Scalable versus Productive Technologies[*]

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

First Version: July 11, 2024—This Version: November 19, 2024

Abstract

Do larger firms have more *productive* technologies, are their technologies more *scalable*, or both? We use administrative data on Canadian and US firms to estimate a joint distribution of output elasticities of capital, labor, and intermediate inputs—thus, returns to scale (RTS)—along with total factor productivity (TFP). We find significant heterogeneity in RTS across firms within industries. Furthermore, larger firms operate technologies with higher RTS, whereas the largest firms do not exhibit the highest TFP. Higher RTS for large firms are entirely driven by higher intermediate input elasticities. Descriptively, these align with higher intermediate input revenue shares. We also show that high-RTS firms grow faster, pay higher wages, and are owned by wealthier households. We then incorporate RTS heterogeneity into the workhorse model of endogenous entrepreneurship that matches the observed heterogeneity in TFP and RTS. We find that the efficiency losses from financial frictions are more than twice as large compared to a conventional calibration that attributes all heterogeneity to TFP and assumes a common RTS parameter.

Keywords: Production function heterogeneity, returns to scale, misallocation.
JEL codes: E22, L11, L25.

[*] 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 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. We thank Zhongdao Wang for his excellent research assistance. Part of this research was performed at a US Federal Statistical Research Data Center under FSRDC Project Number 2019, and at a Statistics Canada Research Data Center. All results based on US Census data have been reviewed to ensure that no confidential information is disclosed. All results based on Canadian administrative data have been screened by Statistics Canada to ensure that no confidential results have been released. The views expressed herein are those of the authors and do not reflect the ones of the Federal Reserve Bank of St. Louis, the Federal Reserve System or the Board of Governors. Ozkan acknowledges financial 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 variation in only TFP (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 and firm-level panel data, 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 in an application to the efficiency costs of misallocation due to financial frictions. Our findings indicate that incorporating RTS heterogeneity has broad implications for a variety of quantitative questions, including optimal capital taxation, firm hiring decisions (as in Gavazza *et al.* (2018)), and firm growth and cyclicality (as in Clymo and Rozsypal (2023)) as well as some of the well-known empirical patterns around firm heterogeneity.

In our main empirical analysis, we use administrative panel data for the universe of incorporated Canadian firms that account 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. Labor and intermediate inputs are measured consistently with Statistics Canada's national accounts, and we construct the capital stock using the perpetual-inventory method, as is standard in the literature. After sample selection, our final dataset comprises 4.3 million firm-year observations. To validate our results, we replicate the analysis for US manufacturing plants using administrative data from the Annual Survey of Manufactures and the Economic Census, finding similar empirical patterns.

In our benchmark approach, we estimate nonparametric production functions building on Gandhi, Navarro and Rivers (2020) (henceforth GNR), which provides the joint distribution of 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 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. Intuitively, high-RTS firms are those with higher expenditure shares, a stronger covariance between inputs and output, or both.

We estimate production functions for each two-digit NAICS industry in our sample. In addition to considerable heterogeneity in TFP (as the previous literature also documented), we find large differences in RTS among firms. The average estimated RTS is 0.96, with considerable variation across industries.[1] More novel, we document significant variation in RTS within industries. The average within-industry difference between the 90^{th} and 10^{th} percentiles (P90-P10) of RTS is 0.08. Interpreted as deviations from constant returns to scale, these differences are large.[2] As we show in Section 5, they are also quantitatively important for the costs of financial constraints.

By construction, the heterogeneity in RTS is explained by the dispersion in output elasticities of inputs. The P90-P10 of estimated output elasticities is 0.36 for intermediates and labor versus 0.08 for capital. These patterns align closely with the corresponding revenue shares of each input. Specifically, intermediate input shares closely track the estimated intermediate input elasticities, as our estimation treats them as a flexible input. For labor and capital, the correlation between revenue shares and elasticities is still strongly positive, though not perfect, reflecting potential adjustment costs, input market power, and the predetermined nature of capital.

[1] A few papers have documented heterogeneity in RTS. For instance, Gao and Kehrig (2017) report RTS across different industries in the US. Demirer (2020) finds heterogeneity in output elasticities and RTS across firms, industries and countries, often with RTS greater than one. Chiavari (2024) documents an increase in the aggregate RTS of the US economy over time. None of these papers, however, look at the relation between RTS, TFP, and firm size within industries.

[2] For example, in an efficient economy with Cobb-Douglas production functions, 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 a firm with RTS of 0.90.

Next, we examine how production technologies vary across the revenue distribution. Our main empirical finding is that RTS increase with firm size, especially for firms above the median revenue. Within two-digit NAICS industries, the average RTS of the largest 5% of firms are 8 percentage points (p.p.) higher than for those in the bottom 50%. Further analysis of the underlying output elasticities reveals that the increase in RTS with revenue is entirely accounted for by higher output elasticities of intermediate inputs.[3] In contrast, labor and capital elasticities tend to decline with firm size, although this result is less consistent across samples and specifications.

Several analyses indicate that a large portion of the RTS heterogeneity is due to persistent differences across firms. First, a 17-year panel regression of RTS at the firm-year level reveals that firm fixed effects account for 75% of the overall variation, even after controlling for firm size and age. Second, a components-of-variance model estimated from the auto-covariance structure of firm RTS shows that only 11% of the total variation is explained by the fully transitory component, whereas permanent fixed effects and the highly persistent component account for 39% and 51% of the differences, respectively. Third, a clustering exercise supports the interpretation that large, high-RTS firms already had high RTS when they were smaller. These results suggest that cross-sectional RTS heterogeneity primarily reflects persistent firm-level characteristics rather than transitory factors.

We find that TFP increases in firm revenue up to the top 10%, after which it flattens out and falls off sharply for the largest firms. In contrast, RTS rise in a convex pattern at the top of the revenue distribution, indicating a more important role for RTS differences in shaping the right tail of distribution. Overall, our results indicate that the largest firms are characterized by the highest RTS rather than the highest TFP, as commonly assumed (e.g., the literature following Hopenhayn (1992)). When counterfactually imposing homogeneous RTS in the estimation, however, the resulting TFP increases monotonically with firm revenue throughout the distribution, indicating the importance of flexible production technologies in TFP estimation.

Our results are robust across several dimensions. We obtain similar findings

[3] In recent work, Mertens and Schoefer (2024) show that firms grow by shifting from labor to intermediate inputs. Our two papers complement each other as they focus on firm growth and the implications for firm and industry labor shares, in a setting with homothetic production functions and imperfect input markets.

whether we estimate the production function for the entire economy or within narrow industries. Methodologically, our main results hold when (i) clustering firms based on their levels and growth rates of inputs and output and estimating separate production functions for each cluster, (ii) including intangibles in the definition of capital, and (iii) imposing homogeneous relative factor elasticities while still allowing for RTS differences. Finally, we also document similar patterns in the US manufacturing sector.

We also revisit 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 over the life cycle and are less likely to exit compared to high-TFP firms. Additionally, we show that high-paying firms tend to have higher RTS. Linking firms to their owners, we show that wealthier households disproportionally invest in more scalable technologies (i.e., firms with higher RTS). These secondary findings highlight the importance of incorporating realistic RTS heterogeneity for a variety of applications, including understanding wage and wealth inequality and designing optimal policies for capital income and wealth taxation (e.g., Guvenen *et al.* (2023); Boar and Midrigan (2022); Gaillard and Wangner (2021)).

To investigate the quantitative implications of our findings, we incorporate heterogeneous RTS into the workhorse model of endogenous entrepreneurship with standard incomplete markets (e.g., Quadrini (2000); Cagetti and De Nardi (2006)). In the model, agents choose between supplying their stochastic efficiency units of labor or operating a private business under a stochastic technology. Input choices are subject to a financial constraint such that the entrepreneur must finance at least a fraction λ of input spending with their own wealth. A novel feature of our model is that an entrepreneur's output depends not only on a standard idiosyncratic TFP term (z) but also on an idiosyncratic RTS term (η).[4]

Our main exercise compares the effects of increasing the financial friction λ in two different economies: the conventional z-economy, where all heterogeneity is driven by variation in only TFP, and the (η, z)-economy, which incorporates the joint first-order Markov process of RTS and TFP from our empirical estimates. We calibrate

[4] We treat RTS as a highly persistent exogenous process to reflect our empirical findings. This approach also allows us to treat RTS symmetrically to TFP. Specific microfoundations for RTS differences, such as scalable expertise (Argente *et al.* (2024)), choice of managerial inputs (Chen *et al.* (2023)), or the industrial revolution in services (Hsieh and Rossi-Hansberg (2023)), complement our analysis.

both economies such that they agree on key observable moments such as the firm-size distribution. We then compare the effects of the financial constraint λ on output and productivity between the two economies.

Our main finding is that in the (η, z)-economy, financial frictions generate more than double the output losses relative to the z-economy. A static misallocation of production factors accounts for the majority of output losses in both economies and is about twice as large in the (η, z)-economy compared to the z-economy. To provide intuition, we derive an analytical result in a static endowment economy. We show that a given marginal input product wedge leads to greater misallocation if the constrained firms have relatively higher RTS—a feature that our dynamic model generates endogenously. Dynamic effects further amplify output losses in the economy with RTS heterogeneity, as a result of an underaccumulation of capital and greater distortions in the selection into entrepreneurship. Intuitively, a highly productive but currently poor potential entrepreneur (i.e., high z) can still achieve profitability at a small scale, making it easier to grow despite the friction. In contrast, a highly scalable but not immediately profitable business (i.e., high η) struggles to outgrow the friction, and the entrepreneur may never enter the market.[5] These results suggest that accounting for RTS heterogeneity is crucial for understanding a broad set of quantitative questions related to misallocation, including wealth inequality and optimal taxation of capital.[6]

2 Empirical Methodology

Our main empirical approach builds on the production function estimation methodology developed by GNR, who estimate a flexible nonparametric gross output production function. We employ this method in different settings and across different samples. This technique provides several advantages over standard methods. First, it allows us to identify output elasticities for the *gross output* production function, whereas other common methods (e.g., Ackerberg *et al.* (2015)) typically identify only

[5] Some entrepreneurs have a high z but a low η, and hence a smaller optimal scale. This can explain why some entrepreneurs do not expect to grow their firms, as in Hurst and Pugsley (2011).

[6] For example, in ongoing work we study the entrepreneurial activity of New Money and Old Money households (à la Hubmer *et al.* (2024)), focusing on differences in their technologies.

value-added production functions. As we show below, variation in the output elasticity of intermediate inputs is a key driver of variation in RTS, making the identification of the gross production function essential. Second, the nonparametric identification strategy minimizes specification error when measuring both output elasticities and the productivity term. Third, this method allows us to estimate a nonhomothetic production function, where output elasticities and RTS vary depending on inputs and input shares, and may differ across firms and over time. These are crucial for understanding the relationship between firm-level TFP, RTS, and firm size.

2.1 Estimating Returns to Scale

We start by introducing our benchmark technique in detail, which closely follows GNR. 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.[7] 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

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

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

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

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

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.[8] This provides us with estimates of the production function as well as $\hat{\omega}_{jt}$, $\hat{\xi}_{jt}$, and $\hat{\omega}_{jt} \equiv \hat{h}(\hat{\omega}_{jt-1})$. 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: $\eta_{jt} \equiv \eta(k_{jt}, \ell_{jt}, m_{jt}) = \varepsilon_K^Y(k_{jt}, \ell_{jt}, m_{jt}) + \varepsilon_L^Y(k_{jt}, \ell_{jt}, m_{jt}) + \varepsilon_M^Y(k_{jt}, \ell_{jt}, m_{jt})$.[9] We use this specification to present the main empirical results in Section 4.

2.2 Identification and Intuition

Although GNR offers a rigorous identification strategy (for which we refer readers to their work for details), we focus here on the intuition behind our estimation results.

[8] 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.

[9] 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.

