Scalable versus Productive Technologies*

Joachim Hubmer ⓡ Mons Chan ⓡ Serdar Ozkan ⓡ Sergio Salgado ⓡ Guangbin Hong

First Version: July 11, 2024—This Version: May 6, 2025

Abstract

Are larger firms more productive, more scalable, or both? We use firm-level panel data from thirteen countries and employ a broad set of methods to estimate factor elasticities—capturing returns to scale (RTS)—and total factor productivity (TFP). We find substantial RTS heterogeneity within industries, with larger firms exhibiting higher RTS driven by greater intermediate input elasticities. TFP, by contrast, rises with firm size only up to the top decile before declining. Incorporating RTS heterogeneity into a standard model of entrepreneurship more than doubles the efficiency losses from financial frictions compared with a conventional calibration with only TFP differences.

Keywords: Production function heterogeneity, returns to scale, firm-size distribution, misallocation.
JEL codes: E22, E23, D24, L11.

---

*Hubmer: University of Pennsylvania; jhubmer@sas.upenn.edu; Chan: Queen's University; mons.chan@queensu.ca; Ozkan: Federal Reserve Bank of St. Louis, University of Toronto; serdar.ozkan@gmail.com, www.serdarozkan.me; Salgado: The Wharton School-University of Pennsylvania; ssalgado@wharton.upenn.edu; Hong: University of Chicago; ghong7@uchicago.edu.

The "ⓡ" symbol indicates certified random order for authors' names. For helpful comments, we thank Paco Buera, Russ Cooper, Joel David, Greg Kaplan, Pete Klenow, Matthias Mertens, Simon Mongey, Sara Moreira, Ezra Oberfield, Sergio Ocampo, Guillermo Ordonez, Diego Restuccia, Maarten De Ridder, Richard Rogerson, Benjamin Schoefer, Yongseok Shin, Venky Venkateswaran, and seminar participants at the SED 2023, NBER SI 2024, Barcelona GSE Summer Forum 2024, CMSG 2024, Bonn, Goethe, Miami, Michigan, Montreal, Wisconsin, Wharton, CEMFI, EUI, Chicago Fed, Indiana, Philadelphia Fed, Temple, Rutgers, St. Louis Fed/Wash U, ASU. We thank Zhongdao Wang for his excellent research assistance. Part of this research was conducted at a U.S. Federal Statistical Research Data Center (FSRDC Project #2019) and a Statistics Canada Research Data Center. All results based on U.S. and Canadian administrative data have been reviewed to ensure confidentiality. The views expressed are those of the authors and do not reflect those of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors. Ozkan acknowledges support from the Canadian Social Sciences and Humanities Research Council. Click here for the latest version.

1 Introduction

The large and persistent firm heterogeneity in *total factor productivity (TFP)* has been extensively documented within industries and for different countries and time periods (see Syverson (2011) for an overview). Seminal models, such as Lucas (1978), Hopenhayn (1992), and Melitz (2003), attribute firm heterogeneity within industries primarily to differences in TFP, assuming homogeneous *returns to scale (RTS)* across firms. Building on these ideas, the misallocation literature (pioneered by Restuccia and Rogerson (2008) and Hsieh and Klenow (2009)) quantifies the efficiency costs of distortions measured from differences in the marginal product of inputs, attributing the technological heterogeneity to only TFP variation (a notable exception is David and Venkateswaran (2019)). Further, models of entrepreneurship, such as Cagetti and De Nardi (2006), incorporate decreasing returns to scale technology with heterogeneous TFP to explain differences in rates of return and wealth inequality.

In this paper, we allow for more general heterogeneity in production technologies among firms by focusing on differences in RTS. Using a broad set of estimation methods, we document substantial heterogeneity in production technologies across firms. We then examine whether larger firms have technologies that are more *productive* (high TFP) or more *scalable* (high RTS). Finally, we demonstrate the importance of this distinction by studying the efficiency costs of misallocation due to financial frictions—an issue central to a variety of quantitative questions.

In our main empirical analysis, we use administrative panel data for the universe of incorporated Canadian firms that accounts for over 90% of private business sector output from 2001 to 2019. This dataset provides detailed balance sheet information, including revenues and the total cost of labor, capital, and intermediate inputs. To validate our results, we replicate the analysis for manufacturing plants using U.S. census data as well as for eleven European countries using the Moody's Orbis dataset.

In our benchmark approach, we estimate nonparametric production functions building on Gandhi, Navarro and Rivers (2020) (henceforth GNR), which recovers output elasticities of labor, capital, and intermediate inputs—thus, RTS—along with TFP at the firm-year level. This technique relies on standard assumptions of profit maximization, adjustment costs, and input choice timing. The common nonhomothetic production function is identified from variation in input expenditure shares and the covariance between input and output levels, controlling for the endogeneity of inputs to TFP.

We estimate production functions for each two-digit NAICS industry. Beyond the well-documented TFP heterogeneity, we notably uncover substantial RTS variation among firms within the same industry.[1] The average within-industry difference between the 90^{th} and 10^{th} percentiles (P90-P10) of RTS is 8 percentage points (p.p.). Interpreted as deviations from constant returns to scale, these differences are large.[2] By construction, the heterogeneity in RTS is explained by the dispersion in output elasticities of inputs. The P90-P10 of elasticities is 0.36 for intermediates and labor versus 0.08 for capital. Output elasticities closely track input revenue shares.[3]

Our key finding is that RTS increases with firm revenue (and alternatively with employment or value added), especially for firms above the median. Within industries, the largest 5% of firms have 8 p.p. higher average RTS than those in the bottom 50%. This pattern is entirely accounted for by higher output elasticities of intermediate inputs for larger firms, consistent with Mertens and Schoefer (2024), who emphasize that firms grow by shifting from labor toward intermediate inputs. In contrast, labor and capital elasticities generally decline with firm size, though with some variation across samples and specifications.

Interestingly, TFP increases in firm revenue only up to the 90th percentile, but then flattens out and declines sharply among the largest firms. In contrast, RTS continues to rise—convexly—for the top 10%, indicating that the largest firms are characterized more by their scalability than by higher productivity—challenging the common assumption in the literature. When counterfactually imposing homogeneous RTS in the estimation, TFP increases monotonically with firm revenue, underscoring the importance of allowing for flexible technologies in measuring TFP.

Our GNR approach allows for adjustment frictions and market power for capi-

[1] While a few papers have documented RTS heterogeneity—across U.S. industries (Gao and Kehrig (2017)), firms and countries (Demirer (2020)), or over time (Chiavari (2024))—none examine the joint relation between RTS, TFP, and firm size within industries, which is the focus of our study.

[2] E.g., in an efficient economy with Cobb-Douglas technology, the elasticity of optimal firm output to TFP ($\frac{1}{1-RTS}$) is five times larger for a firm with RTS of 0.98 compared to a firm with RTS of 0.90. They are also quantitatively important for the costs of financial constraints (Section 5).

[3] As expected, the correlation between revenue shares and elasticities is strongest for intermediate inputs, which we treat as a flexible input in our estimation. For labor and capital, correlations remain positive but weaker, potentially reflecting adjustment costs or input market power.

tal and labor inputs but not firm-specific markups or factor-biased productivity. It recovers flexible nonparametric production functions but only a common one within industries. We address these limitations through complementary approaches. First, we apply Demirer (2020)'s method, which allows for heterogeneous markups and factor-augmenting productivity shocks—at the cost of stronger assumptions on the labor input choice and the functional form of the technology. We find an even steeper RTS–size gradient.[4] Second, since our relatively short panel precludes firm-specific production functions, we cluster firms based on inputs and revenue and estimate cluster-specific technologies. This approach yields similar results, and also reveals that the RTS–firm size gradient is driven by variation across, rather than within, firm clusters, thereby indicating that cross-sectional RTS heterogeneity primarily reflects persistent firm characteristics rather than transitory factors.[5] Third, similar results with homogeneous relative factor elasticities but heterogenous RTS confirm that our findings are robust to the flexible functional form in the GNR method.

Our results are remarkably consistent across countries and data sources as well. While our baseline uses data for the Canadian economy, we find a similarly strong RTS–size gradient among U.S. manufacturing plants and across eleven European countries using firm-level Orbis data—the gradient is positive in every single country and of broadly similar magnitude. The gradient becomes even more pronounced when we include intangible capital in measuring firms' inputs.

We also revisit several well-known empirical patterns in firm heterogeneity that were previously explained by TFP differences. We find that high-RTS firms grow faster over the life cycle and are less likely to exit compared to high-TFP firms. Additionally, high-wage firms tend to have higher RTS. Linking firms to their owners, we show that wealthier households own firms with more scalable technologies. These secondary findings highlight the importance of incorporating realistic RTS heterogeneity for a variety of applications, including wage and wealth inequality.

To investigate the quantitative implications of our findings, we incorporate het-

[4] Our results are also robust to using output market shares as proxies for unobserved price elasticities (à la De Loecker *et al.* (2016)) in our implementation of GNR.

[5] To support this interpretation, we show that firm fixed effects explain 75% of RTS variation in our baseline estimates conditional on firm age and size, while an autocovariance analysis attributes just 11% of the variation to transitory shocks versus 39% and 51% to permanent fixed effects and the highly persistent component, respectively.

erogeneous RTS into a standard incomplete markets model of entrepreneurship (e.g., Quadrini (2000); Cagetti and De Nardi (2006)).[6] In the model, agents choose whether to supply stochastic efficiency units of labor or to operate a private business under a stochastic technology that depends not only on a standard idiosyncratic TFP term (z) but also on an idiosyncratic RTS term (η).[7] Entrepreneurs must finance at least a fraction λ of their input expenditures using their own wealth. Our main exercise compares the effects of increasing the financial friction λ on output and productivity in two different economies: the conventional z-economy, where technological heterogeneity stems from TFP alone, and the (η, z)-economy, which incorporates heterogeneity in both RTS and TFP based on our empirical estimates. We calibrate both economies to match key moments such as the firm-size distribution.

We find that in the (η, z)-economy, financial frictions generate over twice the output losses compared to the z-economy. Static misallocation of inputs accounts for the bulk of output losses in both economies and is about twice as large in the (η, z)-economy. To build intuition, we analytically show in a static endowment economy that a given wedge in marginal products leads to larger misallocation when constrained firms have relatively higher RTS—an endogenous feature of our dynamic model. Dynamic effects further exacerbate output losses in (η, z)-economy, due to under accumulation of capital and distortions in the selection into entrepreneurship. Intuitively, a highly productive (high-z) but poor potential entrepreneur can operate profitably at a small scale, making it easier to grow despite the friction. In contrast, a highly scalable (high-η) but less immediately profitable business struggles to outgrow the friction, and the entrepreneur may never enter the market. These results highlight the critical importance of accounting for RTS heterogeneity for a broad set of quantitative questions related to misallocation, including capital taxation (e.g., Guvenen *et al.* (2023); Boar and Midrigan (2022); Gaillard and Wangner (2021)).

[6] To isolate the novel role of RTS heterogeneity, we use this standard framework that abstracts from several empirically relevant features of production, such as intermediate inputs and pre-determined inputs. We show in model extensions that our findings are robust to introducing these richer model features, including those employed in our empirical framework.

[7] To treat RTS symmetrically to TFP, we model η as a highly persistent exogenous process, consistent with our empirical findings. Various microfoundations for RTS differences complement our analysis, including scalable expertise (Argente *et al.* (2024)), scale-dependent IT intensity (Lashkari *et al.* (2024)), choice of managerial inputs (Chen *et al.* (2023)), or the industrial revolution in services (Hsieh and Rossi-Hansberg (2023)).

2 Empirical Methodology

2.1 The Firm's Problem

We first introduce a general form of the firm's production setting. Each of the methods we employ imposes some identifying restrictions on this general model.

Consider firm j in year t that produces output Y_{jt} using capital K_{jt}, labor L_{jt}, and intermediate inputs M_{jt} according to $Y_{jt} = F_j(K_{jt}, L_{jt}, \omega^M_{jt} M_{jt})e^{\nu_{jt}}$. Hicks-neutral productivity, $\nu_{jt} = \omega_{jt} + \varepsilon_{jt}$, is composed of (i) a persistent component, ω_{jt}, which is known to the firm when it makes input decisions in period t, and (ii) a transitory component, ε_{jt} (i.i.d. across firms and time with $\mathbb{E}[\varepsilon_{jt}] = 0$), which is observed after choosing inputs. Changes in these productivity terms may arise from both technology shocks and market demand shifts, while the transitory component may also reflect measurement error in output. Furthermore, ω^M_{jt} captures intermediate-augmenting productivity *relative to labor*, which is persistent over time and is also known to the firm when it makes input decisions. We assume that the persistent productivity components follow a joint exogenous first-order Markov process: $\mathcal{P}_\omega(\omega_{jt}, \omega^M_{jt} \,|\, \mathcal{I}_{jt-1}) = \mathcal{P}_\omega(\omega_{jt}, \omega^M_{jt} \,|\, \omega_{jt-1}, \omega^M_{jt-1})$, where we define $\mathcal{I}_{jt}$ as the information set available to firm j when it makes its decisions in period t.

Inputs that are functions of the previous period's information set, $X_t(\mathcal{I}_{jt-1})$, are defined as *predetermined*. Inputs that are chosen in period t are defined as *variable*. Capital is predetermined and a state variable ($K_{jt} \in \mathcal{I}_{jt}$). Capital may be subject to arbitrary adjustment costs and financial constraints, and we do not need to assume that firms choose investment optimally.

Firms are also subject to arbitrary labor adjustment costs. We assume labor is a *dynamic* input in that it is a variable input and the firm's choice of L_{jt} may depend on its own lagged value, $L_{jt-1} \in \mathcal{I}_{jt}$. We also allow firms to have wage setting power. Similarly to capital, most of our estimation approaches do not require assumptions on the optimality of the labor choice or the nature of wage setting.

Finally, given the capital and labor choices, firms maximize short-run expected profits by selling their output at a price according to a demand function specific to its industry i, $P_{jt} = P^i_t(Y_{jt})$, and buying intermediate inputs in a perfectly competitive market without adjustment costs or other frictions.

2.2 Estimation Methods

We now outline our estimation methods, specifying how each approach imposes distinct identifying assumptions on the general model presented above.

2.2.1 GNR Method

Our main empirical approach builds on the GNR methodology, which estimates a flexible nonparametric gross output production function. This technique provides several advantages. First, it identifies output elasticities for *gross output*, whereas common alternatives (e.g., Ackerberg *et al.* (2015)) typically only identify value-added technologies. As we show below, variation in the output elasticity of intermediate inputs is a key driver of variation in RTS. Second, the nonparametric GNR approach minimizes specification error when estimating both output elasticities and TFP. Third, it allows us to estimate a nonhomothetic production function, where output elasticities and RTS are functions of inputs, thereby varying across firms and over time—crucial for understanding the relationship between firm-level TFP, RTS, and size.

Identifying restrictions. We introduce our implementation of the GNR technique in Appendix A and refer to Gandhi *et al.* (2020) for technical assumptions. Here, we specify the substantive restrictions imposed on the general firm problem: (i) Firms within an industry i have access to a common but flexible nonhomothetic production function, $F_j(.) = F^i(.)$. (ii) Firms are price takers in the output market, $P_{jt} = P^i_t$. (iii) There are no intermediate input-augmenting productivity shocks, $\omega^M_{jt} = 1$.

Identification and intuition. Although GNR provide a rigorous identification proof (to which we refer readers), we focus here on the intuition behind our estimation. Because the intermediate input is flexible (i.e., variable and static), the first-order condition (FOC) from the firm's short-run expected profit maximization implies that the expected intermediate expenditure share equals its output elasticity. Covariation between the (expected) share and input levels then identifies this output elasticity.[8]

[8] Intuitively, if the production function were Cobb-Douglas, the expenditure share would be uncorrelated with input levels, and its output elasticity would be constant (equal to the mean share). This direct relationship (from the FOC) holds under the assumption that firms are price takers in intermediate input markets and do not face adjustment costs when choosing intermediates.

We thus recover the output elasticity of intermediate inputs as a function of input levels via a nonparametric regression of the revenue share of intermediate expenditure on inputs. This regression also identifies the ex-post transitory shock: for two firms with the same input levels, variation in intermediate expenditure shares arises only from differences in ex-post shocks (through unexpected variation in revenues).

With estimates of the intermediate input elasticity and transitory shocks at hand, we remove the effects of intermediates and ex-post shocks from gross output, leaving a residual "value-added" function to estimate in the next step.[9] Because capital is predetermined and labor is subject to adjustment costs and input market power, (unknown) wedges arise between expenditure shares and output elasticities, preventing identification of those elasticities via the FOC approach used in the first step. Therefore, our second-stage estimation follows the proxy-variable literature (Olley and Pakes (1996)) in exploiting Markov timing assumptions on the persistent shock to form GMM conditions. Intuitively, conditional on the previous period's persistent productivity (ω_{jt-1}), covariations between value added and capital and labor (instrumented with its lagged value) inputs identify their output elasticities. Similarly, conditional on inputs and ω_{jt-1}, variation in value added identifies the persistent shocks. Thus, a high-RTS firm is characterized by a high intermediate input expenditure share, a strong correlation between output and capital or labor, or both, whereas a high-TFP firm exhibits greater value added conditional on inputs and their elasticities.

Cobb-Douglas with RTS heterogeneity. The baseline GNR allows for a nonparametric production function. To investigate whether our findings are sensitive to this flexible functional form, as a special case, we impose homogeneous relative output elasticities—within each industry i—while allowing for heterogeneity in RTS: $F_{jt}(K_{jt}, L_{jt}, M_{jt}) = \left(K_{jt}^{\varepsilon_K^i} L_{jt}^{\varepsilon_L^i} M_{jt}^{\varepsilon_M^i}\right)^{\eta_{jt}}$.[10] The other assumptions are the same as above. We follow the same two-step estimation procedure and identify firm-year level η_{jt} in the first stage from variation in intermediate input expenditure shares.

[9] This is a slight abuse of language: for example, our value-added production function does not contain transitory shocks and is derived by removing the contribution of intermediate inputs to output (including their interactions with capital and labor).

[10] In this specification, a normalization is required between the sum of the three output elasticities, $\varepsilon_K + \varepsilon_L + \varepsilon_M$, and η_{jt}. We normalize $\varepsilon_K + \varepsilon_L + \varepsilon_M = 1$, so that η_{jt} denotes RTS.

2.2.2 Clustering firms

In our baseline estimation method based on GNR, we assume that all firms j within an industry i share the same nonparametric production function, $F_j(.) = F^i(.)$. Firms differ in their factor elasticities, and therefore in RTS, because they operate at different points in the input space. Later, we show that these differences are highly persistent, reflecting persistent technological differences across firms. Ideally, we would estimate a separate production function for each firm, but this is infeasible given the short panel. Instead, we group firms with similar features as a practical alternative.

We use two clustering strategies. First, given the importance of input shares in identifying factor elasticities, we apply a k-means algorithm to group firms into 20 clusters based on their average revenue shares of the three inputs and their average within-industry revenue percentile. Second, to capture persistent technology differences across firms with different growth trajectories, we group firms by their maximum revenue percentile attained over their life cycle. Specifically, within each industry, we compute a firm's revenue rank each year and assign it to one of 11 bins based on its highest lifetime rank: 1-10, 11-20, ..., 91-95, and 96-100. Firms with fewer than 10 years of data are excluded to minimize selection bias. Under both clustering approaches, a firm remains assigned to the same cluster throughout its life.

For each clustering approach, we estimate production functions using the nonparametric GNR method, and using a traditional Cobb-Douglas specification:

1. GNR: For firm j in cluster k, the production function becomes $F_j(.) = F^k(.)$, a nonparametric production function estimated at the cluster level.

2. Cobb-Douglas case: For firms j in cluster k, we impose a Cobb-Douglas functional form, $F_j(.) = K_{jt}^{\alpha_k} L_{jt}^{\beta_k} M_{jt}^{\gamma_k}$, where the elasticities α_k, β_k, and γ_k are common within each cluster. Consequently, firms within a given cluster also share a common RTS, determined by the sum $\alpha_k + \beta_k + \gamma_k$.

2.2.3 Demirer Method

Finally, we estimate TFP and RTS using the method developed in Demirer (2020). Unlike our GNR specification, this approach allows for factor-augmenting productivity shocks and firm-level price-setting (markups) in output markets. Yet, it requires several stronger assumptions on production and input choices: (i) Firms j in industry i share a common, weakly homothetic separable production function $F^i(K_j, h^i(L_{jt}, \omega_{jt}^M M_{jt}))$ where h^i is homogeneous in both inputs and ω_{jt}^M is intermediate-augmenting productivity. (ii) Both M_{jt} and L_{jt} are flexible inputs, optimally chosen, and firms are price takers in both input markets. (iii) Firms have price setting power in output markets via a demand function $P_{jt} = P_t^i(Y_{jt})$. We refer readers to Demirer (2020) for additional technical assumptions.

Identification and intuition. To identify the production function with the additional dimension of heterogeneity relative to GNR, Demirer (2020) assumes homothetic separability between capital and a composite of labor and intermediate inputs, treating both as flexible inputs. He adopts a control-variable approach to address endogeneity in the presence of the two unobserved productivity components. First, a control variable for relative factor-augmenting productivity is constructed using capital and the flexible input ratio: under homothetic separability, this ratio is strictly monotonic in ω_{jt}^M, conditional on capital, and independent of Hicks-neutral productivity. Second, a control variable for Hicks-neutral productivity is constructed using capital, intermediate inputs, and the control variable for intermediate-augmenting productivity.

Conditional on these control variables, capital and composite input elasticities are identified from their covariation with output. Finally, since both labor and intermediates are flexible, Demirer exploits the two FOCs to link the ratio of their output elasticities to the ratio of observed expenditure shares, allowing for separate identification of labor and intermediate elasticities.

3 Data and Sample Selection

Our main dataset is the Canadian Employer-Employee Dynamics Database of Statistics Canada (CEEDD), a set of linkable administrative tax files covering the universe of tax-paying Canadian firms and individuals from 2001 to 2019. We obtain balance sheet and income statement information from the National Accounts Longitudinal

Table I – Summary Statistics For the Baseline Sample

Log of	Mean	Median	St.dev	P10	P50	P90	P99
Revenue	13.73	13.54	1.39	12.13	13.54	15.60	17.75
Intermediates	13.18	12.99	1.52	11.41	12.99	15.21	17.46
Wage bill	12.35	12.19	1.30	10.82	12.19	14.07	16.04
Capital stock	11.29	11.26	1.82	9.02	11.26	13.54	15.97

Notes: Table I shows cross-sectional moments of the distributions of log values for revenue, intermediate inputs, wage bill, and capital. All variables are in 2002 Canadian dollars. The total number of observation is 4.3 million firm-years.

