ECONOMIC RESEARCH
FEDERAL RESERVE BANK OF ST. LOUIS

WORKING PAPER SERIES

Scalable vs. Productive Technologies

Authors	Mons Chan, Guangbin Hong, Joachim Hubmer, Serdar Ozkan, and Sergio Salgado
Working Paper Number	2024-019A
Creation Date	August 2024
Citable Link	https://doi.org/10.20955/wp.2024.019
Suggested Citation	Chan, M., Hong, G., Hubmer, J., Ozkan, S., Salgado, S., 2024; Scalable vs. Productive Technologies, Federal Reserve Bank of St. Louis Working Paper 2024-019. URL https://doi.org/10.20955/wp.2024.019

Federal Reserve Bank of St. Louis, Research Division, P.O. Box 442, St. Louis, MO 63166

The views expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Federal Reserve System, the Board of Governors, or the regional Federal Reserve Banks. Federal Reserve Bank of St. Louis Working Papers are preliminary materials circulated to stimulate discussion and critical comment.

Scalable vs. Productive Technologies[*]

Mons Chan[†] Guangbin Hong[‡] Joachim Hubmer[§] Serdar Ozkan[¶] Sergio Salgado[‖]

July 11, 2024

Abstract

Do larger firms have more *productive* technologies or are their technologies more *scalable*, or both? We use administrative data on Canadian and US firms to estimate flexible nonparametric production functions. Our estimation results in a joint distribution of output elasticities of capital, labor, and intermediate inputs—therefore, returns to scale (RTS)—along with total factor productivity (TFP). We find significant heterogeneity in both RTS and TFP across firms. Larger firms operate technologies with higher RTS, both across and within industries. Higher RTS for large firms are entirely driven by higher intermediate input elasticities. Descriptively, these align with higher intermediate input revenue shares. We then incorporate RTS heterogeneity into an otherwise standard incomplete markets model with endogenous entrepreneurship that matches the observed heterogeneity in TFP and RTS. In this model, we find that the efficiency losses of financial frictions are more than twice as large relative to the conventional calibration that loads all heterogeneity on TFP and imposes a common RTS parameter.

Keywords: Production function heterogeneity, returns to scale, misallocation

[*] For helpful comments, we thank seminar participants at the Barcelona GSE Summer Forum, SED, Bonn University, Goethe University, University of Wisconsin, and Wharton. 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.

[†] Queen's University; mons.chan@queensu.ca

[‡] University of Chicago; ghongecon@gmail.com

[§] University of Pennsylvania; jhubmer@sas.upenn.edu

[¶] St. Louis Fed, University of Toronto; serdar.ozkan@gmail.com

[‖] The Wharton School-University of Pennsylvania; ssalgado@wharton.upenn.edu

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). The earlier literature focused on this dimension of heterogeneity to explain the firm size distribution (at least within industries). For example, the misallocation literature (pioneered by Restuccia and Rogerson (2008) and Hsieh and Klenow (2009)) quantified the efficiency costs of differences in the marginal product of capital across firms with different TFP (a notable exception is David and Venkateswaran (2019)). Moreover, models of entrepreneurship with heterogeneous TFP and decreasing returns to scale production functions have been used to explain persistent differences in rates of return and wealth inequality (e.g., Cagetti and De Nardi (2006)).

In this paper, we allow for more general heterogeneity in production technologies among firms—focusing on differences in returns to scale (RTS)—and investigate how these differences shape the firm size distribution. We employ a broad set of estimation methods and use panel data on firms to document significant heterogeneity in production technologies across firms. Our benchmark method builds on Gandhi, Navarro and Rivers (2020) (henceforth GNR), and provides us with a joint distribution of output elasticities of labor, capital, and intermediate inputs—thereby, RTS—as well as TFP, at the firm-year level. We then study whether larger firms have technologies that are more *productive* (high TFP) or more *scalable* (high RTS). Finally, we show why the answer to this question matters in an application to efficiency costs of misallocation from financial frictions, indicating the potential role of RTS heterogeneity for an array of quantitative questions, such as optimal taxation of capital, firm hiring decisions (as in Gavazza *et al.* (2018)) or firm growth and cyclicality (as in Clymo and Rozsypal (2023)).

