taking stock, returns to scale have risen substantially in the US economy, and this rise is occurring across all sectors. This transformation in the firms' production technology could stem from many things. For instance, since the 1980*s*, and with an acceleration from the beginning of the 1990*s*, a digital revolution took place in the US. New technologies such as the internet, mobile phones, computers, and software were developed. These new technologies brought forth an incredible transformation in the way production and business models could be organized. All of a sudden, firms could share internal information at a higher pace and could reach customers at a speed and on a scale previously not possible. The ability of these new technologies to increase the scale at which firms can operate has been the object of interest among researchers since the beginning of the aforementioned digital revolution.[25] I acknowledge that drawing a clear causal link between the digital revolution in information technology and the rise of returns scale requires better data than what I have. However, in this project, I will nonetheless interpret the rise of returns to scale as a pervasive technological transformation that US firms are experiencing across all sectors.

3 Model

To study the implications of the technological change outlined above for firms' investment in their customer base, business dynamism, and markups, I build a firm dynamics model with search frictions in the product market. Search frictions are a natural choice to microfound (i) the presence of heterogeneous endogenous markups in equilibrium; (ii) firms' expenditures to attract new customers; and (iii) the empirical observation that firms grow over their life span mostly by accumulating new customers.[26] I refer the interested reader to Appendix B.1 for a discussion of the technical features of the model.

3.1 Population and Technology

Time is discrete. The economy is populated by a representative household, comprising a continuum of measure one of potential buyers and by a large number of workers, and by an endogenous measure of firms with free entry.[27] The representative household discounts the future at a rate β. The instantaneous utility of the household is:

[25] A particularly relevant paper is Lashkari, Bauer, and Boussard (2021), which documents, via rich firm-level data from France that investment in ICT allowed French firms to increase their returns to scale in production in recent years. Newman (2014), Agrawal, Gans, and Goldfarb (2018), Begenau, Farboodi, and Veldkamp (2018), Goldfarb and Trefler (2018), Carriere-Swallow and Haksar (2019), and Jones and Tonetti (2020) emphasize the potential role of data, particularly gathering information from the customer base, as a source of increasing returns to scale.

[26] Afrouzi et al. (2020) and Einav, Klenow, Levin, and Murciano-Goroff (2020) show that 70% of firm growth comes from accumulating new customers over their life cycle.

[27] In the text, I refer to buyers and customers interchangeably.

$$uC - v(L), \tag{15}$$

where uC is the utility from the consumption of the frictional good, and $v(L)$ is the disutility of labor.[28] The representative household aggregates consumption C is a bundle of the consumption of each active buyer via the following CES aggregator:

$$C = \int_{i \in \mathcal{I}} c_i \mathrm{d}i, \tag{16}$$

where c_i is buyers' consumption of the frictional good, and $\mathcal{I} \subseteq 1$ is the set of active buyers. Equation (16) assumes that the goods of the different firms are perfect substitutes, so we can interpret the continuum of firms as effectively selling the same product. Moreover, I assume that buyers wish to buy exactly one unit of the firm's good, and hence, their shopping value will be equal to the marginal utility of the household's consumption, $u \geq 0$.[29]

Firms differ in their idiosyncratic productivity z, independent across firms, that lies in the finite set $\mathcal{Z}$ and follows a Markov process $\pi(z'|z)$. A firm with a measure ℓ of workers operates with the production technology:

$$y = e^z F(\ell), \tag{17}$$

where F is a strictly increasing production function with $F(0) = 0$. Upon entry, firms must pay a sunk entry cost κ. Following Hopenhayn (1992), I assume that firms must pay a fixed operating cost $f \geq 0$ every period to use the production technology. This operating cost is crucial in generating endogenous exit in the model. Finally, I also assume that firms exit exogenously with probability $\delta \in (0, 1)$.

3.2 Frictional Product Market

The product market is frictional, and the search is directed on buyers' and firms' sides. Firms announce contracts to attract buyers. Because utility is transferable between buyers and firms, a sufficient statistic for each contract is the utility x that it delivers to the buyer upon matching. Firms offering identical contracts compete in the same market segment; therefore, I describe the product

[28] As a consequence, the labor supply of the household will be given by:

$$\lambda^{BC} w = v'(L),$$

where λ^{BC} is the Lagrange multiplier associated with the household budget constraint, w is the wage, and $v'(L)$ is the marginal disutility of labor. For convenience, I normalize λ^{BC} to 1 without loss of generality.

[29] The fact that buyers wish to purchase exactly one unit implies that only the extensive margin of demand matters in the model, that is, to how many buyers I should sell. This assumption implies that $c_i = 1, \ \forall i \in \mathcal{I}$.

market as a continuum of submarkets indexed by the utility $x \in [\underline{x}, \bar{x}]$ that firms promise to buyers. Firms must pay a cost c for each *ad* they post.[30] Moreover, firms that change their customer base are subject to a convex cost $\mathcal{K}(n_i; n)$, where n_i is the number of new customers that the firm wants to acquire.[31] Buyers can direct their search and choose in which submarket to look.

A standard matching function with constant returns to scale governs match creation in each market segment. I denote by $\theta(x)$ the ads-buyers ratio or tightness of submarket x. In a submarket with tightness θ, buyers find a firm with probability $m(\theta)$, while firms find potential customers with probability $q(\theta) = m(\theta)/\theta$. As standard in the search literature, I assume that m is increasing, while q is decreasing, and that $m(0) = 0$, $q(0) = 1$. Buyers and firms must solve a trade-off between the level of utility of a given contract and the corresponding probability of being matched. The search process takes time, and I assume that firms and buyers can only visit one submarket at a time.

Buyers are allowed to search while already being attached to a firm. The equilibrium market tightness can be written as $\theta(x) = a/\mu$, where a stands for the number of ads posted in submarket x, and μ stands for the corresponding efficiency-weighted number of searching buyers.[32] The number of ads a that a firm posts is not required to be discrete and should be interpreted as a mass. As a result, the law of large numbers applies, and firms do not face uncertainty about the number of buyers they recruit. In particular, a firm that posts a ads exactly meets a measure $aq(\theta) = n_i$ of buyers.

3.3 Contractual Environment and Timing

Contracts specify various elements relevant to the firm and its customers. I assume that contracts are state-contingent, and that there is full commitment from the firm side. 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. Each element at time $t+j$ is contingent on the entire history of shocks (z^{t+j}). A more detailed exposition of the contractual environment and its implications for the model is in Appendix B.1.

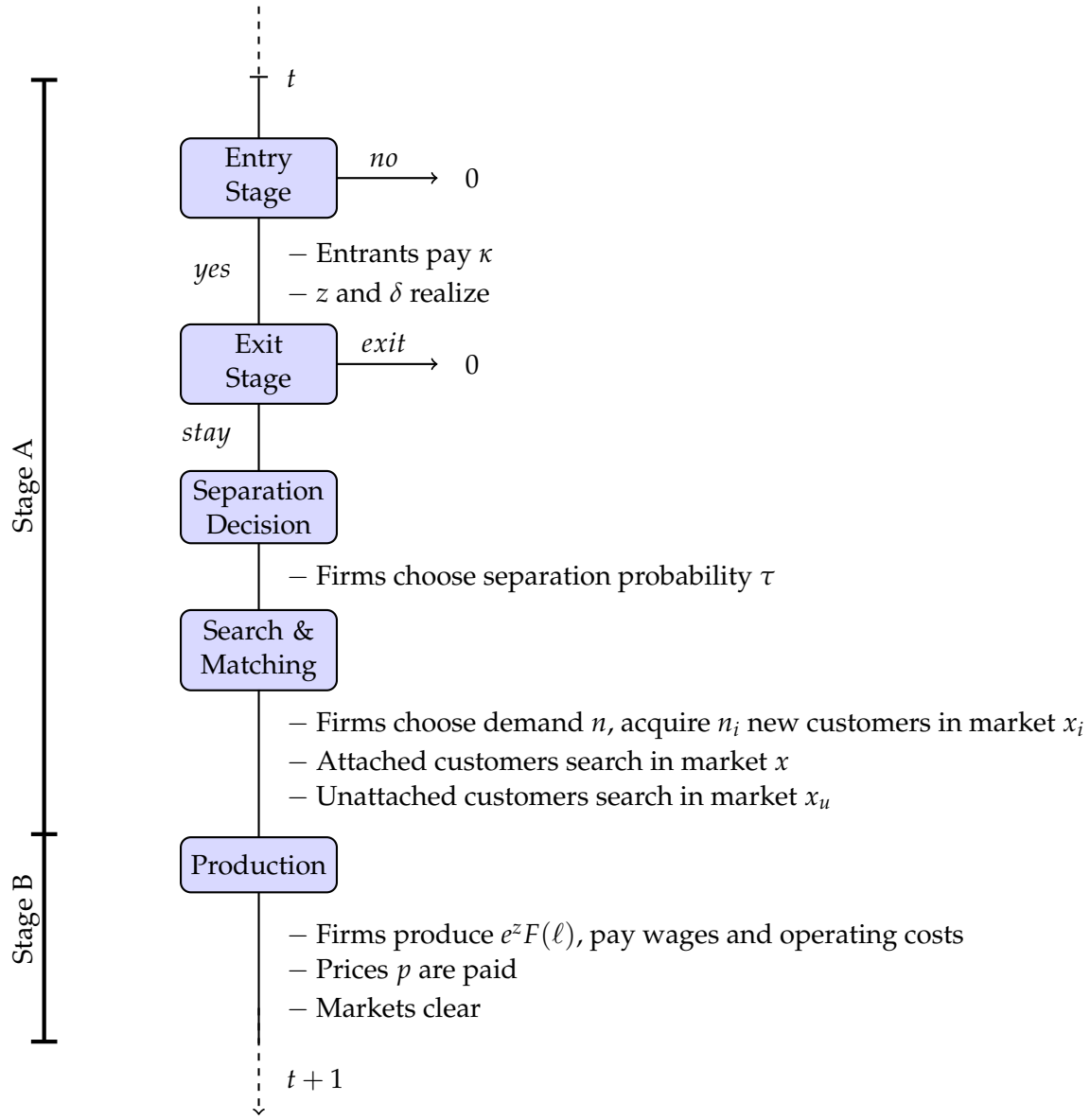
The contracts offered by firms are large objects but can be written in their recursive form. Contracts are rewritten every period after matching occurs and when production takes place (stage B in Figure 4). At this stage, the firm starts with some utility $\mathcal{C}$, promised in the past to its incumbent customers or new ones. A recursive contract $\omega = \{p, \tau, d, \mathcal{C}'\}$ for the current period specifies the current price p and the next period's quantities $\{\tau(z'; w), d(z'; w), \mathcal{C}'(z'; w)\}$, contingent on the next period's state, where $\mathcal{C}'(z'; w)$ is some future promised utility. Because of commitment on the firm

[30] The term *ad* in the model is a stand-in for a broader notion of marketing and selling effort, and will be interpreted as such later on.

[31] The convex cost slows down the adjustment of firms' customer base and is pivotal in generating a realistic endogenous firm life cycle. Moreover, this convex cost is the key friction, together with the exogenous exit shock, preventing the model from settling on a degenerate distribution of firms.

[32] In particular, $\mu = \mu_u + \mu_a$, where μ_u is the number of unattached buyers and, μ_a is the corresponding number of attached customers searching on the market.

Figure 4: Timing of the Model



side, contract ω is required to deliver at least the promised utility $\mathcal{C}$ to the customers.

The timing of the model is illustrated in Figure 4. At the beginning of period t, firms decide whether to enter or not. Immediately afterward, incumbent and entering firms learn their idiosyncratic productivity z and their exogenous exit shock δ. Then, conditional on surviving, they decide whether to exit ($d = 1$) or stay. In the following stage, separation occurs with probability τ. Search and matching follow with new and incumbent firms on one side and unattached/attached customers on the other side. Production takes place in the final stage of the period, and the markets clear.

3.4 Customer's Problem

As conventional in the search literature, the value functions below are expressed at stage B of the period when production takes place. I write the value of an unattached buyer as follows:

$$\mathcal{U} = \max_{x_u} \beta[m(\theta(x_u))x_u + (1 - m(\theta(x_u)))\mathcal{U}']. \tag{18}$$

If a buyer is not attached to a firm, she does not enjoy any utility in that period. In the following period, she chooses a market segment, x_u, where to search. In doing so, she must solve a trade-off between the offered utility, x_u, and the likelihood of getting a job, $m(\theta(x_u))$. When successful, she enjoys the promised utility x_u, but she remains unattached otherwise.

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

$$\begin{aligned}\mathcal{C}(z, \omega; w) = u - p + \beta\mathbb{E}\{&(\delta + (1-\delta)d + (1-\delta)(1-d)\tau)\mathcal{U}' \\ &+ (1-\delta)(1-d)(1-\tau)\max_{x'}[m(\theta(x'))x' + (1 - m(\theta(x')))\mathcal{C}'(z'; w)]\}.\end{aligned} \tag{19}$$

An attached customer buys one unit of the firm's output at a price p and values it at the marginal utility of the representative household, $u \geq 0$. The following period may then lead to three different outcomes, which correspond to the three terms in brackets: (i) in the case of exit, that is, exogenously with $\delta \in (0, 1)$ or endogenously if $d = 1$, or in the case of destructing the relation, $\tau \in (0, 1)$, the customer goes back to the potential buyers' pool with value $\mathcal{U}'$; (ii) she moves to a different firm under a contract with value x' with probability $m(\theta(x'))$; or (iii) she stays in the current firm and receives a promised utility $\mathcal{C}'(z'; w)$ in the following period. Notice that customers entering the pool of potential buyers in the given period cannot search in the same period.

3.5 Firm's Problem

Consider the problem of a firm at the production stage with a measure n of customers. Customers within the same firm may differ in their level of promised utility. Each customer is identified by an index $j \in [0, n]$ and a corresponding level of promised utility $\mathcal{C}(j)$.

The problem of a firm consists of choosing a list of contracts for its customers:

$$\omega(j) = \{p(j), \tau(z'; w, j), d(z'; w), \mathcal{C}'(z'; w, j)\}, \quad \forall j \in [0, n]. \tag{20}$$

In addition, the firm must decide on a submarket $x_i(z'; w)$ in which to search for new potential customers, and it must choose the number of new customers that it wants to acquire $n_i(z'; w)$. I describe the problem faced by firms as follows:

$$
\begin{aligned}
&\boldsymbol{\mathcal{V}}(z,n,\{\mathcal{C}(j)\}_{j\in[0,n]};w) \\
&\quad = \max_{n_i'(z';w),x_i'(z';w),\{\omega(j)\}_{j\in[0,n]}} \int_0^n p(j)\mathrm{d}j - w\ell - wf \\
&\quad + (1-\delta)\beta\mathbb{E}\Big\{ -n_i'\frac{wc}{q(\theta(x_i'))} - w\mathcal{K}(n_i';n) + \boldsymbol{\mathcal{V}}(z',n',\{\widehat{\mathcal{C}}(z';w,j')\}_{j'\in[0,n']};w)\Big\}^+,
\end{aligned} \tag{21}
$$

subject to:

$$
n'(z';w) = \int_0^n (1-\tau(z';w,j))(1-m(\theta(x'(z';w,j))))\mathrm{d}j + n_i'(z';w), \tag{22}
$$

$$
\widehat{\mathcal{C}}(z';w,j') = \begin{cases} \mathcal{C}(z';w,j) & \text{for} \quad j' \in [0, n'(z';w) - n_i'(z';w)] \quad \text{and} \quad j' = \Phi(z';w,j), \\ x_i(z';w) & \text{for} \quad j' \in [n'(z';w) - n_i'(z';w), n'(z';w)], \end{cases} \tag{23}
$$

$$
y = e^z F(\ell), \tag{24}
$$

$$
y = n, \tag{25}
$$

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

In the current period, the firm earns revenue, $\int_0^n p(j)\mathrm{d}j$, minus the cost of labor, $w\ell$, and minus the fixed operating cost, wf. In the following period, the firm survives with probability $(1-\delta)$ and then it chooses whether to exit or not. The $\{.\}^+$ notation, standing for $\max(.,0)$, captures this decision, which I summarize in the dummy $d(z';w) \in \{0,1\}$ ($d=1$ for exit). Following this decision, the firm then chooses a number of new customers to acquire $n_i'(z';w)$ and the submarket $x_i'(z';w)$ in which to direct its selling effort. Because each ad has a probability $q(\theta(x_i'))$ of being successful, the total selling cost incurred for these new customers is $n_i' wc/q(\theta(x_i'))$. Additionally, to slow down the adjustment pace of firms' customer base, I introduce a convex cost, that is, $w\mathcal{K}(n_i';n)$, which each firm must pay to change its customer base. This is one of the two fundamental assumptions that allows the model to produce a realistic life cycle.[33] Moreover, the constraint that this convex cost imposes on the firm's ability to expand its customer base is the key friction, together with the exogenous exit shock, that prevents the economy from settling on a degenerate distribution of firms.

Constraint (22) is the law of motion of total customers. Customers n' in the next period are the sum of the new customers $n_i'(z';w)$ with the remaining customers after the departure of those separated with probability $\tau(z';w,j)$ and of those moving to other jobs with probability $m(\theta(x'(z';w,j)))$. Constraint (23) keeps track of the promised utilities across customers. Because the measure of cus-

[33] The second fundamental assumption, as explained later, is related to the fact that each firm enters with a predetermined measure of initial customers. In the quantitative section of the paper, I will calibrate this to be lower than the average mass of customers attached to incumbent firms.

tomers evolves over time, I use the mapping Φ to re-index the customers that stay and make sure that a customer with an original index $j \in [0, n'(z';w) - n'_i(z';w)]$ is assigned a new index $\Phi(z';w,j) \in [0, n'(z';w) - n'_i(z';w)]$ in the next period. Newly recruited customers with promised utility, $x'_i(z';w)$, are assigned an index in $[n'(z';w) - n'_i(z';w), n'(z';w)]$. Constraint (24) defines the technology with which the firm operates; therefore, this determines the amount of labor ℓ that a firm will hire in each period. Finally, constraint (25) states that the output must be equal to the number of available customers n in the given period.