We first recover the output elasticity of intermediate inputs, $\varepsilon^Y_{M_{jt}} \equiv \varepsilon^Y_M(k_{jt}, \ell_{jt}, m_{jt})$, as a function of input levels from the nonparametric regression of the revenue share of intermediate expenditure on inputs. Since the expected intermediate expenditure share is simply equal to its output elasticity (per the FOC in equation 1), covariation between the (expected) share and input levels identifies this output elasticity.[10] This nonparametric regression also identifies the ex post transitory shock: For two firms with the same input levels, variation in intermediate expenditure shares can only come from differences in the ex post shocks (through unexpected variation in revenues).

With the estimates for intermediate input elasticity and transitory shocks at hand, we can then remove the effect of intermediates and the ex post shock on gross output, which will leave us with a "value-added" production function to estimate in the next step.[11] Recall that we allow for adjustment costs for capital and labor without assuming optimality in the choice of either input, which introduces an (unknown) wedge between their expenditure shares and output elasticities. This means we cannot use the FOC approach from the first step to identify these elasticities. Therefore, as in GNR, our second-stage estimation follows the Olley and Pakes (1996) proxy-variable literature in exploiting the Markov timing assumptions on the persistent shock to form GMM moment conditions. Intuitively, conditional on the previous period's persistent productivity (ω_{jt-1}), the covariation between the value added and capital and labor (instrumented with its lagged value) inputs identifies the component of their output elasticities not recovered in the first step. Similarly, conditional on capital and labor inputs and ω_{jt-1}, variation in value added identifies the persistent shocks. Thus, in the data, a high-RTS firm will be characterized by a high intermediate input expenditure share, a strong correlation between output and capital and/or labor, or both.[12]

[10] Intuitively, if the underlying production function were Cobb-Douglas, then the expenditure share would be uncorrelated with input levels, and its output elasticity would remain constant (and equal to the mean expenditure 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 the level of m_{jt}.

[11] This is a slight abuse of the language as, for example, our value-added production function in this step does not contain transitory productivity shocks and is derived by removing the contribution of intermediate inputs to output.

[12] Note that as a result of the nonhomotheticity of the production function, these correlations are functions of input levels and thus vary across firms.

3 Data and Sample Selection

Our main dataset is the Canadian Employer-Employee Dynamics Database of Statistics Canada (CEEDD), which is a set of linkable administrative tax files covering the universe of tax-paying Canadian firms and individuals between 2001 and 2019. We obtain the balance sheet and income statement information on firms from the National Accounts Longitudinal Microdata File, which covers all incorporated firms.[13] We use the total revenue and total wage bill variables constructed by Statistics Canada based on the corresponding corporate tax return line items. These same variables are used in the calculation of the national income and product accounts. Therefore, our microdata are consistent with aggregate measures. We construct total tangible capital by employing the perpetual-inventory method (PIM), using information on the first book value of tangible capital observed in the dataset, 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 converted to 2002 real Canadian dollars.

To construct the estimation sample, we start from firm-year observations with nonmissing values in total revenue, capital stock, wage bill, intermediate input, and industry code. For the first few firm-year observations of capital stock, the PIM method relies heavily on the initial available book value; therefore, in our estimation we only include those with at least two previous observations of capital. We also drop observations with outlier factor shares. Specifically, we drop firm-year observations with (i) a wage-bill-to-revenue ratio below the 1st percentile or above the 99th percentile, (ii) a wage-bill-to-value-added ratio below the 1st percentile or above the 99th percentile, (iii) an intermediate-input-to-revenue ratio greater than 0.95 or smaller than 0.05, and (iv) a capital-to-revenue ratio above the 99.9th percentile. After selection, our sample contains 4.3 million firm-year observations and 620,000 firms with an average of 6.9 observations per firm. See Table I for summary statistics.[14]

[13] Our CEEDD dataset also covers all unincorporated firms in Canada. Unincorporated firms in Canada are typically small businesses owned by self-employed individuals, which account for 9.5% of the total GDP in the economy in 2005, with the share declining since the mid-1990s (Baldwin and Rispoli (2010)). We do not include these firms because they do not report capital stock.

[14] The firm-level distributions of these variables are similar to economy-wide microdata in other countries. For example, Chan *et al.* (2024) find very similar distributions of log revenues and inputs in administrative data for the entire Danish private sector.

Table I – Summary Statistics

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

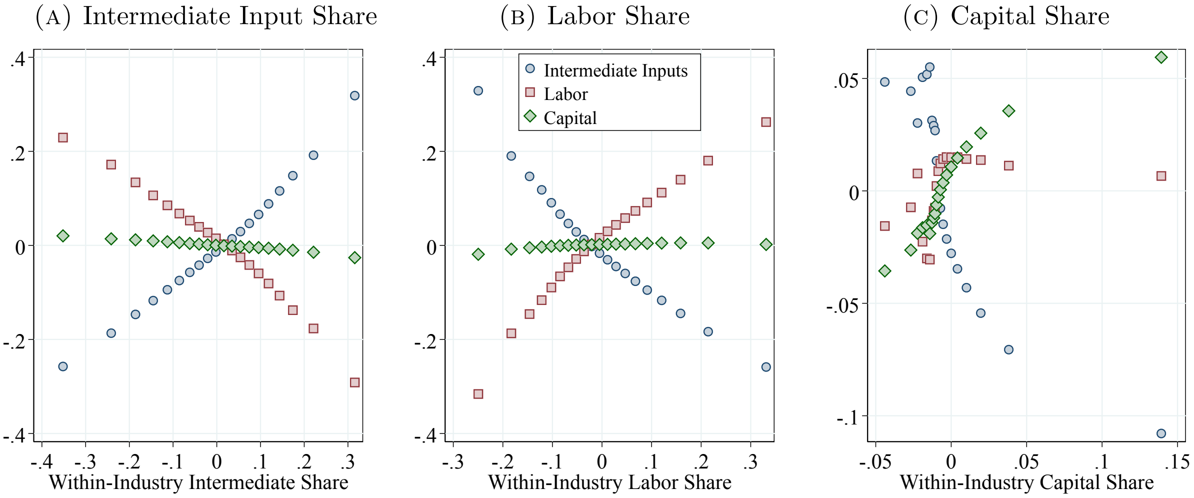
Notes: Table I shows cross-sectional moments of the distributions of log values for revenue, intermediate inputs, wage bill, and capital. All variables are in 2002 Canadian dollars. The total number of observation is 4.3 million firm-years.

US Manufacturing Sector. As a robustness exercise, we perform similar analysis using data from the US Economic Census and the Annual Survey of Manufactures (ASM), which has been extensively used in the literature for the study of firm-level productivity in the US (see, for instance, Foster *et al.* (2001) and Bloom *et al.* (2018a)). This dataset contains detailed information on over 60,000 manufacturing plants between 1974 and 2019. Unlike our Canadian data, this sample does not contain the entire universe of firms but a representative panel of manufacturing plants that is redrawn every five years. We focus on a sample of firms with at least two years of data with nonmissing information of key variables, generating a sample of 3.1 million establishment-year observations. We measure revenue for all plants with information on their total value of shipments. The US Census also provides information on real capital stock (measured used PIM), total wages of all workers in the plant, and expenditures on intermediate inputs measured in 2019 US dollars.

4 Empirical Results

In this section, 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 in our sample (see Table OA.1 in the appendix for the list of industries and summary statistics for their technologies). We begin by presenting the unconditional moments of these estimated parameters. Next, we explore how these estimates vary across the firm-size distribution. We also highlight the key data features that significantly impact our empirical results, thereby demonstrating our identification argument (Section 2.2) in the data. Finally,

Figure 1 – Average Output Elasticities By Factor Shares of Revenue



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.

we connect our findings to broader discussions on the life-cycle growth of firms, as well as wage and wealth inequality.

4.1 Unconditional Heterogeneity in Production Technologies

We start by examining cross-sectional moments of the unconditional distribution of firm technologies. First, we calculate within-industry moments from the distribution of firm-level estimates for each year. We then average these moments across industries and time. Our results, shown in Table II, reveal considerable heterogeneity in the estimated RTS, output elasticities, and TFP across firms.

RTS Heterogeneity. Starting with the within-industry RTS moments, we find an average of 0.96 with a 90^{th}-to-10^{th} percentile gap (P90–P10) of 0.08.[15] 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. More importantly, these differences are substantial when interpreted as deviations from constant returns to scale. For instance, in an efficient economy with Cobb-Douglas

[15] Consistent with earlier literature (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 OA.1), ranging from 0.59 (for Healthcare) to 1.03 (for Management of Companies and Enterprises).

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 the method in Section 2 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.

production function, 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 a firm with RTS of 0.90.[16] Furthermore, above-median differences are larger compared with the below-median dispersion: the average within-industry P50–P10 is only 0.03 compared with 0.05 for P90–P50 and 0.13 for P99–P50. Finally, the average 90th percentile for RTS across industries is 1.00; that is, most firms operate decreasing returns to scale technologies, yet some have annual RTS above 1.[17]

By construction, differences in RTS arise from heterogeneity in output elasticities. As shown in Panel B of Table II, the output elasticity of intermediate inputs has

[16] We indeed find that the revenues of high-RTS firms respond more strongly to aggregate TFP shocks (Table OA.6).

[17] Note that RTS is not fixed over time and firms are subject to adjustment costs. Therefore, the fact that some firms have increasing returns to scale does not necessarily mean that they can increase their supply indefinitely. Furthermore, other studies commonly estimate RTS to be above 1 for some industries or firms as well (e.g., Gandhi *et al.* (2020) and Demirer (2020) find average RTS above 1 across multiple industries and countries).

the highest average value of 0.59, followed by labor at 0.33 and capital at 0.04.[18] Labor and intermediate input elasticities vary more across firms within industries than capital elasticities. For instance, the average within-industry P90-P10 gap for intermediate inputs and labor is 0.36, while for capital it is only 0.08. Variance decompositions further reveal that more than 60% of the overall variation in each output elasticity is explained by within-industry differences (see Table OA.2). This indicates that within-industry heterogeneity accounts for a larger share of the total firm-level variation in output elasticities compared to RTS (for which a quarter of the total variance is accounted for by within-industry differences). This result is partly due to the negative correlation between intermediate input and capital/labor elasticities within industries (Table OA.3).

Output Elasticities and Input Shares. Based on our theoretical identification argument in Section 2.2, we now present the data features that have a pronounced effect on our empirical results. Typically, output elasticities reflect their corresponding revenue input shares. In fact, for Cobb-Douglas production functions, output elasticities are exactly equal to (average) input shares. Our specification is more flexible than Cobb-Douglas, and the GNR method does not solely rely on the FOCs of profit-maximizing firms. Nevertheless, output elasticities tend to be positively correlated with the respective factor shares.

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 three inputs, the corresponding output elasticity is strongly correlated with the respective input share. 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, which is expected since we treat them as flexible input and use the firm's FOC to estimate intermediate input elasticities (see equation (1)).

[18] Our estimated average capital elasticity is lower than typical estimates. This is because, following the literature, we construct the capital stock using the perpetual-inventory method and by only including tangible capital such as structures and equipment. Therefore, we exclude 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 more extensive definition of capital as net values of assets from their balance sheets, we find the average intermediate, labor, and capital elasticities to be 0.58, 0.30, 0.12, respectively.