In our main empirical analysis, we use administrative panel data for the universe of incorporated Canadian firms—which account for more than 90% of private business sector output—from 2001 to 2019 from the National Accounts Longitudinal Microdata file. This dataset contains firms' balance sheet information on revenues as well as the total cost of labor, capital, and intermediate inputs. The measurement of labor and intermediate inputs is consistent with Statistics Canada's national accounts, and we construct the capital stock for each firm using the perpetual-inventory method (as standard in the literature). After sample selection, our sample includes 4.3 million firm-year observations. We also corroborate our main findings from Canada with similar empirical results for manufacturing firms in the US using administrative data from the Annual Survey of Manufactures and Economic Census.

In our benchmark approach, we estimate nonparametric production functions building on the method of GNR.[1] This technique leverages standard assumptions on profit maximization, adjustment costs, and input choice timing to recover firm-level measures of TFP, output elasticities, and returns to scale. The non-homothetic production function is identified from variation in input expenditure shares, as well as from the covariance between in input and output levels, controlling for the endogeneity of inputs to TFP. Intuitively, higher returns-to-scale firms are those with higher expenditure shares and/or a higher covariance of input and output levels.

We apply the GNR method to each 2-digit NAICS industry, and for each year in our sample. In addition to considerable heterogeneity in TFP (as the previous literature also documented), we find large differences in RTS among Canadian firms (even within industries). The mean of our estimated RTS equals 0.97, with considerable variation across industries.[2] More novel, we document significant variation in RTS within 2-digit NAICS industries. The average within-industry differential between the 90^{th} and 10^{th} percentiles (P90-P10) equals 0.07. These differences imply that, for a given level of TFP, when increasing the input bundle by 1%, the output of the firm at the 90^{th} percentile increases by 7% more relative to the increase in output of the firm at the 10^{th} percentile. Interpreted as deviations from constant returns to scale, these differences are large.[3] Furthermore, these differences are mainly due to ex-ante

[1] In particular, our estimation involves two steps. First, we employ the factor share or first-order condition approach to nonparametrically identify the output elasticity of the fully flexible intermediate input, and transitory productivity shocks. In the second step, we use these estimates to identify the output elasticities of capital and labor along with the persistent TFP shocks by exploiting dynamic panel (or proxy) variable conditional moment restrictions based on lagged input decisions.

[2] A few papers have documented heterogeneity in RTS. For instance, Gao and Kehrig (2017) reports 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. None of these papers, however, look at the relation between RTS, TFP, and firm size within industries.

[3] E.g., with Cobb-Douglas production functions in an efficient economy, the elasticity of optimal firm output to firm TFP is $\frac{1}{1-RTS}$. This elasticity is 3.3 times larger for a firm with RTS of 0.97 compared to a firm with RTS of 0.90.

production technology heterogeneity. Using our firm-year level estimates, we run a panel regression for RTS and find that firm fixed effects account for 86% of the overall variation after controlling for firm size and age. Thus, while we estimate production functions at the 2-digit industry-year level, we interpret heterogeneity between firms as mainly stemming from permanent differences.

The dispersion in the output elasticities of capital, labor, and intermediates is larger than dispersion in RTS, meaning that firms' technologies differ both in their relative factor intensities and in the overall RTS. The P90-P10 of our estimated intermediate input elasticity and of the labor elasticity are both equal to 0.35, while the P90-P10 dispersion of the capital elasticity is lower, at 0.07. The estimated output elasticity patterns align with the corresponding revenue shares of the respective input. In particular, the intermediate input shares in firm revenue closely mirror the estimated intermediate input elasticities, which is expected as our estimation treats them as flexible input. For labor and capital, the correlation between observed revenue shares and estimated elasticities is less than perfect but still strongly positive. Again, this is expected as these two inputs are potentially subject to adjustment costs and input market power, and capital is pre-determined.

Next, we investigate the variation in production technologies across the revenue distribution. Our main empirical finding is that RTS is increasing in firm size, especially above the median: When pooling firms across the economy we find that, on average, the largest 5% of firms exhibit RTS that are 10 percentage points (pp) higher compared to those in the bottom 50%. Similarly, within 2-digit NAICS industries, the largest 5% firms within industries exhibit RTS that are 8 pp higher compared to the bottom 50%. When we zoom into the respective output elasticities, we find that the overall increase in RTS along the revenue distribution is entirely accounted for by an increase in the output elasticity of intermediate inputs. In contrast, labor and capital elasticities tend to decline along the revenue distribution, though this finding is less robust across samples and specifications.