In addition to these constraints, and due to commitment on the firm side, the firm is subject to the following *promise-keeping* constraint:

$$\forall j \in [0, n], \quad \mathcal{C}(j) \leq \boldsymbol{\mathcal{C}}(z, \omega(j); w). \tag{26}$$

Constraint (26) ensures that the contract $\omega(j)$, assigned to customer j, delivers at least the promised lifetime utility $\mathcal{C}(j)$. Note that there is no non-negativity constraint on the firm's profits, implying that firms have deep pockets and no limited liability.

3.6 Firm's Pricing

Until now, I have allowed firms to charge different prices to their customers, conditional on their past histories. In this section, I present the optimal prices charged by the firms to their different customers.

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

$$m(\theta(x'))x' + (1 - m(\theta(x')))\mathcal{C}(z') \geq \boldsymbol{\mathcal{U}}, \tag{27}$$

which states that the continuation value for a customer, conditional on remaining matched, given by equation (19), must be higher than the value of being unmatched, given by equation (18). This ensures that the customer does not prefer to be unmatched. Second, the contract must take into account the following *incentive constraint*:

$$x' = \underset{\tilde{x}}{\operatorname{argmax}}\, m(\theta(\tilde{x}))\tilde{x} + (1 - m(\theta(\tilde{x})))\mathcal{C}'(z';w), \tag{28}$$

which states that the submarket in which the customer will search is the one that maximizes the continuation value, conditional on remaining matched, given by equation 19. This verifies that the submarket in which the customers search is the optimal submarket in which they would like to search. A contract satisfying constraints (27) and (28) is said to be an incentive-compatible contract.

It is now easy to derive prices from the promise-keeping constraint (26). The price for a customer j is given by:

$$p(j) = \boldsymbol{\mathcal{C}}(z, \{p = 0, \tau, d, \mathcal{C}'\}; w) - \varkappa(j), \tag{29}$$

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

Notice that the price charged to each customer for the good is the difference between the present value of being attached to a firm evaluated at today's price equal to zero, that is, $\boldsymbol{\mathcal{C}}(z, \{p = 0, \tau, d, \mathcal{C}'\}; w)$, minus the history-dependent promised utility $\varkappa(j)$. Therefore, the higher the value customers get from the match, the higher the price charged by the firm. Conversely, the higher the utility a firm promises, the lower the prices charged to its customers.

Equation (29) captures one of the main trade-offs for the firms in the model. In particular, firms are always subject to two opposite tensions. On the one hand, firms that want to grow need to attract customers; to do so, they must give a high promised utility, meaning low prices. On the other hand, firms want to extract value from their matches, meaning that they want to charge high prices to their customers. Therefore, the evolution of prices, and hence of markups, strictly follow the life cycle of firms: young firms, being small, must invest in their customer base, and hence, charge low prices and markups. On the contrary, old firms—which are on average bigger—want to harvest their customer base, and hence, charge high prices and markups.

3.7 Free Entry and Equilibrium Definition

To close the model, I am left to specify the process of entry. Every period, before the idiosyncratic shock z is realized, the potential entrants decide whether or not to enter. Upon entry, firms must pay an entry cost κ, after which they draw their idiosyncratic productivity from a distribution g_z. Depending on the outcome, firms may decide to exit or stay, in which case they can start searching for customers and producing as any normal firm.

I define the problem faced by an entering firm of type z as follows:

$$\boldsymbol{\mathcal{V}}^e(z; w) = (1 - \delta) \max_{x_e} \left\{ -n_e \frac{wc}{q(\theta(x_e))} + \boldsymbol{\mathcal{V}}(z, n_e, \{\mathcal{C}(j)\}_{j \in [0, n_e]}; w) \right\}^+. \tag{30}$$

Having drawn the idiosyncratic productivity z and surviving the exit shock $\delta \in (0, 1)$, the potential entrant first decides whether or not to exit, a decision captured by the notation $\{.\}^+$ and summarized in the dummy $d_e(z; w)$. If it stays, the firm searches $n_e \in \mathbb{R}^+$ new customers, and chooses a submarket, x_e, to maximize its expected value of operating, minus the total ad cost $n_e wc/q(\theta(x_e))$. I do not allow the entering firms to choose n_e optimally. This is the second necessary ingredient, to-

gether with the convex adjustment cost that firms must pay to change their customer base, to obtain a well-defined notion of life cycle within the model.[34]

Due to the presence of free entry, firms enter as long as expected profits exceed the entry cost κ, driving these expected profits down to κ. Therefore, the expected surplus from entering must be equal to κ in equilibrium:

$$w\kappa = \int \boldsymbol{\mathcal{V}}^e(z;w)g_z(dz). \tag{31}$$

3.8 Firm Distribution Dynamics and Recursive Equilibrium

Using the optimal decision of firms, we may now describe the evolution of customers over time. Let $g(z,n;w)$ be the distribution of customers across firms in stage B of the current period when production takes place. The dynamics of the distribution of customers across firms can be described by:

$$\begin{aligned} g(z',n';w) = &\sum_{z,n} \mathbb{1}\{n'(z';w,n) = n'\}(1-d(z';w,n))(1-\delta)\pi(z'|z)g(z,n;w) \\ &+ m_e\mathbb{1}\{n_e(z';w) = n'\}(1-d_e(z';w))(1-\delta)g_z(z'), \end{aligned} \tag{32}$$

where $\mathbb{1}\{\cdot\}$ denotes an indicator function. Equation (32) defines the mass of firms with an individual 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 on every market segment.

Finally, I define the stationary recursive equilibrium in this economy.

Definition 3.1 (Stationary recursive equilibrium)**.** A *stationary recursive competitive equilibrium* consists of value functions $\{\boldsymbol{\mathcal{U}},\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{V}},\boldsymbol{\mathcal{V}}^e\}$, policy functions $\{x_u,x,p,\tau,d,\mathcal{C}',n_i,x_i,d_e,x_e\}$, a wage $\{w\}$, an invariant measure of incumbents g, and a measure of entrant firms m_e, such that: (i) $\boldsymbol{\mathcal{U}}$ and x_u solve the unattached buyers' problem (18); (ii) $\boldsymbol{\mathcal{C}}$ and x solve the attached buyers' problem (19); (iii) $\boldsymbol{\mathcal{V}}$, τ, d, n_i, and x_i solve the incumbent firms' problem (21); (iv) $\boldsymbol{\mathcal{V}}^e$, d_e, and x_e solve the entrant firms' problem (30); (v) p and $\mathcal{C}'$ solve (28) and (29); (vi) the labor market clears; and (vii) the invariant measure of incumbents g and the measure of entrants firms m_e satisfy the dynamics of the distribution of customers across firms, given by (32) and the free-entry condition (31).

[34] I let entering firms enter with an n_e lower than the average size. Together with the convex adjustment cost described earlier, this implies that new firms start small and grow slowly to reach the average size in the economy.

4 Model Parametrization and Validation

In this section, I bring the model presented in Section 3 to the data. Particularly, the model is estimated to replicate certain salient moments from the cross-section of firms around 1980. First, I present the functional forms and the stochastic processes used in the quantitative analysis. Second, the aforementioned salient moments are used to discipline some deep parameters that are not directly observable to the researcher. Third and finally, I validate the model on non-targeted moments of both the cross-section and the life cycle of firms.

4.1 Functional Forms and Stochastic Processes

The household disutility of labor is given by:

$$v(L) = \vartheta \frac{L^{1+\frac{1}{\psi}}}{1+\frac{1}{\psi}}, \tag{33}$$

where ϑ is a parameter governing the cost of supplying labor for the household, and ψ is the Frisch elasticity.

The firm-level production function is given by:

$$F(\ell) = \ell^{\alpha}, \tag{34}$$

where α governs the firm-level returns to scale of production. Given that time is discrete, I choose a functional form for the probability that a customer finds a firm bounded between 0 and 1, which rules out the Cobb-Douglas matching functions. In particular, I pick the following functional forms:

$$m(\theta) = \theta/(1+\theta)^{-1}, \quad \text{and} \quad q(\theta) = (1+\theta)^{-1}. \tag{35}$$

The convex cost of relaxing the customer base is given by:

$$\mathcal{K}(n_i; n) = \chi_1 \left(\frac{n_i}{n}\right)^2 n^{\chi_2}, \tag{36}$$

with $\chi_1, \chi_2 \geq 0$.The idiosyncratic productivity shock follows an AR(1) process, given by:

$$z_t = \rho z_{t-1} + \sigma \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1), \tag{37}$$

where z_t is the time-varying idiosyncratic productivity, $\rho \in (0,1)$ is the parameter governing the persistence of the process, and σ is the standard deviation of the innovation to the process.

4.2 Parametrization

The model is parametrized in two steps. First, I fix a set of parameters to match the standard targets in the steady state. Second, given the values of those parameters, I choose the remaining parameters to match identifying moments from the data. A model period is one year, and the calibration targets moments from the 1980s.

I set the discount rate β equal to 0.97 so that the annual interest rate is about 3%, a value standard in the literature. The degree of firm-level returns to scale α is set equal to 1. This implies constant returns to scale, a value consistent with the empirical estimates presented in Section 2.3. I set the persistence of the productivity shock ρ equal to 0.8, the value found by Foster, Haltiwanger, and Syverson (2008).[35] The standard deviation of the innovations to the productivity process σ is set to 0.2, a value close to Foster, Haltiwanger, and Syverson (2008) and common in the firm dynamics literature. The marginal utility from consumption u is set equal to 1, implying a unitary evaluation of each extra unit of consumption. Finally, the Frisch elasticity ψ is set equal to 2.84, corresponding to the average aggregate Frisch elasticity of hours reported by Chetty, Guren, Manoli, and Weber (2011).

The parameters left to be internally calibrated are $\{c, \chi_1, \chi_2, n_e, f, \kappa, \delta, \vartheta\}$. All these parameters are disciplined through cross-sectional and life-cycle moments. The linear cost c, paid by firms to search for an extra customer, is disciplined by the average markup in 1980. This is identified because this is a sunk cost that firms must recover—in the long run. Hence the higher this cost is, the higher the markup that a firm must charge to operate. The convex cost of increasing the customer base χ_1 is deeply tight with respect to the life cycle of the firms. Particularly, it influences the speed at which firms increase their size. Hence, I use the average size of firms that are five years old in 1980 to identify this value. The initial mass of customers that each entering firm has, n_e, together with the aforementioned convex cost, completely informs the endogenous life cycle in the model. Specifically, given χ_1, the mass of customers upon entry informs us about the size of the entrant firms, which is indeed used as the identifying moment for this parameter. The operating cost f is used to match the average firm size in the period. This is so because if this cost increases, only relatively more productive firms can operate, meaning that the average firm in the market becomes bigger. The entry cost κ is identified with the entry rate in 1980, as it is standard in the literature. The exit shock probability δ is identified with the aggregate excess reallocation rate, as the higher the exit probability is, the higher the reallocation of labor in the model. The convex cost parameter χ_2 is disciplined with the share of firms that are greater or equal to eleven years. This is because the higher χ_2 is, the more costly it is to grow for larger firms. Hence, the more likely they will exit at a younger ages. Finally,

[35] Foster, Haltiwanger, and Syverson (2008) is an important reference, as they disentangle from firm-level sales the contribution of prices from the contribution of true productivity. This is particularly important in our setting, given that the model differentiates firm-level prices and firm-level productivity.

the labor supply shifter ϑ is set such that the equilibrium wage in 1980 is equal to one.

The parameters are estimated using the following routine. For arbitrary values of the vector of parameters, $\mathcal{P} = (c, \chi_1, \chi_2, n_e, f, \kappa, \delta, \vartheta)$, the dynamic programming problem is solved, and the policy functions are generated. Using these policy functions, the decision rules are simulated until the distribution of firms over $\{n, z\}$ is converged. I draw from this stationary distribution, simulating the economy for many periods, and construct a panel of firms. I compute the aforementioned moments of interest, which I denote as $\mathcal{M}(\mathcal{P})$, whereas the empirical moments are denoted as $\widehat{\mathcal{M}}$. I estimate the fitted parameters $\widehat{\mathcal{P}}$ using a minimum distance criterion, given by:

$$\mathcal{L}(\mathcal{P}) = \min_{\mathcal{P}} \Big(\widehat{\mathcal{M}} - \mathcal{M}(\mathcal{P})\Big)' \boldsymbol{\mathcal{W}} \Big(\widehat{\mathcal{M}} - \mathcal{M}(\mathcal{P})\Big). \tag{38}$$

Following Asker, Collard-Wexler, and De Loecker (2014), I set the weighting matrix $\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{I}}$ and use a grid search algorithm to find the vector $\widehat{\mathcal{P}}$ that minimizes the objective function (38).

Table 1: Estimated Parameters and Moments

Fixed	Value	Description			
β	0.97	Annual interest rate			
α	1	Returns to scale			
ρ	0.8	Autocorrelation idiosyncratic productivity			
σ	0.2	Standard deviation idiosyncratic productivity			
u	1	Marginal utility			
ψ	2.84	Frisch elasticity			
Fitted	**Value**	**Description**	**Moments**	**Model**	**Data**
c	0.45·1e-3	Linear cost of searching	Avg. markup	1.20	1.17
χ_1	0.46	Convex cost of searching 1	Avg. size age 5	12.32	10.16
χ_2	1.91	Convex cost of searching 2	Share of old firms	0.32	0.32
n_e	6.79	Customers' entrant firms	Avg. entrant size	5.98	5.97
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	0.98	Exit shock probability	Reallocation rate	0.29	0.31
ϑ	0.985	Labor supply shifter	Wage	1	–

Note. The table reports the values of the parameters and model-implied moments. All the moments have been calculated from 1977 to 1985. I do this because BDS reports data only from 1977; by 1980, not all moments of interest can be computed accurately. Firms size is measured by the total labor ℓ employed in a given period—which is consistent with the measure reported by BDS. The average markup is calculated with cost weights, as in the data.

Table 1 summarizes the parameter values resulting from the calibration, along with the fit of the model. The fit is, overall, quite satisfactory. In the calibration, I focus on the average cost-weighted markup. However, the model-implied average sales-weights markup is 1.28, very close to the 1.25

value from the data. Finally, the model implies a slope of selling-related activities on sales of 0.15, close to the value of 0.49 documented by Afrouzi, Dernik, and Kim (2020).[36] The next section validates the calibration in deeper detail.

4.3 Validation

To validate the model, I test the overall calibration against two different dimensions of interest. First, I document the model's performance on the cross-section and the life cycle of firms. Second, I test the cross-sectional and life-cycle implications produced by the model for the markups and for the selling ratio—the ratio of non-production to production costs. The reader interested only in the main results can go directly to Section 5. Additional steady-state implications of the model are presented in Appendix B.3.

4.3.1 Cross-Sectional and Life-Cycle Implications

The model is designed to capture some relevant aspects of the cross-sectional differences in the micro-data. Part of this cross-sectional heterogeneity is inherently linked with the life cycle of firms. In particular, firms enter small and, conditional on surviving, slowly expand their size when accumulating new customers. This implies that firms of different cohorts have different sizes, with younger firms exhibiting fewer employees—our measure of size, consistent with the BDS data. Moreover, only a few firms survive and keep operating, making the mass of firms belonging to the old cohorts a decreasing share of the total.

Figure 5 presents the aforementioned facts, both for the model and data. The figure on the left shows the size of each cohort, measured as the average number of employees within each firm of a given age, in the model and the data. It can be seen that the model and data track each other well; this is not surprising, given that average employment for firms of age 0 and age 5 used as a target in the calibration. Nonetheless, the model slightly understates the size of the oldest (11+) firms. Instead, the figure on the right documents the distribution of firms across cohorts in the data and the model. The model manages to track satisfactory data.

Empirical works on firm-level data have established many regularities about the life cycle of firms. Since the seminal work by Dunne, Roberts, and Samuelson (1989), we know that in the US manufacturing sector, establishment growth is unconditionally negatively correlated with age.[37]

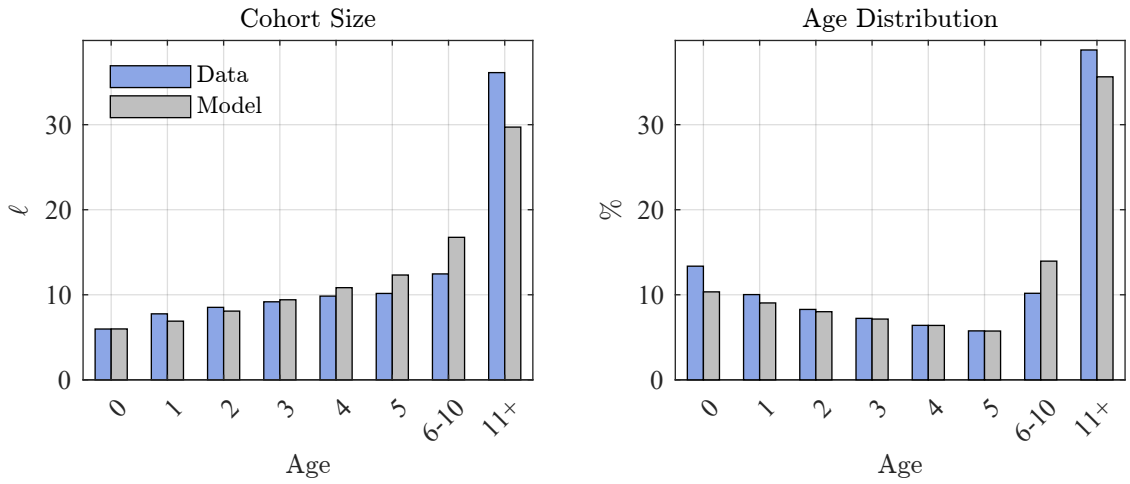
[36] To obtain the slope of selling-related activities on sales, I follow the recent paper by Afrouzi, Dernik, and Kim (2020), and I run the following regression specification:

$$s_j = \beta_1 \int_0^{n_j} p_\kappa d\kappa + \beta_2 w\ell_j + \varepsilon_j,$$

where, in the model, s, the selling-related expenditure, is computed as $wcn_i/q(\theta) + w\chi_1(n_i/n)^2 n^{\chi_2} + wf$, total sales are $\int_0^n p_\kappa d\kappa$, and $w\ell$ is the labor cost. Hence, the coefficient of interest is given by β_1.