Microdata File, which covers all incorporated firms.[11] Revenue and wage bill variables are constructed by Statistics Canada based on corporate tax return line items and are consistent with the national income and product accounts. We construct total tangible capital using the perpetual-inventory method (PIM), starting from the first book value observed in the data, annual tangible capital investment, and amortization. Intermediate inputs are calculated as the sum of operating expenses and costs of goods sold net of capital amortization. All nominal values are deflated to 2002 real Canadian dollars. See Appendix C.1 for further details.

To construct the estimation sample, we start from firm-year observations with nonmissing values for revenue, capital, wage bill, intermediate inputs, and industry code. To ensure reliable PIM capital estimates, we include only observations with at least two prior years of capital data. We further drop observations with outlier factor shares: (i) wage-bill-to-revenue or wage-bill-to-value-added ratios below the 1st or above the 99th percentile; (ii) intermediate-input-to-revenue ratios outside $[0.05, 0.95]$; and (iii) capital-to-revenue ratios above the 99.9th percentile. After sample selection, our dataset comprises 4.3 million firm-year observations and 620,000 firms, with an average of 6.9 observations per firm. Summary statistics are reported in Table I.

U.S. manufacturing sector. As a robustness exercise, we conduct a similar analysis using data from the U.S. Economic Census and the Annual Survey of Manufactures (ASM), widely used in the literature on firm-level productivity in the U.S. (e.g., Foster *et al.* (2001) and Bloom *et al.* (2018)). This dataset contains detailed information on

[11] CEEDD also covers all unincorporated firms—typically small businesses owned by self-employed individuals—accounting for 9.5% of GDP in 2005 (Baldwin and Rispoli (2010)). We exclude unincorporated firms because they do not report capital stock.

over 60,000 manufacturing plants between 1974 and 2019. Unlike our Canadian data, it does not cover the full universe of firms but a representative panel of manufacturing plants, redrawn every five years. We restrict the sample to plants with at least two years of nonmissing data for key variables, resulting in 3.1 million establishment-year observations. Revenue is measured by the total value of shipments. The Census also reports real capital stock (measured using the PIM), total wages of all plant workers, and expenditures on intermediate inputs, all expressed in 2019 U.S. dollars. We discuss additional details in Appendix C.2.

Moody's Orbis dataset. We further complement our analysis using data from eleven European countries including Germany, France, Italy, and Spain. This dataset provides harmonized information on revenues, wage bill, capital stock, and intermediate inputs (all in 2019 prices) for a large sample of private and public firms across industries. For most countries, data coverage spans 2005–2019.[12] See Appendix C.3 for additional details.

4 Empirical Results

We apply our baseline methodology to each of the 23 two-digit NAICS industries in the Canadian administrative data, estimating the output elasticities of inputs and TFP for all firm-year observations (see Table A.1 for the list of industries and summary statistics). We begin by presenting the unconditional moments of these parameters, then explore how they vary across the firm-size distribution. We highlight key data patterns that illustrate the identification argument discussed in Section 2.2.1. In Sections 4.3 and 4.4, we show that our key result on RTS heterogeneity is robust to alternative estimation methods and samples. Finally, we relate our findings to broader debates on firms' life-cycle growth and on wage and wealth inequality in Section 4.5.

[12]We use the 2021 vintage of Orbis, accessed through Wharton Research Data Services. In practice our data cover from the early 1990s to 2019 with substantially better coverage starting in 2005. For detailed information on constructing a consistent dataset, see Kalemli-Özcan *et al.* (2024). Data on intermediate inputs and capital stock are available only for a subset of eleven countries.

Table II – Distribution of Production Function Estimates

	Mean	St. dev	P10	P50	P90	P99
Panel A: Main Estimates						
TFP	—	0.17	–0.18	0.00	0.17	0.52
RTS	0.96	0.04	0.92	0.95	1.00	1.08
Panel B: Output Elasticities						
Intermediates	0.59	0.15	0.42	0.59	0.78	0.99
Labor	0.33	0.15	0.14	0.33	0.50	0.66
Capital	0.04	0.03	0.00	0.03	0.08	0.13
Panel C: Input Shares						
Intermediates	0.61	0.18	0.36	0.61	0.85	0.93
Labor	0.29	0.15	0.11	0.28	0.50	0.72
Capital	0.23	0.48	0.01	0.09	0.51	2.16

Notes: Table II shows cross-sectional moments of the distributions of firm-level log TFP, RTS, and the elasticities of output with respect to intermediate inputs, labor, and capital. To obtain these estimates, we apply our baseline method (GNR) within two-digit NAICS and calculate the cross-sectional moment within the same cell. Then we average across all estimated values weighting by the number of observations in each cell. The total number of observation is 4.3 million firm-years. To compare TFPs across industries, we normalize its median to zero within each industry.

4.1 Unconditional Heterogeneity in Production Technologies

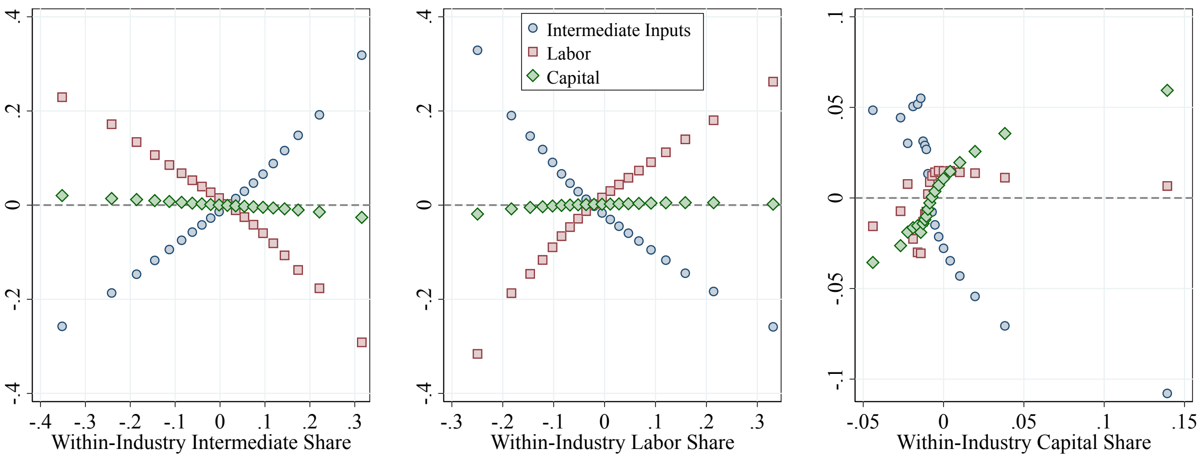
To examine unconditional heterogeneity in firm technologies, we calculate within-industry moments from firm-level estimates for each year, then average across industries and time. Table II reveals considerable heterogeneity in output elasticities, RTS, and TFP across firms.

RTS heterogeneity. Starting with the within-industry RTS distribution, we find an average of 0.96 with a 90^{th}-to-10^{th} percentile gap (P90–P10) of 0.08.[13] This implies that with a 1% larger input bundle, the firm at the 90^{th} percentile produces about 8.3% more output than the firm at the 10^{th} percentile, holding TFP constant. These differences are quantitatively important when interpreted as deviations from constant returns to scale. For instance, in an efficient economy with Cobb-Douglas production,

[13] Consistent with earlier studies (e.g., Basu and Fernald (1997); Ruzic and Ho (2023); Gao and Kehrig (2017)), we also find substantial differences in average RTS across industries (see Table A.1), ranging from 0.59 (for Healthcare) to 1.03 (for Management of Companies and Enterprises).

Figure 1 – Average Output Elasticities By Factor Shares of Revenue

(a) Intermediate Input Share (b) Labor Share (c) Capital Share



Notes: Figure 1 shows the relation between the input revenue shares defined as the ratio between the total cost of intermediate inputs, the total wage bill, and the total value of capital stock, divided by firm revenue, and the estimated output elasticity. Firms are ordered by the respective factor shares on the horizontal axis. The vertical axis shows averages of estimated output elasticities, demeaned within two-digit NAICS industry.

the elasticity of optimal firm output to firm TFP is $\frac{1}{1-RTS}$. This elasticity is five times larger for a firm with RTS of 0.98 compared to one with RTS of 0.90.[14] Dispersion is more pronounced above the median: the average within-industry P50–P10 is 0.03 compared to 0.05 for P90–P50 and 0.13 for P99–P50. The average 90th percentile for RTS across industries is 1.00; that is, most firms operate decreasing returns to scale technologies, yet some exhibit annual RTS above 1.[15]

By construction, differences in RTS arise from heterogeneity in output elasticities. As shown in Panel B of Table II, intermediate inputs have the highest average output elasticity at 0.59, followed by labor at 0.33 and capital at 0.04.[16] Labor and intermediate input elasticities also show larger within industry variation than capital elasticities, with average P90-P10 gaps of 0.36 for intermediates and labor versus 0.08

[14] We validate this prediction in the data by showing that high-RTS firms' revenues respond more strongly to aggregate TFP shocks (Table A.7).

[15] RTS is not fixed over time, and firms are subject to adjustment costs. Thus, increasing returns to scale do not imply unbounded expansion. Furthermore, other studies commonly estimate RTS above 1 for some industries or firms as well (e.g., Gandhi *et al.* (2020) and Demirer (2020)).

[16] Our estimated average capital elasticity is lower than typical estimates because, following the literature, we construct the capital stock using the PIM and include only tangible capital such as structures and equipment. This excludes other forms of capital typically included in the aggregate measure of capital, such as intangible capital and inventories. When we estimate the production function using a broader capital definition based on net asset values from balance sheets, the average intermediate, labor, and capital elasticities become 0.58, 0.30, 0.12, respectively.

for capital. Variance decompositions show that over 60% of the variation in each output elasticity is explained by within-industry differences (Table A.2), whereas only about a quarter of the variance in RTS is accounted for by within-industry differences. This partly reflects a negative correlation between intermediate input and capital/labor elasticities within industries (Table A.3).

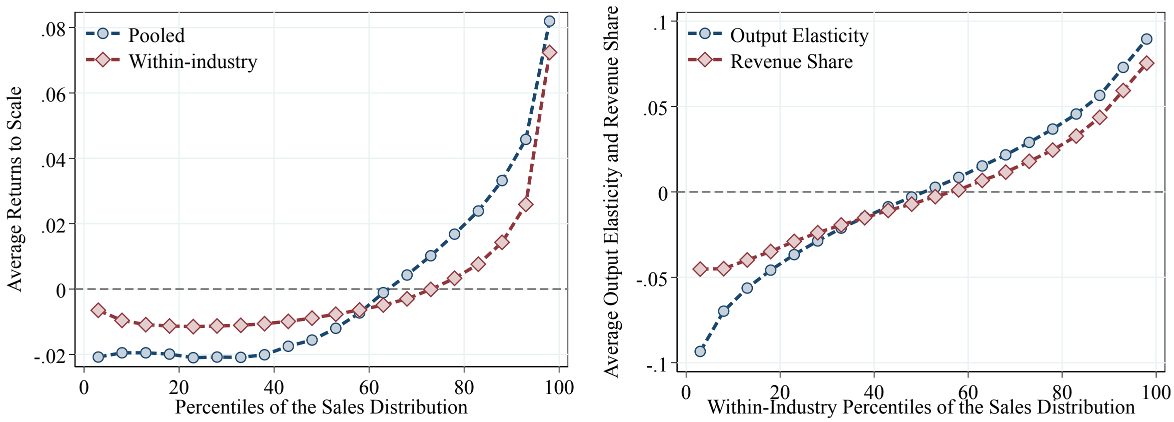
Output elasticities and input shares. Following the identification argument in Section 2.2.1, we now present the key data features driving our empirical results. Typically, output elasticities closely reflect their corresponding revenue input shares. In fact, for profit-maximizing firms with Cobb-Douglas production functions and flexible inputs, output elasticities are exactly equal to (average) input shares. Our specification is more general, and the GNR method does not solely rely on the FOCs of profit-maximizing firms. Nevertheless, output elasticities tend to be close to factor shares (Panels B and C of Table II) and positively correlated with them.

Figure 1 shows a bin scatter of (demeaned) output elasticities for all three inputs on the y-axis conditional on a different input share on the x-axis in each panel. Across all inputs, output elasticities are strongly correlated with their respective input shares: intermediate input-intensive firms have higher intermediate input elasticities, labor-intensive firms have higher labor elasticities, and capital-intensive firms have higher capital elasticities. The correlation is particularly strong for intermediate inputs, as expected, since we treat them as flexible and use the firm's FOC to estimate intermediate input elasticities. In contrast, for labor and capital, our estimation does not rely on FOCs. Nevertheless, we find strong positive correlations for these inputs as well. These patterns support the identification intuition in Section 2.2.1: heterogeneity in output elasticities, and thus in RTS, largely reflects differences in input shares. They also suggest that high-RTS firms should have relatively low profit shares, which we confirm in the data by showing the negative correlation between the EBITDA-to-revenue ratio and RTS across firms (Figure A.5).

TFP dispersion. We find that the P90-P10 gap in firm-level TFP is 0.35, implying that a firm at the 90^{th} percentile produces about 41.9% more output than a firm at the 10^{th} percentile, conditional on inputs and output elasticities. This gap is substantially smaller than previous estimates even for narrow six-digit industries in Canada and

Figure 2 – RTS and Intermediate Input Elasticity Increase With Firm Size

(a) RTS | (b) Intermediate Input Elasticity



Notes: Figure 2a shows the average RTS across quantiles of the firm-revenue distributions, using both pooled and within-industry percentiles. RTS is demeaned by the pooled average (average RTS of 0.96) in the former and by industry averages in the latter. Figure 2b presents the estimated output elasticity and the observed revenue share of intermediate inputs, both demeaned by industry averages, across quantiles of the within-industry revenue distribution.

the U.S., where typical P90-P10 TFP gaps are about twice as large (e.g., De Loecker and Syverson (2021) and Syverson (2011)). Two factors explain the difference: first, we estimate a flexible nonparametric production function that allows for differences in RTS; second, we use the wage bill rather than headcount or hours as the measure of labor input (see Fox and Smeets (2011)).

4.2 Production Technologies over the Firm-Size Distribution

Returns to scale by firm revenue. We now turn to the systematic variation in RTS across the firm revenue distribution. To this end, we pool all firm-year estimates from 23 two-digit NAICS industries. Figure 2a shows bin scatter plots of (demeaned) average RTS by firm revenue using two ranking methods. First, we pool firm-year observations across all industries and rank them into revenue percentiles. We find that RTS is relatively flat across the bottom two-fifths of firms but increases sharply for larger firms in the economy. Average RTS rises by about 10 p.p. from the bottom to the top of the revenue distribution, primarily above the median.

Part of the variation in our pooled ranking may reflect differences across industries. For example, manufacturing firms, which tend to have higher RTS and larger revenues, are overrepresented at the top. Therefore, to isolate within-industry patterns, we calculate within-industry revenue percentile rankings. This approach reveals similar patterns: RTS is roughly constant below the median and increases steeply among larger firms within industries, with an 8 percentage point gap between the top 5% and the bottom half. Thus, most of the observed variation in RTS by firm size is driven by differences within industries, rather than across industries.

Output elasticities by firm revenue. Our analysis shows that the positive relationship between RTS and firm revenue is entirely driven by the intermediate input elasticity (Figure 2b). The intermediate input elasticity increases monotonically from -0.09 (relative to the industry average) for firms in the bottom 5% of the revenue distribution, to approximately zero around the median, and up to 0.09 for firms in the top 5%. This 9 p.p. gap in intermediate input elasticities between the top 5% and median firms fully explains the corresponding 8 p.p. gap in RTS over the same range. Figure 2b also shows that the intermediate input revenue share mirrors this pattern, with larger firms allocating a higher share of their revenue to intermediate inputs compared to smaller firms. This result is expected, as our estimation treats intermediate inputs as a flexible factor.[17] Consistent with our findings, in a contemporaneous study Mertens and Schoefer (2024) use a setting with homothetic production functions and imperfect input markets to show that firms grow by shifting from labor to intermediate inputs. Finally, capital and labor elasticities decline slightly with firm revenue (Figure A.2), underscoring the importance of estimating gross output production functions: relying on value-added specifications may lead to misleading conclusions about firm-level technologies.

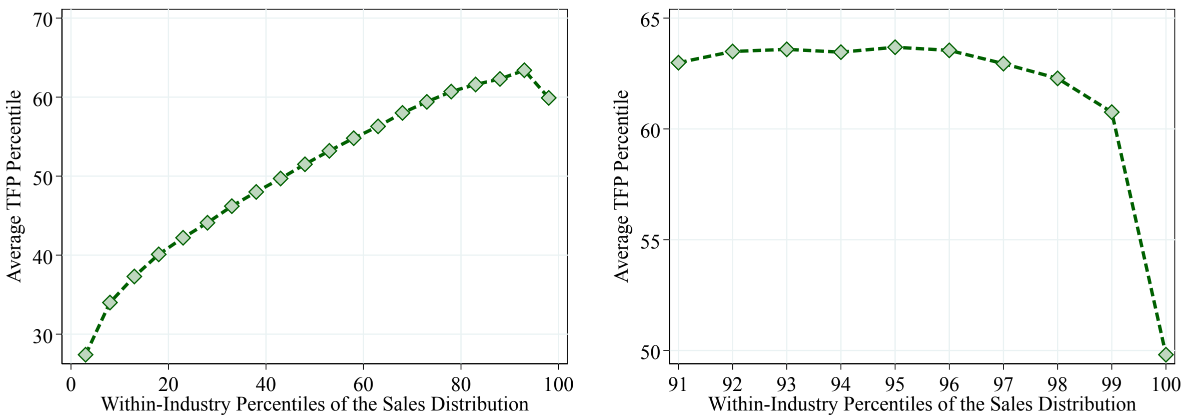
Total factor productivity by firm revenue. We next investigate whether larger firms also exhibit higher TFP. Since TFP levels are not comparable across industries, we focus on firms' relative productivity ranks within industries. Figure 3a displays the average within-industry TFP percentile by within-industry revenue percentile.

We find that relative TFP increases with firm size up to the top decile of the revenue distribution, after which it flattens out. In fact, zooming in on the top 10%, we find that TFP falls off sharply for the largest firms (Figure 3b). In contrast,

[17] Note that the intermediate elasticity does not equal its revenue share because of the ex-post shock ε. See Appendix A for details.

Figure 3 – Firm TFP Flattens out at the Top of the Firm Size Distribution

(a) Entire Revenue Distribution　　(b) Zoom in on Top 10%



Notes: Figure 3a shows the average firm TFP rank within percentiles of the within-industry revenue distribution. The TFP rank is calculated within each industry. Figure 3b zooms in on the top 10% of the revenue distribution.

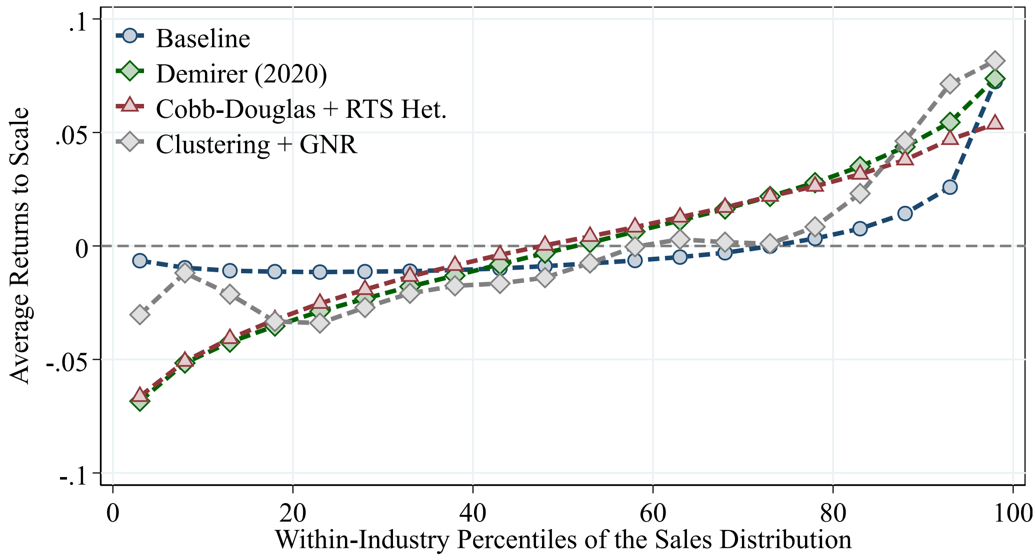
RTS increases even more steeply among the largest firms (Figure A.6). Thus, the largest firms tend to feature the highest RTS—not necessarily the highest TFP—as commonly assumed.

A few papers have studied the TFP-size relation (see Leung *et al.* (2008) or Baldwin *et al.* (2002)). Our results on the TFP-revenue gradient differ from these studies because we allow for heterogeneity in production technologies. To illustrate this, we reestimate a standard Cobb-Douglas production function imposing homogeneous RTS: $Y_{jt} = e^{\nu_{jt}} K_{jt}^{\alpha_i} L_{jt}^{\beta_i} M_{jt}^{\gamma_i}$, where j and i denote firms and industries, respectively. As expected, under this restriction TFP increases monotonically with firm size (Figure A.8). This contrast highlights the importance of allowing for flexible production technologies in understanding the relationship between firm-level TFP, RTS, and size.

4.3 Robustness and Interpretation

Our benchmark method, GNR, relies on several identifying assumptions, such as firms being price takers in output markets. In this section, we relax some of these assumptions and apply alternative methods (described in Section 2.2) to show that our key result—larger firms operate technologies with higher RTS—is robust. Table III summarizes the RTS gap between the largest 5% of firms and bottom 50% for different methods and samples. Figure 4 highlights a subset of these results by showing the RTS–firm size relationship for the most important alternative specifications.

Figure 4 – RTS Increases with Firm Size for Different Specifications



Notes: In the cluster specification, we apply the k-means clustering algorithm (20 clusters) based on revenue percentile and revenue shares of the three factors for firms within each industry and estimate the nonparametric production function following GNR within clusters. In all specifications, we sort firms based on revenue within industry, and RTS is demeaned by industry averages.