We find that TFP increases in firm revenue until the top 10% of the revenue distribution, after which it then flattens out. RTS instead increases in a convex fashion at the top of the revenue distribution, indicating a more important role for RTS differences in shaping the right tail of the firm revenue distribution. In sum, our empirical results indicate that the largest firms tend to feature the highest RTS, and not necessarily the highest TFP as commonly assumed.

Our results are robust along several dimensions. We find similar results whether we estimate a production function across the entire economy or within narrow industry groups. Methodologically, our main findings are unchanged when (i) clustering firms according to input shares and estimating separate production functions within clusters, (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 a positive relation between plant size and RTS in the US manufacturing sector.

We also show that high-RTS firms grow faster over the life-cycle and are less likely to exit, in addition to having a higher output level. Furthermore, high-paying firms on average feature higher RTS. These secondary findings indicate that incorporating realistic RTS heterogeneity is important for a wide range of applications, including wage inequality.

Finally, we link firms to their owners. Many models of wealth inequality (building on Quadrini (2000); Cagetti and De Nardi (2006)) feature entrepreneurs that operate production technologies with heterogeneous TFP but common (and decreasing) RTS. Hence, in these models, the right tail of the wealth distribution is populated by households operating the highest TFP technologies. Testing this assumption, we first document that wealthier households invest in firms with higher RTS. Second, firm TFP increases with owner wealth, but flattens out at the top of the wealth distribution, where RTS increases even more steeply. These findings suggest that wealthier households tend to invest in firms with high RTS (i.e., in more scalable technologies), and not necessarily in high-TFP firms as commonly assumed; therefore, they have important implications for the optimal taxation of wealth and income from wealth (e.g., Guvenen *et al.* (2023); Gaillard and Wangner (2021)).

To investigate the implications of our findings, we incorporate heterogeneous RTS into a standard incomplete markets heterogeneous agents model with endogenous entrepreneurship (e.g., Quadrini (2000); Cagetti and De Nardi (2006)). In the model, agents have an occupational choice between supplying their stochastic efficiency units of labor or operating a private business. The private business technology is stochastic.

Inputs choices are constrained by a friction λ such that, for each dollar spent on inputs, the entrepreneur needs to finance at least a fraction λ with their private wealth. More novel, an entrepreneur's output depends not only on a standard stochastic idiosyncratic TFP term z, but also on a stochastic idiosyncratic RTS term η. We then investigate the output and productivity effects of changing the financial conditions in this economy. That is, we evaluate the impact of raising λ from zero—the unconstrained case—to positive values all the way up to one—the fully constrained case where all expenditures have to be financed with the owner's equity.

Our main exercise compares the effects of increasing the friction λ in two economies: the conventional z-economy, where there is no variation in η and which fundamentally loads all heterogeneity on TFP, and the (η, z)-economy, which features the observed joint distribution of RTS and TFP that we estimate in our empirical work. We calibrate the two economies such that they agree on key observable moments; in particular, both economies feature the same firm size distribution.

Our main finding is that in the (η, z)-economy, financial frictions generate more than double the output losses relative to the z-economy. 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 translates into more misallocation if the constrained firms have relatively higher RTS—a feature that our dynamic model generates endogenously. Dynamic effects also lead to more output losses in the economy with RTS heterogeneity, both because of underaccumulation of capital and because the selection intro entrepreneurship is relatively more distorted. If a potential superstar (but currently poor) entrepreneur has a business idea that is highly productive (i.e., has a high z), then her business is already profitable even at a small scale, and it is relatively easier to outgrow the friction. Instead, if the business idea is highly scaleable (i.e., high η), but not necessarily profitable at a small scale, it is harder to outgrow the friction, and she might never enter entrepreneurship.[4] Given these results, we conclude that taking into account RTS heterogeneity is crucial for a broad set of quantitative questions related to misallocation such as wealth inequality and

[4] 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).

optimal taxation of capital.[5]

2 Empirical Methodology

2.1 Estimating Returns to Scale

Our main empirical approach builds on the production function estimation methodology developed by GNR. We use a slightly modified GNR procedure to estimate a flexible non-parametric revenue production function which allows us to recover firm and time-varying measures of total factor productivity (TFP) and returns to scale (RTS). This production function relates each firm-year pair's output Y_{jt} to its inputs, including capital stock K_{jt}, labor input L_{jt}, and intermediate inputs M_{jt}, with the following assumption:

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 how to separate productivity shocks that affect firm's output from input choices. Following GNR, we use the first-order conditions of the firm as well as timing assumptions on the nature of productivity and input choices to form moment conditions. We illustrate the details below.