[37] This finding was confirmed for a variety of sectors and countries. See Coad (2009) for a recent survey of the literature.

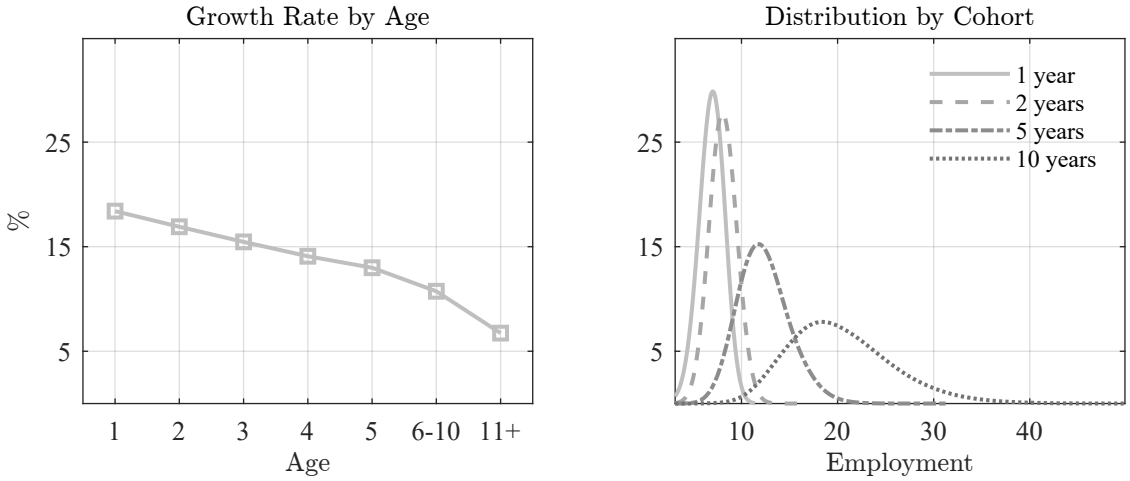
Figure 5: Model Cross-Section



Note. The figure on the left shows the size of each cohort, measured as the number of employees within firms. The figure on the right shows the distribution of firms across cohorts. The light blue bars represent BDS data; the light grey bars show the model predictions.Data reported are between 1977-1985.

Moreover, Cabral and Mata (2003), using a comprehensive data set of Portuguese manufacturing firms, show that the employment distribution shifts to the right and becomes less right-skewed as cohorts age. Figure 6 shows the aforementioned life-cycle facts in the model.

Figure 6: Model Life Cycle



Note. The figure on the left shows the employment growth rate by age, that is, $g^{\ell}_{it} \equiv (\ell_{it} - \ell_{it-1}) / \frac{1}{2}(\ell_{it} + \ell_{it-1})$. The figure on the right shows the employment distribution across cohorts. Both y-axes are in percentage points.

The model aptly captures the life-cycle facts. In the model, firms enter small, with few customers, and grow only slowly, accumulating new customers. Moreover, the accumulation of customers is less costly for young firms; hence, they experience higher growth relative to older firms. The same mechanism explains the results presented in the right figure. In particular, while firms age, they

expand their size, pushing the distribution of their cohort to the right. Overall, the model fits well with many non-targeted moments of the cross-section and the life cycle of firms.

4.3.2 Implications for Markups and Selling-Related Activities

The model produces clear predictions about the evolution of markups and selling-related activities over the life cycle of the firms. In particular, young firms charge lower markups and spend more on selling-related activities (relative to production costs) to grow faster. Therefore, in the data, we should expect to observe a growing profile for markups and a declining profile for selling-related activities over production costs as firms age.

To map the model's expenditures to an empirically meaningful empirical counterpart, I define the *selling ratio* in the model as:

$$\varrho = \frac{f + n_i c/q(\theta) + \chi_1 (n_i/n)^2 n^{\chi_2}}{\ell}, \tag{39}$$

where the numerator is composed of the total non-production costs (which, through the lens of the model, I interpret as selling-related activities), whereas the denominator is composed of the total production costs.[38]

Moreover, to test the aforementioned predictions of the model in the data, I exploit the following regression specification, given by:

$$\log y_{it} = \alpha + \sum_{a=1}^{10} \gamma_a \mathbb{1}\{\text{age}_{it} = a\} + \phi_{st} + \varepsilon_{it}, \tag{40}$$

where $y_{it} \in \{\mu_{it}, \varrho_{it}\}$, the firm-level markup μ_{it} is defined in Appendix A.1.5, the selling-ratio ϱ_{it} is the ratio of selling-related expenditure to the cost of goods sold, where selling-related expenditure is defined in Appendix A.1.4, age_{it} is the firm's age, and ϕ_{st} are sector-year fixed effects. The coefficients γ_a are the parameters of interest that measure the average $\log y_{it}$ for each age group using within sector-year variation.

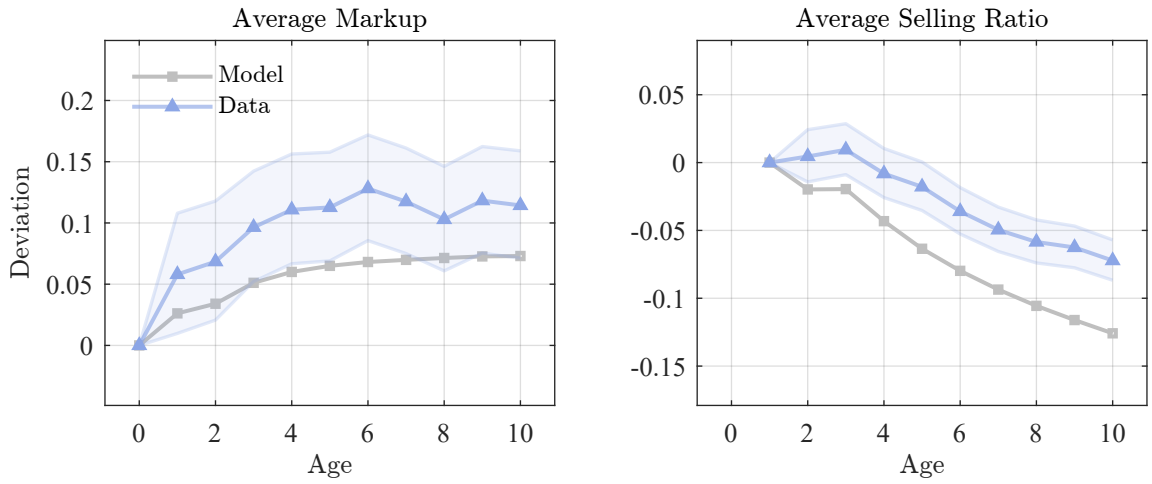
Figure 7 shows the evolution of the average markup and average selling ratio for firms of different ages.[39] The model-implied markups over the life cycle satisfactorily follow the one in the data; if anything, the model-implied one has a slightly flatter profile over the life cycle.[40] The model-implied selling ratio declines over the life cycle of the firm, as we also see in the data. However, in this case,

[38] Notice that both the numerator and the denominator should be multiplied by w, the wage, which however is not reported, as it is canceled out.

[39] I plot the results for the initial part of firms' life cycle; however, the patterns remain similar when the age is more than ten. The selling ratio is plotted only when the age is greater than zero because in the model, entrant firms face a cost composition that is different, as they do not pay the convex cost.

[40] Similar empirical findings have also been documented by Alati (2021) in Compustat and by Peters (2020) in Indonesian data. They both find that markups increase over the firms' life cycle.

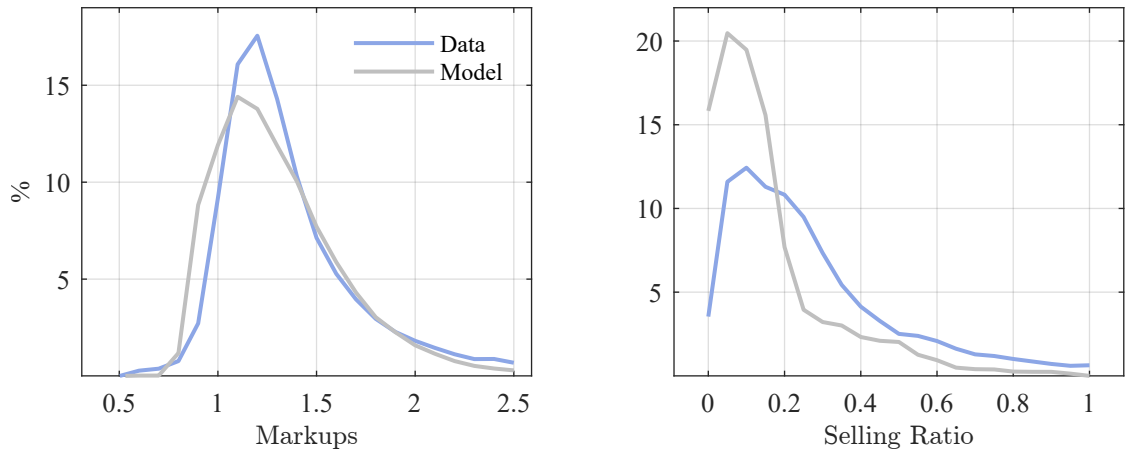
Figure 7: Life Cycle of Markups and Selling Ratio—Model and Data



Note. The figure on the left shows the average markup across firms of different ages, both in the model (light grey line with squares) and in the data (light blue line with triangles); the figure on the right shows the average selling ratio across firms of different ages, both in the model (light grey line with squares) and in the data (light blue line with triangles). The light blue areas are the 90% confidence interval. All variables are reported relative to the initial year, which is normalized to zero.

the model performs quantitatively less well than in the markups case. In the data, the selling ratio declines less compared to the model.

Figure 8: Distributions of Markups and Selling Ratio—Model and Data



Note. The figure on the left shows the markup distribution in the data (light blue) and in the model (light grey). The figure on the right shows the selling ratio distribution in the data (light blue) and in the model (light grey). The distributions in the data are calculated within the period 1977-1985. The distributions of markups are showed within the [0.5, 2.5] range, whereas, the distributions of the selling ratio are shown within the [0, 1] range.

As a final validation exercise, I compare the model-implied distribution of the markups and the selling ratio with their empirical counterparts. Figure 8 shows the comparison. The figure on the right shows the model-implied distribution of markups (light grey) and its empirical counterpart (light blue); the figure on the left shows the model-implied distribution of the selling ratio (light

grey) and its empirical counterpart (light blue). Overall, the qualitative fit is satisfactory.

The distribution of markups implied by the model is very close to the empirical counterpart. This is a successful outcome of the model, as the only targeted moment of that distribution is its cost-weighted average. Moreover, the model aptly captures the right skewness of the empirical distribution of the selling ratio. However, as none of the moments of this distribution has been used to calibrate the model, there are some quantitative differences: (i) the data show a higher mass near zero; and (ii) the empirical distribution of the selling ratio is less dispersed compared to the one implied by the model. Without further data, is impossible to say where these differences come from; however, in Appendix B.3.3, I show that by using an alternative measure of selling-related expenditure, the overall qualitative features of the empirical distribution of the selling ratio remain unchanged.

5 Rising Returns to Scale and the Macroeconomics

Having calibrated and validated the model, in this section, I move forward to study the macroeconomic implications of a rise in the returns to scale, as documented in Section 2.3. To this end, I will analyze, within the model, the effect of rising returns to scale from 1 to 1.05, keeping all the other parameters fixed. First, to shed light on the main mechanism, I discuss the qualitative implications of such a rise in returns to scale in the model. Second, I present suggestive evidence for the mechanism inbuilt in the model. Third, I use the model to study the quantitative implications of this 5% rise in returns to scale, as documented in section 2.3.

5.1 Inspecting the Mechanism

In this section, I explore the qualitative implications of a rise in returns to scale from 1 to 1.05, keeping all the other parameters fixed to the 1980 calibration. First, I link the effect that rising returns to scale have on the marginal costs of production at the firm level. Second, I explain how changes in the marginal cost of production affect markups and business dynamism.[41]

Given the production structure of the model, as specified in Section 3, the firm-level marginal cost of production is given by:

$$\mathcal{MC}(z,n;w) \equiv \frac{1}{\alpha}\left(\frac{n}{e^z}\right)^{\frac{1-\alpha}{\alpha}}\frac{w}{e^z}, \tag{41}$$

where α is the firm-level returns to scale, n is the mass of customers, that is, the firm-level size, e^z is the idiosyncratic productivity, and w is the wage. Notice that, when $\alpha = 1$—in the presence of

[41] Appendix B.2 extends the intuitions provided in this section to a more general case in which firms produce also using capital.

constant returns to scale—the marginal cost of production reduces to the more familiar w/e^z; hence, it is just the ratio of the wage to the idiosyncratic productivity.

However, when $\alpha > 1$, the marginal cost of production not only depends on the firm's size, but also decreases in it—this is under the quantitative-relevant scenario in which $n/e^z \geq 1$.[42] Therefore, this model, once calibrated to the empirical findings presented in Section 2.3, implies that the bigger a firm is, the better it becomes to produce, and hence, the lower its marginal cost of production is. This link can be interpreted as the model microfoundation of a technological change biased toward larger firms. In particular, the negative dependence of firm-level marginal costs of production with size stems from the notions that bigger firms (with bigger economic activities) manage to gather more information about their production processes (and potentially about their customers as well) and use it, owing to new information and communication technologies (ICT), to improve production. This mechanism creates a virtuous circle where bigger firms are better at producing; hence, become even bigger and better at producing, and so on.[43]

Figure 9 on the left shows, for firms with a different number of customers n, that is, for firms of different size, how the firm-level marginal cost of production changes as the returns to scale change—the analysis is performed under the already stated quantitative-relevant scenario in which n/e^z is big enough, that is, is weakly greater than 1.

In the model, the firm-level marginal cost of production declines monotonically from the 1980 steady state to 2014 steady state, meaning that an increase in the returns to scale lowers the marginal cost of production for all firms in the latter economy. Moreover, as shown by the graph, the marginal cost of production, after an increase in the returns to scale above one (the green area), declines much more for bigger firms. This is a well-known feature of increasing returns to scale in the production function (which, as shown in Section 2.3, is the empirically relevant case), where an increase in the input allows firms to produce more than proportionally, effectively lowering the quantity of input needed to achieve a given level of outputs. Therefore, the increase in returns to scale has a differential effect across firms, favoring bigger firms in the economy.

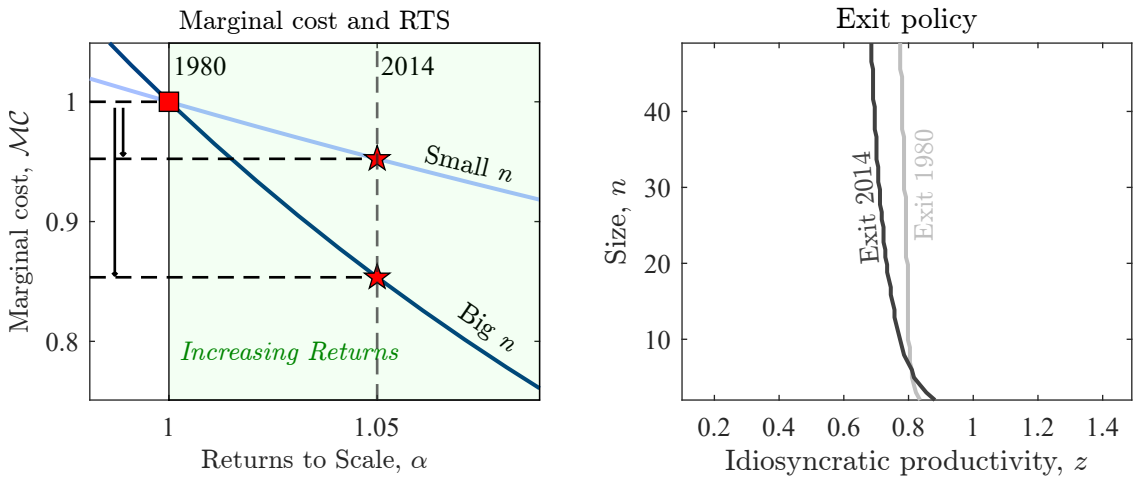
The decline in the marginal cost of production has three direct implications: (i) it increases the willingness of firms to scale up, and hence, their expenditures devoted to customers acquisition; (ii) it raises the firm-level markups; and (iii) it weakens the selection process in the model.

[42] In the model, I do not restrict this ratio to be greater than one, but when I calibrate it to match the firm size distribution, as explained in Section 4.2, I indeed find that this is the case. In this sense, this should not be seen as an assumption but as a quantitative result.

[43] Lashkari, Bauer, and Boussard (2021) document (using rich firm-level data from France) that investment in ICT has allowed French firms to increase their returns to scale in production in recent years.

Newman (2014), Agrawal, Gans, and Goldfarb (2018), Begenau, Farboodi, and Veldkamp (2018), Goldfarb and Trefler (2018), and Carriere-Swallow and Haksar (2019) provide additional microfoundations for the same concept. All of them emphasize the potential role of data, particularly gathering information from the customer base, which can give rise to increasing returns to scale.

Figure 9: Returns to Scale, Marginal Costs, and Selection



Note. The figure on the left shows the relation between the firm-level marginal cost of production and the returns to scale, α, for different levels of customers, that is, size. The dark blue line represents the marginal cost of a Big firm (high customer firm), whereas the light blue line represents the marginal cost for a Small firm (low customer firm).
The figure on the right shows the exit threshold in the 1980 and 2014 steady state over the firms' state space. The 2014 steady state has the same calibration as the 1980 one but with higher returns to scale, that is, $\alpha = 1.05$, The dark light grey line is the 1980 threshold, whereas, the dark grey line is the 2014 threshold.