4.3.1 Markups and Market Power

Our RTS estimates are based on *revenue* elasticities of the three inputs. A key concern in the literature (e.g., Bond *et al.* (2021)) is that identifying revenue-based or physical production functions typically requires either price and quantity data or strong parametric assumptions about demand and technology. In particular, using revenue data alone may lead to unknown biases in estimates of markups and output elasticities. However, recent evidence (e.g., De Ridder *et al.* (2022)) suggests that such biases are modest in practice and that relative variation in markups and elasticities is well identified even with revenue-based data. Consistent with this view, we show that relative variation in RTS—our main object of interest—is robust across multiple estimation methods. Specifically, we present three sets of theoretical and empirical arguments supporting our interpretation that the observed positive relationship between RTS and firm size primarily reflects technological differences, rather than variation in markups (e.g., De Loecker *et al.* (2020)), or monopsony markdowns (e.g., De Loecker *et al.* (2016) and Burstein *et al.* (2024)).

Table III – RTS Increases with Firm Size For Different Specifications

Method	Specification (vs. baseline)	Average RTS	RTS Gap
GNR	Baseline (Canada, by industry)	0.96	0.08
Demirer (2020)	Same as baseline	0.78	0.10
Cobb-Douglas + RTS Het.	Same as baseline	1.00	0.08
GNR	Include intangible capital	0.89	0.20
GNR	Cluster by factor sh. and revenue pct.	0.93	0.10
Cobb-Douglas	Cluster by factor sh. and revenue pct.	0.85	0.14
GNR	Cluster by max. firm size	0.95	0.08
GNR	Only manufacturing industries	1.00	0.04

Notes: Each row reports the average RTS and the average within-industry RTS gap between the largest (revenue-wise) 5% and the bottom 50% of firms for one specification. Intangible capital is constructed using PIM. See Section 2.2 for a discussion of the estimation methods and clustering approaches.

Role of markups. First, if larger firms charge higher markups or markdowns—as implied by models with oligopolistic competition (e.g., Atkeson and Burstein (2008)) or monopolistic competition under log-concave demand systems (e.g., Edmond *et al.* (2023))—then physical RTS would increase even more strongly with firm size than revenue-based RTS. This follows because the physical output elasticity equals the revenue elasticity multiplied by the firm's markup (adjusted for markdowns). We directly estimate firm-level markups following De Loecker and Warzynski (2012) and find that markups increase with firm revenue in our data (Figure A.1)—consistent with De Loecker *et al.* (2020)— indicating that the size gradient is larger for physical RTS than for revenue-based RTS.

Controlling for market power. Second, while our baseline method permits markdowns in capital and labor markets, it does not account for markups in output markets or markdowns on intermediate inputs. We then extend the GNR approach to explicitly control for both types of market power using firms' output market shares as proxies for unobserved price elasticities (following De Loecker *et al.* (2020) and De Loecker *et al.* (2016)). Relaxing the perfect competition assumption, we allow firms to face downward-sloping demand and adjust the FOC for intermediate inputs accordingly.[18] If markups (or markdowns) are a significant determinant of input ex-

[18] This is an exact control if demand takes the common (nested) logit or CES forms. De Loecker *et al.* (2020) use this approach to control for unobserved output prices while De Loecker *et al.* (2016) apply it to control for unobserved intermediate input prices. Following them, we use a cubic function

penditure shares, we should find that our estimates of the intermediate input elasticity are sensitive to the inclusion of these controls. Controlling for market share barely changes the size gradient of the intermediate input elasticity (Figure A.10), the main driver of RTS differences along the firm-size distribution.

Demirer method. Third, we apply Demirer (2020)'s method, which explicitly allows for heterogenous markups across firms (while imposing stronger assumptions on labor input choices and functional form of the production function). Despite these differences, the estimated RTS-size gradient remains robust and, if anything, becomes even steeper: RTS increases by about 10 percentage points from the bottom half to the top 5% of the firm-size distribution (Figure 4). These results are consistent with the view that physical RTS increases more strongly with firm size than revenue-based RTS, as expected if markups increase with firm revenue—a relationship we confirm again using the Demirer methodology (Figure A.1).

Factor-augmenting productivity shocks. Another potential concern is that larger firms may have higher intermediate input elasticities simply because they use intermediate inputs more efficiently, reflecting factor-specific productivity shocks. Since our results remain robust under Demirer (2020)'s method, which accommodates factor-augmenting productivity shocks, this concern is likewise alleviated.

4.3.2 Permanent versus Transitory Differences in RTS

Our benchmark GNR method estimates a flexible but common production function across firms within an industry, allowing firms to differ in their output elasticities based on their positions in the input space. A natural question, therefore, is whether the observed dispersion in RTS reflects permanent differences in production technologies across firms or merely transitory fluctuations around a common structure. The distinction is important because, for example, for models of the firm-size distribution (e.g., the literature that follows Lucas (1978) and Hopenhayn (1992)) not only the magnitude of heterogeneity but also its persistence is crucial (e.g., Sterk *et al.*

of market shares (defined at the two-digit NAICS). Since period-t market shares may be correlated with transitory productivity shocks, we estimate a modified first-stage equation with GMM using lagged market shares as instruments for current shares. See Appendix A.1 for details.

(2021)). We present three complementary sets of results suggesting that the bulk of RTS heterogeneity reflects persistent firm characteristics rather than transitory factors.

Fixed effects regression. First, we regress RTS estimates from our baseline GNR method on firm size, firm age, time dummies, and firm fixed effects. Intuitively, if differences in RTS are largely permanent, firm fixed effects should absorb most of the variation. This is exactly what we find: of the total RTS variance of 0.052^2, firm fixed effects (with a variance of 0.045^2) account for 75% of the variation when controlling for firm age and size. We find similarly strong persistence in the U.S. manufacturing sector: RTS has a variance of 0.058^2 and fixed effects account for 65% of the total variation after controlling for other firm observables.

Autocovariance structure. Second, following the earnings dynamics literature (e.g., Abowd and Card (1989); Karahan and Ozkan (2013)), we exploit the autocovariance structure of the RTS estimates to decompose firm-level RTS into a firm-specific fixed effect, a persistent AR(1) component, and a fully transitory component (see Appendix D for details). Consistent with the fixed effects results, only 10.5% of the total RTS variance is attributable to purely transitory shocks. Firm fixed effects and the highly persistent component (with an estimated persistence parameter of 0.94) account for 38.9% and 50.6% of the total variation, respectively.

Clustering analysis. While estimating firm-specific production functions would be ideal, it is not econometrically feasible given the relatively short panel. As a practical alternative—detailed in Section 2.2.2—we cluster firms within industries based on input shares and revenue percentiles, and estimate cluster-specific production functions using both the nonparametric GNR specification and a more restrictive Cobb-Douglas form. In both cases, the RTS–size gradient remains strong. Under the GNR specification, RTS rises by nearly 10 p.p. from the median to the top 5% of firms within industries, consistent with baseline findings. Under the Cobb-Douglas specification, the increase is even larger—14 p.p. between the bottom half and the top 5%.

To further distinguish persistent type effects from size effects, we group firms based on their maximum revenue attained over their life cycle and track the RTS evolution within and across clusters. If RTS differences purely reflect persistent firm types, we would expect large level differences across clusters but flat within-cluster size gradients. If instead RTS differences are driven by scale-dependent nonhomotheticities, we would observe positive size gradients within clusters. Regressions of RTS on log revenue support the persistent firm type interpretation: without cluster fixed effects, the size gradient is 0.012 (capturing pooled variation); when including cluster fixed effects, the gradient drops to -0.001. Moreover, the pattern of average RTS by firm size remains very similar to our baseline pooled estimates, with a gap of about 8 p.p. between the largest 5% and the bottom 50% of firms (Figure A.9).

Taken together, these three sets of results indicate that the observed heterogeneity in RTS is primarily driven by persistent differences between firms, consistent with the endogenous entrepreneurship model with heterogeneous RTS in Section 5.

4.3.3 Cobb-Douglas with RTS Heterogeneity.

Another potential concern is that our results may be sensitive to the flexible functional form assumed in the GNR method. To address this, we reestimate the production function by imposing homogeneous relative factor elasticities while still allowing for RTS differences, i.e., $F_{jt}(K_{jt}, L_{jt}, M_{jt}) = \left(K_{jt}^{\varepsilon_K} L_{jt}^{\varepsilon_L} M_{jt}^{\varepsilon_M}\right)^{\eta_{jt}}$. Consistent with our baseline findings, the Cobb-Douglas series in Figure 4 shows that RTS increases with firm size by about 10 p.p., with a steeper rise in the bottom half of the revenue distribution and a more moderate increase among the largest firms relative to our baseline.

4.4 International Evidence and Further Robustness

Our results are also robust across multiple samples and alternative specifications. We find similar patterns when analyzing U.S. manufacturing plants, firms in eleven European countries, and when incorporating intangible capital into the measurement of the capital stock. The relationship between RTS and firm size also remains robust under alternative ways of ranking firms within industries.

U.S. manufacturing. Our results are not unique to the Canadian economy but also hold for the U.S. manufacturing sector. Figure 5a shows the RTS–revenue relationship at the plant level, relative to the four-digit NAICS industry average. The pattern is

U-shaped, with a notable steep increase at the top: RTS rises by about 9 p.p. from the 50th percentile to the top 1% of the revenue distribution. For comparison, we include a corresponding series for Canadian manufacturing in the same figure. Both sectors display similar U-shaped patterns, but the increase in RTS among the largest plants is more pronounced in the U.S., consistent with the longer right tail of the U.S. manufacturing size distribution (Leung *et al.*, 2008).

Remarkably, as in Canada, the increase in RTS is primarily driven by a rise in the output elasticity of intermediate inputs, which increases from about 0.35 at the bottom of the size distribution to around 0.55 for the largest U.S. plants (Figure A.2b). Furthermore, labor elasticities decline steadily with firm revenue, while capital elasticities decline up to the 90th percentile and then rise slightly among the very largest plants (Figure A.2). Revenue shares of labor, capital, and intermediate inputs across the revenue distribution also exhibit remarkably similar patterns between U.S. and Canadian manufacturing, and more broadly among all Canadian corporations (Figure A.4).

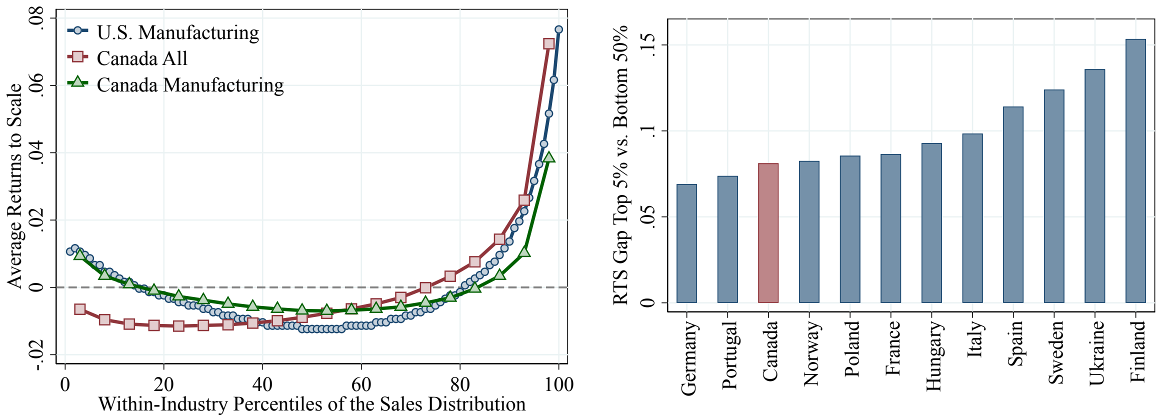
Our U.S. manufacturing production function estimates are at the plant level, whereas the Canadian data are measured at the firm level, which aggregates over multiple plants. In the Canadian data, we find that RTS increases significantly with the number of plants per firm (Figure A.7). However, controlling for the number of plants only slightly attenuates the RTS–revenue gradient: regressing demeaned RTS on log firm revenue yields a coefficient of 0.012, which drops modestly to 0.010 when controlling for plant count. Together with the U.S. manufacturing evidence, these results suggest that variation in RTS by firm revenue is not primarily driven by differences in the number of plants, but rather by systematic differences in production technologies across individual plants.

International evidence. Our results extend to several other countries using firm-level data from the Orbis database. Figure 5b shows estimates based on our baseline GNR method, applied within 2-digit NAICS industries across countries. We summarize the findings by plotting the average RTS difference between the top 5% and bottom 50% of the within-country-industry revenue distribution. The results are remarkably consistent across countries, with RTS differences ranging from about 7 p.p. in Germany to 15 p.p in Finland. Canada falls near the middle of this distribution,

FIGURE 5 – THE POSITIVE RTS-FIRM SIZE GRADIENT IS ROBUST ACROSS COUNTRIES

(A) RTS in U.S. Manufacturing

(B) RTS Gap Across Countries



Notes: Figure 5a plots average RTS (demeaned by industry averages) against within-industry revenue percentiles for Canadian and U.S. manufacturing, as well as for the baseline economy-wide Canadian sample. Canadian results are shown within 5% quantiles of the revenue distribution. Figure 5b compares the within-industry RTS gap between the top 5% and bottom 50% of firms across 11 countries using Orbis data, including Canada for reference.

with similar patterns observed for other large European economies such as France and Italy. Consistent with our main results, we find that the intermediate input elasticity also increases strongly with firm size across countries (Figure A.11). Last, we apply the Demirer methodology to the Orbis data and find similar results (Figure A.12).

Intangible capital. We also analyze the importance of including intangibles in our measure of firms' capital stock and reestimate their production functions using Canadian data. In theory, including intangibles affects measured productivity, the output elasticity of capital, and therefore RTS. In particular, if larger firms invest disproportionally more in intangible capital, omitting intangibles would understate their capital elasticity (and thus RTS) and overstate their TFP. Consistent with this intuition, we find that the positive relationship between firm size and RTS becomes even stronger when intangible capital is included. As shown in Table III, the P95–P50 RTS gap increases from 0.08 in the baseline to 0.20 when intangible capital is incorporated.

Ranking firms by employment or value added. Appendix Figure A.3 presents the results when ranking firms, within industry, by employment or value added rather than by revenue. While the patterns for RTS are similar, the output elasticities display distinct variations: firms with high employment or high value added exhibit higher labor elasticities, whereas the intermediate input elasticity shows only a small increase among the largest firms. This pattern arises mechanically from the ranking criterion: firms with high employment or value added are, by construction, more labor-intensive. Therefore, we prefer to rank firms by revenue—a factor-neutral approach—in our primary analysis.

4.5 Implications for Firm Dynamics and Inequality

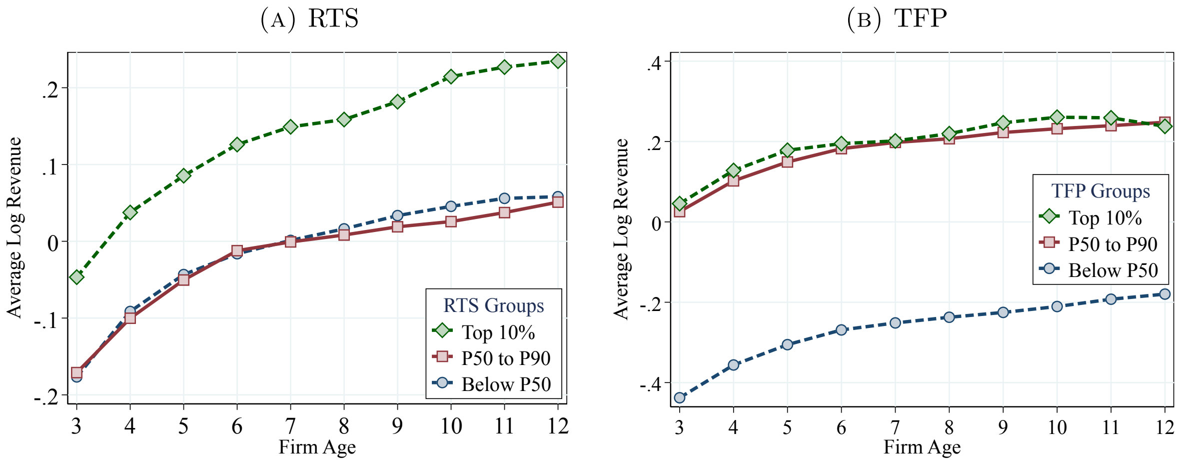
We now revisit several well-known empirical patterns in firm heterogeneity that have traditionally been attributed to TFP differences. For example, the literature has argued that firms with higher TFP grow faster (e.g., Sterk *et al.* (2021)), pay higher wages (e.g., Kline (2024)), and are disproportionately owned by wealthier households (e.g., Quadrini (2000); Cagetti and De Nardi (2006)). In this section, we argue that RTS differences are at least as important in explaining these empirical patterns.

4.5.1 Firm Dynamics

Heterogeneity in RTS has significant implications not only for the firm size distribution but also for firms' growth trajectories over the life cycle. Firms with higher RTS are expected to grow faster to reach their larger optimal sizes, compared to firms with similar TFP but lower RTS. To analyze these life-cycle patterns, we construct a balanced panel of Canadian firms from our baseline estimates and group firms by their initial production function characteristics. Specifically, we focus on firms born between 2002 and 2005 that are observed for 12 consecutive years. We group firms based on their initial RTS and initial TFP, demeaned at the industry level. We then track their average log revenue, again demeaned within industries, over their life cycle (Figure 6).

Firms with higher initial RTS and TFP start with higher revenues relative to their industry peers. More importantly, firms with higher initial RTS (Panel A) exhibit significantly faster growth: firms in the top 10% of the initial RTS distribution grow about 30 log points over 10 years, whereas firms in the bottom 90% grow by only about 20 log points. This evidence supports our interpretation that high-RTS firms operate more scalable technologies, enabling substantially greater life-cycle growth.

Figure 6 – Life Cycle of Firms Starting with Different RTS and TFP



Notes: Figure 6 compares the life-cycle profile of revenue between firms with different initial RTS (Figure 7a) and TFP (Figure 7b). They are constructed using a balanced panel of firms which (i) are born between 2002 and 2005 and (ii) survive for at least 12 years. We demean firms' initial RTS at the two-digit NAICS industry level. We bin firms into three groups based on their initial demeaned RTS in the left panel and three groups based on initial within-industry TFP percentiles in the right panel. Firm log revenue is also demeaned at the two-digit NAICS industry level.

We also rank firms by their average growth rates over 12 years and find that the top 1% fastest-growing firms ("gazelles") exhibit an average RTS of 0.98 compared to 0.95 among those below the 90th percentile. Similarly, Guntin and Kochen (2025) recently show that a firm dynamics model with ex-ante heterogeneity in production functions is required to explain empirical life cycle trajectories of the largest firms.

In contrast, Panel B shows that firms entering with high TFP, while initially larger, do not grow faster than other firms in their industry. Indeed, higher initial TFP is associated with slightly lower subsequent growth, which can be explained by TFP being a mean-reverting process. These finding suggest that highly productive firms might have low RTS, which constrains their growth (Hurst and Pugsley (2011)).

While these results focus on surviving firms, we also examine the effects of RTS and TFP heterogeneity on firm exit. We estimate probit regressions of firm exit on TFP percentile and RTS (Table A.6). Across specifications, higher RTS is associated with a lower probability of firm exit. The effect of TFP on firm exit is smaller, and varies in sign across specifications. We conclude that, from an ex ante perspective, RTS heterogeneity better predicts differences in firm growth and survival over the life cycle than TFP heterogeneity.

Finally, we investigate whether firms with varying RTS respond differently to aggregate shocks. We use two types of shocks: changes in industry-level TFP, and the 2007-2008 global financial crisis. We estimate regressions of firm revenue growth on RTS, the aggregate shock, and their interaction (Table A.7). The results show that firms with higher RTS respond more strongly to aggregate shocks, consistent with greater scalability (see also Smirnyagin (2023) and Argente *et al.* (2024)).

4.5.2 Role of RTS Heterogeneity in Wealth and Wage Inequality

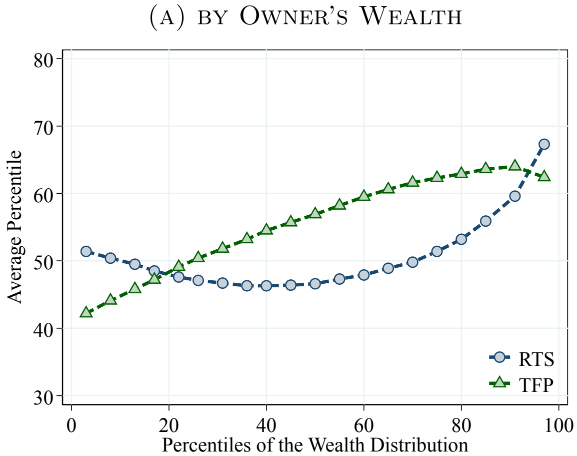
We conclude this section by examining how firm-level RTS relates to the wealth of firm owners and the wages of workers. First, we analyze how production function parameters vary with the equity wealth of business owners. A key advantage of our dataset is that we can link firms to their ultimate individual owners using administrative records from the Shareholder Information in Corporate Tax Files. We calculate each individual's equity wealth by aggregating the value of the firms they own, weighted by ownership shares. For each owner, we then compute an equity-value-weighted average RTS and TFP percentile across the firms they own. Figure 7a shows that wealthier individuals tend to own firms with higher RTS. That is, owners at the top of the wealth distribution are more likely to own firms with more scalable production technologies. In addition, conditional on RTS, TFP is also increasing in owner wealth, but in a concave manner, particularly at the top of the distribution.

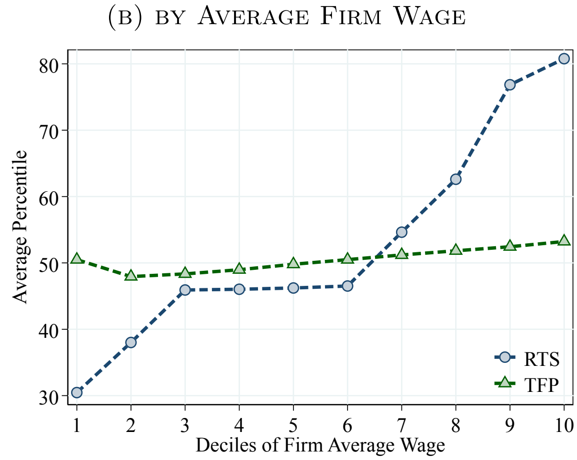
It is well established that large firms tend to pay higher wages than smaller firms for similar workers (Brown and Medoff (1989)). Given our results, a natural question is whether this firm size-wage premium is driven by higher TFP or greater scalability (RTS) among large firms. To disentangle these factors, we rank firms by their average wage and compute the average RTS and TFP across wage deciles. Figure 7b shows that higher-paying firms tend to have significantly higher RTS, while the relationship between wages and TFP is weaker and less systematic. These findings suggest that RTS heterogeneity is an important driver of the firm size-wage premium.