In contrast, for labor and capital, our estimation does not rely on FOCs. Nevertheless, we still find a strong positive correlation between input shares and their respective output elasticities. These findings resonate with our identification intuition in Section 2.2 that heterogeneity in output elasticities, and consequently in RTS, reflects differences in input shares.[19]

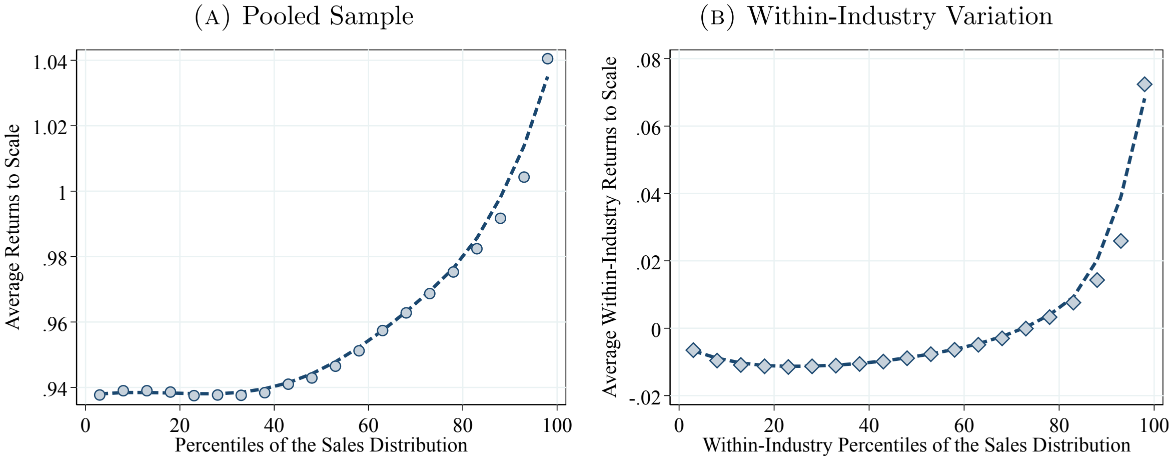
TFP Dispersion. We find that the P90-P10 of firm-level TFP is 0.31. This implies that a firm at the 90^{th} percentile produces about 36.2% more output than the firm at the 10^{th} percentile, with the same inputs and holding output elasticities constant. This gap is substantially lower than previous estimates of productivity dispersion even for narrow six-digit industries in Canada and the US, which typically find P90-P10 TFP gaps closer to 2 (see, for instance, De Loecker and Syverson (2021) and Syverson (2011)). The difference stems from the use of a flexible nonparametric production function estimation—which allows for differences in RTS—and from using the wage bill as our measure of labor input rather than the number of workers or the total number of hours (see Fox and Smeets (2011)). Using a similar method, Chan *et al.* (2024) estimate a P90-P10 TFP gap of 0.43 for the entire Danish private sector. Furthermore, for the manufacturing firms in our Canadian sample, the P90-P10 of TFP is 0.28, and for manufacturing plants in the US, the P90-P10 equals 0.54.

4.2 Production Technologies over the Firm-Size Distribution

Returns to scale by firm revenue. We now turn to the systematic variation of our estimates over the revenue distribution. To this end, we pool all firm-year estimates from our estimation of production functions within 23 two-digit NAICS industries. Figure 2a shows a bin scatter plot of average RTS by firm revenue for this pooled sample of all industries. We find that firms in the bottom two-fifths of the revenue distribution have, on average, similar (decreasing) RTS of about 0.94. As we move to higher percentiles of the revenue distribution, however, RTS increase monotonically and strongly by firm revenue from 0.94 for firms below the 40th percentile to 1.04 for those in the top 5%.

[19] These findings suggest that high-RTS firms should have relatively low profit shares. It is indeed the case that, on average across firms, the EBITDA-revenue ratio correlates negatively with RTS (see Figure OA.5).

Figure 2 – Returns To Scale by Firm Size



(A) Pooled Sample

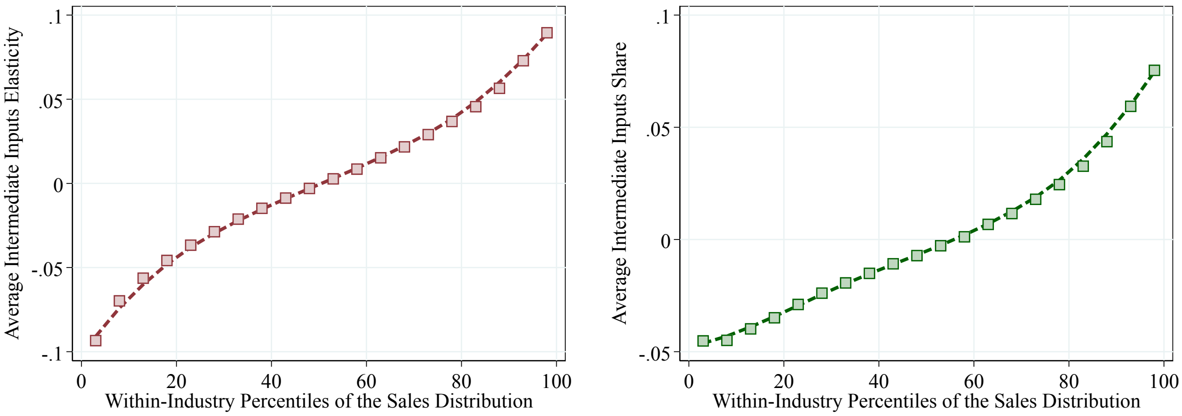
(B) Within-Industry Variation

Notes: Figures 2a and 2b show the average RTS within ventiles of the firm-revenue distributions. In Panel (B), RTS are demeaned by industry averages. The dashed line is obtained from a Lowess smoothing estimation over the points. In Figure 2b, we demean the within-quantile average by two-digit NAICS industry averages.

The variation in Figure 2a reflects both within- and between-industry heterogeneity. For instance, manufacturing firms tend to be larger and are therefore over-represented at the upper end of the revenue distribution. Additionally, manufacturing industries exhibit higher average RTS. Therefore, a portion of the overall variation in RTS is driven by differences across industries. To isolate the role of within-industry differences, we first demean RTS within industries and then rank firms into quantiles within their respective industries based on firm revenue. Figure 2b shows average demeaned RTS across the within-industry revenue distribution. Again, firms below the median have similar average RTS, whereas it increases sharply with firm revenue above the median: the average RTS of firms in the top 5% of the within-industry revenue distribution is 8 p.p. higher than the average RTS of firms in the bottom half of the distribution. This variation is almost as large as the variation in the pooled sample (10 p.p.). Thus, we conclude that most of the variation in RTS by firm size is driven by within-industry differences.

Output elasticities by firm revenue. As discussed in Section 2, we measure RTS as the sum of output elasticities with respect to inputs. Therefore, the significant increase in RTS could be driven by either of these inputs or a combination thereof. Our analysis, however, shows that the intermediate input elasticity entirely accounts for the positive relationship between RTS and firm revenue. Figure 3a shows that

Figure 3 – Intermediate Input Elasticities by Firm Size


(a) Intermediate Inputs Elasticity

(b) Intermediate Inputs Share

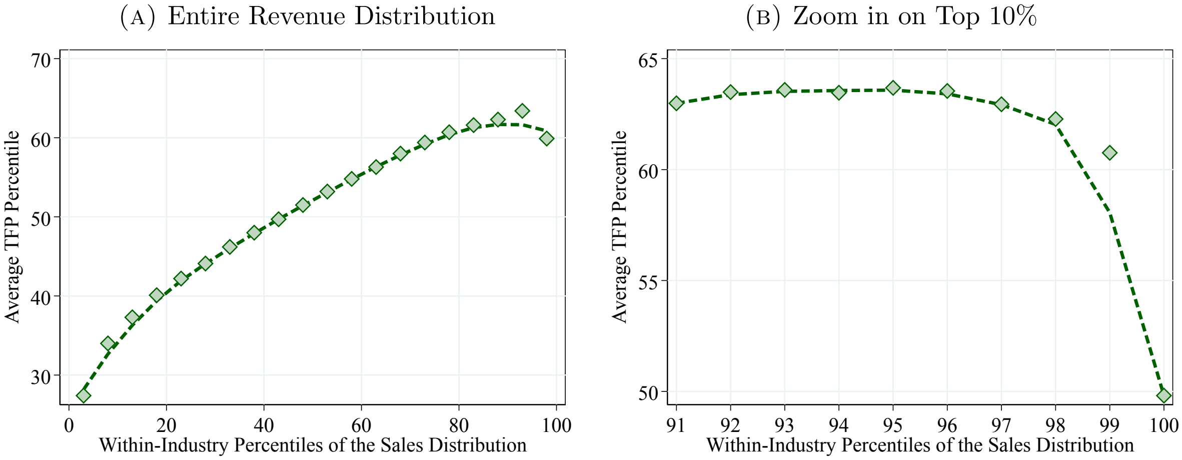
Notes: Figure 3 shows the average estimated factor elasticities within ventiles of the revenue distribution; Figure 3b shows the intermediate inputs revenue share. The dashed line is obtained from a Lowess smoothing estimation over the points. In both panels, we demean the within-quantile average by two-digit NAICS industry averages.

the intermediate input elasticity monotonically increases from -0.09 (relative to the industry average) for firms in the bottom 5% of the revenue distribution, to approximately zero for firms 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 the median firms fully explains the corresponding 8 p.p. gap in RTS over the same range. Furthermore, Figure 3b demonstrates that the variation in 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. These findings underscore the importance of estimating gross output production functions. On average, capital and labor elasticities decline with firm revenue, suggesting that using value added production functions may lead to misleading conclusions (see Figure OA.2).

Total factor productivity by firm revenue. Intuitively, one might expect that the largest firms are also the most productive. Next, we investigate TFP differences across the revenue distribution. Because we are pooling all firms across industries—whose average TFPs are not comparable with each other—in Figure 4 we measure the relative TFP of a firm as its TFP rank within industries to show the average TFP differences across the within-industry firm revenue distribution.

Consistent with prior studies, relative TFP increases with firm size up to the top

Figure 4 – Firm Size and Productivity



Notes: Figure 4 shows the average firm TFP rank within percentiles of the within-industry revenue distribution. The TFP rank is calculated within percentiles of the RTS distribution. The right panel zooms in on the top 10% of the revenue distribution.

decile of the revenue distribution (see Leung *et al.* (2008) or Baldwin *et al.* (2002)), after which, however, we find that it flattens out. In fact, zooming in on the top 10% of the revenue distribution we find that TFP falls off sharply for the largest firms (Figure 4b). In contrast, RTS increase even more steeply at the top of the distribution (see Figure OA.7). Therefore, we conclude that the largest firms tend to feature the highest RTS and not necessarily the highest TFP as commonly assumed.

Our results on the TFP-revenue gradient differ from previous studies because we allow for heterogeneity in production technologies across firms. In contrast, when we restrict our estimation to disallow RTS heterogeneity and estimate a standard homogeneous Cobb-Douglas production function, $Y_{jt} = e^{\nu_{jt}} \cdot K_{jt}^{\gamma_k} L_{jt}^{\gamma_L} M_{jt}^{\gamma_M}$, as expected, we also find that TFP increases monotonically with firm size (see Figure OA.9). This exercise demonstrates the importance of allowing for flexible production technologies in understanding the relationship between firm-level TFP, RTS, and firm size.

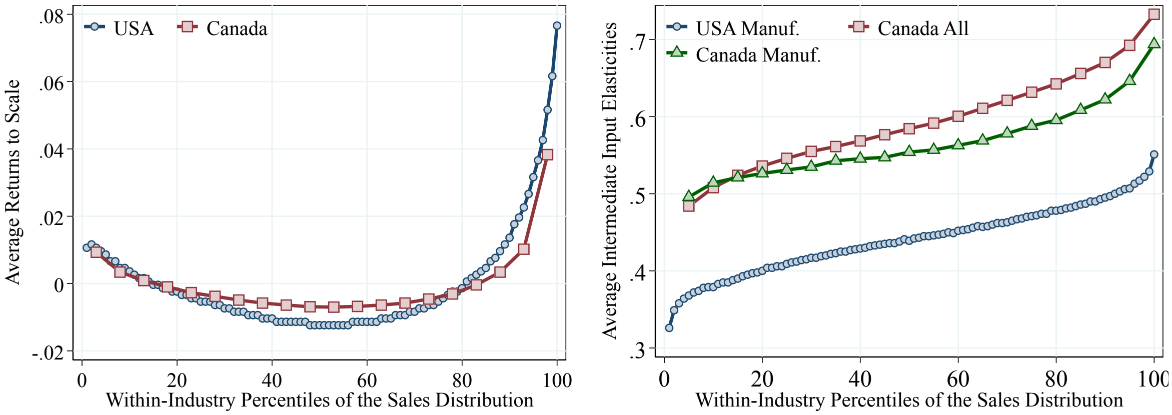
4.2.1 Robustness Checks

US Manufacturing. Our results are not unique to the Canadian economy but also hold within the US manufacturing sector. Figure 5 illustrates plant-level RTS relative to the industry average (four-digit NAICS), showing a U-shaped pattern with respect to revenue and a notable steep increase at the top, where RTS rise by about 9 p.p. from the 50th percentile to the top 1% of the revenue distribution. For comparison,

Figure 5 – Returns To Scale and Elasticities of Intermediate Inputs in Manufacturing

(a) Returns to Scale

(b) Intermediate Input Elasticity



Notes: Figure 5a shows the average RTS within percentiles of the sales growth distribution demeaned by industry averages for the Canadian and US manufacturing sectors. Canadian results are shown within 5% quantiles of the revenue distribution. Figure 5b shows the average intermediate input elasticity for the US and Canadian manufacturing sectors and for the entire Canadian private sector within percentiles of the sales distribution.

we include a corresponding series for the Canadian manufacturing sector in the figure, which reveals a similar U-shaped pattern. However, the increase in RTS among the largest firms is more pronounced in the US, which can be explained by their fatter right tail compared to Canadian manufacturing firms (Leung *et al.*, 2008).