First, with lower marginal costs of production, firms want to achieve a bigger size; as a consequence, they devote more resources to activities related to accumulating new customers. This implies that, in the 2014 steady state, firms will devote relatively more resources to non-production costs compared to production costs. As a consequence, there will be a shift away from production costs toward non-production costs, as observed in the Compustat data by De Loecker, Eeckhout, and Unger (2020).

Second, the decline in the marginal cost of production increases the surplus generated by the customer-firm relation, as firms are effectively better at producing. However, because there is an incomplete pass-through of costs in the model, only a fraction of this increase in the surplus will be passed on customers in the form of lower prices. Firms will retain the remaining fraction in the form of higher markups. Therefore, due to the decline in the marginal cost of production, firms will experience an increase in markups in the 2014 steady state.

Third, the decline in the marginal cost of production weakens, on average, the firms' selection process. This can be seen in Figure 9 on the right. The figure plots the exit threshold over the firms' state space in the 1980 and 2014 steady state. It can be seen that, in the latter steady state, the exit threshold moves, on average, to the left, implying that less productive firms will be able to operate in the economy. This is because, in the 2014 economy, firms are better at producing, which increases their resilience to adverse productivity shocks.

This decline in selection has two direct implications: (i) it lowers the entry rate of firms in the

economy; and (ii) it decreases the churning of firms, which has as a consequence a decline in the reallocation of labor. First, when the selection declines, the exit rate declines as well. In a stationary equilibrium, where the exit rate must equal the entry rate, this translates into a one-to-one decline in the entry rate. Second, the decline in the entry and exit rate translates into a firms' lower attrition rate. This implies that the reallocation of labor between entrant and exciters declines, and hence, the overall labor reallocation declines. Thus, the aforementioned decline in the selection translates into a decline in business dynamism.

As a final remark, I emphasize that, although selection weakens on average, it increases for the smallest firms in the economy, that is, firms with few customers. This is because small firms have to attract new customers. However, this is more costly in the 2014 steady state because these small firms must compete with the biggest firms that can now exploit their scale economies to compete for customers through very low prices.[44] Therefore, only marginally more productive small firms can do so, and consequently, this increases the selection process for small firms. Moreover, given that small firms are mostly new entrant firms, this acts as an entry barrier, which ulteriorly exacerbates the decline in the entry rate in the new economy.

5.2 Mechanism Validation

In this section, I test in the data the main qualitative predictions outlined above. The model predicts that a rise in firm-level returns to scale should increase markups and selling-related expenditures and decrease business dynamism. To this extent, I first show in the data that there has been a rise over time in the firm-level selling-related expenditures relative to production costs.[45] Then, exploiting only cross-sectoral variation in the data, I document that, in sectors where returns to scale are higher, selling expenditures and markups are higher, whereas business dynamism is lower.

5.2.1 The Rise in Selling-Related Expenditures

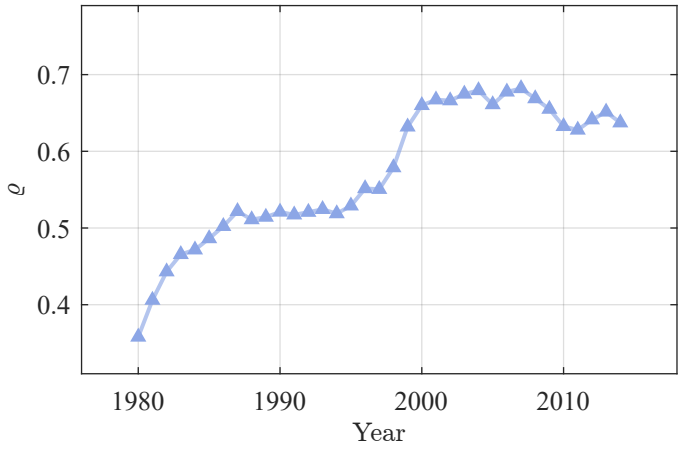
The model predicts that a rise in returns to scale increases firm-level expenditure in selling-related activities at the expense of production costs. Therefore, it is natural to look at the evolution of this ratio over time as a first test of the implications outlined above. To this extent, I look at the evolution over time of the average selling ratio, as defined in Appendix A.1.4.[46]

[44] This kind of behavior has recently received a great deal of attention in the antitrust debate; see Khan (2016). The model rationalizes this behavior as the outcome of the rise of scale economies that big firms, such as Amazon, can take advantage of to set prices lower those of their smaller competitors.

[45] I focus my attention only on the rise over time in firm-level selling-related expenditures relative to production costs because it is relatively less known. The rise in markups and the decline in business dynamism have been extensively documented; see De Loecker, Eeckhout, and Unger (2020) and Decker, Haltiwanger, Jarmin, and Miranda (2014).

[46] The rise in non-production costs over time relative to total costs has already been documented by De Loecker, Eeckhout, and Unger (2020). However, I focus on a different measure, that is, the ratio of selling-related costs (these are similar to the non-production costs analyzed in De Loecker, Eeckhout, and Unger (2020); see Appendix A.1.4 for a more detailed

Figure 10: Average Selling Ratio



Note. The figure plots the evolution of the average selling ratio between 1980 and 2014. The measure is constructed using a simple average.

Figure 10 shows the evolution of the average selling ratio. At the beginning of the sample, the average selling ratio was approximately 0.4, then rose almost up to 0.7 around 2000, and then went back roughly to 0.65 by the end of the period. Hence, the measure has experienced an increase slightly above 62% over the period of analysis. Therefore, I conclude that, with higher returns to scale over time, we should expect to observe higher firm-level expenditures in selling-related activities, relative to production costs, from the data. In Appendix A.3, I show that using an alternative measure of selling-related activities produces similar results.

I finish emphasizing that this is an aspect peculiar to the theory outlined in this paper. Only a model in which the market power is a long-term investment would produce such an empirical pattern in selling-related activities relative to production costs. Models in which the market power is derived from the love for variety (see, for example, Dixit and Stiglitz (1977), Kimball (1995), and Atkeson and Burstein (2008)) or from search frictions with only pricing strategies and no expenses devoted to the acquisition of new customers (see, for example, Paciello, Pozzi, and Trachter (2019) and Roldan-Blanco and Gilbukh (2020)), would not be able to produce an endogenous increase in selling-related activities relative to production costs, as in the one documented above.

5.2.2 The Cross-Sectoral Implications of Higher Returns to Scale

Here, I test in the cross-section of sectors the qualitative predictions of the model outlined above. The model would predict that, in sectors where returns to scale are higher, we should expect to observe lower business dynamism (lower entry and reallocation rates), higher markups, and higher selling-

explanation) to production costs. This has two advantages: (i) it avoids the challenges of computing the costs of holding capital, which requires additional assumptions; and (ii) it focuses directly on the shifts in those particular costs emphasized by the theory in this paper. However, regarding the results, both measures show a clear rise over time.

expenditures relative to production costs (selling ratio). To do so, I regress all these variables against sector-level returns to scale, as estimated in Sections 2.3. The sector-level entry and reallocation rates are from the BDS data; the sector-level cost-weighted markups are computed with the method proposed by Hall (1988) and De Loecker and Warzynski (2012); and the sector-level selling ratio is computed, as described in Appendix A.1.4.

Table 2: Returns to Scale and Cross-Sectoral Correlations

	Business Dynamism			
	(1)	(2)	(3)	(4)
	Entry rate	Reallocation rate	Markups (log)	Selling ratio
Returns to scale	−0.047***	−0.145***	0.354*	0.839***
	(0.010)	(0.020)	(0.213)	(0.113)
Observations	518	518	722	722
R-squared	0.602	0.764	0.144	0.687
Sector-Time FE	✓	✓	✓	✓

Notes. Fixed effects are at the sector-time level, where the sector is at the 1-digit level. Robust standard errors are in parenthesis. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table 2 shows the results.[47] The coefficients are estimated using only within sector-time variation; this is important because most of these variables have time trends, which could give rise to spurious correlations. The regressions clearly show that in sectors where firms produce with higher returns to scale, business dynamism is lower; that is, entry and reallocation rates are lower.[48] All coefficients are significant. In Appendix A.3, I show that using alternative measures of selling-related activities produces similar results.

Although these correlations seem to propose a rise in returns to scale as an underlying factor behind some recent firm-level trends, I should caution the reader from any causal interpretation of these relations. However, the presence of these correlations indeed supports the economic forces outlined by the model.

[47] It is worth noticing that the coefficient related to business dynamism is estimated over a smaller sample. This is because the BDS data merge some sectors; for example, manufacturing, which normally is classified by NAICS codes 31-32-33, in BDS is a unique sector.

[48] In related work, Gao and Kehrig (2017) use Census data to show that, where firms produce with higher returns to scale, the average firm size and concentration are higher. This reinforces the correlations documented above, as it confirms in a different dataset similar patterns compared to the analysis emphasized in this section.

5.3 Quantitative Implications

This section explores the main quantitative implications of the rise in returns to scale. First, it analyzes the effect that rising returns to scale have on business dynamism, markups, and other aggregate trends. Second, it studies the implication of this technological change on the distribution of markups. Third, it examines the consequences of the rise in returns to scale for firm-level responsiveness of employment growth to productivity shocks. Appendix B.4 shows additional quantitative results.

5.3.1 Rising Returns to Scale and Aggregate Trends

Here, I study the quantitative implication of a 5% rise in returns to scale (from 1 to 1.05, as documented in Section 2.3) for the decline in business dynamism, the rise in markups, and the evolution of other trends, such as the rise in concentration and the rise in firm-level selling-related activities. To this end, I compare two steady states, the 1980 one, calibrated as documented in Section 4.2, and the 2014 one, where I only let the returns to scale α rise from 1 to 1.05, keeping all the other parameters fixed.

Table 3: Effect of Rising Returns to Scale

			Change			
	1980 S.S.	2014 S.S.	Model	BDS	Compustat	Model/Data
Business Dynamism						
Entry rate	0.139	0.104	−25%	−40%	–	62%
Reallocation rate	0.294	0.237	−19%	−27%	–	70%
Share of old firms	0.322	0.467	+45%	+50%	–	90%
Employment share of young firms	0.204	0.094	−69%	−56%	–	96%
Markups						
Avg. markup (cost-weighted)	1.202	1.229	+2%	–	+7%	29%
Others						
Avg. selling ratio	0.4	0.65	+9%	–	+62%	14%
Concentration (HHI)	7.003e-06	7.440e-06	+6%	–	+33%	18%

Notes. All variables are calculated coherently with their definitions, as used in the data. The average markup is calculated using cost weights, whereas the average selling ratio is calculated using a simple average across firms. Concentration is calculated as described in Grullon, Larkin, and Michaely (2019). The data sources are BDS and Compustat. To calculate the empirical moments from the 1980s I use the time window 1977-1985, whereas for the empirical moments from the 2014, I use simple the values in that year. The last column shows the fraction of the overall empirical variation explained by the model.

Table 3 shows the quantitative implications of the rise in returns to scale for the aggregate trends.

The model can explain an important share of the decline in business dynamism, as it explains 62% of the decline in the entry rate and 70% of the decline in the reallocation rate. Moreover, because the rise in returns to scale inherently favors the bigger and oldest firms in the economy, the model can explain 90% of the rise in the share of the old firms (firms with 11+ years) and 96% of the decline in the employment share of young firms (firms with less than 5 years).

Moreover, the model is able to explain 29% of the rise in the average cost-weighted markup.[49] I focus on the evolution of the cost-weighted measure, which is the welfare-relevant aggregate measure, as documented by Grassi (2017) and Edmond, Midrigan, and Xu (2018). However, in the next session, I look at the evolution over time of the markup distribution to analyze the features of the rise in the sales-weighted measures that are related to the reallocation of economic activity toward bigger firms. Although the model explains a non-negligible fraction of the rise in the aggregate markups, it cannot explain most of it. This shows that the rise in returns to scale does not seem to be the only force behind the rise in the data, suggesting that there may be additional mechanisms at work in the US economy that can account for the unexplained rise.

Finally, the model is also consistent with the rise in selling-related expenditures and product market concentrations, as observed in the data. In particular, the model can explain 14% of the rise in the benchmark measure of selling-related expenditures and 45% of the increase in the alternative advertisement-based measure, as documented in Appendix A.3. Although the model explains only a fraction of this rise, we can still define this as a success, given that this endogenous rise is a distinctive feature of this model, where firms actively invest in their market power (see Section 5.2.1 for a more detailed explanation of this point). The model also explains 18% of the rise in concentration, which shows that the model captures the reallocation of economic activity toward bigger firms that have been documented empirically by Kehrig and Vincent (2021), Autor, Dorn, Katz, Patterson, and Van Reenen (2020), and De Loecker, Eeckhout, and Unger (2020).

5.3.2 Evolution of the Markup Distribution

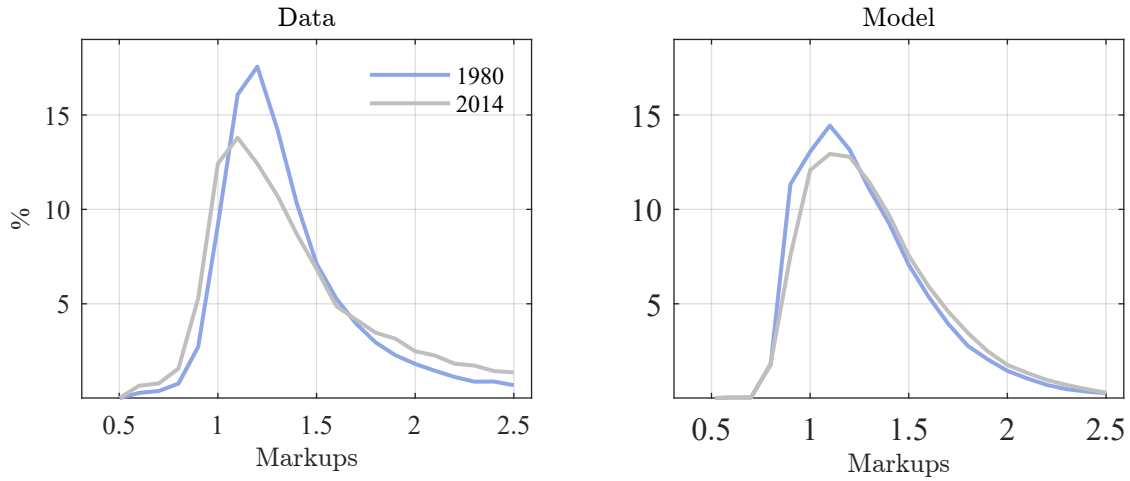
Analyzing the rise in markups, De Loecker, Eeckhout, and Unger (2020) show that there has been a substantial change in their distribution overall. In particular, they notice that much of the rise in the average markup is due to reallocation of the economic activity toward the right tail of the distribution—in their words, there has been a fattening of the right tail of the markup distribution.

In this section, I look at this prediction in the model. Hence, I compare the model-implied dis-

[49] I document a 7% increase in the cost-weighted markup; De Loecker, Eeckhout, and Unger (2020) report a rise of approximately 10%. This difference is mainly due to the way I clean Compustat. In particular, I drop all firms that are not incorporated in the US and all utilities and financial firms. Notice that these choices do not change the qualitative behavior of markups compared to the paper above. However, they lower their rise. Therefore, readers should keep this caveat in mind when interpreting the numbers.

tribution of markups in both steady states, that is, in the 1980s and 2014, with the one in the data as documented by De Loecker, Eeckhout, and Unger (2020).

Figure 11: Distributions of Markups—Model and Data



Note. The figure on the left shows the empirical markup distribution in 1977-1985 (light blue) and in 2005-2014 (light grey). The figure on the right shows the model-implied markup distribution in the 1980 calibration (light blue) and in the 2014 calibration (light grey). Both distributions are shown within the [0.5, 2.5] range.

Figure 11 presents the results. The figure on the left shows the empirical distribution of markups in the 1980s, light blue, and in 2014, light grey. The figure on the right shows the model-implied distribution of markups in the 1980s (light blue) and in 2014 (light grey). It can be seen that the model qualitatively captures the overall change in the distribution of markups. Specifically, in the 2014 steady state, the model exhibits a considerable fattening of the right tail, compared to the 1980s steady state, as the one portrayed in the data and emphasized in De Loecker, Eeckhout, and Unger (2020).

This, in the model, happens because the rise in returns to scale reallocates the economic activity toward bigger firms, which are also the ones the higher markups. This reallocation toward bigger firms translates into a fatter right tail of the markup distribution. Therefore, the model produces the rise in the average markup jointly with the distributional changes emphasized by the empirical works of Kehrig and Vincent (2021), Autor, Dorn, Katz, Patterson, and Van Reenen (2020), and De Loecker, Eeckhout, and Unger (2020).[50]

[50] Edmond, Midrigan, and Xu (2018) demonstrated with an exact decomposition that this rise in the variance of the markups distribution is the main reason why the sales-weighted average markup rose by more compared to the cost-weighted one. The model captures this qualitatively, as it produces a rise in sales-weighted markup of 3%, which is approximately 17% of what I observe in my calculations, compared to a 2% rise in the cost-weighted markup.

5.3.3 Declining Responsiveness

In this section, I look into additional facts highlighted by the empirical literature related to the decline in business in the US. In particular, Decker, Haltiwanger, Jarmin, and Miranda (2020) show that an important component of the decline in business dynamism is the fact that firms in recent decades have responded less to productivity shocks, that is, conditional to a productivity shock, they expand (or contract) less.

To analyze this feature in the model and in the Compustat data, I proceed in two steps: (i) I replicate the spirit of their empirical investigation, both in the model and in the data; and (ii) I propose an exact decomposition to shed light on the forces behind this decline in responsiveness.