5 Misallocation with RTS Heterogeneity

So far, we have documented substantial heterogeneity in RTS across firms and examined its implications for a broad set of empirical patterns related to the firm size

FIGURE 7 – RTS AND TFP BY OWNER'S WEALTH AND AVERAGE FIRM WAGE





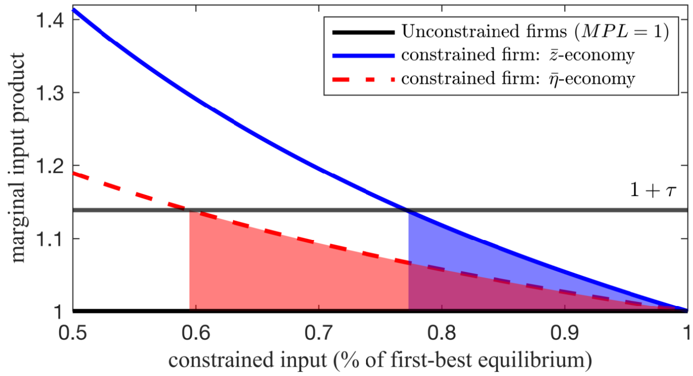
Notes: Figure 7a shows the average percentiles of RTS and TFP by percentiles of owners' equity wealth. Figure 7b shows the average percentiles of RTS and TFP by deciles of firms' average wage. Average wage is demeaned at the industry level. RTS and TFP percentiles are calculated within industry using our baseline (GNR) method.

distribution. In this section, we argue this heterogeneity is also important theoretically and quantitatively by focusing on a fundamental question in macroeconomics: the efficiency costs of financial frictions. We employ an off-the-shelf quantitative model of entrepreneurship (i.e., Quadrini (2000); Cagetti and De Nardi (2006)) to compare the misallocation arising from financial frictions in a calibration with heterogeneity in both RTS (η) and TFP (z)—the (η, z)-economy—against a standard setting with heterogeneity only in z—the z-economy. Our main conclusions remain robust in richer settings—those we consider in the empirical analysis—with intermediate inputs and pre-determined capital. Our analysis suggests that RTS heterogeneity has broad implications for various quantitative questions, such as optimal taxation of capital (Boar and Midrigan, 2022), firm recruiting intensity (Gavazza *et al.*, 2018), or firm cyclicality (Clymo and Rozsypal, 2023; Smirnyagin, 2023). To build intuition, we first derive an analytical result in a static endowment economy and then quantify the mechanism in a dynamic setting.

5.1 Analytical Result in an Endowment Economy

We consider an endowment economy with aggregate factor supply normalized to one, $X = 1$. There is a continuum of firms $i \in [0, 1]$, producing perfectly substitutable goods. A fraction $\chi \in (0, 1)$ of these firms face an input price wedge $\tau \geq 0$ and are

Figure 8 – Efficiency costs in endowment economy



Notes: The figure provides a qualitative illustration of efficiency losses from input wedges for a representative constrained firm. In the $\bar{z}$-economy, the firm has high TFP ($z = \bar{z}, \eta = \eta_0$); in the $\bar{\eta}$-economy, the firm has high returns to scale ($z = z_0, \eta = \bar{\eta}$), with $0 < z_0 < \bar{z}$ and $0 < \eta_0 < \bar{\eta} < 1$. The shaded areas capture the implied output loss under each scenario.

thus constrained in their production. Each constrained firm is characterized by a pair of parameters (η, z), where $\eta \in (0, 1)$ indicates decreasing returns to scale and z is the firm's TFP. The output of a constrained firm is given by $y = f(x; z, \eta) = z \cdot x^\eta$. The remaining fraction of firms $1 - \chi$ is unconstrained and has constant returns to scale ($\eta = 1$).[19] The following proposition characterizes misallocation in terms of the output share of constrained firms and the RTS of constrained firms:

Proposition 1. *Consider an interior equilibrium where the output share of constrained firms is below one. Then, up to a second order approximation around the first best ($\tau = 0$), the percent output loss associated with τ is given by*

$$\Delta \ln Y(\tau) = \underbrace{\frac{\tau^2}{2}}_{size\ of\ friction} \cdot \underbrace{\int_0^\chi w_i \cdot di}_{output\ share\ of\ constrained\ firms} \cdot \underbrace{\int_0^\chi \frac{w_i}{\int_0^\chi w_j dj} \cdot \frac{\eta_i}{1-\eta_i} \cdot di}_{avg.\ \frac{RTS}{1-RTS}\ constrained\ firms}$$

where $w_i \equiv \frac{y_i^\star}{Y^\star}$ denotes the relative output of firm i in the first-best equilibrium.

Proof. See Appendix E.1 for the proof of the proposition. □

The proposition states that misallocation is proportional to the size of the friction and the output share of constrained firms and, more importantly, is increasing and

[19] Alternatively, one could assume that unconstrained firms also exhibit decreasing RTS, but the presence of free entry ensures constant RTS at the sectoral level.

convex in the (weighted-average) RTS of constrained firms (see also Atkeson *et al.* (1996) and Guner *et al.* (2008) for related points). Consequently, for a given friction, misallocation becomes more severe when constrained firms have higher RTS. Furthermore, as a result of the convexity of misallocation in RTS, greater dispersion in RTS among constrained firms also leads to more severe misallocation.

Intuitively, a given input price wedge results in a larger quantity adjustment when RTS is high, as marginal products decline more slowly. This causes constrained firms to reduce their inputs more, leading to greater misallocation. In contrast, firm TFP affects misallocation only indirectly through its influence on the output share of constrained firms. We illustrate this in Figure 8, which depicts the marginal input product of firms that would be "large" in the first-best equilibrium and contribute most to misallocation. The solid blue line represents the conventional setting where large firms have high TFP ($\bar{z}$), while the dashed red line represents an economy where large firms have high RTS ($\bar{\eta}$). For a given wedge τ, misallocation—represented by the area under the curve—is larger in the $\bar{\eta}$-economy.

5.2 Quantitative Dynamic Model

We now consider a dynamic workhorse model of entrepreneurship in the tradition of Quadrini (2000) and Cagetti and De Nardi (2006), in which the set of constrained firms emerges endogenously. We use this model to quantify misallocation in an economy where firms differ in both RTS and TFP, as in our empirical findings, and compare it with the misallocation in an economy where firms differ only in TFP. Apart from the introduction of RTS differences, our framework remains deliberately simple and closely follows these standard models of entrepreneurship in quantitative macroeconomics. At the end of this section, we show that our main finding remains robust in richer settings, including those employed in our empirical framework (e.g., if we explicitly introduce intermediate inputs): accounting for RTS heterogeneity substantially amplifies the misallocation losses from financial frictions.

5.2.1 Model Setup

Time is discrete and there is a continuum of agents of mass one, who derive log utility from consumption. They discount the future at rate $\tilde{\beta}$ and face a constant death probability $p \in [0,1)$. Thus, their effective discount factor is $\beta = (1-p)\cdot\tilde{\beta}$, and they maximize $\mathbb{E}\left[\sum_{t\geq 0}\beta^t \ln(c_t)\right]$. Agents face an occupational choice between employment as a worker and entrepreneurship, $o \in \{W, E\}$. A worker's labor income equals $w \cdot h$, where w denotes the wage rate and h the efficiency units of labor supply, which follow a first-order Markov process. Entrepreneurs are price takers in input and output markets, using labor ℓ and capital k at rental rates w and R, respectively, to produce output $z \cdot f(k,\ell)^\eta$, where $f(\cdot)$ is a constant RTS production function. The pair (z,η) denotes entrepreneurial productivity z and scalability of their project η, which follow a joint first-order Markov process.

Asset markets are incomplete, and agents can invest their wealth $a \geq 0$ in an annuity that pays an interest rate r. Upon death, individuals are replaced by an equal number of newborn households who start with zero wealth. We parameterize financial frictions by $\lambda \in [0,1]$ and assume that a fraction λ of total input expenditures must be financed by the entrepreneur's own wealth. As a result, static profit maximization yields a net profit of

$$\pi(a,z,\eta) = \max_{k\geq 0,\ell\geq 0} z\cdot f(k,\ell)^\eta - w\cdot\ell - R\cdot k$$
$$\text{s.t. } w\cdot\ell + R\cdot k \leq \frac{a}{\lambda},$$

implying input choices $k(a,z,\eta), \ell(a,z,\eta)$ and output $y(a,z,\eta)$.[20] Thus, the agent's dynamic problem can be written in recursive form as

$$V(a,h,z,\eta) = \max_{a'\geq 0, c\geq 0, o\in\{W,E\}} u(c) + \beta\cdot\mathbb{E}[V(a',h',z',\eta')]$$
$$\text{s.t.} \quad c + a' = \mathbb{I}_{o=W}\cdot w\cdot h + \mathbb{I}_{o=E}\cdot\pi(a,z,\eta) + (1+r)\cdot a.$$

We assume that there is a competitive financial intermediary, investing in physical capital with depreciation rate δ, and issuing the annuities.

Equilibrium. We relegate the standard definition of equilibrium to Appendix E.2.

[20] We assume that the friction affects all inputs symmetrically to focus on overall firm size distortions, without introducing additional distortions on relative input use (as would be the case, for example, with a collateral constraint on k only). In the extension with intermediate inputs in Section 5.2.4, we also consider asymmetric constraints.

5.2.2 Calibration

The main idea is to calibrate both the (η, z)- and the z-economy to the same set of observable moments of the firm size distribution and entrepreneurship dynamics. We employ the standard calibration strategy in this literature. First, we briefly discuss fixed common parameters. We set the death probability to $\frac{1}{80}$, corresponding to an expected life expectancy of 80 years.[21] We use a Cobb-Douglas production function (f) with capital share $\alpha = 0.4$ and depreciation rate $\delta = 0.05$. Labor efficiency units h follow a log-normal AR(1) process with an autocorrelation of 0.9 and a cross-sectional standard deviation of 1.3, with the mean normalized to $\mu_h = -\frac{\sigma_h^2}{2}$. This process is estimated directly from the data on individual post-tax earnings. We calibrate both economies at $\lambda = 0.3$, indicating that 30% of input expenditures must be financed with the owner's wealth, and then vary λ in counterfactuals.[22]

(z)-Economy: We jointly calibrate a set of five parameters $(\beta, \eta, \sigma_z, \rho_z, \xi_z)$ to match a set of six empirical moments as summarized in the middle column of Table IV. We provide intuition on how these parameters are identified. The effective discount factor β primarily influences the aggregate capital-output ratio. The (common) RTS parameter η is closely tied to the fraction of the population engaged in entrepreneurship, as it determines the share of income entrepreneurs receive. We model the z-process as log-normal AR(1) with normalized mean $\mu_z = -\frac{\sigma_z^2}{2}$. Its autocorrelation (ρ_z) affects the transition rates into (and out of) entrepreneurship. The cross-sectional dispersion of z, captured by σ_z, plays a crucial role in shaping the firm size distribution. Additionally, we model the top 1% of the z-distribution with a Pareto tail, where ξ_z denotes the tail coefficient, enabling the model to better match the right tail of the firm size distribution. Overall, the model achieves an excellent fit with the targeted empirical moments.

(η, z)-Economy: In essence, we replicate the calibration of the z-economy, but also account for heterogeneity in η by matching the observed dispersion of RTS along the

[21] The death rate affects in particular wealth accumulation at the bottom of the wealth distribution, as newborns enter with zero wealth. The bottom 50% wealth share equals 3.3% in the (η, z)-model and 2.2% in the z-model, in the ballpark of the value for Canada of 4.9%.

[22] Defining the debt d of entrepreneurs as $d = \max\{0, k - a\}$, the aggregate debt-to-capital ratio is 81% in the (η, z)-model and 71% in the z-model, both in line with Canada's ratio of roughly 70%.

Table IV – Dynamic model: Targeted moments and calibrated parameters

	Data	Model	
		z-economy	*(η, z)-economy*
A. Targeted moments			
Fraction entrepreneurs	0.117	0.117	0.117
Transition rate W→E	0.021	0.021	0.021
Top 10% revenue share	0.799	0.804	0.796
Top 1% revenue share	0.522	0.515	0.524
Top 0.1% revenue share	0.282	0.284	0.283
RTS: Top 5% vs bottom 50% (by revenue)	0.083	0*	0.083
Capital-output ratio	2.970	2.970	2.971
B. Internally calibrated parameters			
Mean RTS	μ_η	0.683	0.782
Standard deviation RTS	σ_η	—	0.054
Standard deviation TFP	σ_z	0.910	0.614
Persistence TFP	ρ_z	0.970	0.954
Pareto tail TFP	ξ_z	2.880	—
Correlation (z, η)	$\sigma_{z,\eta}$	—	-0.262
Discount factor	β	0.902	0.890

Notes: Steady-state calibration of the (η, z)- and z-economy (both at $\lambda = 0.3$). * not targeted. Data moments are derived from Canadian data, and the RTS gap corresponds to our baseline estimation.

revenue distribution (rightmost column of Table IV). Specifically, we model η as a truncated normal AR(1) process in the interval $(0, 1)$ with parameters $(\mu_\eta, \sigma_\eta, \rho_\eta)$. We ex ante fix the autocorrelation to a high value of $\rho_\eta = 0.98$, which equals the persistence of RTS in our empirical analysis. The mean μ_η determines the fraction of entrepreneurs, while the cross-sectional standard deviation σ_η is closely linked to the difference in average RTS between the top 5% and the bottom 50% of firms, ordered by revenue. We also allow z and η to be correlated by setting log TFP $\ln z = \tilde{z} + \sigma_{\eta,z} \cdot \frac{\sigma_z}{\sigma_\eta} \cdot (\eta - \mu_\eta)$, where $\tilde{z}$ follows a normal AR(1) process with parameters $(\sigma_z, \rho_z, \mu_z = -\frac{\sigma_z^2}{2})$. Intuitively, both $\sigma_{\eta,z}$ and σ_z influence moments of the firm size distribution: If the empirical dispersion in RTS is small, a high residual TFP dispersion σ_z is needed to match the observed concentration of revenue among firms. Conversely, if the observed dispersion in RTS is large, the calibration would infer a more negative correlation parameter $\sigma_{\eta,z}$. Rather than directly using the estimated joint distribution of η and z in the model, we calibrate the TFP parameters residually in this manner. This approach is necessary because when firms operate under different production functions with varying η, the inferred relative TFPs are not comparable across firms.[23] This is the case in our model as well as in some of our empirical approaches (for instance, when we cluster firms such that firms within an industry do not share a common production function). In summary, we calibrate six parameters to match seven empirical moments. This model version also fits with the data perfectly. Notably, it does not require a Pareto tail in z to replicate the right tail of the firm-size distribution; the observed heterogeneity in RTS, combined with a log-normal z, is sufficient.

5.2.3 Quantitative Findings

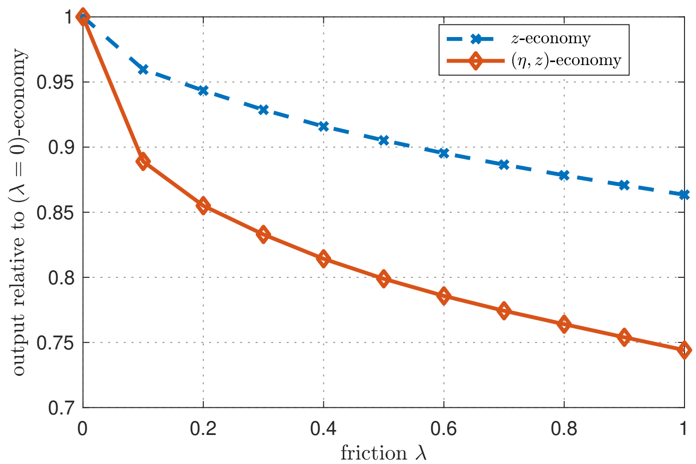
The two model economies are observationally equivalent in terms of the fraction of entrepreneurs, the persistence of entrepreneurship, the firm-size distribution, and the ratio of wealth (capital) to output. We now evaluate the efficiency losses associated with the same financial friction in both economies. Figure 9 compares the output losses induced by increasing the financial friction parameter λ from the unconstrained case of $\lambda = 0$ up to $\lambda = 1$, across stationary equilibria of the two models. For example, if entrepreneurs need to use 30 cents of their own wealth to finance each dollar of input expenditure ($\lambda = 0.3$), the (η, z)-economy—with heterogeneity in both TFP and RTS, disciplined by our empirical estimates—features an output loss of 18.3 log points relative to the frictionless case. In contrast, the conventional z-economy, which imposes homogeneous RTS, incurs a significantly smaller output loss of 7.4 log points. Thus, incorporating realistic heterogeneity in RTS, while otherwise matching

[23] To see this, consider a simple example of two firms, $j = 1, 2$, that differ in their RTS ($\eta_1 > \eta_2$) and TFP. Assume their production function is given by $y_j = z_j \cdot \ell_j^{\eta_j}$, where RTS (a unit-free elasticity), as well as input and output levels, are known. Then, the ratio of their measured TFP is given by

$$\frac{z_1}{z_2} = \frac{y_1}{y_2} \cdot \left(\frac{\ell_1}{\ell_2}\right)^{-\eta_1} \cdot \underbrace{\frac{1}{\ell_2^{\eta_1 - \eta_2}}}_{\text{unit dependence}},$$

which depends on the level of the input ℓ and, therefore, on the unit of measurement. In particular, the relative TFP of the higher-RTS firm is inversely proportional to the unit of measurement. Therefore, depending on the choice of unit (e.g., hours vs. full-time equivalents), one can find any relationship (in both sign and magnitude) between the TFPs of these two firms using the same data. As a result, when firms operate different production functions with varying RTS, relative TFP lacks the usual cardinal interpretation. For a similar discussion on unit dependence in the context of house price elasticities, see Greaney (2019).

Figure 9 – Output losses from financial friction in dynamic model



Notes: The figure plots output as a function of the financial friction λ, for both the (z)- and the (η, z)-economy. Output in both cases is normalized to one at $\lambda = 0$.

the same observables, amplifies the output losses due to financial frictions by 147%.

To understand this finding, we decompose the total log output loss into three terms: (i) static misallocation of production factors, holding fixed occupational choice, (ii) misallocation of talent across occupations, and (iii) under-accumulation of capital. Panel A of Table V shows that static misallocation of production factors across firms contributes 10.6 log points in the (η, z)-economy—more than half of the total GDP loss and twice as much as in the conventional z-economy. This is the channel highlighted in our analytical discussion in Section 5.1, and our quantitative findings are in line with Proposition 1: a given wedge distorts input choices more for high-η businesses. The dynamic setting magnifies this effect: Consider two hypothetical superstar entrepreneurs that are currently poor. The one distinguished by high productivity (high z) finds it easier to operate profitably at small scale, outgrowing the friction rather quickly. In contrast, the one distinguished by high scalability (high η) is less profitable at small scale, and struggles to outgrow the friction.

The majority of the remaining output loss is due to the under-accumulation of capital. Misallocation of talent across occupations also contributes slightly more to the output loss in the (η, z)-economy but remains relatively small in both economies. The λ-friction primarily misallocates production factors across firms rather than distorting the decision to become an entrepreneur. We chose a simple and transparent calibration strategy with a small number of parameters, deliberately avoiding addi-

Table V – Decomposition of output losses and other comparisons

A. Decomposition of output losses	*z-economy*	*(η, z)-economy*
Total log GDP loss	7.4	18.3
… due to misallocation of production factors	5.0	10.6
… due to misallocation of talent	0.5	0.6
… due to K accumulation	1.9	7.1
B. Alternative comparisons, total log GDP loss		
1. Equating aggregate debt/capital ratio	7.4	26.9
2. Equating dispersion in log marginal products	7.4	21.6

Notes: Panel A additively decomposes the total (steady-state) log GDP loss going from $\lambda = 0$ to $\lambda = 0.3$ into (i) misallocation of production factors (starting from the $\lambda = 0.3$ steady state, fixing K, L, and occupational status, allowing for efficient reallocation of K, L across firms); (ii) misallocation of talent (in addition allowing for efficient change of occupational status), and (iii) dynamic under accumulation of capital. Panel B reports the total GDP loss from the financial friction λ, in log points, in alternative scenarios. We raise λ from 0 to 0.3 in the z-economy, and from 0 to x in the (η, z)-economy, where x is chosen to match the debt ratio (row 1), respectively marginal input product dispersion (row 2), of the z-economy with $\lambda = 0.3$.

tional elements such as fixed costs of entry and exit that could magnify the importance of the occupational choice channel.

In the benchmark scenario, we increase λ from 0 to 0.3 in both economies. Panel B Table V shows that our results are even stronger when we instead equate observable moments, such as the aggregate debt-to-capital ratio or the dispersion in log marginal input products. For these exercises, we continue to raise λ from 0 to 0.3 in the z-economy, which generates an aggregate debt-to-capital ratio of 0.708 and a cross-sectional standard deviation of log marginal products of 0.144. We then adjust λ in the (η, z)-economy—raising it from 0 to 0.797 to replicate the debt ratio, or to 0.454 to match the marginal product dispersion. In these scenarios, the (η, z)-economy generates output losses that are $192 - 264\%$ larger than those in the z-economy.

Our findings are related to results in the macro-development literature (Buera *et al.* (2011); Midrigan and Xu (2014); Moll (2014)). A key quantitative finding in these studies is that misallocation losses due to financial frictions are relatively small when firms differ only in TFP but otherwise share the same, homothetic production technology. Output and efficiency costs are larger when taking into account technology choice. In particular, a choice between a high fixed cost, low marginal cost technology versus a low fixed cost, high marginal cost technology locally generates an increase in RTS across the firm size distribution. Our framework does not feature technology choice, and as such the entry margin contributes little to the output losses from financial frictions (as shown by the small contribution of the misallocation of entrepreneurial talent in Table V). However, static misallocation is greatly amplified when allowing for differences in RTS among existing firms.

5.2.4 Intermediate Inputs and Pre-determined Capital

Our baseline model deliberately adopts a streamlined entrepreneurial framework that omits several empirically relevant features of production, in order to isolate the quantitative importance of heterogeneity in RTS. Here, we show that while incorporating these richer features affects the overall level of misallocation—as documented in related literature—it does not overturn our main result: heterogeneity in RTS amplifies misallocation far beyond what is implied by TFP heterogeneity alone.

We first modify the production function in line with our empirical findings: $z \cdot k^{\alpha_K} \cdot \ell^{\alpha_L} \cdot m^{\eta-\alpha_K-\alpha_L}$, where α_K and α_L are fixed across firms, and η governs RTS. In this formulation, differences in RTS arise entirely from heterogeneity in intermediate input elasticities. Consistent with the data, larger firms have higher RTS due to greater use of intermediate inputs.