Remarkably, most of the increase is again due to a significant rise in the output elasticity of intermediate inputs, which rises from 0.4 at the bottom of the establishment-size distribution to around 0.55 for the largest plants in US manufacturing. Furthermore, similar to our results from Canada, the labor elasticity is also declining over the revenue distribution, while the capital elasticity is only declining in firm revenue up to the 90th percentile of the size distribution, after which it increases slightly (see Figure OA.2). Relatedly, labor, capital, and intermediate input revenue shares also exhibit patterns across the revenue distribution that are similar between US and Canadian manufacturing sectors and Canadian corporations (see Figure OA.4).

Our production function estimates using the US manufacturing data are at the plant level. In contrast, the Canadian data are at the firm level and include the number of plants per firm. Figure OA.8 shows that, in the Canadian data, RTS increase significantly with the number of plants, which is also strongly correlated with firm revenue. A regression of demeaned RTS on log firm revenue yields a coefficient of 0.012, and controlling for the number of plants per firm only slightly reduces this size gradient to 0.010. Together with the US manufacturing results, these findings suggest that variation in RTS by firm revenue is not primarily driven by plant count but rather by differences in production technologies across individual plants.

Cobb-Douglas specification. One potential concern is that our results may be sensitive to the specific estimation method we use. To address this, we we reestimate the production function by imposing homogeneous relative factor elasticities while still allowing for RTS differences. This approach isolates heterogeneity in RTS independent of variations in relative output elasticities. For this exercise, we specify the production function as $Y_{jt} = e^{\nu_{jt}} \cdot \left(M_{jt}^{\gamma_M} K_{jt}^{\gamma_K} L_{jt}^{\gamma_L}\right)^{\eta_{jt}}$. In this specification, the RTS parameter η_{jt} is estimated in the first stage of the procedure described in Section 2. Consistent with our baseline findings, the Cobb-Douglas series in Figure 6 shows that RTS increases with firm size by about 10 p.p., with a stronger increase in the bottom half and a more moderate rise in the top half of the revenue distribution relative to our baseline results.

Clustering specification. In another robustness exercise, we reestimate firms' production technology within clusters of firms with similar characteristics. The motivation behind this approach is that while estimating a separate production function for each firm would be ideal, it is not econometrically feasible. Clustering firms with similar features offers a practical alternative to approximate this ideal. We cluster firms based on their average levels and growth rates of output, capital stock, labor expenditure, and intermediate input expenditures. We standardize these firm-level variables and then apply the k-means clustering algorithm with 20 clusters. Each firm remains in the same cluster throughout its life cycle. We then estimate the nonparametric production function separately for each cluster. After estimation, we rank firms by revenue within their respective industry, as in our baseline results. The cluster series in Figure 6 shows that while these estimates are less smooth along the revenue distribution, the main patterns are consistent with our baseline results: RTS are relatively stable within the bottom half of firms but increase by close to 10 p.p. from the median to the top 5% of firms within an industry.

Intangible capital specification. We also include intangibles in our measure of the capital stock and reestimate firms' production functions. 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, then excluding it from the capital stock measure can lead to underestimation of the capital elasticity (and thus RTS) and overestimation of TFP for these firms. Consistent with this intuition, the intangible capital series in Figure 6 shows that the positive relationship between firm size and RTS becomes even stronger when including intangible capital.[20]

Ranking firms by employment or value added. Appendix Figure OA.3 presents our findings when ranking firms, within industry, by employment or value added instead of by revenue. Although the patterns for RTS are similar, the output elasticities display distinct variations: firms with high employment or high value added exhibit higher labor elasticities, while the intermediate input elasticity shows only a small increase for the largest firms. This pattern is somewhat mechanical, as we expect high employment or high value added firms to be labor intensive by construction of the ranking. Therefore, we prefer to rank firms by revenue—a factor-neutral approach—in our primary analysis.

Markups and Market Power. Our RTS estimates are based on revenue elasticities of the three inputs. One may be concerned that the positive relation between RTS and firm size could be driven by markup variation (e.g., De Loecker *et al.* (2020)), unobserved variation in prices, or monopsony markdowns for intermediate inputs (e.g., De Loecker *et al.* (2016) and Burstein *et al.* (2024)).[21] Ideally, we would want to separately estimate physical output elasticities, markups, and markdowns. However, doing so would require firm-level information on input and output prices and physical quantities, which we do not have in the Canadian or US datasets. Even with such price and physical quantity information, it is challenging to estimate the physical elasticity for multi-product firms in the absence of information on product-

[20] The positive relationship between firm size and the TFP percentile remains unchanged, likely because firms with similar RTS tend to have comparable levels of intangible capital intensity.

[21] Our estimation method is robust to market power in capital and labor markets.

specific inputs. It is important to note that if larger firms charge higher markups or markdowns—as implied by models with oligopolistic competition (e.g., Atkeson and Burstein (2008)) or with monopolistic competition and log-concave demand systems (e.g., Edmond *et al.* (2023))—then physical RTS would be increasing even more with firm size compared to measured revenue RTS.[22] Indeed, when we estimate firm-level markups in our data following the approach of De Loecker and Warzynski (2012), we find that markups increase with firm size. Specifically, we adopt the value-added and translog production function specification of De Loecker and Warzynski (2012) and conduct the estimation by industry. Figure OA.1 shows that, on average, markups monotonically increase with firm revenue, consistent with De Loecker *et al.* (2020).

In another robustness check, we control for firm-level markups in the first stage to explicitly account for market power. To do this, we relax Assumption 5 and allow firms to face a downward-sloping demand curve such 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. We follow De Loecker *et al.* (2020) and De Loecker *et al.* (2016) by using functions of output market shares to proxy for unobserved price elasticities (ε_P^Y).[23] If markups (or markdowns) are a significant determinant of input expenditure shares, we should find that our estimates of the intermediate input elasticity are sensitive to the inclusion of these controls.[24] We show in Figure OA.11 that controlling for market shares barely changes the size gradient of intermediate input elasticity, the main driver of RTS differences along the firm-size distribution.[25]

These theoretical and empirical considerations reinforce our interpretation of measured RTS differences along the firm-size distribution as representing differences in production technologies.[26]

[22] We can easily show that the physical output elasticity is the product of the revenue elasticity and markup, divided by the markdown.

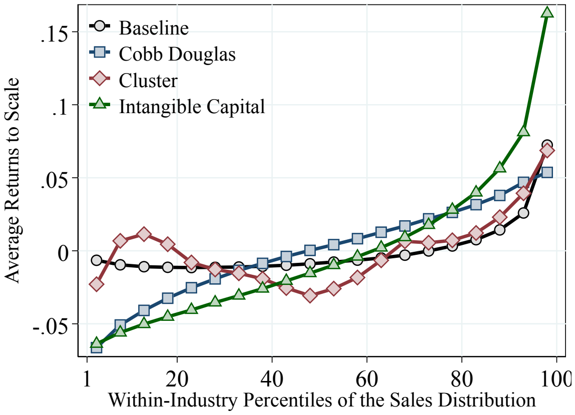
[23] This is an exact control if demand takes common (nested) logit or CES forms. See De Loecker *et al.* (2016) for further discussion.

[24] De Loecker *et al.* (2020) use this approach to control for unobserved output prices whileDe Loecker *et al.* (2016) apply it to control for unobserved intermediate input prices.

[25] Following De Loecker *et al.* (2020) and De Loecker *et al.* (2016), we proxy μ^P with 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.

[26] One concern raised by the literature (e.g., Bond *et al.* (2021)) is that identification of revenue

Figure 6 – Returns to Scale and Firm Size for Different Specifications



Notes: In the Cobb-Douglas specification, we estimate the production function by restricting to homogeneous relative output elasticities while allowing for heterogeneity in RTS. In the cluster specification, we apply the k-means clustering algorithm (20 clusters) and estimate the nonparametric production function within clusters. Intangible capital is constructed using PIM. In all specifications, we sort firms based on sales within industry, and RTS is demeaned by industry averages.

4.3 Permanent versus Transitory Differences in RTS

The key input for models that account for the firm-size distribution (e.g., literature that follows Lucas (1978) and Hopenhayn (1992)) is not only the extent of heterogeneity across firms but also the degree of persistence in their characteristics (e.g., Sterk *et al.* (2021)). Therefore, it is important to determine whether the observed dispersion in RTS is primarily due to fixed differences in production technologies across firms or to transitory fluctuations of individual firms around a common production function. We provide three sets of results that collectively suggest a significant portion of the observed heterogeneity is driven by permanent differences between firms.

Fixed effects regression. First, we run a panel regression of RTS on firm size, firm age, time dummies, and firm fixed effects. Intuitively, if the dispersion is primarily driven by fixed differences between firms, the firm fixed effects should absorb most of the variation in our estimates. This is indeed 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

or physical production functions requires either price and quantity data or very strong parametric assumptions on demand and production. In particular, the level of estimated markups and output elasticities may be biased in unknown ways. However, recent evidence has shown that this bias is small in practice and that *relative* variation in markups and output elasticities is well identified using revenue data (De Ridder *et al.*, 2022). We find that our estimates of this relative variation (our primary interest) is robust across multiple methods and data.

when controlling for firm age and size. We find remarkably similar results for the US manufacturing sector: RTS has a variance of 0.058^2 and fixed effects account for 65%.

Autocovariance structure. We take inspiration from the earnings dynamics literature (e.g., Abowd and Card (1989); Karahan and Ozkan (2013)) and exploit the autocovariance structure of the RTS estimates to estimate a components-of-variance model. In particular, we decompose firm-level RTS into a firm-specific fixed effect, an AR(1) persistent component, and a fully transitory component (see Appendix C for details). Consistent with our fixed effect regression result, we find that only 10.5% of total variation in RTS across firms is explained by the transitory component, whereas fixed effects and the highly persistent component (with a persistence parameter of 0.94) account for 38.9% and 50.6% of the differences, respectively.

Clustering. The second approach involves clustering firms based on their maximum size over their life cycle and then comparing the estimated RTS–firm size gradients both across and within clusters. In this case, within-cluster variation reflects the changes in firms' RTS over their life cycle, while between-cluster differences are mainly driven by permanent firm characteristics. Then, intuitively, if RTS differences are purely due to permanent firm types, we would expect significant level differences across clusters, with an average within-cluster RTS–firm size gradient of zero. For example, firms that are large and exhibit high RTS in 2019 would have displayed high RTS even when they were small in 2001. Conversely, if RTS variation is driven purely by scale effects (nonhomotheticities), we would observe overlapping profiles and a positive within-cluster gradient equal to the pooled RTS–firm size gradient. This implies that firms that are large and have high RTS in 2019 would have shown low RTS when they were smaller in 2001.

To implement this analysis, we calculate each firm's revenue percentile within its industry and year and take the maximum over its life cycle. We then group the firms, within industry, into 11 clusters based on the maximum attained rank: 1-10, 11-20, . . . , 91-95, and 96-100. Each firm remains in the same cluster throughout its life cycle. To reduce selection bias, we exclude firms with fewer than 10 years of data. We then estimate the nonparametric production function separately for each cluster.