Therefore, in the data, I implement the following regression:

$$g^{\ell}_{it+1} = \alpha + \beta\, a_{it} \otimes \mathcal{F}(t) + \boldsymbol{X}'_{it}\boldsymbol{\gamma} + \phi_{st} + \varepsilon_{it}, \tag{42}$$

where $g^{\ell}_{it+1} \equiv 2 \times (\ell_{it} - \ell_{it-1})/(\ell_{it} + \ell_{it-1})$ is the growth rate of employment, a_{it} is the empirical measure of total factor productivity revenue (TFPR), that is, the residual from the production function in Section 2.3, $\mathcal{F}(t)$ is a flexible function of time, $\boldsymbol{X}_{it}$ is a vector of controls, and ϕ_{st} are sector-time fixed effects. The symbol $\otimes$ represents the full interaction between the two variables. Therefore, the coefficient of interest will be the β associated with the interaction between a_{it} and $\mathcal{F}(t)$, which captures the evolution over time of the marginal effect of changes in productivity.[51]

Results are presented in Table 4. The first three columns show the decline in firm-level responsiveness in the data with three different specifications: (i) a linear trend; (ii) a dummy capturing responsiveness after 2000; and (iii) a set of dummies that captures the responsiveness in each decade having as a benchmark the first decade. Both the first (parametric) and last two (semi-parametric) regressions show a statistically significant decline in firm-level responsiveness over time. In particular, the first specification shows a decline in responsiveness, between 1980 and 2014, of 0.035, whereas the last specification shows a decline of 0.011.

The last column shows the evolution of responsiveness in the model. In the model, firm-level responsiveness also declines between the two steady states. In particular, we can see that this decline of 0.027 lies in the empirical range reported above.

Finally, to understand which forces lie behind the above decline in the model, I define firm-level

[51] In the model, as I only have a simulated panel for the two distinct steady states, I have to run a different regression. In particular, I run the following regression in both steady states:

$$g^{\ell}_{it+1} = \alpha + \beta\, a_{it} + \boldsymbol{X}'_{it}\boldsymbol{\gamma} + \phi_{st} + \varepsilon_{it}, \tag{43}$$

where I do not allow for time-dependent functions. However, the regression follows the same spirit and allows for a very similar interpretation. Therefore, to analyze the decline in responsiveness, within the model, I look at the difference of the estimated coefficients in the two steady states, that is, $\beta^{2014} - \beta^{1980}$.

Table 4: Declining Firm-Level Responsiveness

	(1) Compustat	(2) Compustat	(3) Compustat	(4) Model
$\widehat{\beta}^{2014} - \widehat{\beta}^{1980}$				-0.027
$a_{it} \times Year$	-0.001^{***}			
	(0.000)			
$a_{it} \times \mathcal{I}_{t\geq 2000}$		-0.006^{**}		
		(0.002)		
$a_{it} \times \mathcal{I}_{t\in[1990,2000)}$			-0.007^{**}	
			(0.003)	
$a_{it} \times \mathcal{I}_{t\in[2000,2010)}$			-0.009^{***}	
			(0.003)	
$a_{it} \times \mathcal{I}_{t\in[2010,2015)}$			-0.011^{***}	
			(0.004)	
Controls	✓	✓	✓	✓
Sector-Time FE	✓	✓	✓	
Observations	143,771	143,771	143,771	
R-squared	0.037	0.038	0.038	

Note. The table reports the change in firm-level responsiveness to productivity shocks. The controls are size, the interaction of employment with the time function, and past productivity. In column (1), I allow for a simple linear trend. In columns (2) and (3), I instead allow for a more flexible set of dummies, where $\mathcal{I}_{t\in T}$ equals 1 when $t \in T$.

responsiveness as:

$$\frac{\Delta \log \ell_{it}}{\Delta z_{it}} = \frac{1}{\alpha} \times \left[\frac{\Delta \log y_{it}}{\Delta z_{it}} - 1 \right], \tag{44}$$

where α is the returns to scale, and $\Delta \log y_{it} / \Delta z_{it}$ is the output growth associated with productivity growth.[52] Equation (44) shows that the rise in returns to scale translates directly into a decline in firm-level responsiveness.[53] Moreover, the rise in returns to scale can also affect responsiveness indirectly through its effect on the output growth associated with productivity growth. Taking stocks, in the model, the direct effect of rising returns to scale dominates, and hence, firm-level responsiveness declines after the aforementioned technological change, making the model consistent with the

[52]This definition of firm-level responsiveness is slightly different from the one implied by the regressions above. However, notice that controlling in the regressions for past productivity allows for a similar interpretation of firm-level responsiveness: the growth in employment associated with productivity growth. In light of this and consistent with Decker, Haltiwanger, Jarmin, and Miranda (2020), I stick with the above regression analysis as the benchmark measure of firm-level responsiveness. However, equation (44) is still useful in understanding which mechanism is behind the decline in the model.

[53]The difference in the brackets is always positive. In particular, it can be shown that $\Delta \log y_{it} / \Delta z_{it} = 1 + \alpha \Delta \log \ell_{it} / \Delta z_{it}$, where $\Delta \log \ell_{it} / \Delta z_{it} > 0$.

findings documented by Decker, Haltiwanger, Jarmin, and Miranda (2020).

6 Conclusion

In this paper, I documented empirically that US firms have undergone a technological change biased toward higher returns to scale. In particular, leveraging the Compustat data and state-of-the-art production function estimators, I document that firm-level returns to scale experienced a 5% increase, going from 1 in 1980 to 1.05 in 2014. Moreover, I find that this rise is happening within all sectors—suggesting a technological interpretation—and is not the outcome of a reallocation of economic activity toward high returns to scale sectors.

To understand the implications of this technological change for some of the main trends in the US economy, I propose a novel heterogeneous firms model grounded in search frictions in the product market. Search frictions make the model consistent with several features of the microdata: (i) they microfound endogenous heterogeneous markups; (ii) entail firms' active expenditures to attract customers, and (iii) imply that firms grow through the accumulation of new customers, which empirically accounts for 70% of their life-cycle growth. In the model, because of the central role of prices for attracting and retaining customers, changes in returns to scale, affecting firm-level marginal costs, influence the firm-level ability to price, grow, and charge markups.

I calibrate the model with firm-level data and use it to quantify the effect of the 5% rise in returns to scale. In the model, such a technological change can explain between 62-70% of the decline in business dynamism, 29% of the increase in the average cost-weighted markup, and between 14-45% of the rise in expenditures devoted to customer acquisition. The model captures all these, while being consistent with additional micro-facts, such as the aging of US firms, the reallocation of economic activity toward high-markup firms, and the decline in firm-level responsiveness to productivity shocks.

Several potential directions are left unexplored. It would be interesting to study the implications of the increase in returns to scale for the increase in merger and acquisition activities, as witnessed in recent decades. Moreover, it would be valuable to introduce in the model horizontal product differentiation as an additional source of market power. I leave these questions to future research.

A Empirical Analysis Appendix

A.1 Data

This section presents the construction of the main sample and main variables, providing summary statistics for the final sample. Then, it shows how to construct the user cost of capital and explains which variable is used in the production function estimation as labor (variable) input. Finally, it shows the construction of the measures of selling-related activities and markups.

A.1.1 Main Sample, 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 (SIC codes between 4900-4999) because they are heavily price regulated, and I also exclude financial firms (SIC codes between 6000-6999) because their balance sheets are dramatically different from other 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:

$$k_{it} = (1 - \delta)k_{it-1} + x_{it}, \tag{45}$$

where $x_{it} - \delta k_{it-1} = \texttt{PPENT}_{i,t} - \texttt{PPENT}_{it-1}$ is the net investment, and the initial capital stock, k_{i0}, is initialized using the first available entry of PPEGT.[54]

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.[55] The table below presents a few basic summary statistics for a few leading variables used in the analysis.

A.1.2 User Cost of Capital

As mentioned in the main body of the paper, one of the challenges of using the cost shares approach is that it requires a measure of the user cost of capital. To this end, I define the user cost of capital as:

[54] Given that a measure of real capital is needed for the analysis, I deflate the measure of net investment with the appropriate deflator.

[55] Deflators are taken from the NIPA tables.

Table A.1: 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
25th Percentile	22,553	13,896	115	5,756	3,552	5
Median	139,060	84,909	638	36,079	24,323	11
75th 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 but Age are in thousands US$. Sales and Costs of Goods Sold are deflated with the GDP deflator using the base year 2012, whereas both capital stocks are deflated using the investment deflator with the base year 2012.

$$r_t = i_t - \mathbb{E}_t \pi_{t+1} + \delta, \tag{46}$$

where i_t is equal to the nominal interest rate, $\mathbb{E}_t \pi_{t+1}$ is expected inflation at time t, and δ is the depreciation rate of capital. I take the annual Moody's Seasoned Aaa Corporate Bond Yield as an empirical proxy for the nominal interest rate, the annual growth rate of the Investment Nonresidencial Price Deflator to calculate the expected inflation, and the depreciation rate is calibrated to $\delta = 0.1$, as in the rest of the paper.[56,57,58]

A.1.3 Variable Input in Production

Recent work based on Compustat, particularly since De Loecker, Eeckhout, and Unger (2020), has used the item Cost of Goods Sold (`COGS`) as the preferred measure of variable input in production. This choice was motivated by the need for a bundle of variable input expenditures to calculate firm-level markups. However, despite being an unavoidable choice, using it as a measure of variable input imposes an additional assumption in the estimation, as it assumes that labor and materials are perfectly substitutable.

However, as the primary goal of this paper is to estimate the returns to scale, and hence output elasticities, and not the markups, I favor a direct measure of the firm-level variable input. In particular, I use as a benchmark measure the variable `EMP`, which represents the number of employees in a given firm, and show robustness exercises using `COGS`. Therefore, to be consistent with this approach, when I calculate cost shares, I need to construct a measure of labor cost, $w_{it}\ell_{it}$. To do so, I use the labor cost expenditure (`XLR`) reported by a subsample of firms. For the firms that report it, I calculate

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

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

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

the labor cost per worker defined as $w_{it} \equiv \texttt{XLR}/\texttt{EMP}$, and then I calculate its within-sector median and use it to impute the labor cost for the firms that do not report it as $w_{it}\ell_{it} = \widehat{w}_{st} \cdot \texttt{EMP}_{it}$.

A.1.4 Selling-Related Expenditure

In this section, I present the two main approaches used to compute firm-level selling-related activities. Unfortunately, in Compustat, there is no perfect way to compute firm-level selling-related activities; therefore, while presenting the two approaches, I will emphasize their virtues and their weaknesses.

Benchmark measure. To measure firm-level selling-related expenditures, I use Selling General and Administrative (XSGA). This item in Compustat has been the focus of many recent studies such as: Gourio and Rudanko (2014), Ptok, Jindal, and Reinartz (2018), Afrouzi, Dernik, and Kim (2020), and Morlacco and Zeke (2021).[59] However, despite the acknowledged ability of Selling General and Administrative to capture firm-level selling-related expenditure, it is well known that this item reports many expenditures that are not directly related to selling efforts, such as bad debt expenses, expenditure in pensions and retirement, rents, and expenditure in research and development.[60] Therefore, to partially overcome the aforementioned limitations, my adjusted measure of *selling-related expenditure* is defined as:

$$\texttt{S}_{it} = \texttt{XSGA}_{it} - \texttt{XRENT}_{it} - \texttt{XPR}_{it} - \texttt{RECD}_{it} - \texttt{XRD}_{it}, \tag{47}$$

where XSGA is an expenditure in Selling General and Administrative, XRENT is an expenditure in Rents, XPR is an expenditure in Pensions and Retirement, RECD is an expenditure due to Bad Debts, and XRD is an expenditure in Research and Development.

Alternative measure. As an alternative measure to the above measure, I use the Compustat variable XAD, which reports the firm-level expenditure in advertisements. This is the only available item in Compustat that measures only (and somehow cleanly) selling-related costs; however, this measure suffers from two main drawbacks: (i) it reports the cost of advertising media (radio, television, newspapers, periodicals) and promotional expenses but excludes selling and marketing expenses, and (ii) half of the observations are missing.

[59] In particular, Ptok, Jindal, and Reinartz (2018) document that Selling General and Administrative is particularly good at capturing firm-level sales force spending.

[60] For a more exhaustive discussion on how research and development are accounted for in Compustat, see Peters and Taylor (2017). For an extensive list of items reported in Selling General and Administrative, see Afrouzi, Dernik, and Kim (2020). In my list, I reported to the best of my knowledge only the items reported in Compustat that are accounted for in Selling General and Administrative.

A.1.5 Firm-Level Markups

Throughout the paper, markups are constructed following Hall (1988) and De Loecker and Warzynski (2012); hence, the firm-level markup is given by:

$$\mu_{it} = \widehat{\beta}_{st}^{cogs} \cdot \frac{\texttt{SALE}_{it}}{\texttt{COGS}_{it}}, \tag{48}$$

where the $\widehat{\beta}_{st}^{cogs}$ is the output elasticity to COGS. To ease the comparability between this paper and the seminal work by De Loecker, Eeckhout, and Unger (2020), I use their measure of this elasticity. However, the results are robust to using the alternative measure of $\widehat{\beta}_{st}^{cogs}$ presented in Appendix A.2.

A.2 Additional Robustness Production Function

Here, I document the robustness of the results in Section 2.3. To this end, first, I present the alternative specification that I will use. Second, I present the results from these specifications, both for the average returns to scale and for the within-between sectors decomposition.

Alternative Control Function: Investment. Here, I document the robustness of the rise in returns to scale to alternative control functions such as investment. This particular control function has been pioneered by Olley and Pakes (1996) and discussed extensively by Ackerberg, Caves, and Frazer (2015).[61] To apply the methodology presented in Section 2.2.1 to the case in which investment is used as a control function, equation (4) has to be modified as:

$$q_{it} = \mathcal{P}(k_{it}, \ell_{it}, x_{it}, \mathbf{d}_{it}) + \varepsilon_{it}, \tag{49}$$

where x_{it} is now the firm's investment. Given this new augmented equation, the rest of the procedure is the same as the one outlined in Section 2.2.1.

Alternative Variable Input: Cost of Goods Sold. Here, I adopt an alternative specification of the production function, as used in the recent paper by De Loecker, Eeckhout, and Unger (2020). To this end, I use COGS instead of EMP as the variable input. This is a necessary shortcut to estimate firm-level markups in Compustat. However, it imposes an alternative set of assumptions as the true estimated production function is:

[61] A known drawback of using this alternative measure as a control function for the estimation is the presence of many zeros in investment (see, Levinsohn and Petrin (2003)). However, in Compustat, this is a minor issue, as the number of observations that are zero is particularly small relative to most of the dataset—this is due to the fact that Compustat is a firm-level dataset containing mostly big firms.

$$q_{it} = \beta^k k_{it} + \beta^{cogs}(\ell_{it} + m_{it}) + \omega_{it} + \varepsilon_{it}, \tag{50}$$

where m_{it} is the firm's materials. Equation (50) implicitly entails two additional assumptions: (i) first, now the production function is defined as the gross output, and hence, is partially subject to the identification criticisms laid out in Gandhi, Navarro, and Rivers (2020); and (ii) second and last, given that `COGS` is the sum of all production costs (particularly labor and materials), its adoption as an input in production implicitly assumes that labor and material are perfect substitutes within the production process.

Additional Dynamic Input: Intangible Capital. Recently, there has been a particular emphasis on the role played by the rise of intangible capital at the firm level.[62] Therefore, this could generate some concerns as that the rise in returns to scale could potentially be partially driven by the rise in unmeasured intangible capital as input in production. To address this concern, I estimate a new production function, augmented by intangible capital, given by:

$$q_{it} = \beta^k k_{it} + \beta^\iota \iota_{it} + \beta^\ell \ell_{it} + \omega_{it} + \varepsilon_{it}, \tag{51}$$

where ι_{it} is the intangible capital in production. This new specification entails an additional challenge: namely, the firm-level measurement of intangible capital. To this end, I take advantage of the balance sheet intangible capital and capitalize firm-level knowledge capital, as done in Chiavari and Goraya (2021). The balance sheet intangible capital is given by:

$$\iota_{it}^{balance\ sheet} = \texttt{INTAN}_{it} + \texttt{AM}_{it} - \texttt{GDWL}_{it}, \tag{52}$$

where `INTAN` is the net balance sheet intangible capital, `AM` is the ammortization of the balance sheet intangible capital, and `GDWL` is goodwill. Knowledge capital is given by:

$$\iota_{it}^{knowledge} = (1 - 0.30)\iota_{it-1}^{knowledge} + \texttt{XRD}_{it}, \tag{53}$$

where the depreciation rate is set to 30%, close to the empirical estimates by Ewens, Peters, and Wang (2019), `XRD` is the firm-level expenditure in research and development, and $\iota_{i0}^{knowledge}$ is set equal to zero. Finally, the total firm-level intangible capital is given by:

$$\iota_{it} = \iota_{it}^{balance\ sheet} + \iota_{it}^{knowledge}. \tag{54}$$

[62] In particular, Chiavari and Goraya (2021) show that intangible capital is rising dramatically as an input in production.