We consider three alternative model setups that differ in the formulation of the financial constraint, and in the timing of input choices. Our calibration strategy closely follows the baseline model. A detailed description of these extensions, along with their calibration and results, is provided in Appendix E.3.

Symmetric constraint on all inputs. First, we maintain that the financial constraint is symmetric across inputs: $w \cdot \ell + R \cdot k + m \leq \frac{a}{\lambda}$. Misallocation losses increase relative to the baseline—both with and without RTS heterogeneity—reflecting that the inclusion of intermediate inputs magnifies distortions (see, e.g., Baqaee and Farhi (2019)). Importantly, RTS heterogeneity continues to generate much larger misallocation losses from financial frictions: 46.3 vs. 9.3 log points, an amplification of 398%, compared to +112% in the baseline (row 2 of Table A.14 in Appendix E.3).

Flexible intermediates and a constraint on capital and labor only. Next, in line with our empirical approach, we treat intermediate inputs as fully flexible, and impose the financial constraint only on capital and labor: $w \cdot \ell + R \cdot k \leq \frac{a}{\lambda}$. Firms with higher η face relatively smaller effective financial frictions since the constraint applies to a smaller fraction of their inputs, weighted by factor elasticities ($\frac{\alpha_L+\alpha_K}{\eta}$). Compared to the baseline, two offsetting effects emerge: intermediates still amplify distortions, but the financial constraint becomes less restrictive. More importantly, RTS heterogeneity continues to magnify misallocation, with losses more than tripling (+214%) relative to the homogeneous RTS case (row 3 of Table A.14).

Pre-determined capital. Finally, we modify the timing of input choices by assuming that capital is chosen one period in advance—i.e., period t capital is predetermined in period $t-1$, prior to observing current shocks. With this added assumption, the model satisfies the empirical assumptions underlying the GNR estimation approach.[24] The overall level of misallocation caused by the financial friction λ is lower. This reflects findings in the literature that when input choices are risky, the financial constraint on inputs (λ) is secondary to the risk wedge as a source of misallocation (David *et al.* (2022); Boar *et al.* (2022)). Even so, the economy with RTS heterogeneity continues to exhibit misallocation losses from λ that are 81% higher as the one without RTS heterogeneity (row 4 of Table A.14).

Summary. Introducing intermediate inputs and varying the timing of input choices affects the overall level of misallocation in ways consistent with standard models. Nonetheless, across all these extensions, the key insight remains unchanged: heterogeneity in RTS significantly amplifies the misallocation induced by financial frictions—well beyond what is captured by TFP differences alone.

6 Conclusion

In this paper, we have documented significant heterogeneity in firms' scalability (RTS), even within narrowly defined industries. RTS heterogeneity is substantial,

[24] We have verified that, when using model-simulated data and estimating production functions using the GNR approach with sufficiently many clusters, we recover the true parameter values. Intermediate inputs are fully flexible, allowing the first stage of GNR to consistently estimate their elasticity. Capital is pre-determined, while labor is endogenous to current TFP. However, due to the persistence of wealth a and the structure of the financial constraint, labor inputs are autocorrelated—making lagged labor a valid instrument in the second stage. Clustering is necessary because firms with different η operate distinct production technologies.

highly persistent, and systematically related to firm size: larger firms tend to exhibit higher RTS. A significant portion of this heterogeneity is driven by persistent differences across firms, rather than by temporary factors or nonhomotheticities.

Accounting for RTS heterogeneity not only attenuates the positive correlation between TFP and firm size but also causes this relationship to break down for the largest firms. The largest firms are distinguished more by their high scalability than by their productivity levels. The positive relation between firm size and RTS is primarily driven by differences in the output elasticity of intermediate inputs, while labor and capital elasticities are jointly decreasing with firm size. We have also revisited some of the well-known empirical patterns around firm heterogeneity that were previously explained by differences in TFP. We find that high-RTS firms grow faster, are owned by wealthier households, and pay higher average wages.

The documented RTS heterogeneity has important implications for understanding the interaction between firm growth, the firm-size distribution, and the distributional impact of financial constraints and taxes, to note a few examples. To illustrate this, we employed an off-the-shelf quantitative model that incorporates firm heterogeneity not only in TFP—as in standard models of entrepreneurship and firm dynamics—but also in RTS. When large firms are characterized by high RTS—as we documented empirically—rather than by high TFP (the conventional view), the efficiency costs of financial frictions are significantly magnified. We provide intuition for this result in a static setting and then quantify the mechanism within a dynamic model. Our results show that the same financial friction generates over twice the efficiency and output costs in an economy with both RTS and TFP heterogeneity, compared to a conventional calibration that attributes all observed firm heterogeneity to TFP dispersion. These findings indicate that incorporating realistic RTS heterogeneity has important implications for related questions, including the optimal design of wealth and capital income taxation.

References

Abowd, J. M. and Card, D. (1989). On the covariance structure of earnings and hours changes. *Econometrica*, **57** (2), 411–445. 4.3.2

Ackerberg, D. A., Caves, K. and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, **83** (6), 2411–2451. 2.2.1

Argente, D., Moreira, S., Oberfield, E. and Venkateswaran, V. (2024). *Scalable Expertise*. Tech. rep. 7, 4.5.1

Atkeson, A. and Burstein, A. (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, **98** (5), 1998–2031. 4.3.1

—, Khan, A. and Ohanian, L. (1996). Are data on industry evolution and gross job turnover relevant for macroeconomics? In *Carnegie-Rochester Conference Series on Public Policy*, Elsevier, vol. 44, pp. 215–250. 5.1

Baldwin, J. R., Jarmin, R. S. and Tang, J. (2002). The trend to smaller producers in manufacturing: A canada/us comparison. *Statistics Canada, Analytical Studies-Economic Analysis, Series 1F0027MIE*, (003). 4.2

— and Rispoli, L. (2010). *Productivity Trends of Unincorporated Enterprises in the Canadian Economy, 1987 to 2005*. Statistics Canada. 11

Baqaee, D. R. and Farhi, E. (2019). Productivity and Misallocation in General Equilibrium*. *The Quarterly Journal of Economics*, **135** (1), 105–163. 5.2.4

Basu, S. and Fernald, J. G. (1997). Returns to scale in us production: Estimates and implications. *Journal of political economy*, **105** (2), 249–283. 13

Bloom, N., Floetotto, M., Jaimovich, N., Saporta-Eksten, I. and Terry, S. J. (2018). Really uncertain business cycles. *Econometrica*, **86** (3), 1031–1065. 3

Boar, C., Gorea, D. and Midrigan, V. (2022). *Why Are Returns to Private Business Wealth So Dispersed?* Tech. rep., National Bureau of Economic Research. 5.2.4

— and Midrigan, V. (2022). Should we tax capital income or wealth? 1, 5

Bond, S., Hashemi, A., Kaplan, G. and Zoch, P. (2021). Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data. *Journal of Monetary Economics*, **121**, 1–14. 4.3.1

Brown, C. and Medoff, J. (1989). The employer size-wage effect. *Journal of* *political Economy*, **97** (5), 1027–1059. 4.5.2

Buera, F. J., Kaboski, J. P. and Shin, Y. (2011). Finance and development: A tale of two sectors. *American Economic Review*, **101** (5), 1964–2002. 5.2.3

Burstein, A., Cravino, J. and Rojas, M. (2024). *Input price dispersion across buyers and misallocation*. Tech. rep., National Bureau of Economic Research. 4.3.1

Cagetti, M. and De Nardi, M. (2006). Entrepreneurship, frictions, and wealth. *Journal of political Economy*, **114** (5), 835–870. 1, 4.5, 5, 5.2

Chen, C., Habib, A. and Zhu, X. (2023). Finance, managerial inputs, and misallocation. *American Economic Review: Insights*, **5** (3), 409–26. 7

Chiavari, A. (2024). *Customer Accumulation, Returns to Scale, and Secular Trends*. Tech. rep., University of Oxford. 1

Clymo, A. and Rozsypal, F. (2023). *Firm cyclicality and financial frictions*. Tech. rep., Danmarks Nationalbank Working Papers. 5

David, J. M., Schmid, L. and Zeke, D. (2022). Risk-adjusted capital allocation and misallocation. *Journal of Financial Economics*, **145** (3), 684–705. 5.2.4

— and Venkateswaran, V. (2019). The sources of capital misallocation. *American Economic Review*, **109** (7), 2531–2567. 1

De Loecker, J., Eeckhout, J. and Unger, G. (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, **135** (2), 561–644. 4.3.1, 4.3.1, 4.3.1, 18, A.1

—, Goldberg, P. K., Khandelwal, A. K. and Pavcnik, N. (2016). Prices, markups, and trade reform. *Econometrica*, **84** (2), 445–510. 4, 4.3.1, 4.3.1, 18, A.1

— and Syverson, C. (2021). An industrial organization perspective on productivity. In *Handbook of industrial organization*, vol. 4, Elsevier, pp. 141–223. 4.1

— and Warzynski, F. (2012). Markups and firm-level export status. *American Economic Review*, **102** (6), 2437–71. 4.3.1, A.1

De Ridder, M., Grassi, B., Morzenti, G. *et al.* (2022). The hitchhiker's guide to markup estimation. 4.3.1

Demirer, M. (2020). Production function estimation with factor-augmenting technology: An application to markups. *Job Market Paper*. 1, 2.2.3, 2.2.3, 15, 4.3.1, 4.3.1, A.1, C.3

Edmond, C., Midrigan, V. and Xu, D. Y. (2023). How costly are markups? *Journal of Political Economy*, **131** (7), 1619–1675. 4.3.1

Foster, L., Haltiwanger, J. C. and Krizan, C. J. (2001). Aggregate productivity growth: Lessons from microeconomic evidence. In *New developments in productivity analysis*, University of Chicago Press, pp. 303–372. 3

Fox, J. T. and Smeets, V. (2011). Does input quality drive measured differences in firm productivity? *International Economic Review*, **52** (4), 961–989. 4.1

Gaillard, A. and Wangner, P. (2021). Wealth, returns, and taxation: A tale of two dependencies. *Available at SSRN*, **3966130**. 1

Gandhi, A., Navarro, S. and Rivers, D. A. (2020). On the identification of gross output production functions. *Journal of Political Economy*, **128** (8), 2973–3016. 1, 2.2.1, 15, A

Gao, W. and Kehrig, M. (2017). Returns to scale, productivity and competition: Empirical evidence from us manufacturing and construction establishments. *Productivity and Competition: Empirical Evidence from US Manufacturing and Construction Establishments (May 1, 2017)*. 1, 13

Gavazza, A., Mongey, S. and Violante, G. L. (2018). Aggregate recruiting intensity. *American Economic Review*, **108** (8), 2088–2127. 5

Greaney, B. (2019). Housing constraints and spatial misallocation: Comment. *Working paper*. 23

Guner, N., Ventura, G. and Xu, Y. (2008). Macroeconomic implications of size-dependent policies. *Review of economic Dynamics*, **11** (4), 721–744. 5.1

Guntin, R. and Kochen, F. (2025). *The Origins of Top Firms*. Tech. rep. 4.5.1

Guvenen, F., Kambourov, G., Kuruscu, B., Ocampo, S. and Chen, D. (2023). Use it or lose it: Efficiency and redistributional effects of wealth taxation. *The Quarterly Journal of Economics*. 1

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica: Journal of the Econometric Society*, pp. 1127–1150. 1, 4.3.2

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, **124** (4), 1403–1448. 1

— and Rossi-Hansberg, E. (2023). The industrial revolution in services. *Journal of Political Economy Macroeconomics*, **1** (1), 3–42. 7

Hurst, E. and Pugsley, B. W. (2011). *What do small businesses do?* Tech. rep., National Bureau of Economic Research. 4.5.1

Kalemli-Özcan, Ş., Sørensen, B. E., Villegas-Sanchez, C., Volosovych, V. and Yeşiltaş, S. (2024). How to construct nationally representative firm-level data from the orbis global database: New facts on smes and aggregate implications for industry concentration. *American Economic Journal: Macroeconomics*, **16** (2), 353–374. 12, C.3

Karahan, F. and Ozkan, S. (2013). On the persistence of income shocks over the life cycle: Evidence, theory, and implications. *Review of Economic Dynamics*, **16** (3), 452–476. 4.3.2

Kline, P. M. (2024). *Firm Wage Effects*. Working Paper 33084, National Bureau of Economic Research. 4.5

Lashkari, D., Bauer, A. and Boussard, J. (2024). Information technology and returns to scale. *American Economic Review*, **114** (6), 1769–1815. 7

Leung, D., Meh, C. and Terajima, Y. (2008). Productivity in canada: Does firm size matter? *Bank of Canada Review*, **2008** (Autumn), 7–16. 4.2, 4.4

Lucas, R. E. (1978). On the size distribution of business firms. *The Bell Journal of Economics*, **9** (2), 508–523. 1, 4.3.2

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, **71** (6), 1695–1725. 1

Mertens, M. and Schoefer, B. (2024). *From Labor to Intermediates: Firm Growth, Input Substitution, and Monopsony*. Working Paper 33172, National Bureau of Economic Research. 1, 4.2

Midrigan, V. and Xu, D. Y. (2014). Finance and misallocation: Evidence from plant-level data. *American Economic Review*, **104** (2), 422–58. 5.2.3

Moll, B. (2014). Productivity losses from financial frictions: Can self-financing undo capital misallocation? *American Economic Review*, **104** (10), 3186–3221. 5.2.3

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, **64** (6), 1263–1297. 2.2.1

Quadrini, V. (2000). Entrepreneurship, saving, and social mobility. *Review of Economic Dynamics*, **3** (1), 1–40. 1, 4.5, 5, 5.2

Restuccia, D. and Rogerson, R. (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic dynamics*, **11** (4), 707–720. 1

Ruzic, D. and Ho, S.-J. (2023). Returns to scale, productivity, measurement, and trends in us manufacturing misallocation. *Review of Economics and Statistics*, **105** (5), 1287–1303. 13

Smirnyagin, V. (2023). Returns to scale, firm entry, and the business cycle. *Journal of Monetary Economics*, **134**, 118–134. 4.5.1, 5

Sterk, V., Sedláček, P. and Pugsley, B. (2021). The nature of firm growth. *American Economic Review*, **111** (2), 547–579. 4.3.2, 4.5

Syverson, C. (2011). What determines productivity? *Journal of Economic literature*, **49** (2), 326–365. 1, 4.1

Appendix for "Scalable versus Productive Technologies"

Joachim Hubmer ⓡ Mons Chan ⓡ Serdar Ozkan

ⓡ Sergio Salgado ⓡ Guangbin Hong

A Details of Our GNR Method Implementation

We introduce our benchmark technique in detail, which closely follows Gandhi *et al.* (2020). We assume that output Y_{jt} of firm j in year t is produced using the firm's capital stock K_{jt}, labor input L_{jt}, and intermediate inputs M_{jt}, in the following way:

Assumption 1. *The firm's production function takes the following general form in levels* $Y_{jt} = F(K_{jt}, L_{jt}, M_{jt})e^{\nu_{jt}}$ *and in logs* $y_{jt} = f(k_{jt}, \ell_{jt}, m_{jt}) + \nu_{jt}$ *where* f *is a continuous and differentiable function which is strictly concave in* m_{jt} *and* ν_{jt} *is Hicks-neutral productivity.*

The traditional challenge in the production function estimation literature is separating productivity shocks that influence a firm's output from its input choices. To address this challenge, we leverage the firm's first-order conditions (FOC) and make timing assumptions regarding the nature of productivity and input choices to form moment conditions. We illustrate the details below.

Define $\mathcal{I}_{jt}$ as the information set available to firm j when it enters period t. The set $\mathcal{I}_{jt}$ includes all relevant information (e.g., firm productivity, current capital stock, and so on) that the firm uses to make its period-t decisions. We define any input $X_t \in \mathcal{I}_{jt}$ as *predetermined*. Predetermined inputs are thus functions of the previous period's information set, $X_t(\mathcal{I}_{jt-1})$. We treat capital as a predetermined input. Inputs that are not predetermined (i.e., those chosen in period t) are defined as *variable*. If the optimal choice of a variable input X_t depends on its own lagged values X_{t-1}, we refer to it as *dynamic* input. We depart from GNR by allowing labor to be a dynamic input. Finally, we define an input that is variable but not dynamic as *flexible*. Intermediate inputs are treated as flexible in our framework. As a result, both K_{jt} and $L_{j,t-1}$ are elements of $\mathcal{I}_{jt}$, but L_{jt} and M_{jt} are not.

Assumption 2. *Capital ($K_{jt} \in \mathcal{I}_{jt}$) is predetermined and a state variable. Labor input ($L_{jt} \notin \mathcal{I}_{jt}$) is dynamic, such that $L_{jt-1} \in \mathcal{I}_{jt}$ is a state variable. Intermediate inputs ($M_{jt} \notin \mathcal{I}_{jt}$) are flexible, so that $M_{jt-1} \notin \mathcal{I}_{jt}$.*

The Hicks-neutral productivity term ν_{jt} is composed of two components: (1) a persistent component, ω_{jt}, which is known to the firm when it makes input decisions, and (2) a transitory component, ε_{jt}, which is unknown to the firm when making input decisions in period t. Changes in these productivity terms may arise from both technology shocks and market demand shifts, while the transitory component may also reflect measurement error in output.

Assumption 3. *The persistent productivity component, $\omega_{jt} \in \mathcal{I}_{jt}$, is observed by the firm prior to making period-t decisions and is first-order Markov, such that $\mathbb{E}[\omega_{jt}|\mathcal{I}_{jt-1}] = \mathbb{E}[\omega_{jt}|\omega_{jt-1}] = h(\omega_{jt-1})$ for some continuous function $h(.)$. The transitory productivity innovation, $\varepsilon_{jt} \notin \mathcal{I}_{jt}$, is i.i.d. across firms and time with $\mathbb{E}[\varepsilon_{jt}] = 0$ and is not observed by the firm prior to period-t decisions, with $P_\varepsilon(\varepsilon_{jt}|\mathcal{I}_{jt}) = P_\varepsilon(\varepsilon_{jt})$.*

Assumption 4. *We assume that demand for intermediate input $m_{jt} = M(k_{jt}, \ell_{jt}, \omega_{jt})$ is strictly monotone in ω_{jt}.*

Note that this intermediate input demand function (conditional on period-t labor and capital inputs) is critical in identifying the production function while allowing labor to be a dynamic (and not predetermined) input. We also make the following assumption about the firm's profit-maximizing behavior and environment:

Assumption 5. *Firms maximize short-run expected profits and are price takers in both output and intermediate input markets. Denote the common output price index for period t as P_t and the common intermediate price index as ρ_t.*

Assumptions 1 to 5 give us the FOC for the firm's profit maximization problem in period t with respect to M_{jt}, $P_t \frac{\partial}{\partial M_{jt}} F(K_{jt}, L_{jt}, M_{jt}) e^{\omega_{jt}} \mathcal{E} = \rho_t$, where $\mathcal{E} \equiv \mathbb{E}[e^{\varepsilon_{jt}}]$

is a constant. Our first estimating equation is provided by multiplying both sides by M_{jt}/Y_{jt}, plugging in the production function, and rearranging the above FOC:

$$s_{jt} = \ln \mathcal{E} + \ln D(k_{jt}, \ell_{jt}, m_{jt}) - \varepsilon_{jt} \equiv \ln(D^{\mathcal{E}}(k_{jt}, \ell_{jt}, m_{jt})) - \varepsilon_{jt}, \tag{1}$$

where $s_{jt} \equiv \ln(\rho_t M_{jt}/P_t Y_{jt})$ is the log revenue share of intermediate input expenditure and $D(k_{jt}, \ell_{jt}, m_{jt}) \equiv \frac{\partial}{\partial m_{jt}} f(k_{jt}, \ell_{jt}, m_{jt})$ is the output elasticity of intermediate inputs. Since we assume $\mathbb{E}[\varepsilon_{jt}] = 0$, we can use equation 1 to identify ε_{jt} and $D^{\mathcal{E}}$.

Given that $\varepsilon_{jt} = \ln\left(D^{\mathcal{E}}(k_{jt}, \ell_{jt}, m_{jt})\right) - s_{jt}$, we can identify the constant $\mathcal{E}$, which subsequently provides the elasticity $D(k_{jt}, \ell_{jt}, m_{jt}) = D^{\mathcal{E}}(k_{jt}, \ell_{jt}, m_{jt})/\mathcal{E}$. Once we know $D(k_{jt}, \ell_{jt}, m_{jt})$ and ε_{jt}, we can integrate the elasticity up to estimate the rest of the production function nonparametrically.[1] In particular, we have

$$\mathcal{D}(k_{jt}, \ell_{jt}, m_{jt}) \equiv \int \frac{\partial}{\partial m_{jt}} f(k_{jt}, \ell_{jt}, m_{jt}) dm_{jt} = f(k_{jt}, \ell_{jt}, m_{jt}) - \mathit{\Psi}(k_{jt}, \ell_{jt}), \tag{2}$$

where $\Psi(k_{jt}, \ell_{jt})$ is the constant of integration (the component of the production function unrelated to m_{jt}). We can then define the residual output as $\tilde{y}_{jt} \equiv y_{jt} - \varepsilon_{jt} - \mathcal{D}(k_{jt}, \ell_{jt}, m_{jt}) = \mathit{\Psi}(k_{jt}, \ell_{jt}) + \omega_{jt}$. Plugging in the structure of ω_{jt} from Assumption 3 and defining $\xi_{jt} = \omega_{jt} - \mathbb{E}[\omega_{jt}|\omega_{jt-1}]$, we get our second estimating equation,

$$\tilde{y}_{jt} = \Psi\left(k_{jt}, \ell_{jt}\right) + h\left(\tilde{y}_{jt-1} - \Psi\left(k_{jt-1}, \ell_{jt-1}\right)\right) + \xi_{jt}, \tag{3}$$

where $\tilde{y}_{jt}$ is observable given the first-stage estimates of ε_{jt} and $\mathcal{D}(k_{jt}, \ell_{jt}, m_{jt})$. Our assumptions on the firm's information set give us $\mathbb{E}[\xi_{jt}|k_{jt}, \ell_{jt-1}, k_{jt-1}, \tilde{y}_{jt-1}, \ell_{jt-2}] = 0$ (i.e., $\mathbb{E}[\xi_{jt}|\mathcal{I}_{jt-1}] = 0$), which we use with equation 3 to identify Ψ, h, and thus ξ_{jt}.