The estimation technique of GNR provides several advantages relative to other more standard methods. First, it offers a robust estimation method of *gross* production functions which is important if we want to obtain precise estimates of output elasticities and thus returns to scale. Second, the non-parametric identification strategy minimizes specification error when measuring both input elasticities and the productivity term which is crucial if we aim to understand the relation between firm-level productivity, returns to scale, and firm size.

Following the literature, we rely on standard timing assumptions to identify the parameters of the firm production function. Define $\mathcal{I}_{jt}$ as the information set available

[5] E.g., in ongoing work we study the entrepeneurial activity of New Money and Old Money households (à la Hubmer *et al.* (2024)), focusing on differences in their production technologies.

to firm j when it enters period t. $\mathcal{I}_{jt}$ contains all the information relevant to the firm (i.e., firm productivity, current capital stock, and so on) when it makes its period-t decisions. Following GNR, we define any input $X_t \in \mathcal{I}_{jt}$ as *predetermined.* Any such input is thus a function of the previous period's information set: $X_t(\mathcal{I}_{jt-1})$. We will treat capital as a predetermined input. Inputs which are not predetermined (and thus are set in period t) we define as *variable.* We define any input which is variable and where the optimal choice of X_t is a function of lagged values of itself as *dynamic.* We will depart from GNR in assuming that labor is a dynamic input. Finally, we define an input which is variable but not dynamic as *flexible.* Intermediate inputs will be treated as flexible in our framework. This implies that 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} can be decomposed into a persistent component, ω_{jt}, which is known to the firm when it makes input decisions, and a transitory component, ε_{jt}, which is unknown to the firm when making input decisions in period t. Notice, the changes in the productivity terms can be due to both technology shocks and market demand shifts.

Assumption 3. *The permanent 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 purely transitory productivity innovation, $\varepsilon_{jt} \notin \mathcal{I}_{jt}$, is i.i.d. across firms and time and it is not observed by the firm prior to making period-t decisions, with $P_\varepsilon(\varepsilon_{jt}|\mathcal{I}_{jt}) = P_\varepsilon(\varepsilon_{jt})$.*

We normalize $\mathbb{E}[\varepsilon_{jt}] = 0$ and define $\xi_{jt} = \omega_{jt} - \mathbb{E}[\omega_{jt}|\omega_{jt-1}]$ implying $\mathbb{E}[\xi_{jt}|\mathcal{I}_{jt-1}] = 0$.

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 conditional (on period-t labor) demand function 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 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.*

Following GNR, the assumptions 1 to 5 give us the following first-order condition 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. Multiplying both sides by M_{jt}/Y_{jt}, plugging in the production function, and rearranging provides our first estimating equation:

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

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 materials. Since by assumption 3 we have $\mathbb{E}[\varepsilon_{jt}] = 0$, we can use equation 1 to identify ε_{jt} and $D^{\mathcal{E}}$.

Given $\varepsilon_{jt} = \ln \left(D^{\mathcal{E}}(k_{jt}, \ell_{jt}, m_{jt})\right) - s_{jt}$, 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 non-parametrically.[6] In particular we have

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

Define $\tilde{y}_{jt} \equiv y_{jt} - \varepsilon_{jt} - \mathcal{D}(k_{jt}, \ell_{jt}, m_{jt}) = \mathit{\Psi}(k_{jt}, \ell_{jt}) + \omega_{jt}$. Plugging in the structure of ω_{jt} from assumption 3, we get our second estimating equation:

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

[6] This result requires one further technical assumption on the support of (k_{jt}, ℓ_{jt}) – see Assumption 5 in GNR.