Alternative Production Function: Translog. Finally, I explore the robustness of the rise in returns to scale to an alternative production function specification. In particular, in this section, I adopt the following translog specification given by:

$$q_{it} = \theta_1^k k_{it} + \theta_1^\ell \ell_{it} + \theta_2^k k_{it}^2 + \theta_2^\ell \ell_{it}^2 + \theta_3^{k\ell} k_{it} \ell_{it} + \omega_{it} + \varepsilon_{it}. \tag{55}$$

To estimate the translog production function, I follow the methodology outlined in Section 2.2.1. However, the output elasticities are now given by:

$$\beta^k = \text{median}\Big\{\theta_1^k + 2\theta_2^k k_{it} + \theta_3^{k\ell} \ell_{it}\Big\}, \tag{56}$$

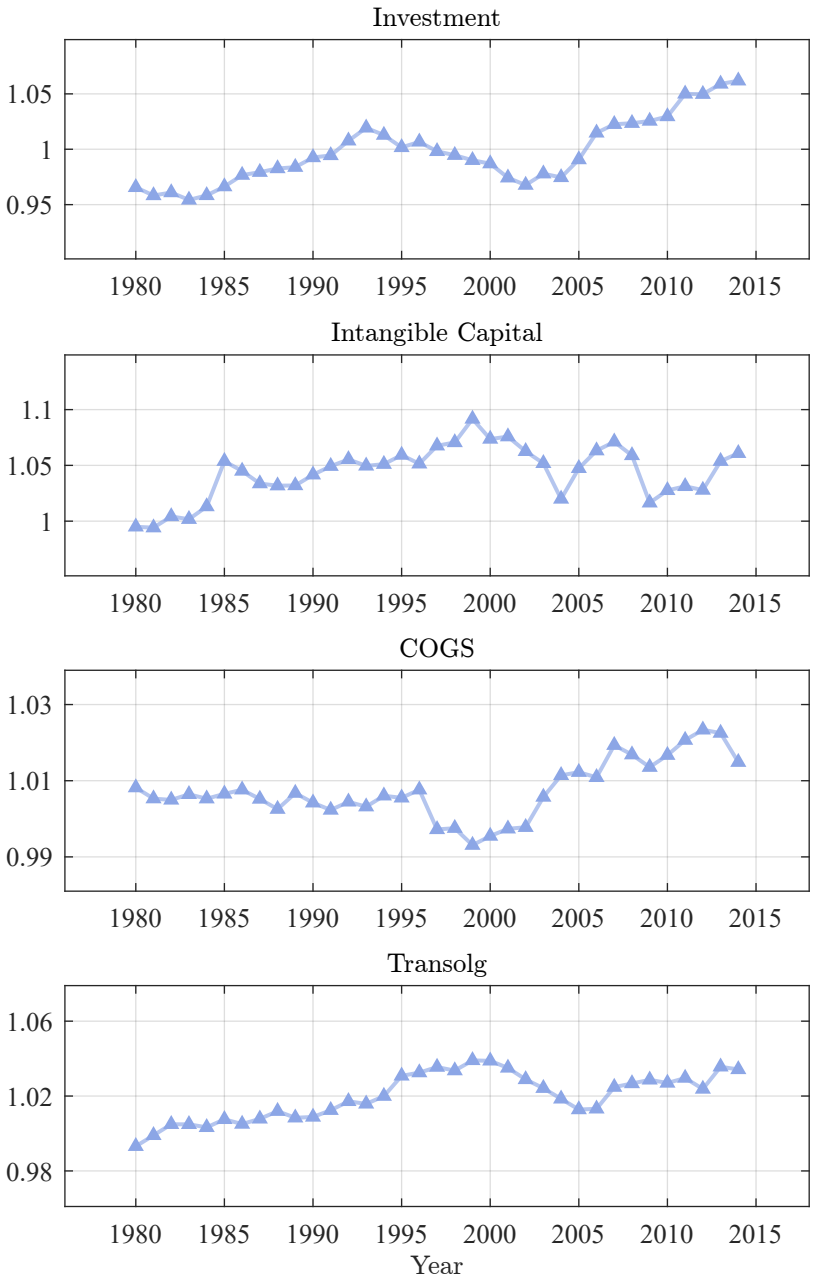
$$\beta^\ell = \text{median}\Big\{\theta_1^\ell + 2\theta_2^\ell \ell_{it} + \theta_3^{k\ell} k_{it}\Big\}. \tag{57}$$

Therefore, the returns to scale implied by the production technology from equation (55) is given by $\alpha = \beta^k + \beta^\ell$.

Results from Alternative Specifications. The results from the above specifications are presented in Figures A.1 and A.2. Figure A.1 shows the evolution of the sales-weighted average returns to scale in production from 1980 to 2014 for the different alternative specifications. The first graph shows the robustness exercise when we use investment as a proxy variable. The second graph shows the robustness exercise when we augment the production function with intangible capital as an additional dynamic input. The third graph shows the robustness exercise when we use the cost of goods sold (COGS) as the variable input. Finally, the fourth graph shows the robustness exercise when we adopt a translog specification for the production function.

Figure A.2 plots the counterfactual evolution of the within and between components implied by the decomposition from equation (14); that is, it shows the evolution of the average returns to scale only if the Δwithin component is at play and the evolution of the average returns to scale only if the Δbetween component is at play. The first graph shows the robustness exercise when we use investment as a proxy variable. The second graph shows the robustness exercise when we augment the production function with intangible capital as an additional dynamic input. The third graph shows the robustness exercise when we use the cost of goods sold (COGS) as the variable input. The fourth graph shows the robustness exercise when we adopt a translog specification for the production function.

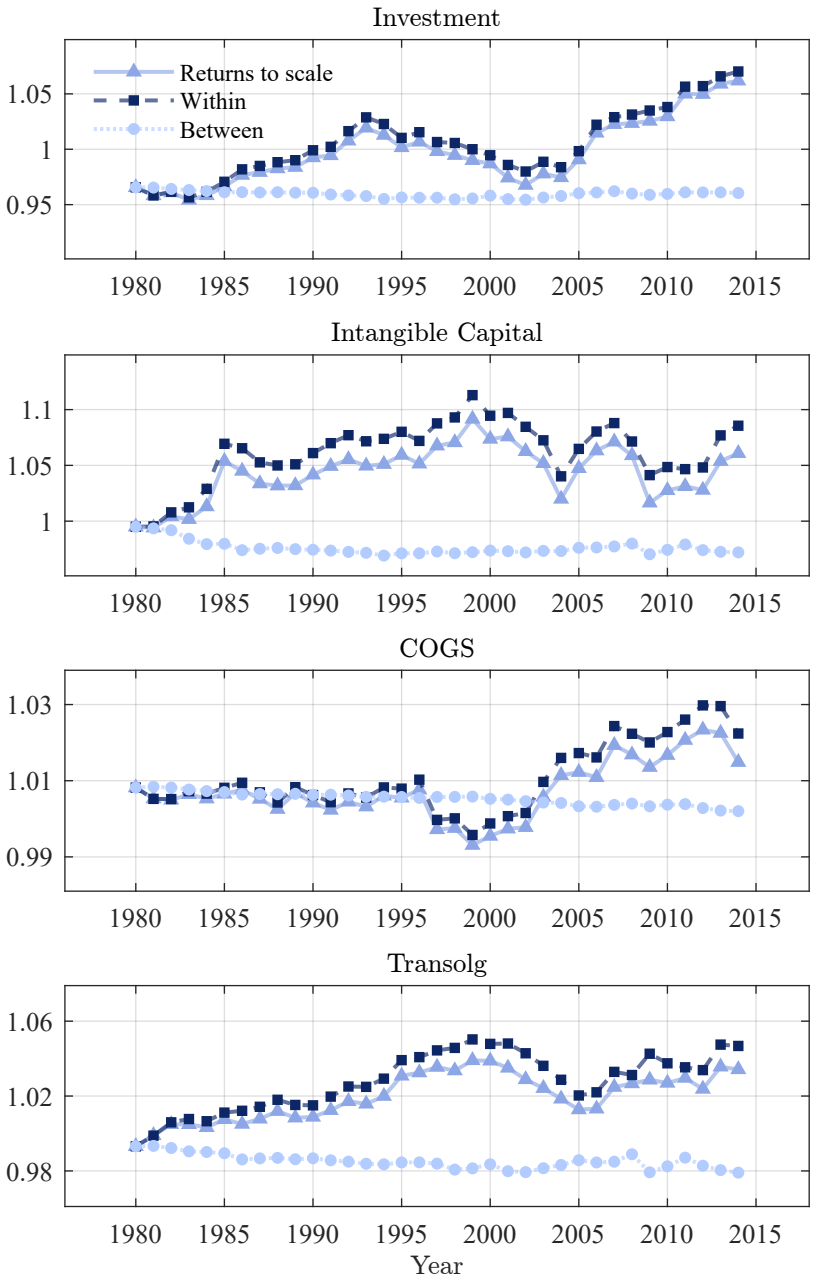
Figure A.1: Alternative Specifications – Robustness 1



Note. The figures above show the evolutions of the average returns to scale for all four robustness specifications. The first figure shows the evolution of the average returns to scale when we use investment as a proxy variable. The second figure shows the evolution of the average returns to scale when we augment the production function with intangible capital as an additional dynamic input. The third figure shows the evolution of the average returns when we use the cost of goods sold (COGS) as the variable input. The fourth figure shows the evolution of the average returns to scale when we adopt a translog specification for the production function.

Overall, these robustness exercises show qualitative patterns that are similar to the benchmark specification presented in Section 2.3. In the 1980s, the average returns to scale are very close to 1 in all specifications, and by 2014, reaches a value between 1.02-1.06. This implies a rise in line with the benchmark specification. Therefore, regardless of the preferred specification, returns to scale in recent

Figure A.2: Alternative Specifications – Robustness 2



Note. The figures above show the results of the decomposition (14) for all four robustness specifications. The first figure shows the evolution of the average returns to scale, the within component, and the between component when we use investment as a proxy variable. The second figure shows the evolution of the average returns to scale, the within component, and the between component when we augment the production function with intangible capital as an additional dynamic input. The third figure shows the evolution of the average returns, the within component, and the between component when we use the cost of goods sold (COGS) as the variable input. The fourth figure shows the evolution of the average returns to scale, the within component, and the between component when we adopt a translog specification for the production function.

years exhibit an increasing trend. Moreover, when we look at the outcome of the decomposition for all the alternative specifications, we can see that, in all cases, the total increase in the average returns to scale is due to the within component. This reinforces the view that returns to scale are increasing

within all sectors of the US economy, regardless of the specification at hand. Taking stocks, we can see from these additional exercises that the main results are a solid feature of the data, suggesting a technological change that is shaping firms' production processes in all sectors of the US economy.

A.3 Selling-Related Activity Robustness

In this section, I explore the extent of the robustness of the results concerning the selling-related expenditure measure. In particular, I check whether using an alternative measure based on the firm-level advertisement expenditure, as reported in Appendix A.1.4, has any effect on the results and the conclusions from the main text. To do so, first, I look at the evolution of this alternative measure over time. Second, I look at the cross-sectoral correlation between this measure and the sector-level measure of returns to scale.

A.3.1 Trend

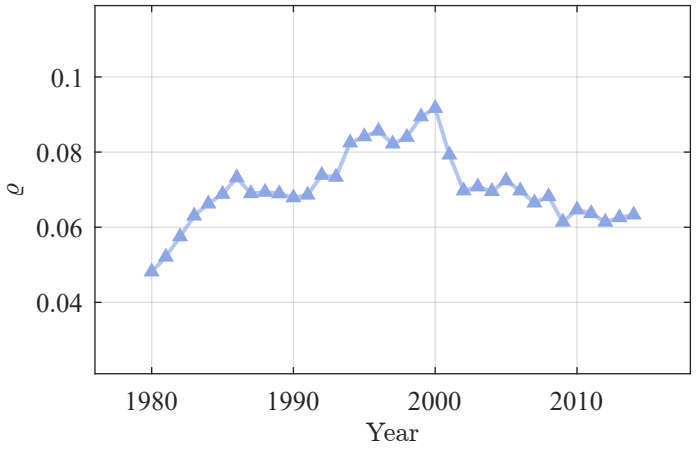
One main prediction of the theory is that the rise in returns to scale implies that the firms spend more on selling-related activities relative to production costs. Therefore, even using the alternative measure of selling-related expenditure, we should observe a rise over time— although we should expect to observe different levels, as explained in Appendix A.1.4. With this alternative specification, the selling ratio becomes:

$$\varrho_{i,t} = \frac{\texttt{XAD}_{i,t}}{\texttt{COGS}_{i,t}}. \tag{58}$$

Figure A.3 shows the evolution of the selling ratio, as defined in equation (58), between 1980 and 2014. This alternative measure of the selling ratio shows a qualitative pattern that is reasonably similar to the benchmark specification. In particular, it increases since 2000 and then declines slightly until the end of the sample, but, overall, it shows an increase over time, as predicted by the theory.

However, the quantitative behavior is very different compared to the benchmark measure. This is not surprising, as this alternative measure (as explained in Appendix A.1.4) reports only the costs related to advertising media (radio, television, newspapers, periodicals) and promotional expenses, but it excludes selling and marketing expenses. Therefore, despite being highly related to the firm's selling activities, it underrepresents the true costs incurred by the firm to attract and retain customers. Overall, the main takeaway, regardless of the preferred measure to calculate the selling ratio, is that in the US, over the last thirty years, there has been a sizeable increase in selling-related activities relative to production costs.

Figure A.3: Average Selling Ratio



Note. The figure shows the evolution of the unweighted average selling ratio, as defined in equation (58), between 1980 and 2014.

A.3.2 Cross-Sectoral Correlation

Here, I show that using this alternative measure of firm-level selling-related expenditure yields a similar sign in the correlation between returns to scale and the measure itself.

Table A.2: Effect of Rising Returns to Scale

	Selling ratio − alternative
Returns to scale	0.086***
	(0.025)
Observations	722
R-squared	0.361
Sector-Time FE	✓

Notes. Fixed effects are at the sector-time level, where the sector is at the 1-digit level. Robust standard errors are in parenthesis. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table A.2 shows the cross-sectional correlation between sector-level returns to scale and the selling ratio. The presence of sector-time-level fixed effects is necessary to ensure that the variation that informs the coefficient estimate does not come from common time trends. The table shows a clear positive correlation between the two variables. Therefore, I can conclude that, regardless of the preferred measure to calculate the selling ratio, the sectors in which firms operate with higher returns to scale are correlated with a higher average selling ratio, as predicted by the theory.

B Model Appendix

B.1 Model Details

In this section, I go through additional details of the model, emphasizing important concepts related to its solution method. Most of the discussion follows the logic developed in Schaal (2017). First, I present a less general contractual environment relative to the one presented in Section 3.3. In this environment, I can solve the model without taking care of the distribution of promised utilities—which is an infinite-dimensional object. This allows the characterization of the real allocation in the economy with standard recursive methods. Second, I comment on how the prices from Section 3.6 implement the same allocation as the one characterized under the less realistic contractual environment.

B.1.1 Alternative Contractual Environment

Here, I assume that contracts are complete, state-contingent, and that there is full commitment on both the customer and firm side. Relative to Section 3.3, the 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 (the submarket in which a firm's customer must search) is a feature of completeness.

B.1.2 Joint Surplus

The additional assumptions embedded in the alternative contractual environment allow the simplification of the problem of the firm. 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:

$$\begin{aligned}\boldsymbol{\mathcal{S}}(z,n) = \max_{\ell,d,n_i',x_i',\tau,x'} \; & nu - w\ell - wf \\ & + \beta\mathbb{E}\Big\{(\delta + (1-\delta)d)n\boldsymbol{\mathcal{U}}' + (1-\delta)(1-d)\Big[\tau n\boldsymbol{\mathcal{U}}' \\ & + (1-\tau)m(\theta(x'))nx' - \left(\frac{wc}{q(\theta(x_i'))} + x_i'\right)n_i' - w\mathcal{K}(n_i';n) + \boldsymbol{\mathcal{S}}(z',n')\Big]\Big\},\end{aligned} \tag{59}$$

subject to:

$$n' = (1-\tau)(1-m(\theta(x')))n + n_i', \tag{60}$$

$$y = e^z F(\ell), \tag{61}$$

$$y = n. \tag{62}$$

The surplus maximization problem characterizes the optimal allocation of physical resources within a firm: the optimal amount of separations, firm-to-firm transitions, the number of new customers, and the decision of whether to exit or not. Because the utility is transferable, transfers between the firms and their 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.

The first element in the surplus maximization problem is the total utility of the customers followed by the wages and operating cost wf paid by the firm. In the next period, conditional on surviving the exit shock δ, the firm chooses whether to exit or not, 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 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 $w\mathcal{K}(n_i';n)$.

B.1.3 Free Entry

Under this different contractual environment, the free entry condition stated in (30) can be restated in terms of joint surplus maximization. I redefine the problem faced by an entering firm of type z as

follows:

$$\boldsymbol{\mathcal{V}}_e(z) = (1-\delta)\max_{x_e}\left[\boldsymbol{\mathcal{S}}(z,n_e) - n_e\left(x_e + \frac{wc}{q(\theta(x_e))}\right)\right]^+. \tag{63}$$

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 firms acquire a measure of customers, $n_e \in \mathbf{R}^+$, and choose a market x_e in which to search, to maximize the joint surplus minus the total advertisement cost $n_e wc/q(\theta(x_e))$ and the total utility $n_e x_e$ 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 first term, $wc/q(\theta(x_e))$, captures the total advertisement cost of acquiring exactly one customer. The second term, x_e, is the utility price that firms offer to their new customers. Firms choose submarkets that minimize the advertisement cost per customer. Define the minimal advertisement cost as:

$$\text{cost} = \min_x \left[x + \frac{wc}{q(\theta(x))}\right]. \tag{64}$$

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

$$\forall x, \quad \theta(x)\left[x + \frac{wc}{q(\theta(x))} - \text{cost}\right] = 0. \tag{65}$$

This condition means that submarkets wether minimize the advertisement cost, $\text{cost} = x + c/q(\theta(x))$, or remain unvisited, $\theta(x) = 0$. In equilibrium, active submarkets will have the same hiring cost, and firms will be indifferent between them. Therefore, the equilibrium market tightness on every active market is:

$$\theta(x) = q^{-1}\left(\frac{wc}{\text{cost} - x}\right). \tag{66}$$

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.

B.1.4 Prices and the Main Model

Once the real allocation of the economy is solved under the contractual environment specified in Section B.1.1, building on the results in Schaal (2017), one can solve equations (28) and (**??**) to construct the prices (equation (29)) that implement the exact same allocation from (59).