The estimation procedure uses a standard sieve-series estimator to nonparametrically identify the output elasticities and production function. We proceed in two steps. First, we estimate equation 1 with a complete second-degree polynomial in k_{jt}, ℓ_{jt}, and m_{jt} using nonlinear least squares. This estimator solves

$$\min_{\gamma'} \sum_{j,t} \varepsilon_{jt}^2 = \sum_{j,t} \left[s_{jt} - \ln\left(\sum_{r_k + r_\ell + r_m \leq 2} \gamma'_{r_k, r_\ell, r_m} k_{jt}^{r_k} \ell_{jt}^{r_\ell} m_{jt}^{r_m} \right) \right]^2, \tag{4}$$

[1] We need one more technical assumption (Assumption 5 in GNR) on the support of (k_{jt}, ℓ_{jt}).

which gives us estimates of $\hat{\varepsilon}_{jt}$ and $\widehat{D^{\mathcal{E}}}(k_{jt}, \ell_{jt}, m_{jt}) = \sum_{r_k+r_\ell+r_m \leq 2}(\hat{\gamma}'_{r_k,r_\ell,r_m} k_{jt}^{r_k} \ell_{jt}^{r_\ell} m_{jt}^{r_m})$. We can then recover $\widehat{\mathcal{E}} = \mathbb{E}[e^{\hat{\varepsilon}_{jt}}]$ and the input elasticity

$$\widehat{D}(k_{jt}, \ell_{jt}, m_{jt}) = \sum_{r_k+r_\ell+r_m \leq 2} \left(\hat{\gamma}_{r_k,r_\ell,r_m} k_{jt}^{r_k} \ell_{jt}^{r_\ell} m_{jt}^{r_m}\right),$$

where $\hat{\gamma} \equiv \hat{\gamma}'/\widehat{\mathcal{E}}$. We then integrate the estimated flexible input elasticity to recover

$$\widehat{\mathcal{D}}(k_{jt}, \ell_{jt}, m_{jt}) = \sum_{r_k+r_\ell+r_m \leq 2} \left(\frac{m_{jt}}{r_m+1} \hat{\gamma}_{r_k,r_\ell,r_m} k_{jt}^{r_k} \ell_{jt}^{r_\ell} m_{jt}^{r_m}\right),$$

which allows us to recover $\hat{\tilde{y}}_{jt} = y_{jt} - \hat{\varepsilon}_{jt} - \widehat{\mathcal{D}}(k_{jt}, \ell_{jt}, m_{jt})$, that is, the component of output unrelated to variation in intermediate inputs.

In the second step, we estimate equation 3 using GMM, by approximating $\Psi(k_{jt}, \ell_{jt})$ and $h(\omega_{jt-1})$ using complete (separate) second- and third-degree polynomials, respectively. Since we can identify both $\Psi(k_{jt}, \ell_{jt})$ and TFP only up to an additive constant, Ψ is normalized to have mean zero, which implies that any fixed component of $\Psi(k_{jt}, \ell_{jt})$ will show up in the firm productivity level. This gives us the following second-stage estimating equation:

$$\tilde{y}_{jt} = -\sum_{0<\tau_k+\tau_\ell \leq 2} \alpha_{\tau_k,\tau_\ell} k_{jt}^{\tau_k} \ell_{tj}^{\tau_\ell} + \sum_{0 \leq a \leq 2} \delta_a \left(\tilde{y}_{jt-1} + \sum_{0<\tau_k+\tau_\ell \leq 2} \alpha_{\tau_k,\tau_\ell} k_{jt-1}^{\tau_k} \ell_{tj-1}^{\tau_\ell}\right)^a + \xi_{jt}, \quad (5)$$

where a is the degree of the polynomial. Since $E[\xi_{jt}|k_{jt}, \ell_{jt-1}, \mathcal{I}_{jt-1}] = 0$, the only endogenous variable is ℓ_{jt}. Thus, we can use functions of the set $\{k_{jt}, k_{jt-1}, \ell_{jt-1}, m_{jt-1}, \tilde{y}_{jt-1}\}$ as instruments. In particular, our moments are $E[\xi_{jt}\tilde{y}_{jt-1}^a]$ and $E[\xi_{jt} k_{jt}^{\tau_k} \ell_{jt-1}^{\tau_\ell}]$ for all $0 \leq a \leq 2$ and $0 < \tau_k + \tau_\ell \leq 2$, leaving us exactly identified.[2] This provides us with estimates of the production function as well as $\hat{\omega}_{jt}$, $\hat{\xi}_{jt}$, and $\hat{\bar{\omega}}_{jt} \equiv \hat{h}(\hat{\omega}_{jt-1})$. We then obtain the firm-level measure of RTS as sum of the output elasticities of capital and labor, combined with the previously estimated intermediate input elasticity: $\eta_{jt} \equiv \eta(k_{jt}, \ell_{jt}, m_{jt}) = \varepsilon_K^Y(k_{jt}, \ell_{jt}, m_{jt}) + \varepsilon_L^Y(k_{jt}, \ell_{jt}, m_{jt}) + \varepsilon_M^Y(k_{jt}, \ell_{jt}, m_{jt})$.[3]

[2]As pointed out by GNR, this implies that the estimator is a sieve-M estimator, which allows us to treat the polynomials as if they were the true parametric structure.

[3]While the notation in this section assumes a common production function for all firms, in practice we allow the production function to vary across different groupings, such as two-digit NAICS industries and clusters of firms with similar combinations of inputs and output.

A.1 Controlling for Market Power

We partially extend the GNR approach to control for variation in firm-level markups by estimating a modified first-step revenue share equation as follows. Relaxing the perfect competition assumption 5, we allow firms to face a downward-sloping demand curve, so that $\frac{\partial P_{jt}}{\partial Y_{jt}} < 0$. The FOC for intermediate inputs (Equation 1) then becomes $s_{jt} = \ln \mathcal{E} + \ln D(k_{jt}, \ell_{jt}, m_{jt}) - \ln \mu^p - \varepsilon_{jt}$, where $\mu^p = \frac{\varepsilon_P^Y}{\varepsilon_P^Y - 1}$ is the firm's price markup over marginal costs. Following De Loecker *et al.* (2020) and De Loecker *et al.* (2016), we use functions of firms' output market shares to proxy for unobserved price elasticities (ε_P^Y). In particular, we use a cubic function of market shares (defined at the two-digit NAICS level). Since period-t market shares may be correlated with transitory productivity shocks, we then estimate the modified first-stage equation with GMM using lagged market shares as instruments for current shares. This allows us to recover the output elasticity of intermediate inputs while controlling for market power, though the remaining output elasticities cannot be identified without price data or stronger parametric assumptions.

B Additional Figures and Tables

Table A.1 – Average Production Function Estimates by Industry

Industry	NAICS	N	RTS	M-elas	L-elas	K-elas
Agriculture	11	37,600	1.00	0.53	0.41	0.05
Mining	21	16,500	1.00	0.46	0.44	0.10
Energy	22	2,500	1.00	0.59	0.34	0.07
Construction	23	738,300	1.00	0.55	0.41	0.04
	31	69,100	1.01	0.61	0.37	0.03
Manufacturing	32	119,700	1.01	0.59	0.38	0.03
	33	247,100	1.00	0.55	0.42	0.03
Wholesale Trade	41	366,400	0.99	0.71	0.26	0.02
Retail Trade	44	614,400	1.00	0.75	0.22	0.02
	45	185,400	1.00	0.71	0.27	0.02
Transportation and warehousing	48	109,300	0.99	0.58	0.36	0.05
	49	13,300	1.01	0.63	0.33	0.04
Information and cultural	51	39,200	1.00	0.56	0.41	0.04
Finance and insurance	52	33,600	0.65	0.57	-0.05	0.13
Real estate	53	69,100	1.01	0.54	0.40	0.07
Professional Services	54	260,000	0.98	0.48	0.47	0.03
Management of companies and enterprises	55	27,700	1.03	0.59	0.39	0.05
Administrative and support	56	186,800	1.00	0.53	0.42	0.04
Education	61	26,700	0.98	0.51	0.45	0.03
Healthcare	62	111,300	0.59	0.40	0.05	0.14
Arts, entertainment and recreation	71	66,000	0.98	0.51	0.44	0.03
Accommodation and food services	72	552,500	0.99	0.59	0.37	0.04
Other Services	81	427,600	0.77	0.54	0.16	0.06

Notes: The numbers of observations are rounded to the nearest hundreds.

Table A.2 – Within-Industry Variance of Elasticity Estimates

	RTS	K-elasticity	L-elasticity	I-elasticity
Fraction of variation (variance) within industry				
Two-digit NAICS	23.3%	61.9%	65.9%	72.7%
Four-digit NAICS	22.0%	57.8%	58.6%	63.6%
Standard deviation within industry				
Two-digit NAICS	0.052	0.031	0.152	0.149
Four-digit NAICS	0.051	0.030	0.143	0.139

Notes: Table A.2 shows the within-industry variations for the three output elasticities and RTS estimates. It includes both the within-industry fraction of total variance and the within-industry standard deviation.

Table A.3 – Correlation of Output Elasticity Estimates

	Between-Industry Variation		Within-Industry Variation	
	Labor	Capital	Labor	Capital
Intermediates	-0.3	-0.7	-0.9	-0.4
Labor	1.0	-0.4	1.0	0.0

Notes: Table A.3 shows the correlation coefficients of the output elasticity estimates of the three inputs. The between-industry results show the weighted correlation of the average output elasticities of each two-digit NAICS industry, and the within-industry results demean the output elasticities at the two-digit NAICS level.

Table A.4 – Summary statistics for Manufacturing Firms

	Mean	Median	St.dev	P50-P10	P90-P50	P99-P50
Revenue	14.15	13.95	1.58	1.67	2.31	4.67
Intermediates	13.56	13.35	1.68	1.76	2.46	4.93
Labor	12.91	12.77	1.49	1.67	2.12	4.10
Capital	12.03	11.98	1.99	2.39	1.87	5.27

Notes: This table shows the moments of the distribution of revenues, intermediate inputs, labor, and capital stock in log real Canadian dollars for the Canadian manufacturing sector. The total number of observations is 436,000.

Table A.5 – Distribution of Production Function Parameters for Manufacturing Firms

	Mean	Median	St.dev	P50-P10	P90-P50	P99-P50
Returns to scale	1.00	1.00	0.02	0.02	0.02	0.07
		Output Elasticities				
Intermediates	0.57	0.56	0.14	0.16	0.18	0.37
Labor	0.40	0.41	0.13	0.17	0.15	0.28
Capital	0.03	0.03	0.03	0.03	0.04	0.09

Notes: This table shows the moments of the distribution of estimates for RTS and output elasticities for the Canadian manufacturing sector. The total number of observations is 436,000.

Table A.6 – Probit Regressions of Firm Exits

	(1)	(2)
RTS	–0.056***	–0.539***
	(0.002)	(0.013)
TFP $Percentile$	-0.020***	0.142***
	(0.001)	(0.002)
N	4.1M	3.4M
Constant	Y	Y
Industry FE	Y	Y
First difference		Y
Pseudo R2	0.010	0.018

Notes: The table reports two probit regressions of firm exit on RTS and within-industry TFP percentiles. To facilitate comparison, both regressors are standardized to have a mean of 0 and a standard deviation of 1. Firm exit is an indicator equal to 1 if a firm is present in the data in one year but not in the following year. Robust standard errors are clustered at the firm level. We first-difference both regressors in column (2). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.7 – High RTS Firms Respond More to Aggregate Shocks

Dependent Variable	Δy_{jt}					
	(1)	(2)	(3)	(4)	(5)	(6)
	Industry-level TFP shock			**Global Financial Crisis**		
$Shock_t$	-2.01***	-1.69***	-8.70***	0.02***	-0.02***	-0.57***
	(0.13)	(0.13)	(0.77)	(0.00)	(0.00)	(0.14)
$RTS_{j,t-1}$	0.02***	-0.28***	-0.28***	0.02***	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$RTS_{j,t-1} \times Shock_t$	4.58***	4.23***	4.46***	-0.02***	-0.02***	-0.01***
	(0.15)	(0.15)	(0.16)	(0.00)	(0.00)	(0.00)
Observations	3.6M	3.6M	3.6M	3.6M	3.6M	3.6M
Constant	Y	Y	Y	Y	Y	Y
Control:						
Revenue and Age		Y	Y		Y	Y
Revenue and Age $\times Shock_t$			Y			Y
R^2	0.01	0.05	0.05	0.00	0.00	0.05

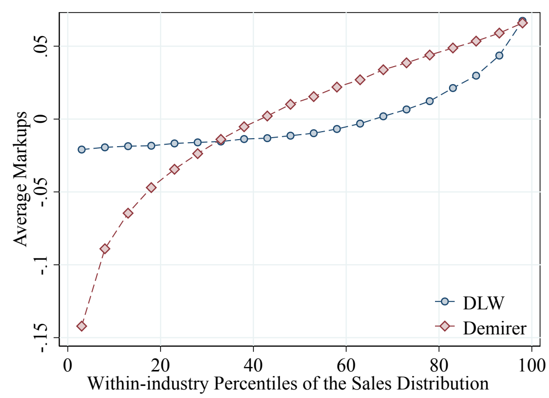
Notes: Robust standard errors are clustered at the firm level reported. In columns (1)-(3), we use the industry-level change in TFP as the aggregate shock, which is calculated as the average firm-level TFP, ν_{jt}, for all firms in the industry in that year. In columns (4)-(6), we use a time dummy for the 2007-2008 global financial crisis as the aggregate shock. We control for log revenue and log firm age and the interaction between the two in columns (2) and (5), and control for their interactions with the aggregate shock in columns (3) and (6). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.8 – Regression of Firm RTS on Size: Specification with Clustering by Size

Dependent Variable	RTS_{jt}	
	(1)	(2)
$\log Y_{jt}$	0.012***	-0.001***
	(0.000)	(0.000)
Observations	2.6M	2.6M
Constant	Y	Y
Industry FE	Y	Y
Cluster FE		Y
R^2	0.210	0.267

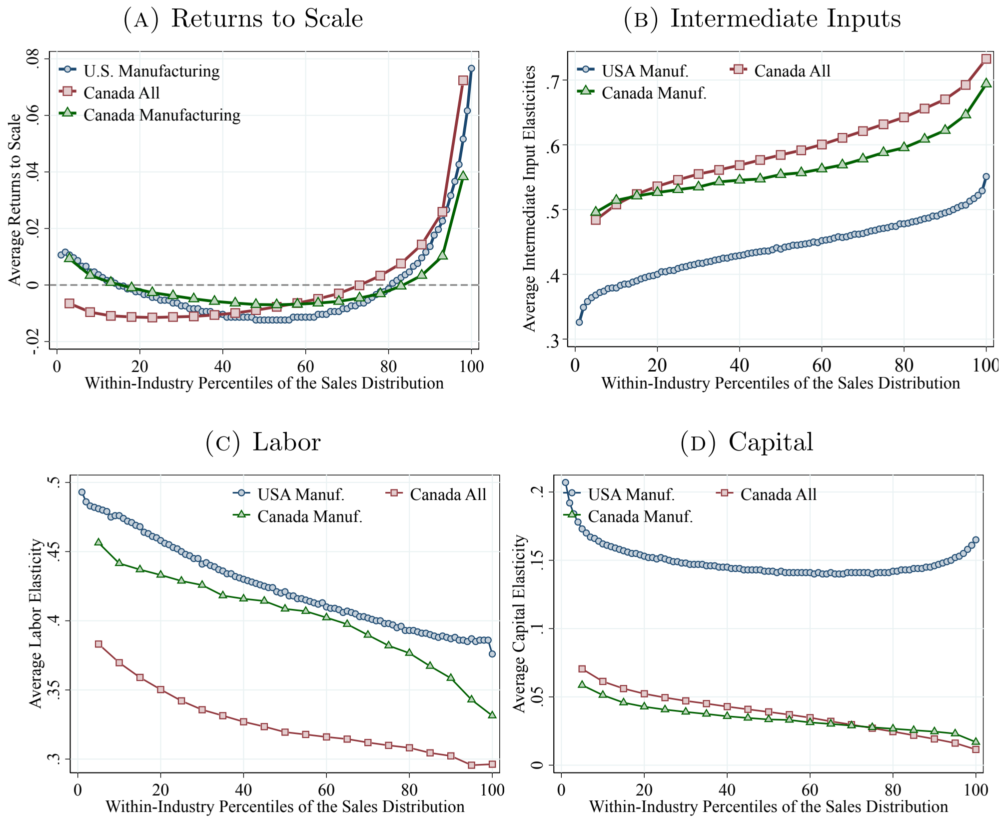
Notes: Table A.8 reports the regressions of firm RTS on log firm revenue at the firm-year level. Estimation results are from the specification where we cluster firms by the maximum attained size (see Section 4.3 for more details). Column (1) includes industry fixed effects, and column (2) further includes cluster fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Figure A.1 – Estimated Markups and Firm Revenue



Notes: Figure A.1 presents estimated markups across the firm-size distribution. We report estimates based on the value-added translog production function approach following De Loecker and Warzynski (2012) (DLW) and those obtained using the Demirer method (Demirer, 2020). In both cases, production functions are estimated separately by industry. Firms are sorted by their within-industry revenue ranks, and the figure plots the average markup within ranks. Markups are demeaned relative to the industry average.

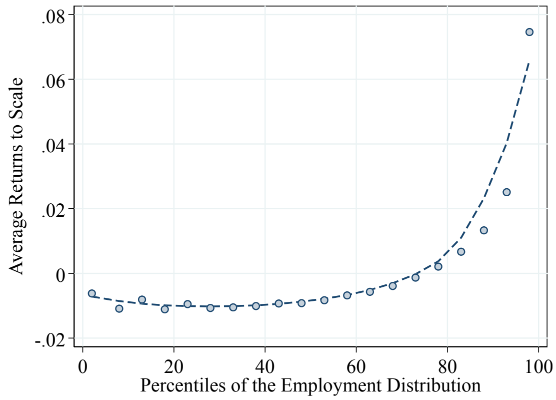
Figure A.2 – RTS and Output Elasticities for Canada and the US



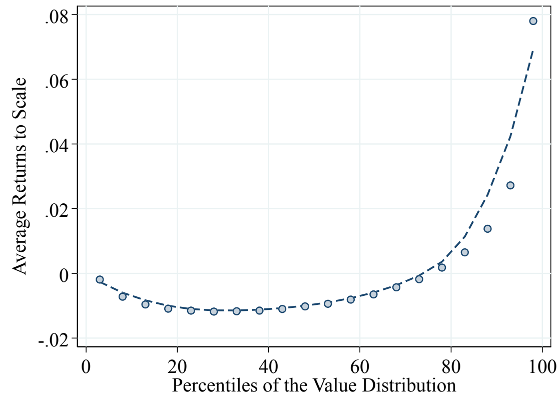
Notes: Figure A.2 shows returns to scale and output elasticities for the US manufacturing sector, for the Canadian private sector, and for the Canadian manufacturing sector. In all figures, we sort firms by within-industry revenue ranks and plot the average within ranks. Panel A shows the returns to scale relative to the industry average.

Figure A.3 – Results by Employment and Value Added

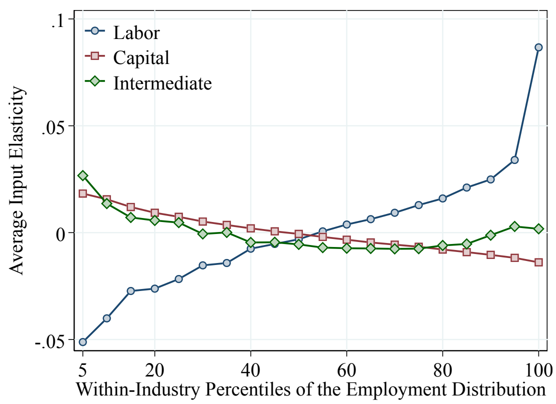
(a) RTS and Employment



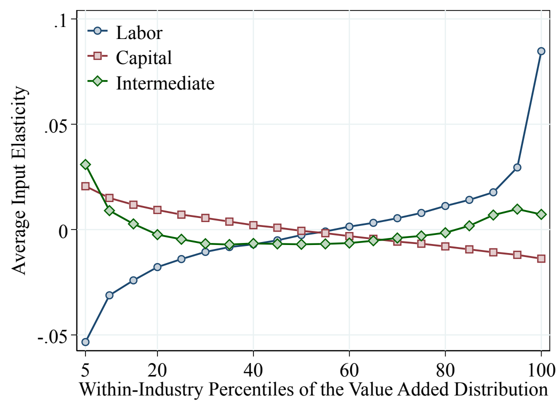
(b) RTS and Value Added



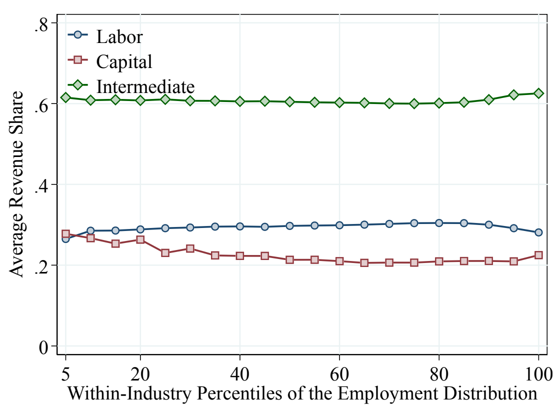
(c) Elasticities and Employment



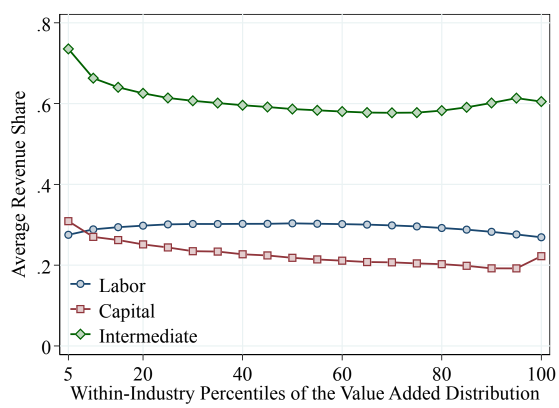
(d) Elasticities and Value Added



(e) Revenue Shares and Employment

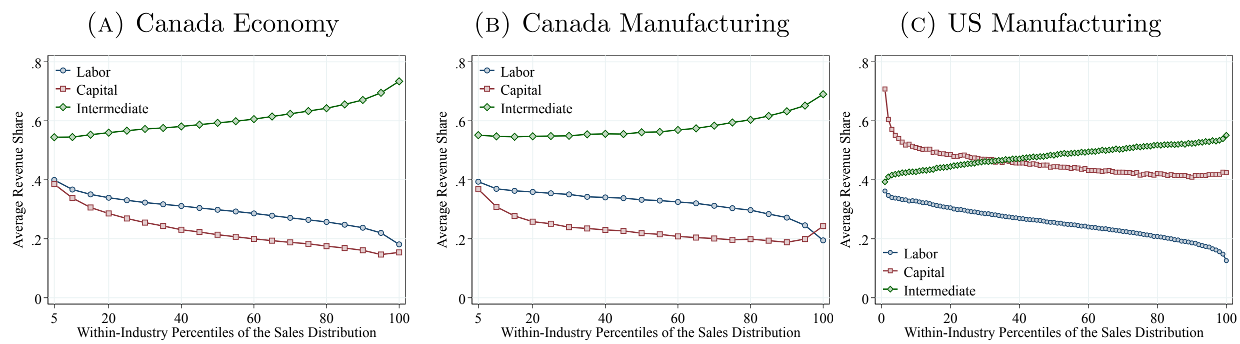


(f) Revenue Shares and Value Added



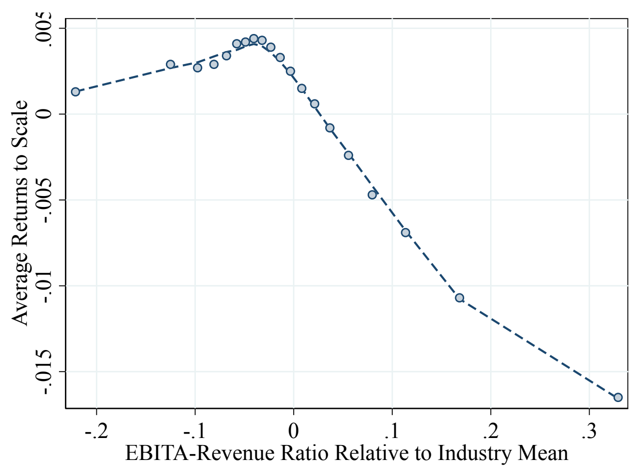
Notes: Figure A.3 shows results sorting firms by within-industry employment ranks (left panels) and within-industry value added ranks (right panels). We use the intermediate input and labor costs and the value of the capital stock to construct the revenue shares.