Figure OA.10 displays the estimated RTS. Reassuringly, when pooling all firm-year estimates across clusters and industries, the pattern of average RTS by firm size is very similar to the one of our baseline pooled estimation (Figure 2b). To distinguish between type dependence and scale dependence, we run regressions of RTS on log firm revenue, controlling for industry fixed effects. Without cluster fixed effects, the size gradient is 0.012, capturing the pooled variation. However, when we include cluster fixed effects, the average size gradient drops to 0.00. These findings suggest that the systematic variation of RTS by firm size primarily reflects persistent firm differences.

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 heterogeneous RTS model in Section 5.

4.4 Does RTS Heterogeneity Matter? Revisiting Classical Patterns Regarding Firm-Level Differences

In this section, we revisit some of the well-known empirical patterns around firm heterogeneity that were previously explained by differences in TFP. For example, the previous literature has argued that firms with high TFP grow faster (e.g., Sterk *et al.* (2021)) and pay higher wages (e.g., references in Kline (2024)) and the wealthiest households own the most productive firms (e.g., the literature that follows Quadrini (2000); Cagetti and De Nardi (2006)). In this section, we argue that RTS differences appear at least as important in explaining these empirical patterns.

4.4.1 Firm Dynamics

Intuitively, heterogeneity in RTS has significant implications not only for the firm size distribution but also for a firm's growth trajectory over its life cycle. Firms with high RTS are expected to grow faster to reach their larger optimal size compared to firms with similar TFP but lower RTS. To analyze these life-cycle patterns, we construct a balanced panel of firms and group them based on their production function characteristics at entry. Specifically, we focus on firms that (i) are born between 2002 and 2005 and (ii) are observed for 12 consecutive years. We group firms based on their initial RTS and initial TFP (both of which are demeaned at the industry level). Figure 7 plots the average log revenue, also demeaned at the industry level, against firm age for different firm groups.

Firms with high initial RTS and TFP start with higher revenues relative to other firms of the same age within the same industry. More importantly, firms with higher initial RTS exhibit significantly faster growth than those with lower initial RTS. Firms in the top 10% of the initial RTS distribution grow about 30 log points over the next 10 years, while those in the bottom 90% grow, on average, only about 20 log points. This evidence supports our interpretation that firms with high RTS operate more scalable technologies, enabling them to achieve substantially greater growth over their lifecycle.[27]

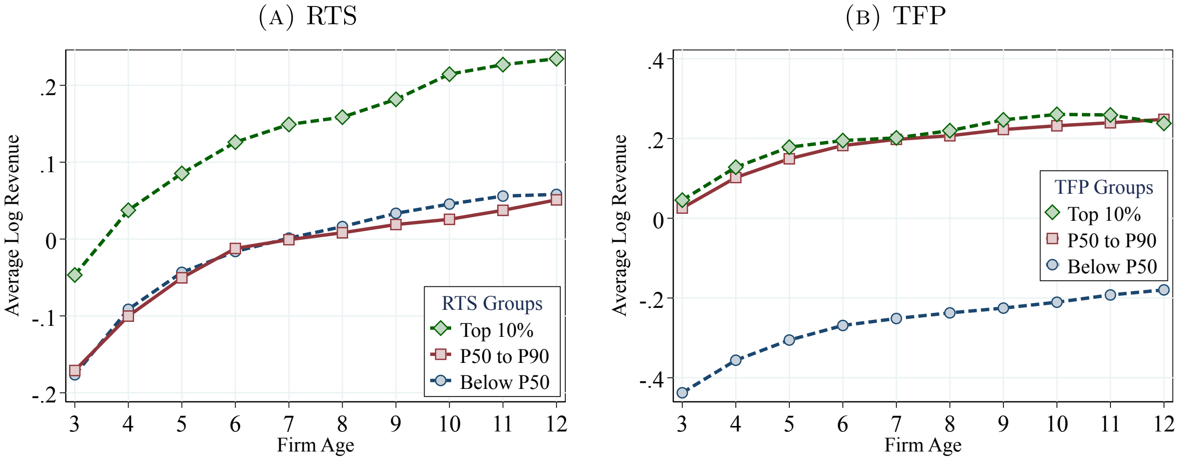
In contrast, Figure 8b shows that firms that enter with high TFP, while initially larger, do not grow faster than other firms in the same industry. In fact, higher initial TFP is associated with slightly lower subsequent firm growth rates. This pattern is consistent with TFP being a mean-reverting process and could explain why some firms do not expect to grow significantly, as documented by Hurst and Pugsley (2011). Our results suggest that these highly productive firms might have low RTS, which limits their scalability and their optimal size.

The previous results focus on the life-cycle patterns of surviving firms. We now examine how RTS and TFP heterogeneity affect firm exit. To do so, we estimate a probit regression of firm exit on TFP percentile and RTS. The results are reported in Table III. Column (1) uses the levels of production parameters, while column (2) uses first-differenced parameters. Both specifications include two-digit NAICS industry fixed effects. Across specifications, a higher RTS is associated with a lower probability of firm exit. The effect of TFP on firm exit, however, is smaller and varies in sign across specifications. We conclude that, from an ex ante perspective, RTS rather than TFP heterogeneity better predicts differences in firm growth and survival over the life cycle.

In addition, we investigate whether firms with varying RTS respond differently to aggregate shocks. We use two types of aggregate shocks: one is the change in industry-level TFP, the other is the 2007-2008 global financial crisis. We estimate regressions of firm revenue growth on RTS, the aggregate shock, and the interaction

[27] We also ranked firms according to their average growth over 12 years. The top 1% fastest-growing firms (e.g., gazelles) have technologies with an average RTS of 0.98 versus 0.95 for those below the 90th percentile.

Figure 7 – Life cycle of firms starting with different RTS and TFP


(a) RTS

(b) TFP

Notes: Left and right panels of Figure 7 compare the life-cycle profile of revenue between firms with different initial RTS and TFP. It is constructed using a balance 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 and construct the TFP as described before. We bin firms into three groups based on their initial demeaned RTS in the left panel and three groups based on initial demeaned TFP in the right panel. Firm log revenue is also demeaned at the two-digit NAICS industry level.

between the two. The results are reported in Table OA.6. The results show that firms with larger RTS respond more to aggregate shocks (see also Smirnyagin (2023)).

4.4.2 Role of RTS Heterogeneity in Wealth and Wage Inequality

We conclude this section by examining the relationship between the wealth of firm owners, the wages of workers, and the RTS of firms. 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 owners using administrative records from the Shareholder Information in Corporate Tax Files.[28] 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 their RTS and TFP percentile by taking an equity-value-weighted average of the firms they own. The results, shown in Figure 8a indicate that high-wealth individuals tend to own firms with higher average

[28] In particular, Schedule 50 of Form T2 provides information on the percentage of shares owned by each shareholder above the 10% threshold and the type of shares owned (common or preferred). Statistics Canada tracks chained ownership by individuals (e.g., if individual A owns some shares of firm B and firm B owns some of firm C) and constructs a dataset of ultimate individual shareholders. We use this information to calculate equity wealth for owners. In Canada, self-employed individuals and business owners constitute 11.7% of individual tax filers. On average, each business owner has 1.96 firms in her portfolio, with a standard deviation of 2.34.

Table III – 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: Robust standard errors are clustered at the firm level. TFP percentile is calculated within each industry. Both RTS and TFP percentiles are standardized to have a mean of 0 and a standard deviation of 1. We first-difference both regressors in Column (2). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

RTS. In other words, wealthier firm owners' production technologies are more scalable. In addition, conditional on within-sector RTS, the TFP rank is also increasing in owners' wealth but in a concave manner, especially at the top of the distribution.

It is well established that large firms tend to pay higher wages to similar workers relative to smaller firms (Bloom *et al.*, 2018b; Brown and Medoff, 1989). Given our results, however, it is unclear whether this relationship is driven by higher TFP or higher RTS in large firms. To disentangle these two explanations, we rank firms by their average wage and compute average RTS and TFP within firm wage deciles. Figure 8b shows that higher-paying firms tend to have higher RTS—and thus tend to be larger. It shows a similar but weaker association between TFP and average wages. These findings suggest that RTS heterogeneity is an important driver of the wage–firm size relation.

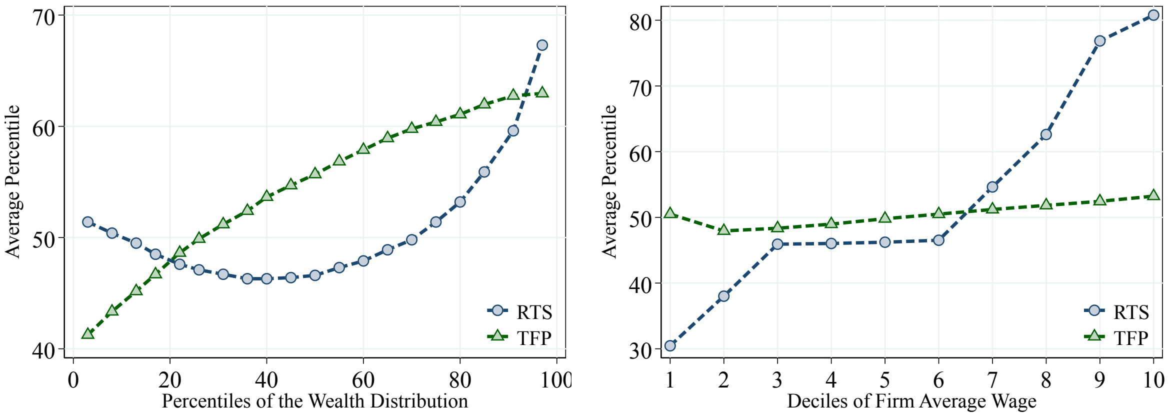
5 Misallocation with RTS Heterogeneity

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 distribution. In this section, we demonstrate why this heterogeneity is also important theoretically and quantitatively by focusing on a fundamental question in macroeconomics: the efficiency costs of misallocation arising from financial frictions. Our

Figure 8 – RTS and TFP by Owner's Wealth and Average Firm Wage

(a) by Owner's Wealth

(b) by Average Firm Wage



Notes: Left panel of Figure 8 shows the average RTS and TFP by percentiles of owners' equity wealth distribution. Right panel of Figure 8 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.

goal is to demonstrate the significance of RTS heterogeneity compared with TFP differences, so we incorporate firm heterogeneity in RTS (η) and TFP (z) into an off-the-shelf quantitative model of entrepreneurship. Our main application compares the aggregate effects of financial frictions in a model with heterogeneity in both η and z—the (η, z)-economy—against a standard setting with heterogeneity only in z—the z-economy. 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 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 ($\eta = 1$).[29] 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*

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

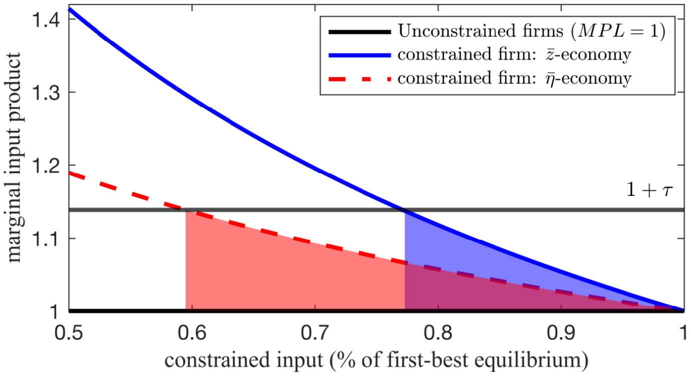
where $w_i \equiv \frac{y_i^{\star}}{Y^{\star}}$ denotes the relative output of firm i in the first-best equilibrium.

Proof. See Appendix D.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 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 are 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 9, 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.

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

Figure 9 – Efficiency costs in endowment 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. In robustness exercises at the end of this section, we show that explicitly modeling intermediate inputs as the driver of RTS differences (in line with our empirical findings) does not alter our main findings.

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

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

implying input choices $k(a, z, \eta), \ell(a, z, \eta)$ and output $y(a, z, \eta)$.[30] Thus, the dynamic agent problem can be written in recursive form as

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

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 D.2.