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 $\mathbb{E}[\xi_{jt}|k_{jt}, \ell_{jt-1}, k_{jt-1}, \tilde{y}_{jt-1}, \ell_{jt-2}] = 0$, which we use with equation 3 to identify Ψ, h, and thus ξ_{jt}.[7]

Our estimation procedure follows GNR in using a standard sieve-series estimator to non-parametrically identify the input 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 $\widehat{\varepsilon_{jt}}$ 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 $\widehat{\mathcal{E}} = E[\exp(\hat{\varepsilon}_{jt})]$ and the input elasticity

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

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 the constant of integration and TFP only up to an additive constant, we follow GNR in normalizing Ψ to have mean zero, which implies that the constant of integration will show up in the firm productivity level. This gives us the following second-stage estimating equation:

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

[7] This differs from the moments used by GNR, who assume that labor is predetermined.

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 $E[\xi_{jt}\tilde{y}^a_{jt-1}]$ and $E[\xi_{jt}k^{\tau_k}_{jt}\ell^{\tau_\ell}_{jt-1}]$ 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 $\hat{\hat{\omega}}_{jt} \equiv \hat{h}(\hat{\omega}_{jt-1})$. The estimated production function also provides estimates of the output elasticities of capital and labor, which combined with the previously estimated output elasticity for intermediate inputs, gives us a firm-level measure of returns to scale. This is the specification we use to obtain the main empirical results in the next section.

2.2 Identification and Intuition

While GNR provides a rigorous identification argument for their strategy (and we defer to them for the details), here we discuss the intuition behind our estimation results. We use the non-parametric identification strategy of GNR to estimate returns to scale for two reasons. First, this method allows us to identify the output elasticities for a gross production function, while other common estimation methods (such as Ackerberg *et al.* (2015)) are typically only able to identify value-added production functions. As we show below, variation in the output elasticity of intermediate inputs is a key driver of variation in RTS, necessitating identification of the gross production function to learn about RTS.

Second, the non-parametric approach provided by GNR allows us to estimate a non-homothetic production function whereby the output elasticities and overall RTS vary by inputs and input shares, thereby, potentially varying across the firm size distribution. In particular, we first recover the output elasticity of intermediate inputs, $\varepsilon^Y_{M_{jt}} = \varepsilon^Y_M(k_{jt}, \ell_{jt}, m_{jt})$, which is a function of input levels through the shape of the production function. This elasticity is identified from the mean intermediate expenditure share of revenue, as well as from covariation between expenditure shares and input levels. Intuitively, if the true underlying production function is Cobb-Douglas, then the expenditure share will be uncorrelated with input levels and the output elasticity will be a constant equal to the mean expenditure share. This direct relationship is implied by the assumption that firms are price-takers in intermediate

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

input markets and do not face adjustment costs when choosing the level of m_{jt}.

We also allow for adjustment costs for capital and labor, and make no assumptions about the optimality of either input choice, driving an (unknown) wedge between expenditure shares and output elasticities. Instead, the output elasticities of capital $\varepsilon^Y_{K_{jt}} = \varepsilon^Y_K(k_{jt}, \ell_{jt}, m_{jt})$, and labor $\varepsilon^Y_{L_{jt}} = \varepsilon^Y_L(k_{jt}, \ell_{jt}, m_{jt})$ are identified jointly from their contribution to the variation in intermediate expenditure shares (e.g., $\frac{\partial}{\partial \ell_{jt}} \mathcal{D}(k_{jt}, \ell_{jt}, m_{jt})$ for labor) and their covariation with (residualized) output (e.g., $\frac{\partial}{\partial \ell_{jt}} \tilde{y}_{jt}$). We define the firm's returns to scale as $\eta_{jt} = \eta(k_{jt}, \ell_{jt}, m_{jt}) = \varepsilon^Y_{K_{jt}} + \varepsilon^Y_{L_{jt}} + \varepsilon^Y_{M_{jt}}$.[9] Thus, in the data, a high RTS firm will be a firm which has a high intermediate input expenditure share of revenue, or a high correlation between output and capital and/or labor.[10]

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.[11] 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 calculation of the national account system. Therefore, our micro data is 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

[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 detailed industry and capital-ratio clusters. See Section 4 for details. Conditional on a firm grouping, our procedure still allows output elasticities and RTS to vary by firms' input levels.

[10] Note that due to the non-homotheticity of the production function, these correlations are functions of input levels and thus vary across firms.

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

and costs of goods sold net of capital amortization. We measure the wage bill as the total worker compensation used to construct the National Account. All nominal monetary values are converted to 2002 real Canadian dollars.