B.2 Capital, Marginal Costs, and Labor Share

Here, I discuss a potentially useful extension that allows the meaningful disjoint analysis of both the labor share and markups in the model. To do so, I augment the model with physical capital. For the sake of exposition, I assume that firms do not own their own capital but borrow it in every period. The firm's problem would then be:

$$\begin{aligned}
&\boldsymbol{\mathcal{V}}(z,n,\{\mathcal{C}(j)\}_{j\in[0,n]};w) \\
&\quad= \max_{n_i'(z';w),x_i'(z';w),\{\omega(j)\}_{j\in[0,n]}} \int_0^n p(j)\mathrm{d}j - w\ell - (r+\delta^k)k - wf \\
&\qquad + (1-\delta)\beta\mathbb{E}\left\{ - n_i'\frac{c}{q(\theta(x_i'))} - w\mathcal{K}(n_i';n) + \boldsymbol{\mathcal{V}}(z',n',\{\widehat{\mathcal{C}}(z';w,j)\}_{j\in[0,n']};w)\right\}^+,
\end{aligned} \tag{67}$$

subject to:

$$n'(z';w) = \int_0^n (1-\tau(z';w,j))(1-m(\theta(x'(z';w,j))))\mathrm{d}j + n_i'(z';w), \tag{68}$$

$$\hat{\mathcal{C}}(z';w,j) = \begin{cases} \mathcal{C}(z';w,j) & \text{for} \quad j' \in [0, n'(z';w) - n_i'(z';w)] \quad \text{and} \quad j' = \Phi(z';w,j), \\ x_i(z',w) & \text{for} \quad j' \in [n'(z';w) - n_i'(z';w), n'(z';w)], \end{cases} \tag{69}$$

$$y = e^z k^{\alpha\omega}\ell^{(1-\alpha)\omega}, \tag{70}$$

$$y = n, \tag{71}$$

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

The firm now borrows the capital at a rental rate given by $r+\delta^k$, where r is the interest rate, and δ^k is the depreciation of physical capital. Therefore, the production function (70) takes both capital and labor as inputs. Moreover, now the production functions directly distinguish the output elasticity to labor from the returns to scale, which are given by ω. Therefore, the marginal product in this *augmented* economy is given by:

$$\mathcal{MC} = \left(\frac{1}{\omega}\right)\left(\frac{1}{\alpha}\right)^{\alpha}\left(\frac{1}{e^z}\right)^{\frac{1}{\omega}} n^{\frac{1-\omega}{\omega}}(r+\delta^k)^{\alpha}w^{1-\alpha}. \tag{72}$$

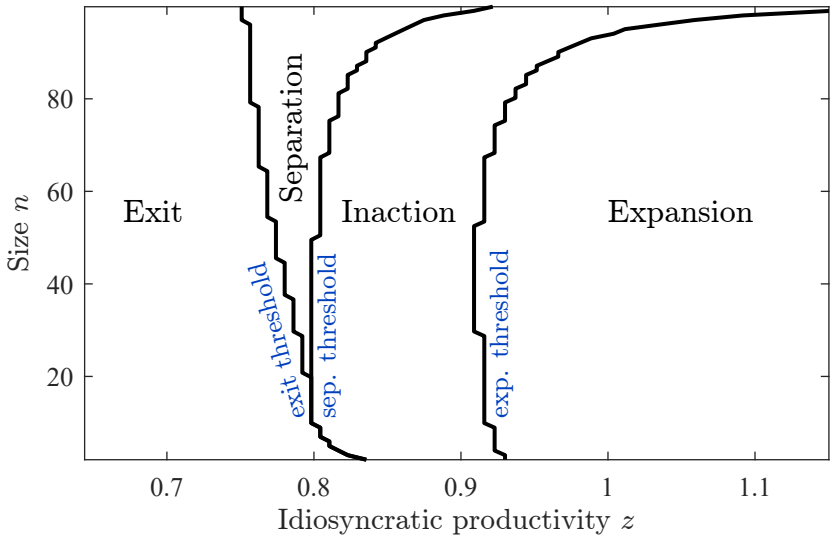
As can be seen from the equation, the marginal cost is still decreasing in the returns to scale; that is, it declines in ω. Therefore, the main mechanism described in the main text is preserved in this case as well. Moreover, because the output elasticity to labor is now governed by a different parameter relative to the one that governs the returns to scale, one can accommodate both an increase in returns to scale and a decline in the labor share. This is indeed consistent with the empirical evidence presented in Section 2.3 and in Chiavari and Goraya (2021). Taking stocks, this model extension should be able to obtain both an increase in markups—as in the main text—and a quantitative-relevant decline in the labor share.

B.3 Additional Validation Exercises

B.3.1 Customer Base Policy at the Firm-Level

Firms, in the model, can use various margins—acquisitions, separations, or exit—to adjust employment. I examine here how the decision of firms to use these margins varies as a function of their individual characteristics (z, n) at the beginning of a period.

Figure B.4: Firms' Action Threshold in the Space (n, z)



Note. The optimal policies depicted in this figure correspond to the baseline calibration. The areas corresponding to the different margins of adjustment are distinct and do not overlap. Notice that customer acquisitions and separations never occur at the same time because it is more costly for firms to acquire new customers than to retain the current ones.

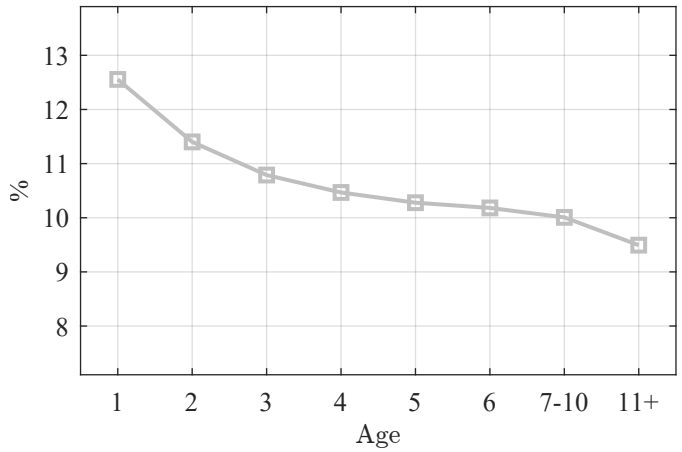
Figure B.4 displays the optimal policy of firms as it appears in the baseline calibration. As expected, customer acquisitions take place in small productive firms, whose marginal value of adding customers is high, while separations occur in unproductive firms. Interestingly, because search frictions show up in the surplus (59) as a linear advertisement cost, cost $= wc/q(\theta(x_i)) + x_i$, a wedge

appears in the adjustment cost faced by firms at $n' = n$. More specifically, separating from a customer earns a value of $\mathcal{U}$ to the customer-firm group, while acquiring new customers incurs the above cost, strictly greater than the value of being an unattached customer in equilibrium. Arising from this kink in adjustment costs, a band of inaction emerges between two thresholds: an expansion threshold, and a separation threshold. Whenever a firm falls in the expansion region, its optimal strategy consists of acquiring new customers until it slowly reaches the expansion threshold—a point at which the marginal value of adding a customer equals the overall cost of acquiring extra customers. Similarly, whenever a firm finds itself in the separation region, its optimal decision is to separate from its customers until it slowly reaches the separation threshold. There, the marginal value of a customer equals the marginal value of separation. The presence of an inaction region implies the existence of a nonnegligible mass of firms that do not adjust their customer base within a period. Exit takes place in unproductive firms. Indeed, due to the presence of a fixed operating cost wf, the decision to exit mostly affects low productivity and low customer firms, as their current production and expected future surplus fall short of the total operating costs.

B.3.2 Additional Life Cycle and Cross-Section Implications

Another implication of the model is that firms with higher productivity and customers are less likely to exit the market; therefore, older firms are also less likely to exit. This feature of the model can be seen from Figure B.4, which shows the exit threshold implied by the baseline calibration. Clearly, the exit region, conditional on having low productivity, is wider when firms have fewer customers than when they have many.

Figure B.5: Exit Rate by Age



Note. The figure shows the exit rate by age group.

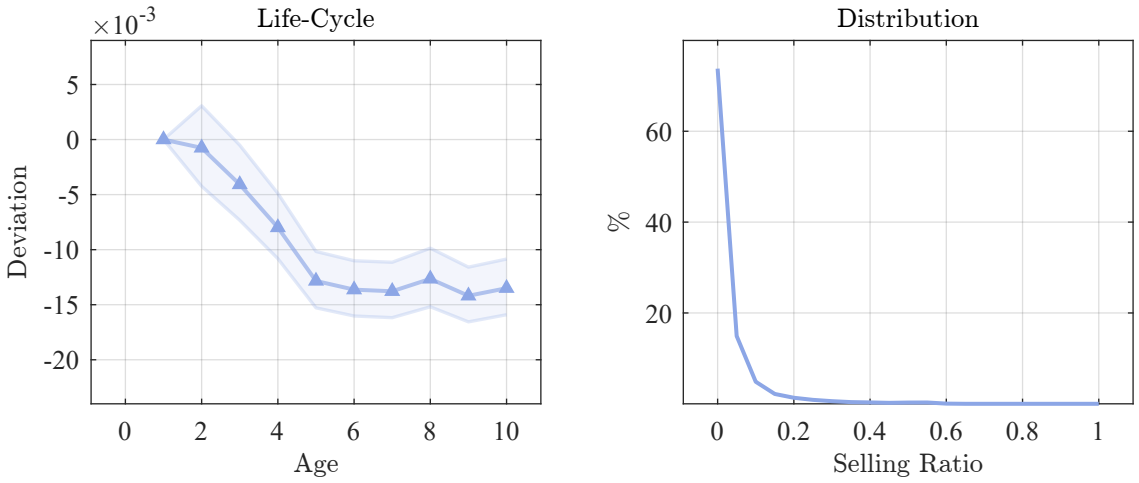
Figure B.5 shows the exit rate for different age groups. As expected, the model produces a neg-

ative correlation between the exit rate and age, meaning that, on average, older firms are less likely to exit the market than younger ones. In the model, this happens because the bigger a firm is, the higher the demand it faces, and hence, the higher its ability to pay its fixed costs. This is an important prediction of the model, as this negative correlation is an empirical finding documented in many empirical papers, such as Haltiwanger, Jarmin, and Miranda (2013).

B.3.3 Robustness on Selling Ratio Implications

In this section, I test the robustness of the patterns documented in Section 4.3.2 regarding the selling ratio. This is particularly important, as already explained in Appendix A.1.4, because Compustat does not offer any ideal measure of selling-related expenditure at the firm level. Therefore, I review the life cycle and distributional patterns of the selling ratio using the alternative measure of selling expenditure defined in Section A.1.4.

Figure B.6: Selling Ratio Robustness



Note. The figure on the left shows the estimated age profile of the selling ratio from equation (40) together with the 90% confidence interval. The figure on the right shows the distribution of the selling ratio. The time frame is 1977-1985.

Figure B.6 shows the results of this robustness exercise. Overall, the main patterns highlighted with the benchmark measure are robust to alternative definitions of selling expenditures. The life-cycle profile of the selling ratio is very similar, aside from the obvious level difference, to the one obtained with the benchmark measure. In the data, firms have a high selling ratio when they are young, which declines with their age.

Moreover, the selling ratio distribution with the alternative measure is very similar to the one obtained with the benchmark measure. In particular, both distributions are right-skewed with a long right tail. Both graphs show that the model predictions regarding firm-level selling expenditures (relative to production costs) are a robust feature of the microdata, regardless of the measure adopted

in the data for this ratio.

B.3.4 Prices and Customers Implications

Prices are one of the main tools that firms have to attract, or retain, customers. In the model, firms that want to grow will charge lower prices to attract and retain customers, and vice versa, firms that are already big will charge higher prices to extract value from their existing customers. Moreover, in the model, firms can discriminate across different customers, as explained in Section 3.6. Therefore, the model has two main sources of price dispersion: first, different firms charge different average prices, and second, within the same firm, customers are also charged different prices. Finally, it is worth emphasizing that the model has clear predictions on the customer side as well. Customers, as previously emphasized, will move from firms charging higher prices to firms charging lower prices. Therefore, the model produces an endogenous turnover over customers in equilibrium.

To look at the price dispersion generated by the model, I compare the standard deviation of the price distribution in the model with the one reported by Kaplan and Menzio (2015). This is particularly sensible, as they look at customer-level prices within a very narrow geography and product category, which maps very close to the model setup where output is identical and homogeneous. The model produces a standard deviation of 2.1e-4, which explains approximately 6% of what is observed in the data by Kaplan and Menzio (2015). This is, of course, only a partial success, but should not come as a surprise because it is well known from the work of Hornstein, Krusell, and Violante (2011) that this class of models struggles in generating the empirically observed dispersion in prices.

Finally, the model produces an endogenous average customer turnover rate of around 11% a year. This is in the range of the estimates from the previous literature. In particular, Gourio and Rudanko (2014) find a customer depreciation rate of 0.15.[63] Hence, the model is within the range found by the literature.

B.3.5 Size and Markups

In the data, firms that have bigger sales within a sector tend to have also higher markups; for instance, this has been documented in India by De Loecker, Goldberg, Khandelwal, and Pavcnik (2016). Here, I look at this prediction in the Compustat data and the model. To do so, I run the following regression specification:

$$\log \mu_{it} = \alpha + \beta_1 \log s_{it} + \beta_2 \log s_{it}^2 + \phi_{st} + \varepsilon_{it}, \tag{73}$$

[63] Significant customer inertia has also been documented empirically by Dubé, Hitsch, and Rossi (2010) and Bronnenberg, Dubé, and Gentzkow (2012).

where $\log \mu_{it}$ is the log-markup, $\log s_{it}$ is the log-sale, and ϕ_{st} is the sector-time fixed effect. I allow for a quadratic specification to permit a nonlinear relation between the two variables.

The regression estimates a positive relation, both in the model and in the data between the log-sale and log-markups. In particular, in the model, the regression estimates a $\beta_1 = 0.43$ and a $\beta_2 = -0.06$, whereas, in the data, the regression estimates a $\beta_1 = 0.30$ and a $\beta_2 = -0.01$. All coefficients are statistically significant, and the time frame is 1977-1985.

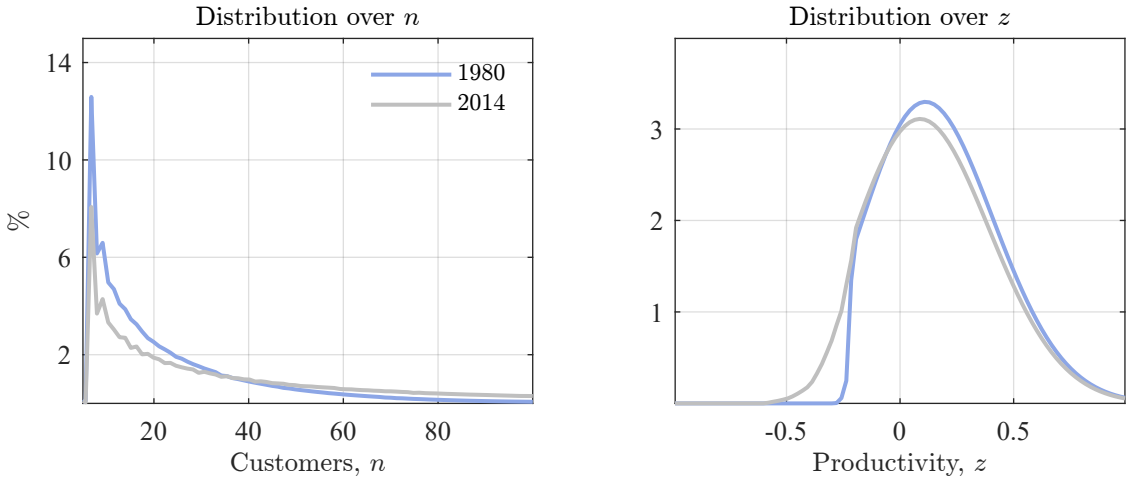
The model's estimates are close to the ones from the data. This is an important result as, in the model, these elasticities have not been a target in the calibration strategy. The model is hence able to replicate moments from the joint distribution of firms' size and markups. This is the case because, in the model, the biggest firms are the most productive, and hence, the ones that face a lower marginal cost of production. In turn, this implies that they are the ones that charge the lowest prices, and hence, are the ones that face a more inelastic demand, which allows them to charge the highest markups.

B.4 Additional Quantitative Exercises

B.4.1 Evolution of the Firms' Distribution

The explored rise in returns to scale has direct implications for the distribution of firms across customers and productivity levels, as explained in Section 5.1. In particular, a rise in returns to scale (i) gives rise to some big firms that, exploiting their scale economies, can attract many customers and to a lot of small firms with instead few customers facing the competition of these big firms; (ii) lower the selection process in the economy, implying that more firms with lower productivity are indeed able to operate in the new equilibrium.

Figure B.7: Firms' Distribution Across Customers and Productivities



Note. The figure on the left shows the distribution of firms across customers, n. The figure on the right shows the distribution of firms across the productivity levels.

Figure B.7 shows the distribution of firms across customers and productivity levels in the 1980 and 2014 steady states. The figure on the right shows the distribution of firms across customers; it shows an increase in its right-skewness and a fattening of the right tail. This is the outcome of the presence—in equilibrium—of big firms that can exploit their scale economies to attract new customers. The fact that the distribution is more right-skewed speaks directly to the literature emphasizing the rise of *superstar* firms (see, for example, Autor, Dorn, Katz, Patterson, and Van Reenen (2020)).

The figure on the left shows the distribution of firms across productivity levels; it shows lower selection. In particular, in the new steady state, there is a fattening of the left tail, which is the outcome of the presence of new firms that, exploiting their scale economies, can operate even after adverse productivity shocks.

B.4.2 Aggregate Output and Welfare

In this final section of the Appendix, I look at the implications of the rise in returns to scale for aggregate output and welfare. To study the effect of this technological change in aggregate output and to highlight which factors are behind its changes, I present the following decomposition of its rise over time:

$$\Delta \log Y_t = \Delta \log \mathcal{Z}_t + \Delta \log L_t/m_t + \Delta \log m_t, \tag{74}$$

where $Y = \int_i y_i di$ is the aggregate output, $\mathcal{Z} = \int_i (y_i/\ell_i)(\ell_i/L)di$ is the aggregate productivity, L is the total labor, and m is the mass of firms. This decomposition helps us understand when the aggregate output changes because of a change in (i) the aggregate productivity, or (ii) the average firm size, or (iii) the mass of firms in the economy.