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.[31] We use a Cobb-Douglas production function

[30] 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).

[31] 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

(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.[32]

(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 rightmost 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 revenue distribution (middle 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 again 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

(η, z)-model and 2.2% in the z-model, in the ballpark of the value for Canada of 4.9%.

[32] 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**	
		(η, z)-*economy*	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	0.796	0.804
Top 1% revenue share	0.522	0.524	0.515
Top 0.1% revenue share	0.282	0.283	0.284
RTS: Top 5% vs bottom 50% (by revenue)	0.083	0.083	0*
Capital-output ratio	2.970	2.971	2.970
B. Internally calibrated parameters		(η, z)-*economy*	z-*economy*
Mean RTS	μ_η	0.782	0.683
Standard deviation RTS	σ_η	0.054	—
Standard deviation TFP	σ_z	0.614	0.910
Persistence TFP	ρ_z	0.954	0.970
Pareto tail TFP	ξ_z	—	2.880
Correlation (z, η)	$\sigma_{z,\eta}$	-0.262	—
Discount factor	β	0.890	0.902

Notes: Steady-state calibration of the (η, z)- and z-economy (both at $\lambda = 0.3$). * not targeted.

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.[33] This is the case in our model as well as in some of our empirical approaches (for instance, when we cluster firms in Section 5.2.4, such that firms within

[33] 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

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

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 10 compares the output losses induced by increasing the financial friction parameter λ from the unconstrained case of $\lambda = 0$ up to $\lambda = 1$, across the 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 the same observables, amplifies the output losses due to financial frictions by 147%.

To understand this finding, we decompose the total log output loss additively into three terms: (i) the static misallocation of production factors, holding fixed occupational choice, (ii) the misallocation of talent across occupations, and (iii) the under-accumulation of capital. 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

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 10 – Output losses from financial friction in dynamic model

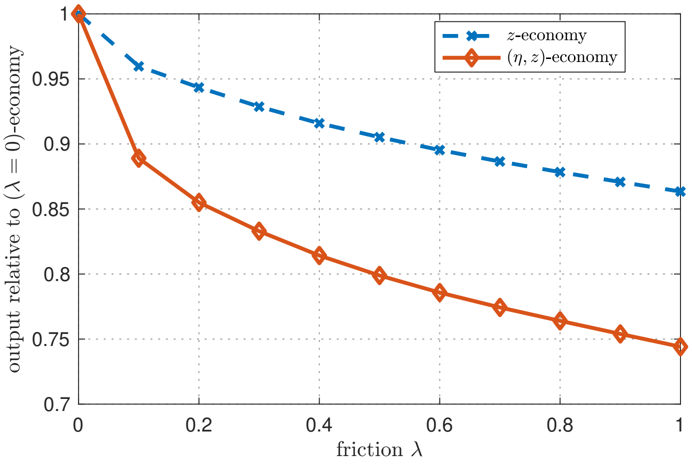


Table V – Dynamic model: Decomposition of output losses

	(η, z)-*economy*	z-*economy*
Total log GDP loss	18.3	7.4
. . . due to misallocation of production factors	10.6	5.0
. . . due to misallocation of talent	0.6	0.5
. . . due to K accumulation	7.1	1.9

Notes: This table 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 underaccumulation of capital.

findings are in line with Proposition 1. 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 additional elements such as fixed costs of entry and exit that could magnify the importance of the occupational choice channel.

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 the 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 the choice of technology and sector. In particular, a choice between a high fixed cost, low marginal cost and a low fixed cost, high marginal cost technology locally generates an increase in RTS across the size distribution. Our framework does not feature the choice of technology or sector, and as such the entry margin contributes little to the output losses from financial frictions. However, static misallocation is greatly amplified when allowing for differences in RTS among existing firms.

5.2.4 Robustness Checks

Our findings are robust to alternative approaches of making the financial friction comparable in both economies. In our benchmark scenario, we increase λ from 0 to 0.3 in both economies. Panel A of Table VI 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 conventional z-economy.

Intermediate Inputs. Our findings are also robust to including intermediate inputs in the production function and different specifications of the financial constraint. For these exercises, we modify the production function to $z \cdot k^{\alpha_K} \cdot \ell^{\alpha_L} \cdot m^{\eta-\alpha_K-\alpha_L}$,where α_K and α_L are constants and η denotes RTS as before. With this formulation, the capital and labor elasticities are common across firms. Variation in RTS is entirely driven by variation in intermediate inputs, reproducing the empirical finding that larger firms have higher RTS because of higher intermediate elasticities. We consider two different sets of models that differ in the formulation of the financial constraint. The calibration strategy closely mimics the baseline model, and we delegate these details to Appendix D.3.

First, we maintain that the financial constraint is symmetric across inputs: $w \cdot \ell + R \cdot k + m \leq \frac{a}{\lambda}$. As row 2 in Panel B of Table VI shows, GDP losses are magnified in

Table VI – Dynamic model: Robustness

	(η, z)-economy	*z-economy*
A. Log GDP loss: robustness to alternative comparisons		
1. Baseline: Equating λ	18.3	7.4
2. Equating aggregate debt/capital ratio	26.9	7.4
3. Equating dispersion in log marginal products	21.6	7.4
B. Log GDP loss: robustness to including intermediate inputs		
1. Baseline: w/o intermediate inputs	18.3	7.4
2. W/ intermediates: constraint on K,L,M	72.3	11.8
3. W/ intermediates: constraint on K,L	14.4	4.8

Notes: Panel A reports the total log GDP loss in alternative scenarios where 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 2), respectively marginal input product dispersion (row 3), of the z-economy with $\lambda = 0.3$. Panel B reports total log GDP losses when raising λ from 0 to 0.3 in alternative model versions. Row 1 corresponds to the baseline model without intermediate inputs. Rows 2 and 3 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, intermediate inputs are assumed to be fully flexible: $w \cdot \ell + R \cdot k \leq \frac{a}{\lambda}$.

both the version with and the one without RTS heterogeneity. This result is expected since the presence of intermediate inputs magnifies distortions (see, e.g., Baqaee and Farhi (2019)). However, it is still the case that the economy with RTS heterogeneity generates far larger GDP losses from financial frictions (72.3 vs. 11.8 log points).

Second, more in line with our empirical approach, we treat intermediate inputs as fully flexible, such that the constraint is $w \cdot \ell + R \cdot k \leq \frac{a}{\lambda}$. Thus, there is a sense in which high-η firms face less severe financial frictions, since the financial constraint applies to a smaller fraction $\frac{\alpha_L + \alpha_K}{\eta}$ of their inputs weighted by factor elasticities. Yet, as row 3 of Table VI (Panel B) shows, the economy with RTS heterogeneity still generates GDP losses that are three times as high as the one without RTS heterogeneity (14.4 vs. 4.8 log points).[34]We conclude from these exercises that incorporating intermediate inputs does not alter our main quantitative result: accounting for RTS heterogeneity significantly amplifies output losses from financial frictions.

6 Conclusion

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

[34] Less of our focus, the level of efficiency losses is slightly lower than in the baseline (row 3 vs. row 1). On the one hand, the presence of intermediates magnifies efficiency losses. On the other hand, the financial constraint is less severe as it only applies to a subset of inputs.

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 more than 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

Ackerberg, D. A., Caves, K. and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, **83** (6), 2411–2451. 2

Argente, D., Moreira, S., Oberfield, E. and Venkateswaran, V. (2024). *Scalable Expertise*. Tech. rep. 4

Atkeson, A. and Burstein, A. (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, **98** (5), 1998–2031. 4.2.1

—, 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. 13

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. 15

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

—, Guvenen, F., Smith, B. S., Song, J. and von Wachter, T. (2018b). The disappearing large-firm wage premium. In *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 108, pp. 317–322. 4.4.2

Boar, C. 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. 26

Brown, C. and Medoff, J. (1989). The employer size-wage effect. *Journal of political Economy*, **97** (5), 1027–1059. 4.4.2

Buera, F. J., Kaboski, J. P. and Shin, Y. (2011). Finance and development: A tale of two sectors. *American Economic Review*, **101** (5), 1964â2002. 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. 4.2.1

Cagetti, M. and De Nardi, M. (2006). Entrepreneurship, frictions, and wealth. *Journal of political Economy*, **114** (5), 835–870. 1, 4.4, 5.2

Chan, M., Mattana, E., Salgado, S. and Xu, M. (2024). *Dynamic Wage Setting: The Role of Monopsony Power and Adjustment Costs*. Working paper. 14, 4.1

Chen, C., Habib, A. and Zhu, X. (2023). Finance, managerial inputs, and misallocation. *American Economic Review: Insights*, **5** (3), 409â26. 4

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. 1, 5

David, J. M. 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. 4.2.1, 24, 25

—, Goldberg, P. K., Khandelwal, A. K. and Pavcnik, N. (2016). Prices, markups, and trade reform. *Econometrica*, **84** (2), 445–510. 4.2.1, 23, 24, 25

— 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. 4.2.1, OA.1

De Ridder, M., Grassi, B., Morzenti, G. *et al.* (2022). The hitchhiker's guide to markup estimation. 26

Demirer, M. (2020). Production function estimation with factor-augmenting technology: An application to markups. *Job Market Paper*. 1, 17

Edmond, C., Midrigan, V. and Xu, D. Y. (2023). How costly are markups? *Journal of Political Economy*, **131** (7), 1619–1675. 4.2.1

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, 17

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, 15

Gavazza, A., Mongey, S. and Violante, G. L. (2018). Aggregate recruiting intensity. *American Economic Review*, **108** (8), 2088–2127. 1, 5

Greaney, B. (2019). Housing constraints and spatial misallocation: Comment. *Working paper*. 33

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

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. 4

Hubmer, J., Halvorsen, E., Salgado, S. and Ozkan, S. (2024). *Why Are the Wealthiest So Wealthy? New Longitudinal Empirical Evidence and Implications for Theories of Wealth Inequality*. Tech. rep. 6

Hurst, E. and Pugsley, B. W. (2011). *What do small businesses do?* Tech. rep., National Bureau of Economic Research. 5, 4.4.1

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

Kline, P. M. (2024). *Firm Wage Effects*. Working Paper 33084, National Bureau of Economic Research. 4.4

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, 4.2.1

Lucas, R. E. (1978). On the size distribution of business firms. *The Bell Journal of Economics*, **9** (2), 508–523. 1, 4.3

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. 3

Midrigan, V. and Xu, D. Y. (2014). Finance and misallocation: Evidence from plant-level data. *American Economic Review*, **104** (2), 422â58. 5.2.3

Moll, B. (2014). Productivity losses from financial frictions: Can self-financing undo capital misallocation? *American Economic Review*, **104** (10), 3186â3221. 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

Quadrini, V. (2000). Entrepreneurship, saving, and social mobility. *Review of Economic Dynamics*, **3** (1), 1–40. 1, 4.4, 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. 15

Smirnyagin, V. (2023). Returns to scale, firm entry, and the business cycle. *Journal of Monetary Economics*, **134**, 118–134. 4.4.1, 5

Sterk, V., Sedláček, P. and Pugsley, B. (2021). The nature of firm growth. *American Economic Review*, **111** (2), 547–579. 4.3, 4.4

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[1] Mons Chan[2] Serdar Ozkan[3]
Sergio Salgado[4] Guangbin Hong[5]

A Appendix for the Canadian Data

We describe how we construct the variables and the estimation sample in this section.

A.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 terms.

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 $K_{i,t} = K_{i,t-1} + Invest_{i,t} - Amort_{i,t}, t \geq t_i^0$, and 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

[1] University of Pennsylvania; jhubmer@sas.upenn.edu
[2] Queen’s University; mons.chan@queensu.ca
[3] Federal Reserve Bank of St. Louis, University of Toronto; serdar.ozkan@gmail.com
[4] The Wharton School-University of Pennsylvania; ssalgado@wharton.upenn.edu
[5] University of Chicago; ghong7@uchicago.edu

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 Renumeration 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.