We link firms to their owners using administrative records from the Shareholder Information in Corporate Tax Files. In particular, Schedule 50 of T2 form provides information on shareholders who own at least 10% of shares, the percentage of shares owned by each shareholder above the 10% threshold, and the type of shares owned (common or preferred). Statistics Canada also 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 shareholder. We merge the ownership information and demographic characteristics (i.e., age and gender) to the firm panel dataset and calculate total equity wealth for each individual, measured as the share weighted sum of the book value of all holding firms.

To construct the estimation sample, we start from firm-year observations with non-missing 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. We then further drop the observations with outlier factor shares. Specifically, we drop firm-year observations with (i) wage bill-to-revenue ratio below the 1st percentile or above the 99th percentile, (ii) wage bill-to-value added ratio below the 1st percentile or above the 99th percentile, (iii) intermediate input-to-revenue ratio greater than 0.95 or smaller than 0.05, and (iv) capital-to-revenue ratio above the 99.9th percentile. This sample selection leaves us with 4.3 million firm-year observations and 620 thousand firms with an average of 6.9 observations per firm. Summary statistics are reported in Table I. [12]

US Manufacturing Sector. As a robustness exercise, we perform similar analysis using data from from the US Economic Census and the Annual Survey of Manufactures (ASM). Our sample considers all manufacturing plants available in the dataset between 1978 and 2019. We measure (revenue) productivity for all plants with information on total value of shipments, real capital stock, wages of plant workers, and information on intermediate inputs and materials in 2019 US dollars.

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

Table I – Summary Statistics

	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

Note: Table I shows cross-sectional moments of the distribution of log revenues, log intermediate inputs, log wage bill, and log capital stock. All variables in 2002 Canadian dollars. The total number of observation is 4.3 million firm-years.

4 Empirical Results

In this section, we apply the GNR method to each of the 23 2-digit NAICS industries in our Canadian administrative data (see Table OA.3 for a list of these industries), which provide estimates of output elasticities of inputs for all the firm-year observations in our sample along with their TFPs. Below, we first present unconditional moments of these estimated parameters from this pooled sample. We then show how they vary over the firm size distribution as well as over the life cycle. Although GNR analytically proves identification for their method, we present the features of the data that have pronounced effects on our parameter estimates as an informal identification argument. We further relate our results to wage and wealth inequality as well as lifecycle growth of firms.

4.1 Unconditional Heterogeneity in Production Technologies

We start by discussing cross-sectional moments of the distribution of firm technologies. In particular, we first calculate within-industry moments from the distribution of firm-level estimates and then we average across industries. Our results, shown in Table II, reveal considerable heterogeneity in the estimated productivity, output elasticities, and returns to scale across firms. Starting with within-industry TFP differences, we find that the 90^{th}-to-10^{th} percentile gap (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 is substantially lower than previous estimates of productivity dispersion even for narrow 6-digit industries in Canada and the USA,

Table II – Distribution of Production Function Estimates

	Mean	St. dev	P10	P50	P90	P99
Panel A: Main Estimates						
TFP	—	0.16	-1.71	-1.55	-1.40	-1.06
RTS	0.97	0.04	0.93	0.97	1.01	1.07
Panel B: Output Elasticities						
Intermediates	0.61	0.14	0.43	0.60	0.79	0.99
Labor	0.33	0.14	0.15	0.33	0.49	0.64
Capital	0.03	0.03	0.00	0.03	0.07	0.12
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

Note: Table II shows cross-sectional moments of the distribution of firm-level log TFP, returns to scale (RTS), and the elasticities of output with respect to Intermediate inputs, labor and capital. To obtain these estimates, we apply the method of Section 2 within two-digits NAICS and calculate 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.

which typically find P90-P10 TFP ratios closer to 2 (see for instance, De Loecker and Syverson (2021) and Syverson (2011)). The difference stems from the use of a flexible non-parametric production function estimation, 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 TFP for the entire Danish private sector and find a P90-P10 ratio of 0.54. We find similar results for manufacturing plants in the US.

The average return-to-scale (RTS) is estimated to be 0.97 with a P90-P10 of 0.07. This implies that with a 1% higher input bundle, the firm at the 90^{th} percentile produces 7.3% ($= e^{0.07} - 1$) more output than the firm at the 10^{th} percentile, holding productivity constant. Interpreted as deviations from constant returns to scale, these differences are large. For example, with Cobb-Douglas production functions in an efficient economy, the elasticity of optimal firm output to firm TFP is $\frac{1}{1-RTS}$. This elasticity is 3.3 times larger for a firm with RTS of 0.97 compared to a firm with

Table III – Within-Industry Variance of Elasticity Estimates

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

Note: Table III shows the within-industry (NAICS2 and NAICS4) variations for the three output elasticities and return-to-scale estimates. It includes both the within-industry fraction of total variance and the within-industry standard deviation.