Table B.3: Evolution of Aggregate Output

	Productivity $\log \mathcal{Z}$	Avg. Firm Size $\log L/m$	Mass of Firms $\log m$	Output $\log Y$
$100 \times \Delta_{2014-1980}$	-28.61	45.63	-15.23	1.79

Note. This table shows the percentage change in the aggregate output and its components, as highlighted in equation (74) between the 1980 and 2014 steady states. The first column reports the percentage change in the aggregate productivity, the second column reports the percentage change in the average firm size, the third column reports the change in the mass of firms, and the fourth column reports the percentage change in the aggregate output. Notice that columns one to three must sum up to column four by construction.

Table B.3 reports the results from the decomposition highlighted in equation (74). In the model,

after a 5% rise in returns to scale, the output increases by almost to 2% relative to the trend.[64] However, this moderate rise in the aggregate output masks sizeable changes in its different components: in particular, I observe a decline in aggregate productivity of approximately 28%, a rise in the average firm size of approximately 45%, and a decline in the mass of firms of approximately 15%.

The decline in labor productivity is the outcome of a rich set of forces. On the one hand, a rise in returns to scale, all else being equal, increases the firm-level average product of labor, y_i/ℓ_i; on the other hand, it weakens the selection process in the economy, allowing less productive firms to operate. Quantitatively this second force dominates and produces the decline in the aggregate productivity documented above.[65] The weakening of the selection process also produces the decline in the mass of firms, as shown in Table B.3. As explained in Section 5.1, with lower selection, entry rates and reallocation rated decline, leading to a steady state with fewer firms.

The rise in the average firm size follows from similar forces as the one outlined above: the returns to scale rise increases the firm size and lowers the selection—that is, it allows smaller and less productive firms to operate. However, in this case, the first effect dominates. Overall, the rise in the average firm size dominates the other two factors, translating into an aggregate output rise.

I now turn my attention to the evolution of aggregate welfare. In this economy, aggregate welfare is the representative household utility. Therefore, the change in welfare measured in consumption-equivalent terms is given by:

$$U(C_{1980}(1+\lambda), L_{1980}) = U(C_{2014}, L_{2014}), \tag{75}$$

where λ measures how much more (or less) the consumption in percentage terms makes the representative household indifferent between the 1980 and 2014 steady states. Given the specific functional form of the representative household preferences, λ is given by:

$$\lambda = \frac{C_{2014} - \vartheta\left(1+1/\psi\right)^{-1}\left(L_{2014}^{1+1/\psi} - L_{1980}^{1+1/\psi}\right)}{C_{1980}} - 1. \tag{76}$$

I find that welfare is approximately 37% below the trend. This decline is due to (i) the lower selection and (ii) higher firm-level selling-related expenditure. Lower selection translates into lower average productivity, which, together with the fact that the average firm becomes bigger, implies

[64] It is noteworthy to acknowledge that we cannot compare the two steady states, as we should view the model as a detrended version of an underlying framework with balanced growth. Therefore, this 2% output rise is an increase relative to a counterfactual experiment in which the output would have only increased due to balance growth, at a 3% rate, for example.

[65] Interpreting the model outlined in this paper as a detrended version of a model featuring balanced growth, we can think of the decline in the aggregate productivity as the model counterpart to the facts highlighted by Fernald (2015). This proves that the technological change documented in the paper can be consistent with the recent US productivity decline. Of course, one should exercise caution with such an interpretation. The model has not been designed to capture the growth phenomena fully, and thus does not allow for a straightforward mapping with the data in this aspect.

that the representative household must supply additional labor to sustain production. Higher firm-level selling-related expenditure, devoted to firm size expansion, is a deadweight loss that must be financed by the representative household with additional labor. These two forces together increase labor, and hence, labor disutility, which being convex, dominates the moderate linear increase in utility from consumption due to the rise in output.

I conclude with a few remarks related to the results on aggregate welfare. First, this does not imply that welfare is lower relative to the 1980s, but only that it is below the trend due to this technological change. To see this, we can compute the level of welfare in 2014, assuming that the economy has grown by 3% a year. In this case, welfare in 2014 would be about 27% higher than in 1980.[66] Second, in the model, consumer welfare, as measured by aggregate consumption and aggregate welfare, moves in the opposite direction. This fact illustrates the tension associated with the common practice of antitrust authorities of using aggregate consumer welfare as a shortcut for the overall welfare.[67]

[66] To see this, consider that absent any change, assuming a 3% growth, the aggregate welfare increased by 2014, which is given by $\log(C_{1980}(1.03)^{34}) - \log(C_{1980})$. Hence, the counterfactual level of welfare after the rise in returns to scale is $(1 - 0.37) \times (\log(C_{1980}(1.03)^{34}) - \log(C_{1980}))$.

[67] The inability of the consumer welfare paradigm to fully capture stakeholders' interests has recently been a highly debated topic among antitrust scholars (Hovenkamp (2019, 2020a,b) and Marinescu and Hovenkamp (2019)).

References

ACKERBERG, D. A., K. CAVES, AND G. FRAZER (2015): "Identification properties of recent production function estimators," *Econometrica*, 83, 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*.

AGHION, P., A. BERGEAUD, T. BOPPART, P. J. KLENOW, AND H. LI (2019): "A theory of falling growth and rising rents," *NBER Working Paper*.

AGRAWAL, A., J. GANS, AND A. GOLDFARB (2018): *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Press.

AKCIGIT, U. AND S. T. ATES (2021): "Ten facts on declining business dynamism and lessons from endogenous growth theory," *American Economic Journal: Macroeconomics*, 13, 257–98.

ALATI, A. (2021): "Initial aggregate conditions and heterogeneity infirm-level markups," *Working Paper*.

ASKER, J., A. COLLARD-WEXLER, AND J. DE LOECKER (2014): "Dynamic inputs and resource (mis) allocation," *Journal of Political Economy*, 122, 1013–1063.

ATKESON, A. AND A. BURSTEIN (2008): "Pricing-to-market, trade costs, and international relative prices," *American Economic Review*, 98, 1998–2031.

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, 645–709.

BEGENAU, J., M. FARBOODI, AND L. VELDKAMP (2018): "Big data in finance and the growth of large firms," *Journal of Monetary Economics*, 97, 71–87.

BLOOM, N., L. GARICANO, R. SADUN, AND J. VAN REENEN (2014): "The distinct effects of information technology and communication technology on firm organization," *Management Science*, 60, 2859–2885.

BORNSTEIN, G. (2018): "Entry and profits in an aging economy: the role of consumer inertia," *Working Paper*.

BRONNENBERG, B. J., J.-P. H. DUBÉ, AND M. GENTZKOW (2012): "The evolution of brand preferences: evidence from consumer migration," *American Economic Review*, 102, 2472–2508.

BURDETT, K. AND M. G. COLES (1997): "Steady state price distributions in a noisy search equilibrium," *Journal of Economic Theory*, 72, 1–32.

BURDETT, K. AND K. L. JUDD (1983): "Equilibrium price dispersion," *Econometrica: Journal of the Econometric Society*, 955–969.

BURDETT, K. AND G. MENZIO (2018): "The (q, s, s) pricing rule," *The Review of Economic Studies*, 85, 892–928.

CABRAL, L. AND J. MATA (2003): "On the evolution of the firm size distribution: facts and theory," *The American Economic Review*, 93, 1075–1090.

CARRIERE-SWALLOW, M. Y. AND M. V. HAKSAR (2019): *The economics and implications of data: an integrated perspective*, International Monetary Fund.

CHETTY, R., A. GUREN, D. MANOLI, AND A. WEBER (2011): "Are micro and macro labor supply elasticities consistent? A review of evidence on the intensive and extensive margins," *American Economic Review*, 101, 471–75.

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.

CROUZET, N. AND J. C. EBERLY (2019): "Intangible capital and the investment-q relation," *Proceedings of the 2018 Jackson Hole Symposium*, 87–148.

DAVIS, S. J., J. HALTIWANGER, R. JARMIN, J. MIRANDA, C. FOOTE, AND E. NAGYPAL (2006): "Volatility and dispersion in business growth rates: publicly traded versus privately held firms," *NBER macroeconomics annual*, 21, 107–179.

DE LOECKER, J., J. EECKHOUT, AND S. MONGEY (2021): "Quantifying market power and business dynamism in the macroeconomy," *NBER Working Paper*.

DE LOECKER, J., J. EECKHOUT, AND G. UNGER (2020): "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 135, 561–644.

DE LOECKER, J., P. K. GOLDBERG, A. K. KHANDELWAL, AND N. PAVCNIK (2016): "Prices, markups, and trade reform," *Econometrica*, 84, 445–510.

DE LOECKER, J. AND P. T. SCOTT (2016): "Estimating market power evidence from the US brewing industry," *NBER Working Paper*.

DE LOECKER, J. AND F. WARZYNSKI (2012): "Markups and firm-level export status," *American Economic Review*, 102, 2437–71.

DE RIDDER, M. (2019): "Market power and innovation in the intangible economy," *Working Paper*.

DECKER, R., J. HALTIWANGER, R. JARMIN, AND J. MIRANDA (2014): "The role of entrepreneurship in US job

creation and economic dynamism," *Journal of Economic Perspectives*, 28, 3–24.
DECKER, R. A., J. HALTIWANGER, R. S. JARMIN, AND J. MIRANDA (2016): "Declining business dynamism: what we know and the way forward," *American Economic Review*, 106, 203–07.
——— (2020): "Changing business dynamism and productivity: shocks versus responsiveness," *American Economic Review*, 110, 3952–90.
DINLERSOZ, E. M. AND M. YORUKOGLU (2012): "Information and industry dynamics," *American Economic Review*, 102, 884–913.
DIXIT, A. K. AND J. E. STIGLITZ (1977): "Monopolistic competition and optimum product diversity," *The American Economic Review*, 67, 297–308.
DUBÉ, J.-P., G. J. HITSCH, AND P. E. ROSSI (2010): "State dependence and alternative explanations for consumer inertia," *The RAND Journal of Economics*, 41, 417–445.
DUNNE, T., M. J. ROBERTS, AND L. SAMUELSON (1989): "The growth and failure of US manufacturing plants," *The Quarterly Journal of Economics*, 104, 671–698.
EATON, J., M. ESLAVA, M. KUGLER, AND J. R. TYBOUT (2009): *8. Export dynamics in colombia: firm-level evidence*, Harvard University Press.
EDMOND, C., V. MIDRIGAN, AND D. Y. XU (2018): "How costly are markups?" *NBER Working Paper*.
EECKHOUT, J. (2021): *The profit paradox: how thriving firms threaten the future of work*, Princeton University Press.
EINAV, L., P. J. KLENOW, J. D. LEVIN, AND R. MURCIANO-GOROFF (2020): "Customers and retail growth," *Working Paper*.
EWENS, M., R. H. PETERS, AND S. WANG (2019): "Acquisition prices and the measurement of intangible capital," *NBER Working Paper*.
FERNALD, J. G. (2015): "Productivity and potential output before, during, and after the great recession," *NBER macroeconomics annual*, 29, 1–51.
FOSTER, L., J. HALTIWANGER, AND C. SYVERSON (2008): "Reallocation, firm turnover, and efficiency: selection on productivity or profitability?" *American Economic Review*, 98, 394–425.
GANDHI, A., S. NAVARRO, AND D. A. RIVERS (2020): "On the identification of gross output production functions," *Journal of Political Economy*, 128, 2973–3016.
GAO, W. AND M. KEHRIG (2017): "Returns to scale, productivity and competition: empirical evidence from US manufacturing and construction establishments," *Working Paper*.
GOLDFARB, A. AND D. TREFLER (2018): "AI and international trade," *NBER Working Paper*.
GOURIO, F. AND L. RUDANKO (2014): "Customer capital," *Review of Economic Studies*, 81, 1102–1136.
GRASSI, B. (2017): "I-O in IO: size, industrial organization, and the input-output network make a firm structurally important," *Working Paper*.
GRULLON, G., Y. LARKIN, AND R. MICHAELY (2019): "Are US industries becoming more concentrated?" *Review of Finance*, 23, 697–743.
HALL, R. E. (1988): "The relation between price and marginal cost in US industry," *Journal of Political Economy*, 96, 921–947.
HALTIWANGER, J., R. S. JARMIN, AND J. MIRANDA (2013): "Who creates jobs? Small versus large versus young," *Review of Economics and Statistics*, 95, 347–361.
HASKEL, J. AND S. WESTLAKE (2018): *Capitalism without capital: the rise of the intangible economy*, Princeton University Press.
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*.
HOPENHAYN, H. A. (1992): "Entry, exit, and firm dynamics in long run equilibrium," *Econometrica: Journal of the Econometric Society*, 1127–1150.
HORNSTEIN, A., P. KRUSELL, AND G. L. VIOLANTE (2011): "Frictional wage dispersion in search models: a quantitative assessment," *American Economic Review*, 101, 2873–98.
HOVENKAMP, H. (2019): "Is antitrust's consumer welfare principle imperiled?" *J. Corp. L.*, 45, 65.
——— (2020a): "Antitrust: what counts as consumer welfare," *Working Paper*.
——— (2020b): "On the meaning of antitrust's consumer welfare principle," *Revue Concurrentialiste (Jan. 17, 2020), U of Penn, Inst for Law & Econ Research Paper*.
HSIEH, C.-T. AND E. ROSSI-HANSBERG (2019): "The industrial revolution in services," *NBER Working Papers*.
JONES, C. I. AND C. TONETTI (2020): "Nonrivalry and the economics of data," *American Economic Review*, 110, 2819–58.
KAPLAN, G. AND G. MENZIO (2015): "The morphology of price dispersion," *International Economic Review*, 56, 1165–1206.
KARAHAN, F., B. PUGSLEY, AND A. ŞAHIN (2019): "Demographic origins of the startup deficit," *NBER Working Paper*.

KEHRIG, M. AND N. VINCENT (2021): "The micro-level anatomy of the labor share decline," *The Quarterly Journal of Economics*, 136, 1031–1087.
KHAN, L. M. (2016): "Amazon's antitrust paradox," *Yale lJ*, 126, 710.
KIMBALL, M. S. (1995): "The quantitative analytics of the basic neomonetarist model," *Journal of Money, Credit and Banking*, 27, 1241–1277.
KORINEK, A., D. X. NG, AND J. HOPKINS (2018): "Digitization and the macro-economics of superstars," *Working Paper*.
KOST, K., J. PEARCE, AND L. WU (2019): "Market power through the lens of trademarks," *Working Paper*.
LASHKARI, D., A. BAUER, AND J. BOUSSARD (2021): "Information technology and returns to scale," *Working Paper*.
LEVINSOHN, J. AND A. PETRIN (2003): "Estimating production functions using inputs to control for unobservables," *The Review of Economic Studies*, 70, 317–341.
LIU, E., A. MIAN, AND A. SUFI (2020): "Low interest rates, market power, and productivity growth," *Econometrica forthcoming*.
MARINESCU, I. AND H. HOVENKAMP (2019): "Anticompetitive mergers in labor markets," *Ind. LJ*, 94, 1031.
MARTINEZ, J. (2018): "Automation, growth and factor shares," *Working Paper*.
MENZIO, G. AND S. SHI (2010): "Block recursive equilibria for stochastic models of search on the job," *Journal of Economic Theory*, 145, 1453–1494.
——— (2011): "Efficient search on the job and the business cycle," *Journal of Political Economy*, 119, 468–510.
MENZIO, G. AND N. TRACHTER (2015): "Equilibrium price dispersion with sequential search," *Journal of Economic Theory*, 160, 188–215.
——— (2018): "Equilibrium price dispersion across and within stores," *Review of Economic Dynamics*, 28, 205–220.
MOEN, E. R. (1997): "Competitive search equilibrium," *Journal of Political Economy*, 105, 385–411.
MORLACCO, M. AND D. ZEKE (2021): "Monetary policy, customer capital, and market power," *Journal of Monetary Economics*.
NEWMAN, N. (2014): "Search, antitrust, and the economics of the control of user data," *Yale J. on Reg.*, 31, 401.
OLLEY, G. S. AND A. PAKES (1996): "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 65, 1263–1297.
OLMSTEAD-RUMSEY, J. (2019): "Market concentration and the productivity slowdown," *Working Paper*.
PACIELLO, L., A. POZZI, AND N. TRACHTER (2019): "Price dynamics with customer markets," *International Economic Review*, 60, 413–446.
PAKES, A. (1994): *Dynamic structural models, problems and prospects: mixed continuous discrete controls and market interactions*, Cambridge University Press, vol. 2 of *Econometric Society Monographs*, 171–274.
PETERS, M. (2020): "Heterogeneous markups, growth, and endogenous misallocation," *Econometrica*, 88, 2037–2073.
PETERS, M. AND C. WALSH (2019): "Declining dynamism, increasing markups and missing growth: the role of the labor force," *Working Paper*.
PETERS, R. H. AND L. A. TAYLOR (2017): "Intangible capital and the investment-q relation," *Journal of Financial Economics*, 123, 251–272.
PHILIPPON, T. (2019): *The great reversal*, Harvard University Press.
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, 987–1011.
ROLDAN-BLANCO, P. AND S. GILBUKH (2020): "Firm dynamics and pricing under customer capital accumulation," *Journal of Monetary Economics*.
RUHL, K. J. AND J. L. WILLIS (2008): "Convexities, nonconvexities, and firm export behavior," in *Manuscript, Midwest Macro Conference, Philadelphia*.
SCHAAL, E. (2017): "Uncertainty and unemployment," *Econometrica*, 85, 1675–1721.
SYVERSON, C. (2004): "Market structure and productivity: a concrete example," *Journal of Political Economy*, 112, 1181–1222.
WEISS, J. (2019): "Intangible investment and market concentration," *Working Paper*.
ZHANG, L. (2019): "Intangible-investment-specific technical change, concentration and labor share," *Working Paper*.

UniCredit Foundation
Piazza Gae Aulenti, 3
UniCredit Tower A
20154 Milan
Italy

Giannantonio De Roni – *Secretary General*
e-mail: giannantonio.deroni@unicredit.eu

Annalisa Aleati - *Scientific Director*
e-mail: annalisa.aleati@unicredit.eu

Info at:
www.unicreditfoundation.org


UniCredit Foundation