A.2 Sample Selection

We convert all the monetary variables to be denominated in 2002 Canadian dollars. Several steps are taken to construct the estimation sample. First, we drop firms with missing industry information. Second, we drop the first-year observation that we observe a firm's book value of tangible capital and the observations before, 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.

A.3 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 ended 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 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 is detailed by state, 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 covered by the CM above a certain threshold and a smaller sample of small and medium size firms. The number of firms in the raw data ios around 50,000 establishments per year. The merged CM/ASM contains consistent data in 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 trough the Census RDC. All the 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 year-firm observations with non missing values in real value of shipments (revenue), the real wage bill of workers in the establishment (employment), the real expenditure in intermediate inputs and materials (material), 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 trim the distribution and the 0.1%. 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. It is important to notice that this skews our sample to larger establishments that tend to stay in the ASM over small firms that tend to be replaced every 5 years.

B Additional Figures and Tables

Table OA.1 – Average production estimates by industry

Industry	NAICS	N	rtscale	melast	lelast	kelast
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
Manufacturing	31	69,100	1.01	0.61	0.37	0.03
	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 OA.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 OA.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 OA.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 OA.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 NAICS2 industry, and the within-industry results demean the output elasticities at two-digit NAICS level.

Table OA.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 OA.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 OA.6 – Regression of Changes in Log Revenue on 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

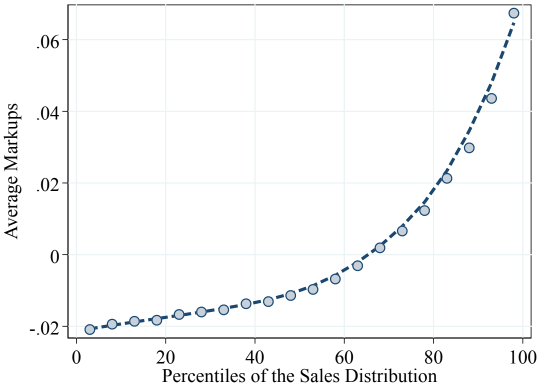
Notes: Robust standard error are clustered at the firm level reported. In columns (1)-(3), we use the industry-level change in TFP as the aggregate shock, which is calculated as the average firm-level TFP, ν_{jt}, for all firms in the industry in that year. In columns (4)-(6), we use a time dummy for 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 OA.7 – 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

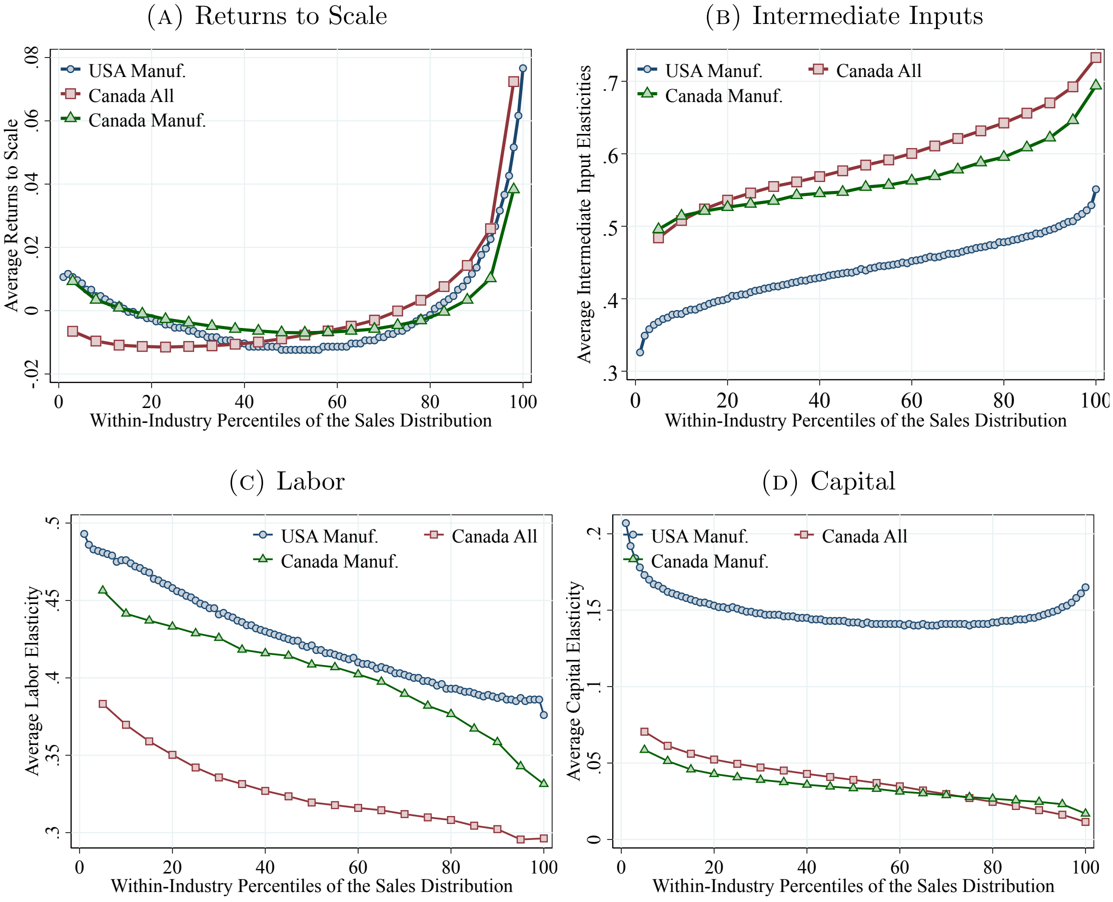
Notes: Table OA.7 reports the regressions of firm RTS on log firm revenue at the firm-year level. Estimation results are from the specification where we cluster firms by the maximum attained size (see Section 4.2.1 for more details). Column (1) includes industry fixed effects, and column (2) further includes cluster fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Figure OA.1 – Estimated Markups and Firm Revenue



Notes: Figure OA.1 shows estimated markup across the firm-size distribution. We follow the value-added translog production function method as in De Loecker and Warzynski (2012). We estimate the production function by industry. In all figures, we sort firms by within-industry revenue ranks and plot the average within ranks. The figure shows the markup relative to the industry average.

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


(a) Returns to Scale

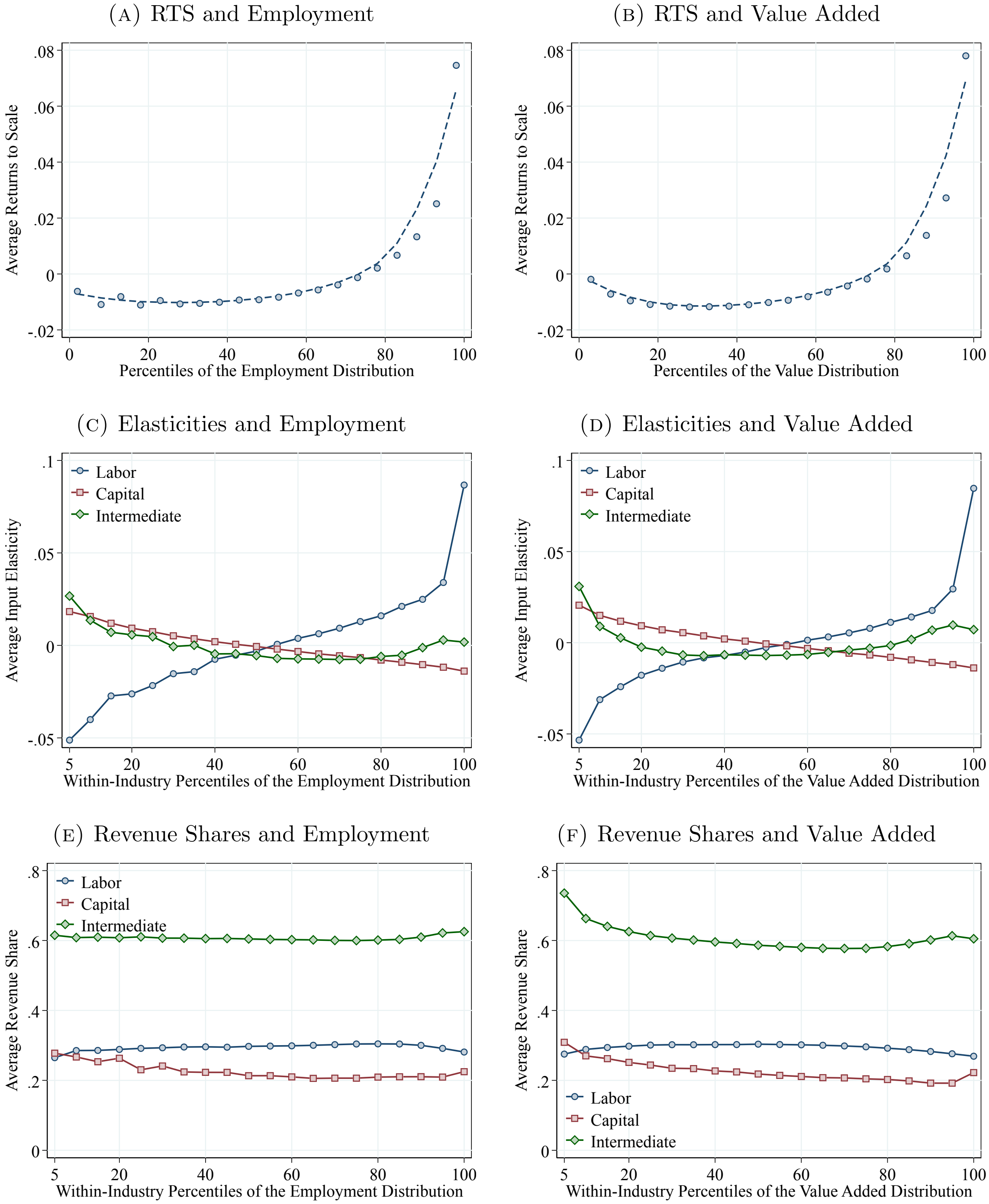
(b) Intermediate Inputs

(c) Labor

(d) Capital

Notes: Figure OA.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 OA.3 – Results by Employment and Value Added