RTS of 0.90. Around 75% of firms have return-to-scale below 1, that is, they operate decreasing return-to-scale technologies. Consistent with earlier literature (e.g., Basu and Fernald (1997); Ruzic and Ho (2023); Gao and Kehrig (2017)) we find large differences in average RTS across industries (Table OA.3), ranging from 0.59 (for healthcare) to 1.03 (for Management of Companies and Enterprises). On top of these large between-industry differences, we further find significant within-industry variation. For example, average within industry P50-P10 and P90-P50 are around 0.03 and 0.04, respectively. In fact, as shown in Table III, a quarter of the total RTS variance in our data is accounted for by within-industry differences.

By construction, heterogeneity in RTS is explained by the significant heterogeneity in estimated output elasticities as shown in Panel B of II. The output elasticity with respect to intermediate inputs has the highest average value of 0.61, followed by labor with an average elasticity of 0.33, and capital with an average elasticity of 0.03.[13] Moreover, labor and intermediate input elasticities are also more dispersed across firms, compared to the capital elasticity. For example, the average within industry P90-P10 for intermediate, labor, and capital inputs are 0.35, 0.35, and 0.07, respectively. Furthermore, variance decompositions reveal that more than 60% of the overall variation for each output elasticity is explained by within-industry differences, meaning that within-industry heterogeneity accounts for a larger fraction of the overall

[13] We find an elasticity of capital that is much lower than typical estimates. This is because our measure of capital stock only considers plants, equipment, and buildings (tangible capital) and ignores any other form of capital such as intangible capital, inventories, cash and so on, which are included in the aggregate measure of capital.

Table IV – Correlation of Output Elasticity Estimates

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

Note: Table IV 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, the within-industry results demean the output elasticities at NAICS2 level.

firm-level variation for output elasticities compared to RTS. This is partly because the intermediate input elasticity is negatively correlated with the capital and the labor elasticity within industries (see Table IV).

One potential concern is that the observed dispersion in TFP, output elasticities, and RTS is driven by transitory movements of the same firm over the production function rather than fixed firm characteristics. To evaluate whether this is the case, we run a panel regression of each of our estimates on firm size, firm age, time dummies, and firm fixed effects. Intuitively, if the dispersion is mostly due to fixed differences between firms, the firm fixed effects should absorb most of the variation in our estimates. We find that this is the case: of a standard deviation of RTS of 0.052, 86% is explained by firm fixed effects after controlling for firm age and size (the standard deviation of the FE is 0.045). We find similar results for the rest of our estimates (elasticities and TFP) and in US manufacturing.[14] We conclude that the observed heterogeneity in RTS is mostly explained by permanent between-firm differences, consistent with the model of heterogenous RTS that we describe in Section 5.

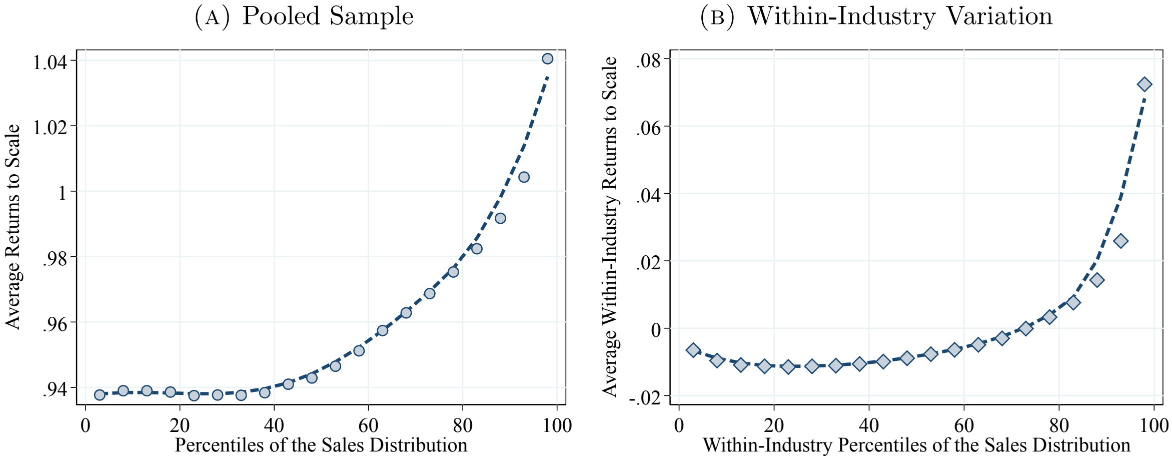
4.2 Production Technologies over the Firm-Size Distribution