Figure A.4 – Input Revenue Shares across the Firm Revenue Distribution



Notes: Figure A.4 shows revenue shares across the firm-size distribution for Canada and for the US manufacturing sector. In each plot, we sort firms by within-industry revenue ranks and then average the revenue share across all firms within corresponding percentiles. We use the intermediate input and labor costs and the value of the capital stock to construct the revenue shares. Results for Canada are presented in ventiles.

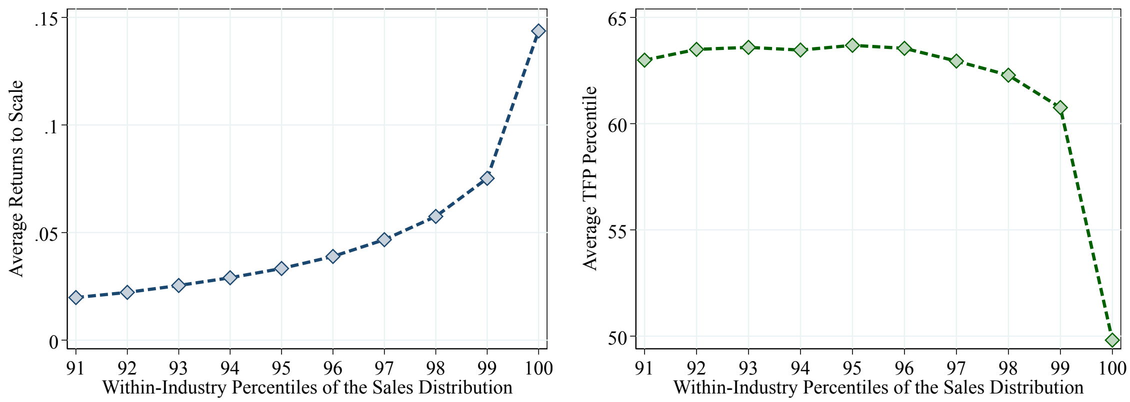
Figure A.5 – Profits and Returns to Scale



Notes: Figure A.5 plots the relationship between the returns to scale and the ratio of EBITA-revenue ratio. EBITA is computed as total revenue net of intermediate inputs and labor costs. Both variables are demeaned at the industry level.

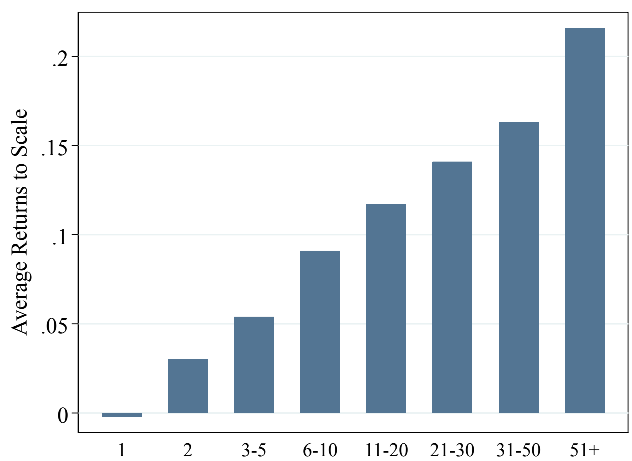
Figure A.6 – RTS and TFP Estimates for Top 10% Firms

(a) RTS　　(b) TFP Percentile



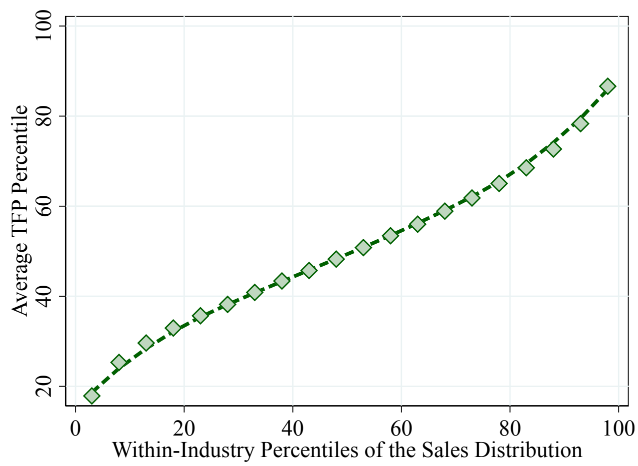
Notes: Figure A.6 plots the RTS and TFP estimates against the firm sales percentile for the top 10% firms. In both panels, we sort firms by within-industry revenue ranks and plot the average within ranks. Panel A shows the returns to scale relative to the industry average. Panel B shows the TFP percentile calculated within each industry.

Figure A.7 – RTS and the Number of Establishments



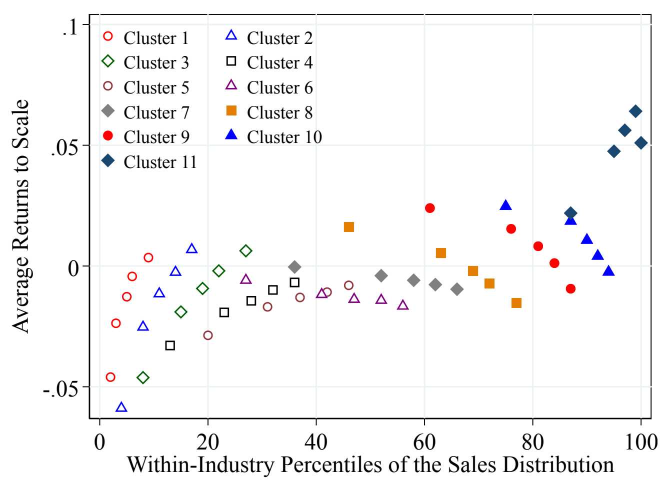
Notes: Figure A.7 plots the average RTS for eight groups of firms with a different number of establishments. RTS is demeaned at the industry level.

Figure A.8 – Robustness: TFP Percentile across the Firm Revenue Distribution, Cobb-Douglas Production Function



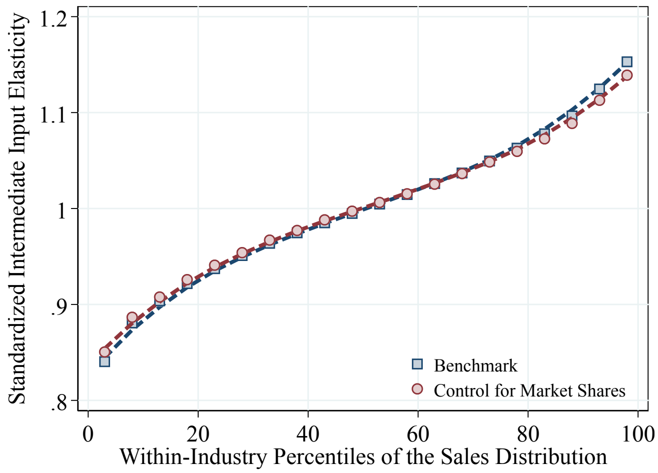
Notes: We re-estimate a Cobb-Douglas production function for each industry. We plot the relationship between TFP percentile and revenue percentile. Both TFP and revenue percentiles are calculated within industry.

Figure A.9 – Robustness: RTS across the Firm Revenue Distribution, Clustered by Maximum Size



Notes: Figure A.9 shows estimated average RTS when firms are clustered by maximum size. We cluster firms within each industry into 11 groups based on each firm's maximized within-industry-year revenue percentile throughout its life cycle. We exclude firms with fewer than 10 years of data and estimate the nonparametric production function separately for each cluster and industry. We pool all observations of firms that belong to the same cluster across industries. Then, we plot, for each cluster separately, the demeaned RTS against the within-industry revenue percentile. Each dot in the figure represents 20% of all the firm-year observations in one cluster.

Figure A.10 – Robustness: Estimation of Intermediate Input Elasticity, Controlling for Market Shares



Notes: Figure A.10 presents the intermediate input elasticity estimates from a specification that controls for firm market shares (as proxy for market power), compared to the benchmark estimates. Specifically, we run $s_{it} = \ln(D^{\mathcal{E}}(k_{jt}, \ell_{jt}, m_{jt})) + \tau^1 x^y_{it} + \tau^2 \left(x^y_{it}\right)^2 + \tau^3 \left(x^y_{it}\right)^3 - \varepsilon_{jt}$, where x^y_{it} represents firm i's revenue share in its industry at time t. We instrument the market share using its one-period lags. We note that the intercept coefficient of the regression contains information on both the average intermediate elasticity and the average markup, and we cannot separately identify these two components. We thus normalize the median intermediate elasticity to one for both versions of the estimates and plot the normalized elasticities across the firm-size distribution.

C Data Appendix

We describe the construction of variables and sample selection for our main dataset of Canadian corporations, as well as for the U.S. manufacturing data from the Census and the international firm-level data from Orbis.

C.1 Canadian Administrative Data

C.1.1 Variable Construction

Revenue We use the revenue measure that is computed by Statistics Canada for constructing the National Account. This measure is derived by summing up relevant terms from the T2 Corporate Income Tax Return Form.

Labor: We use the total worker compensation, which is also computed by Statistics Canada for constructing the National Account. This measure includes wages, salaries, and commissions paid to all the workers employed within a year.

Capital: We employ the perpetual-inventory method (PIM) to construct the capital stock. We make use of information on the first book value of tangible capital observed in the dataset, annual tangible capital investment, and amortization. Specifically, the capital stock K of firm i in year t is computed as $K_{i,t} = K_{i,t-1} + Invest_{i,t} - Amort_{i,t}, t > t_i^0$, where t_i^0 is the first year we observe the book value of the tangible capital of firm i. The initial year capital stock K is calculated as the book value of tangible capital net of accumulated tangible capital amortization. Tangible investment includes investments in building and land, computers, and machines and equipment. In addition, we construct a capital stock measure that includes intangible capital. We also follow the PIM for intangibles and make use of information on the book value of intangible capital, annual intangible capital investment, and amortization.

Intermediates: We measure intermediate inputs as the total expenses not related to capital and labor. Specifically, the measure is computed as the sum of operating expenses and costs of goods sold net of capital amortization. The operating expenses and costs of good sold variables are also constructed by Statistics Canada to replicate the National Account, and neither of them encompasses worker compensation.

Firm owner and wealth information: We obtain ownership information from the Schedule 50 Shareholder Information of T2 Corporate Tax Files. Schedule 50 provides information of the filing firms on their shareholders with at least 10% of shares, the percentage of shares owned by each shareholder, and the type of shares owned (common or preferred). Statistics Canada tracks chained ownership by individuals (e.g., individual A owns a share of firm B, and firm B owns a share of firm C) and constructs a tracked share of ownership of firms by each ultimate individual shareholder. We merge the ownership information with the firm panel dataset and calculate total individual equity wealth as the ownership share weighted sum of the value of all holding firms. Firm value is calculated as total assets net of total liabilities.

Linked employer-employee information: We obtain linked employer-employee and earnings information from the T4 Statement of Remuneration Paid form. The T4 files provide job-level earnings information with individual and firm identifiers, where a job is defined as a worker-firm pairing. A worker can have multiple T4 records in a year if she works for more than one firm. For multiple job holders, we keep the job that offers the highest earnings of the year and call it the main job. In addition, we drop workers with annual earnings from the main job that are lower than 5,000 CAD.

C.1.2 Sample Selection

Several steps are taken to construct the estimation sample. First, we drop firms with missing industry information. Second, we exclude the initial year in which a firm's book value of tangible capital is observed, along with all prior observations, as we cannot use the PIM to construct the capital stock for these observations. Third, we drop firm-year observations with missing and nonpositive revenue, labor, capital, and intermediate input values. We further drop the observations whose one-year lagged revenue or inputs are missing or non-positive, as our identification strategy requires using lagged labor input as the instrument. Fourth, we drop the observations with extreme factor shares, that is, the ones with a ratio of wage-bill-to-revenue below the 1st percentile or above the 99th percentile, with a ratio of wage bill-to-value-added below the 1st percentile or above the 99th percentile, with a ratio of intermediate-input-to-revenue above 0.95 or below 0.05, and with a ratio of capital-stock-to-revenue above the 99.9th percentile. This sample selection procedure leaves us with around 4.3 million firm-year observations. We convert all monetary variables to 2002 Canadian dollars.

C.2 US Census and Survey of Manufacturing

Here we describe the sample selection and moment construction using data from the US Census of Manufacturing (CM) and the Survey of Manufacturing firms (ASM). The CM, which is part of the Economic Census, is conducted every five years, in every year ending in 2 or 7, and was first implemented in 1963. It covers all establishments with at least one paid employee in the manufacturing sector (NAICS 31-33) for a total sample between 300,000 and 400,000 establishments per Census. Information is delivered by firms at the establishment level, and the Census provides a unique identifier (lbdnum) which we use to follow establishments over time. The CM provides information on Employment, Payroll, Value of Shipments, Costs of Material, and Inventories. It also provides information on investment in machinery, equipment, and structures. Furthermore, it contains information on the location of the establishment (state and county), and industry classification (NAICS).

The Census Bureau complements the CM data with the ASM every year the Economic Census is not conducted since 1973. Relative to the CM, the ASM is skewed towards large firms as it covers all establishments of firms considered by the CM above a certain threshold, and a smaller sample of small and medium sized firms. The number of observations in the raw data is around 50,000 establishments per year. The merged CM/ASM dataset contains consistent information on industry, sales, employment, capital expenditures, materials, and others. Beyond the information available in the CM, the ASM also contains information on R&D expenditures, and measures of capacity utilization, and capital investment, which is used by the Census to calculate the real value of capital stock using the PIM method.

We access the US Census information through the Census RDC. All results presented in this paper have been approved by the US Census and do not reveal any firm-level information. Our starting base is the panel data available in the ASM. We impose similar selection criteria as we do with the data from Canada. In particular, we select establishment-year observations with non missing values in real value of shipments (revenue), the real wage bill of workers in the establishment (labor), the real expenditure in intermediate inputs and materials (intermediates), and the real value of the capital stock (capital) which is calculated by the Census using PIM. All nominal values are deflated to 2018 prices. We then calculate the revenue shares of each of these components, and we drop observations below the 0.1 and above the 99.9th percentiles within each distribution. Finally, since our estimation method relies on lagged input values, we drop the first two observations of each establishment in our dataset. This sample selection generates a panel of 3.1 million establishment-year observations.

C.3 International Evidence from ORBIS

In this appendix, we provide additional details for the construction of our measure of firm-level TFP using data from Orbis. Moody's Orbis (formerly Bureau van Dijk's Orbis) is a large firm-level dataset providing harmonized information on private and public firms across several countries. It aggregates and standardizes data from thousands of sources—national registries, regulatory filings, rating agencies, and press releases—into a single dataset. In our analysis, we use information for European countries (the subsample called Amadeus) containing over 150 million public and private European companies. Our sample contains information from the early 1990 to 2019 with substantially better coverage starting in 2005. See Kalemli-Özcan *et al.* (2024) for additional details about constructing a representative dataset using Moody's Orbis data.

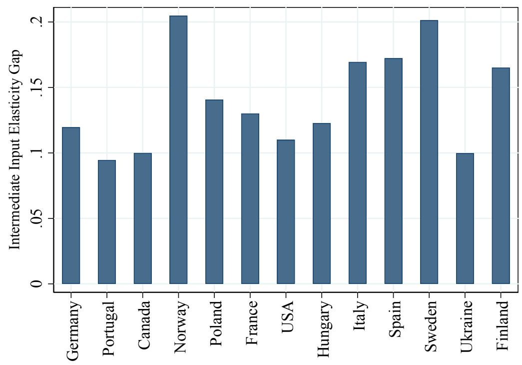
We consider 11 countries in our analysis including Finland, France, Germany, Hungary, Italy, Norway, Poland, Portugal, Spain, Sweden, and Ukraine, for which firm-level information is available for enough industries and sectors. For each country in the sample, we retrieve firm-level panel data from Amadeus through WRDS. Our data contains a large range of firms, from small to very large firms (the V+L+M+S: plus Small Companies dataset), both publicly traded and privately held. Revenues are measured by sales (TURN); if sales data are unavailable, we use operating revenues (OPRE). Intermediate input costs are captured by material expenses (MATE), while the value of the capital stock is taken from total fixed assets (FIAS). Labor costs are measured using the firm's wage bill, as reported in cost of employees (STAF). Firms are classified according to two-digit NAICS industry codes, and all financial variables denominated in local currency are converted to euros using the exchange rate provided by Orbis (EXCHANGE2).

In order to estimate firm-level productivity for a large number of firms within each country, we perform a simple sample selection. For each country, we drop duplicates, observations without information on industry (NAICS), and firms with discrepancies between the country identifier and the firm identifier (INDR). We also drop all observations with missing, zero, or negative values in either of the following variables: OPRE, MATE, FIAS, and STAF. All monetary values are transformed to Euros using the exchange rate supplied by Moody's and deflated by country-specific CPI to 2019 prices (obtained from the World Bank' s WDI database).

After sample selection, our sample contains about 16.9 million country-firm-year observations, with Spain (3.8M), France (4.3M), and Italy (3.5M) having the largest samples. Then, for each country, we estimate industry-level production functions (NAICS2 industries) using the GNR method and the method developed by Demirer (2020).

Table A.10 shows cross-sectional moments of the (log) revenue distribution, intermediate inputs, wage bill, and capital stock, all in real terms. Table A.10 shows unconditional cross-sectional moments of the distribution of revenue shares and output elasticity estimates within each country estimated using the GNR method. Similarly to our results based on administrative data from Canada and the US, there is significant within country-industry dispersion in revenue shares (columns 1 to 3) and therefore in input elasticities (columns 4 to 6). Column 7 shows cross-sectional moments of the distribution of RTS that displays large dispersion as well, with P90-P10 of around 8 percentage points across countries. This is similar to the within country-industry differences in RTS across the firm-size distribution as shown in the main text. Importantly, and similarly to our baseline results, the increase in RTS along the firm size distribution is driven by an increase in the output elasticity of intermediate inputs, as show in Figure A.11. Figure A.12 shows that these findings are robust to applying the Demirer method.

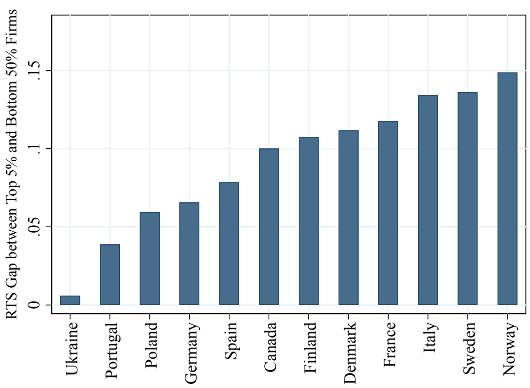
Figure A.11 – Intermediate Input Elasticity Increases with Firm Size



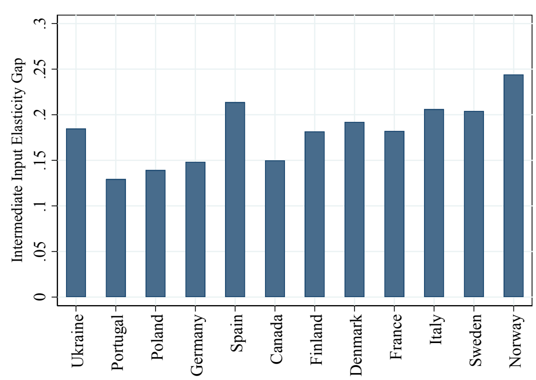
Notes: Figure A.11 shows the difference between the average intermediate input elasticity among firms in the top 5% of the within country-industry-year revenue distribution and the average at the bottom 50% percent. Results calculated using the GNR method.

Figure A.12 – RTS and Intermediate Input Elasticity Using Demirer (2020)

(a) RTS



(b) Intermediate Input Elasticity



Notes: Figure A.12 shows the difference between the average intermediate input elasticity among firms in the top 5% of the within country-industry-year revenue distribution and the average at the bottom 50% percent. Results calculated using the Demirer method.

Table A.9 – Cross-Sectional Moments of TFP Distribution by Country

Country	SD	P10	P50	P90
Germany	0.38	−0.37	−0.02	0.42
Denmark	0.39	−0.41	−0.01	0.45
Spain	0.34	−0.40	0.00	0.36
Finland	0.32	−0.33	0.00	0.34
France	0.33	−0.35	−0.01	0.37
Italy	0.44	−0.44	0.00	0.49
Norway	0.29	−0.29	0.00	0.28
Poland	0.43	−0.43	−0.01	0.47
Portugal	0.39	−0.46	0.00	0.45
Sweden	0.31	−0.26	0.01	0.30
Ukraine	0.63	−0.68	−0.03	0.76
Total	0.38	−0.40	−0.01	0.42

Notes: Table shows within-country cross-sectional moments of the TFP distribution calculated using GNR. TFP is estimated within each country-industry defined as two-digit NAICS. We then demean each distribution by country-industry and calculate within industry cross sectional moments. We then average across year-industries within a country. Total is the grand average across all countries and years.