(a) RTS and Employment

(b) RTS and Value Added

(c) Elasticities and Employment

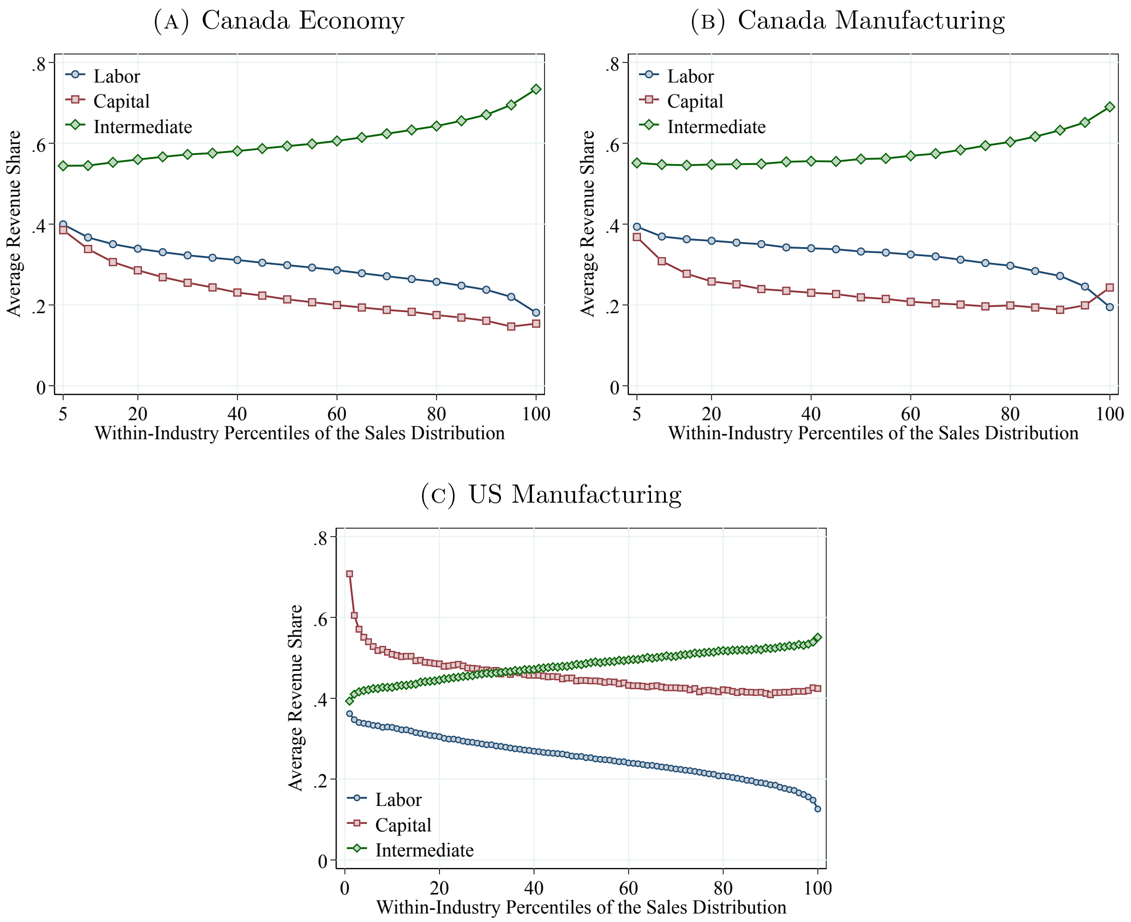
(d) Elasticities and Value Added

(e) Revenue Shares and Employment

(f) Revenue Shares and Value Added

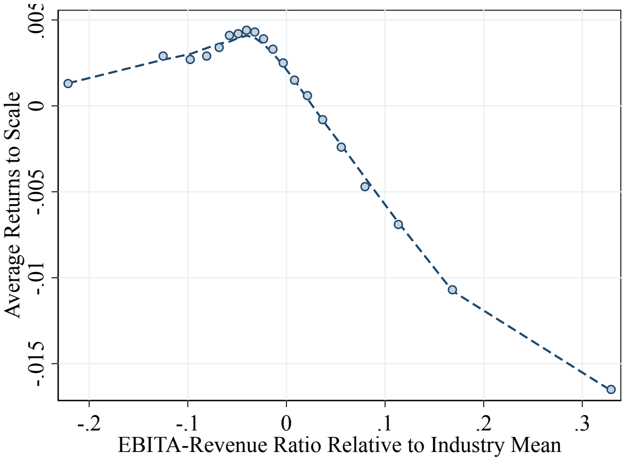
Notes: Figure OA.3 shows results sorting firms by within-industry employment ranks (left panels) and within-industry value added ranks (right panels). We use the intermediate input and labor costs and the value of the capital stock to construct the revenue shares.

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



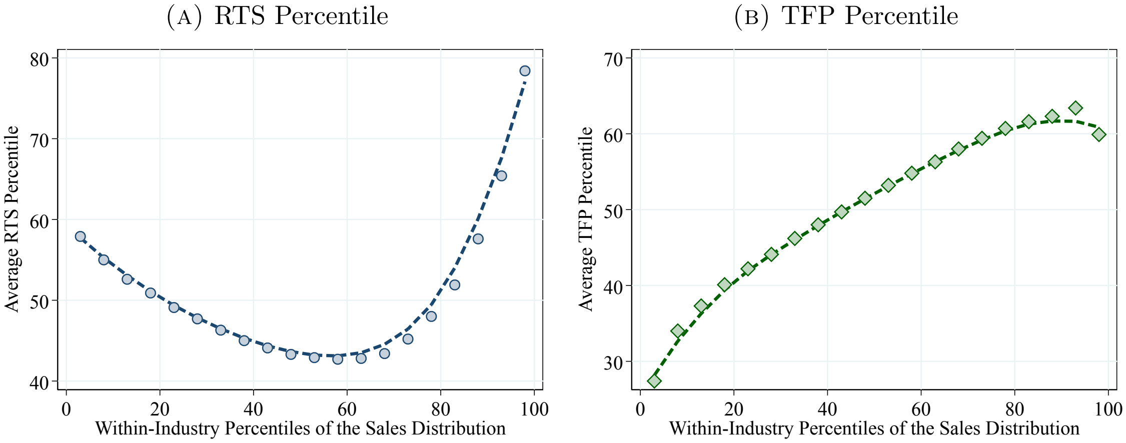
Notes: Figure OA.4 shows revenue shares across the firm-size distribution for Canada and for the US manufacturing sector. In each plot, we sort firms by within-industry revenue ranks and then average the revenue share across all firms within corresponding percentiles. We use the intermediate input and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

Figure OA.5 – Profits and Returns to Scale



Notes: Figure OA.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.

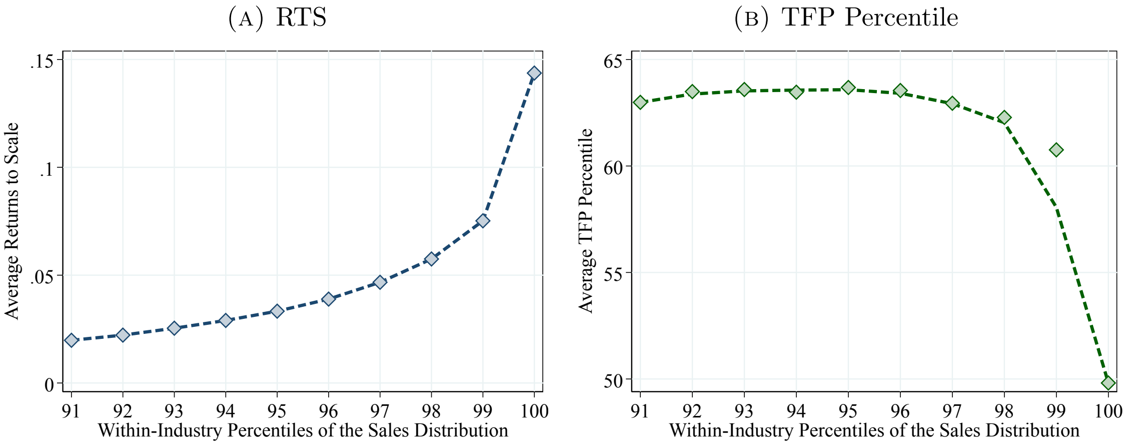
Figure OA.6 – RTS and TFP Percentile across the Firm Revenue Distribution


(a) RTS Percentile

(b) TFP Percentile

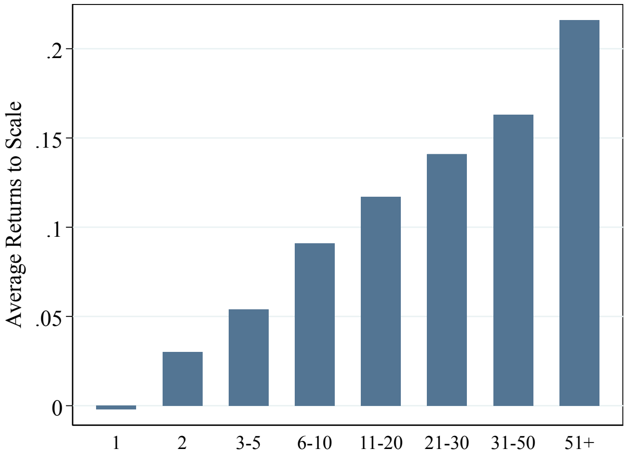
Notes: Figure OA.6 plots the TFP percentile across the firm-size distribution. We calculate the RTS and TFP percentiles for each firm-year observation within an industry. We sort firms by within-industry revenue ranks and plot the average RTS and TFP percentiles within revenue ranks.

Figure OA.7 – RTS and TFP Estimates for Top 10% Firms



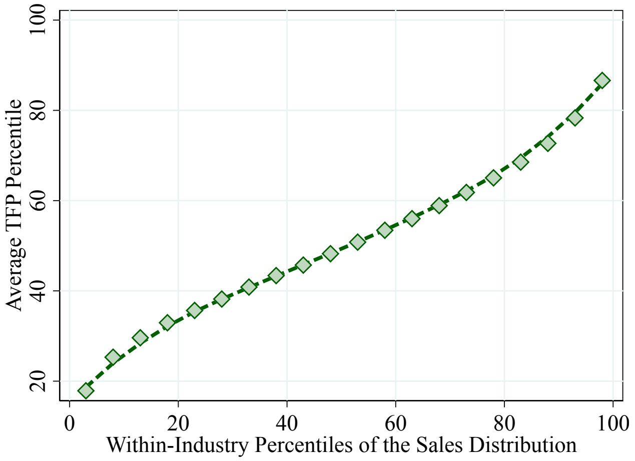
Notes: Figure OA.7 plots the RTS and TFP estimates against the firm sales percentile for the top 10% firms. In both panels, we sort firms by within-industry revenue ranks and plot the average within ranks. Panel A shows the returns to scale relative to the industry average. Panel B shows the TFP percentile calculated within RTS-industry bins.

Figure OA.8 – RTS and the Number of Establishments



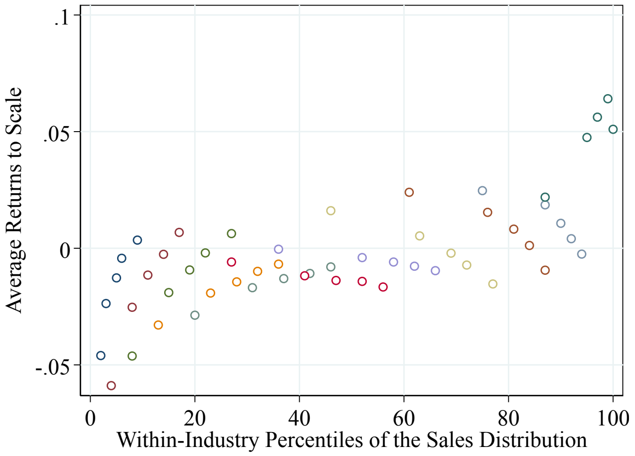
Notes: Figure OA.8 plots the average RTS for eight groups of firms with a different number of establishments. RTS is demeaned at the industry level.

Figure OA.9 – Robustness: TFP Percentile across the Firm Revenue Distribution, Cobb-Douglas Production Function



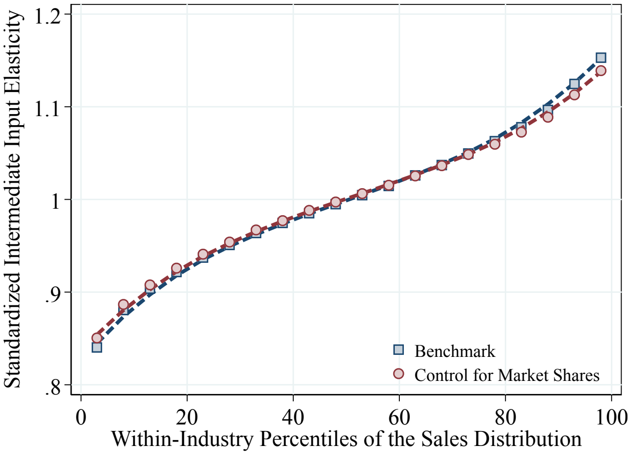
Notes: We re-estimate a Cobb-Douglas production function for each industry. We plot the relationship between TFP percentile and sales percentile. Both TFP and sales percentiles are calculated within industry.

Figure OA.10 – Robustness: RTS across the Firm Revenue Distribution, Clustered by Maximum Size



Notes: Figure OA.10 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 OA.11 – Robustness: Estimation of Intermediate Input Elasticity, Controlling for Market Shares



Notes: Figure OA.11 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 RTS Variance-Component Model

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 firm 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 the 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 OA.8 – 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%

D Model Appendix

D.1 Proof of Proposition 1

Without loss of generality, set the productivity of the constant-returns-to-scale (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:

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

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

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

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:

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

D.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:

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

D.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), choosing 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,[6] and estimate the parameters in Table OA.9 using the exact same strategy as in our baseline model versions.

[6] 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 OA.9 – Dynamic model w/ intermediates: Calibration

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

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