Table A.11 – Input Shares and Elasticities Within Countries

			Revenue Shares			Elasticities and RTS			
Country	Stat.	Obs.	(1) Intermediate	(2) Labor	(3) Capital	(4) Intermediate	(5) Labor	(6) Capital	(7) RTS
Germany	Mean	271,202	0.45	0.24	0.25	0.42	0.47	0.06	0.96
	P10		0.13	0.06	0.01	0.14	0.21	0.01	0.88
	P50		0.43	0.22	0.08	0.39	0.49	0.05	0.96
	P90		0.79	0.46	0.58	0.74	0.70	0.12	1.00
Denmark	Mean	26,057	0.55	0.34	1.80	0.51	0.42	0.082	1.00
	P10		0.20	0.06	0.01	0.23	0.13	−0.05	0.82
	P50		0.57	0.27	0.11	0.50	0.42	0.03	1.00
	P90		0.87	0.67	6.20	0.83	0.73	0.32	1.20
Spain	Mean	3,833,412	0.48	0.31	0.69	0.45	0.46	0.04	0.95
	P10		0.15	0.08	0.01	0.16	0.21	0.01	0.85
	P50		0.48	0.27	0.15	0.43	0.47	0.04	0.95
	P90		0.79	0.57	1.30	0.76	0.69	0.07	1.00
Finland	Mean	563,541	0.37	0.29	0.32	0.34	0.50	0.08	0.91
	P10		0.10	0.08	0.01	0.11	0.25	0.01	0.78
	P50		0.33	0.27	0.09	0.30	0.52	0.07	0.92
	P90		0.69	0.52	0.62	0.63	0.71	0.14	1.00
France	Mean	4,389,186	0.39	0.28	0.14	0.37	0.49	0.06	0.92
	P10		0.12	0.09	0.01	0.14	0.28	0.02	0.82
	P50		0.36	0.27	0.05	0.33	0.51	0.05	0.92
	P90		0.70	0.49	0.28	0.63	0.69	0.10	1.00
Italy	Mean	3,492,067	0.43	0.21	0.52	0.39	0.43	0.05	0.87
	P10		0.12	0.04	0.01	0.14	0.21	0.01	0.73
	P50		0.41	0.18	0.09	0.36	0.45	0.05	0.89
	P90		0.77	0.42	0.86	0.69	0.63	0.10	1.00
Norway	Mean	575,877	0.42	0.30	0.30	0.40	0.48	0.05	0.93
	P10		0.12	0.09	0.01	0.13	0.26	0.00	0.82
	P50		0.41	0.29	0.05	0.37	0.50	0.04	0.94
	P90		0.74	0.54	0.54	0.70	0.69	0.10	1.00
Poland	Mean	546,413	0.48	0.17	0.92	0.44	0.43	0.06	0.93
	P10		0.10	0.03	0.01	0.12	0.19	0.00	0.79
	P50		0.49	0.12	0.12	0.41	0.43	0.05	0.95
	P90		0.84	0.38	1.80	0.79	0.67	0.15	1.10
Portugal	Mean	1,342,545	0.49	0.25	0.44	0.46	0.45	0.05	0.96
	P10		0.15	0.07	0.01	0.17	0.21	0.01	0.88
	P50		0.50	0.22	0.12	0.44	0.46	0.05	0.96
	P90		0.80	0.48	0.91	0.75	0.67	0.09	1.00
Sweden	Mean	949,905	0.41	0.30	0.27	0.39	0.48	0.05	0.91
	P10		0.13	0.09	0.00	0.14	0.26	0.01	0.78
	P50		0.39	0.29	0.05	0.36	0.49	0.04	0.92
	P90		0.71	0.53	0.60	0.66	0.68	0.10	1.10
Ukraine	Mean	394,734	0.40	0.30	1.90	0.30	0.60	0.10	1.00
	P10		0.10	0.00	0.00	0.10	0.30	0.00	0.80
	P50		0.40	0.20	0.20	0.30	0.60	0.10	1.00
	P90		0.80	0.60	2.40	0.60	0.80	0.20	1.10
Total	Mean	16,960,816	0.43	0.27	0.47	0.40	0.46	0.05	0.92
	P10		0.12	0.06	0.01	0.14	0.23	0.01	0.80
	P50		0.41	0.24	0.09	0.37	0.48	0.05	0.93
	P90		0.76	0.50	0.78	0.69	0.68	0.10	1.00

Notes: Table shows within-country cross-sectional moments of the corresponding distribution. Elasticities and returns to scale calculated using GNR applied within country-two digits NAICS industries. Firm-level revenue is either sales or operating revenue, if sales variables is missing.

Table A.10 – Firm-Level Distributional Statistics by Country in Orbis

Country	Stat	Revenue	Intermediates	Wage Bill	Capital Stock	Country	Stat	Revenue	Intermediates	Wage Bill	Capital Stock
	Mean	16.17	15.17	14.48	13.49		Mean	13.87	12.81	12.44	10.90
	Std. Dev.	2.021	2.250	2.005	2.874		Std. Dev.	1.633	1.970	1.690	2.188
	P10	13.64	12.33	12.06	9.977		P10	11.98	10.39	10.61	8.217
Germany	P50	16.14	15.11	14.40	13.43	Norway	P50	13.71	12.70	12.44	10.75
	P90	18.75	18.06	17.09	17.18		P90	15.98	15.38	14.42	13.65
	P99	20.96	20.36	19.11	19.54		P99	18.54	17.97	16.78	17.04
	Mean	14.67	13.93	13.21	12.57		Mean	14.07	13.10	11.83	11.90
	Std. Dev.	2.768	3.005	2.706	3.275		Std. Dev.	1.788	2.167	1.779	2.582
	P10	11.33	10.27	9.894	8.374		P10	11.74	10.13	9.590	8.387
Denmark	P50	14.32	13.47	12.99	12.40	Poland	P50	14.13	13.28	11.85	12.03
	P90	18.42	18.08	16.73	16.99		P90	16.25	15.71	14.03	15.16
	P99	20.56	20.18	18.76	19.74		P99	18.34	17.91	16.05	17.48
	Mean	13.08	12.17	11.64	11.07		Mean	12.50	11.62	10.86	10.27
	Std. Dev.	1.556	1.841	1.466	2.204		Std. Dev.	1.523	1.757	1.398	2.257
	P10	11.27	9.922	9.897	8.245		P10	10.77	9.512	9.216	7.351
Spain	P50	12.92	12.06	11.58	11.12	Portugal	P50	12.31	11.49	10.73	10.30
	P90	15.08	14.53	13.43	13.76		P90	14.48	13.89	12.64	13.05
	P99	17.57	17.08	15.74	16.21		P99	16.95	16.41	14.87	15.65
	Mean	13.14	11.92	11.65	10.65		Mean	13.29	12.21	11.85	10.28
	Std. Dev.	1.671	1.989	1.761	2.072		Std. Dev.	1.555	1.875	1.710	2.230
	P10	11.22	9.488	9.480	8.068		P10	11.49	9.885	9.922	7.522
Finland	P50	12.99	11.79	11.68	10.52	Sweden	P50	13.18	12.15	11.92	10.16
	P90	15.26	14.48	13.73	13.29		P90	15.26	14.60	13.81	13.20
	P99	18.01	17.34	16.23	16.25		P99	17.63	17.12	15.89	15.84
	Mean	13.30	12.18	11.86	10.24		Mean	11.67	10.50	10.04	10.14
	Std. Dev.	1.557	1.834	1.559	2.044		Std. Dev.	2.419	2.681	2.197	3.003
	P10	11.54	10.02	10.05	7.715		P10	8.721	7.138	7.421	6.314
France	P50	13.10	12.01	11.83	10.22	Ukraine	P50	11.41	10.33	9.779	10.10
	P90	15.32	14.54	13.74	12.71		P90	14.93	14.09	13.01	14.02
	P99	17.90	17.33	16.08	15.68		P99	17.66	16.90	15.48	17.05
	Mean	13.70	12.65	11.73	11.26		Mean	13.32	12.29	11.71	10.81
	Std. Dev.	1.619	1.926	1.865	2.325		Std. Dev.	1.722	1.995	1.740	2.323
	P10	11.77	10.26	9.654	8.385		P10	11.35	9.897	9.704	7.931
Italy	P50	13.60	12.59	11.83	11.14	Total	P50	13.17	12.17	11.70	10.72
	P90	15.75	15.10	13.84	14.31		P90	15.51	14.85	13.77	13.76
	P99	18.15	17.59	16.12	16.79		P99	18.21	17.63	16.37	16.70

Notes: Table shows within-country cross-sectional moments of the corresponding distribution. All variable in 2019 Euros. Firm-level revenue is either sales or operating revenue, if sales variables is missing.

D RTS Variance-Component Model

The RTS process has three components:

$$RTS_{ih} = \underbrace{\alpha_i}_{\text{permanent}} + \underbrace{z_{ih}}_{\text{AR(1)}} + \epsilon_{ih},$$

where $\alpha_i \sim N(0, \sigma_\alpha^2)$ is the fixed effect of firm i, $\epsilon_{ih} \sim N(0, \sigma_\epsilon^2)$ is a fully transitory i.i.d. shock at age h, and z_{ih} is a persistent component that follows the process

$$z_{ih} = \rho_z z_{i,h-1} + \eta_{ih}, \;\; z_{i,0} = 0.0,$$

where η_{ih} is an i.i.d. innovation with mean zero and variance σ_η^2. So, we estimate four parameters, $(\sigma_\alpha^2, \sigma_\eta^2, \rho, \sigma_\epsilon^2)$ by targeting the autovariance matrix of firm-level RTS. We compute the autocovariance matrix of RTS over the life cycle in levels in the data. We then estimate these parameters by minimizing the distance between empirical values and the corresponding simulated values. For this purpose we employ the multi-start global minimization algorithm, TikTak, which can be found at https://github.com/serdarozkan/TikTak.

Table A.12 – Parameter Estimates

σ_α^2	ρ	σ_η^2	σ_ϵ^2
0.001	0.937	0.00025	0.00027

σ_α	ρ	σ_η	σ_ϵ
0.0319	0.937	0.0158	0.0165

Variance decomposition			
RTS	α	ϵ	z
0.00257	0.001	0.00027	0.0013
1	38.9%	10.5%	50.6%

E Model Appendix

E.1 Proof of Proposition 1

Without loss of generality, set the productivity of the unconstrained (CRTS) sector to 1. Then, the equilibrium input price equals 1. Given $\tau \geq 0$, the input choice and output of constrained firm i are, respectively:

$$x_i(\tau) = \left(\frac{\eta_i \cdot z_i}{1+\tau}\right)^{\frac{1}{1-\eta_i}} \quad \text{and} \quad y_i(\tau) = z_i^{\frac{1}{1-\eta_i}} \cdot \left(\frac{\eta_i}{1+\tau}\right)^{\frac{\eta_i}{1-\eta_i}}.$$

By market clearing, the aggregate input and output of unconstrained firms both equal

$$1 - \int_0^\chi x_i(\tau) di.$$

Thus, we can write the aggregate misallocation loss as

$$\begin{aligned}
\Delta Y(\tau) = Y^\star - Y(\tau) &= \int_0^\chi (y_i(0) - y_i(\tau))\, di - \left(\int_0^\chi x_i(0) di - \int_0^\chi x_i(\tau) di\right) \\
&= \int_0^\chi (y_i(0) - y_i(\tau)) - (x_i(0) - x_i(\tau))\, di \\
&= \int_0^\chi y_i^\star \cdot \underbrace{\left[\left(1 - \left(\frac{1}{1+\tau}\right)^{\frac{\eta_i}{1-\eta_i}}\right) - \eta \cdot \left(1 - \left(\frac{1}{1+\tau}\right)^{\frac{1}{1-\eta_i}}\right)\right]}_{\equiv L_i(\tau)} di
\end{aligned}$$

Perform a second-order approximation of $L_i(\tau)$ around $\tau = 0$. Since $L_i(0) = L_i'(0) = 0$ and $L_i''(0) = \frac{\eta_i}{1-\eta_i}$, it follows that $L_i(\tau) \approx \frac{\tau^2}{2}\frac{\eta_i}{1-\eta_i}$. Using the definition $w_i \equiv \frac{y_i^\star}{Y^\star}$, the proof follows:

$$\begin{aligned}
\Delta \ln Y(\tau) = \frac{\Delta Y(\tau)}{Y^*} &\approx \frac{1}{Y^*} \cdot \int_0^\chi y_i^\star \cdot \frac{\tau^2}{2}\frac{\eta_i}{1-\eta_i} di \\
&= \frac{\tau^2}{2} \cdot \int_0^\chi w_i \cdot \frac{\eta_i}{1-\eta_i} di \\
&= \frac{\tau^2}{2} \cdot \int_0^\chi w_i \cdot di \cdot \int_0^\chi \frac{w_i}{\int_0^\chi w_j \cdot dj} \cdot \frac{\eta_i}{1-\eta_i} di.
\end{aligned}$$

E.2 Equilibrium Definition

We consider the stationary equilibrium of this model, which is described by a set of prices (r, R, w) such that:

1. Agents optimize, giving rise to decision rules $a'(\theta), c(\theta), o(\theta), k(\theta), \ell(\theta), y(\theta)$, where $\theta = (a, z, h, \eta)$ summarizes the individual's state, as well as an ergodic distribution $G(\theta)$.

2. The financial intermediary maximizes profits, implying $R = r + \delta - p \cdot (1 + r)$.

3. Given $G(\theta)$, all markets clear:

$$\begin{aligned} L &\equiv \int_{o=W} h \cdot dG(\theta) = \int_{o=E} \ell(\theta) \cdot dG(\theta) \quad \text{(labor market)} \\ K &\equiv \frac{1}{1-p} \int a \cdot dG(\theta) = \int_{o=E} k(\theta) \cdot dG(\theta) \quad \text{(capital market)} \\ Y &\equiv \int c(\theta) \cdot dG(\theta) + \delta \cdot K = \int_{o=E} y(\theta) \cdot dG(\theta) \quad \text{(goods market)} \end{aligned}$$

E.3 Model Robustness

Here, we discuss calibration details for the extended model versions with intermediate inputs in Section 5.2.4.

We introduce intermediate inputs as follows: an entrepreneur with technology (η, z), and inputs capital k, labor ℓ, and intermediates m, produces output

$$z \cdot k^{\alpha_K} \cdot \ell^{\alpha_L} \cdot m^{\eta - \alpha_K - \alpha_L}.$$

We assume a simple round-about production network, such that gross output Y is used for consumption, investment, and intermediate inputs, $Y = C + I + M$, with $GDP \equiv C + I$.

We fix $\alpha_K = 0.13$ and $\alpha_L = 0.29$, corresponding to our estimated mean output elasticities.[4]

[4] These values correspond to an estimation that expanded the definition of K as total assets, more in line with conventional macroeconomic aggregates that imply a capital share of value added of around one-third.

Table A.13 – Dynamic model w/ intermediates: Calibration

	Data	Model with intermediate inputs			
		Constraint on K,L,M		Constraint on K,L	
		z-econ.	*(η, z)-econ.*	*z-econ.*	*(η, z)-econ.*
A. Targeted moments					
Fraction entrepreneurs	0.117	0.117	0.121	0.116	0.116
Transition rate W→E	0.021	0.021	0.022	0.021	0.021
Top 10% revenue share	0.799	0.811	0.779	0.811	0.790
Top 1% revenue share	0.522	0.523	0.555	0.511	0.539
Top 0.1% revenue share	0.282	0.281	0.278	0.285	0.280
RTS: Top 5% vs Bottom 50%	0.083	0*	0.082	0*	0.083
Capital-output ratio	2.970	2.969	2.962	2.972	2.979
B. Internally calibrated parameters					
Mean RTS	μ_η	0.776	0.841	0.732	0.695
Standard deviation RTS	σ_η	—	0.070	—	0.079
Standard Deviation TFP	σ_z	0.653	0.573	0.823	1.097
Persistence TFP	ρ_z	0.971	0.948	0.970	0.950
Pareto tail TFP	ξ_z	3.944	—	3.557	—
Correlation (z, η)	$\sigma_{z,\eta}$	—	-0.712	—	-0.380
Discount factor	β	0.915	0.908	0.916	0.907

Notes: Steady state calibration of the (η, z)- and z-economy (both at $\lambda = 0.3$), in the model versions with intermediate inputs (and ex-post capital choice). * not targeted.

E.3.1 Adding intermediates, ex-post capital choice as in baseline

First, we maintain the baseline timing of input choices: capital is chosen after observing the realization of shocks, as are labor and intermediate inputs. We estimate the parameters in Table A.13 using the exact same strategy as in our baseline model versions. Rows 2 and 3 in Table A.14 show the resulting misallocation costs from raising the financial friction λ from 0 to 0.3, in both the calibration with and without RTS heterogeneity. Row 2 contains the results from the model that imposes the financial constraint on intermediates as well, while row 3 treats intermediates as fully flexible in line with the assumptions of our empirical approach (only capital and labor are subject to the constraint).

E.3.2 Adding intermediates, and pre-determined capital

Finally, we also change the timing of input choices, in addition to adding intermediate inputs in production: period t capital is chosen in period $t-1$, so prior to the realization of period t shocks. This specification is rich enough such that the identification assumptions of GNR hold. The computation becomes more involved, as an

Table A.14 – Misallocation: Different assumptions on production

	z-economy	*(η, z)-economy*	*Amplification*
1. Baseline: no intermediate inputs (M)	5.0	10.6	+112%
Including intermediates inputs (M):			
2. Constraint on K,L,M	9.3	46.3	+398%
3. Constraint on K,L	3.6	11.3	+214%
4. Constraint on K,L; pre-determined K	1.0	1.9	+81%

Notes: This table reports static misallocation from the financial friction λ, in log points, in alternative model versions (lowering λ from 0.3 to 0 when holding fixed occupational choice and factor supply). Row 1 corresponds to the baseline model without intermediate inputs. Rows 2, 3, and 4 add intermediate inputs in the production function. In row 2, there is a symmetric constraint on the three production factors: $w \cdot \ell + R \cdot k + m \leq \frac{a}{\lambda}$. In row 3 and 4, intermediate inputs are assumed to be fully flexible: $w \cdot \ell + R \cdot k \leq \frac{a}{\lambda}$. In row 4, period t capital is chosen in period $t-1$, prior to the realization of period t shocks.

agent's inter-temporal choice includes (i) net savings a', (ii) capital k' (part of net savings), (iii) and occupation o'. In our numerical solution, we exploit that when using resources after production x ("cash-on-hand") as endogenous state variable, there is no need to keep track of the two assets separately; instead, the problem becomes a portfolio choice problem conditional on occupational choice. The agent's dynamic problem is:

$$V(x,h,z,\eta) = \max_{c\geq 0, a'\geq 0, k'\geq 0, o'\in\{W,E\}} u(c) + \beta \cdot \mathbb{E}[V(x'_{o'}(a',k',h',z',\eta'),h',z',\eta')]$$
$$\text{s.t.} \quad c + a' = x,$$
$$x_W(a,k,h,z,\eta) = w \cdot h + (1+r) \cdot a - R \cdot k,$$
$$x_E(a,k,h,z,\eta) = \pi(a,k,z,\eta) + (1+r) \cdot a - R \cdot k,$$

where x_W (x_E) denotes cash-on-hand of workers (entrepreneurs),[5] and the variable profit of entrepreneurs is given by

$$\pi(a,k,z,\eta) = \max_{\ell\geq 0, m\geq 0} z \cdot k^{\alpha_K} \cdot \ell^{\alpha_L} \cdot m^{\eta-\alpha_K-\alpha_L} - w \cdot \ell - m$$
$$\text{s.t.} \quad w \cdot \ell + R \cdot k \leq \frac{a}{\lambda}.$$

This formulation is the natural extension of our general setup to pre-determined capital. The financial constraint applies to capital and labor, for every shock realization. The interpretation is, as before, that a fraction λ of the expenditures on capital $R \cdot k_t$ and labor $w \cdot \ell_t$ required for production in period t need to be financed with the

[5]Prospective workers will always optimally set the capital choice to zero, $k' = 0$.

Table A.15 – Dynamic model w/ pre-determined capital: Calibration

	Data	Model	
		z-economy	*(η, z)-economy*
A. Targeted moments			
Fraction entrepreneurs	0.117	0.117	0.129
Transition rate W→E	0.021	0.021	0.021
Top 10% revenue share	0.799	0.812	0.827
Top 1% revenue share	0.522	0.504	0.544
Top 0.1% revenue share	0.282	0.287	0.284
RTS: Top 5% vs Bottom 50%	0.083	0*	0.081
Capital-output ratio	2.970	2.970	2.969
B. Internally calibrated parameters			
Mean RTS	μ_η	0.703	0.632
Standard deviation RTS	σ_η	—	0.063
Standard Deviation TFP	σ_z	0.968	1.171
Persistence TFP	ρ_z	0.967	0.958
Pareto tail TFP	ξ_z	3.350	—
Correlation (z, η)	$\sigma_{z,\eta}$	—	-0.103
Discount factor	β	0.905	0.906

Notes: Steady state calibration of the (η, z)- and z-economy (both at $\lambda = 0.3$), in the model versions with intermediate inputs, pre-determined capital choice, and the financial constraint imposed on labor and capital expenditures (not on intermediates). * not targeted.

owner's period t net wealth, a_t. The difference to before is that capital k_t is chosen before the realization of period t shocks, while the labor choice ℓ_t is made (as before) after observing current shocks.

Table A.15 displays the calibration, following again the same strategy as in the previously discussed model versions. Row 4 of Table A.14 shows the resulting static misallocation costs from raising the financial friction λ from 0 to 0.3. The overall level of static misallocation associated with λ is now much smaller, because capital is still chosen ex-ante, and so even with $\lambda = 0$, marginal products of capital are not equalized. Instead, eliminating the λ friction only removes dispersion in marginal labor input products (intermediates are fully flexible, and hence marginal intermediate input products are fully equalized, regardless of the value of λ). We re-iterate that our main point is not the overall level of misallocation associated with financial frictions, but rather the additional amount of misallocation (+81% in this model version) generated by allowing for realistic RTS heterogeneity in line with our empirical results.