We now turn to the systematic variation of our estimates over the revenue distribution.

Returns to scale by firm size. Figure 1a shows a binned scatter plot of average RTS by firm revenue for a pooled sample of all firm-year observations across industries. For this figure, we use the firm-year estimates that result from estimating

[14] In US data, we find a standard deviation of RTS of 0.058, of which 0.048 is accounted for by firm fixed effects.

Figure 1 – Returns To Scale by Firm Size


(A) Pooled Sample

(B) Within-Industry Variation

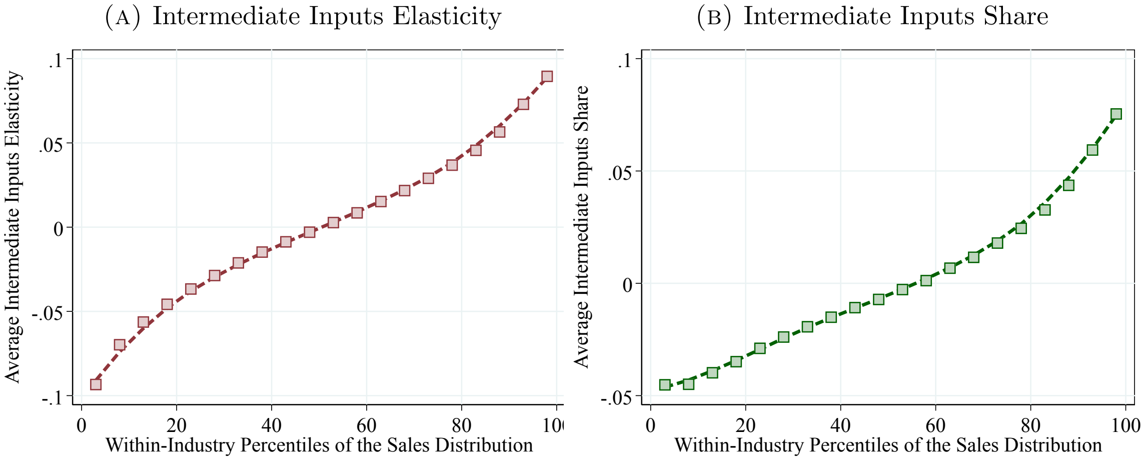
Notes: Figures 1a and 1b show the average RTS within 5%-quantiles of the firm-revenue distributions. In Panel (B), RTS are demeaned by industry averages.

production functions within 2-digit NAICS. 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 increases monotonically and strongly by firm revenue from 0.94 for firms below the 40th percentile to 1.04 for those in the top 5%. These differences imply that firms in the top 5% of the revenue distribution produce 11% more with a 1% higher input bundle compared with those below the 40^{th} percentile, holding TFP constant.[15]

The variation in Figure 1a reflects within- and between-industries heterogeneity. For example, manufacturing firms are, on average, larger and therefore overrepresented in the upper end of the revenue distribution, and manufacturing industries also have higher average RTS. Therefore, some of the overall variation is driven by heterogeneity across industries. To isolate the role of within-industry differences, we rank firms into 20 quantiles within their industries according to their revenues and demean firm RTS by industry averages. Figure 1b shows average demeaned RTS over the within-industry revenue distribution. Again, firms below the median have technologies with similar RTS, whereas above the median RTS increases steeply by firm revenue: the average RTS of firms in the top 5% of the revenue distribution is

[15] Note that RTS is not fixed over time and firms are still 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 (e.g., Gandhi *et al.* (2020) and Demirer (2020) find average RTS above 1 across multiple industries and countries).

Figure 2 – Intermediate Input Elasticities by Firm Size


(a) Intermediate Inputs Elasticity

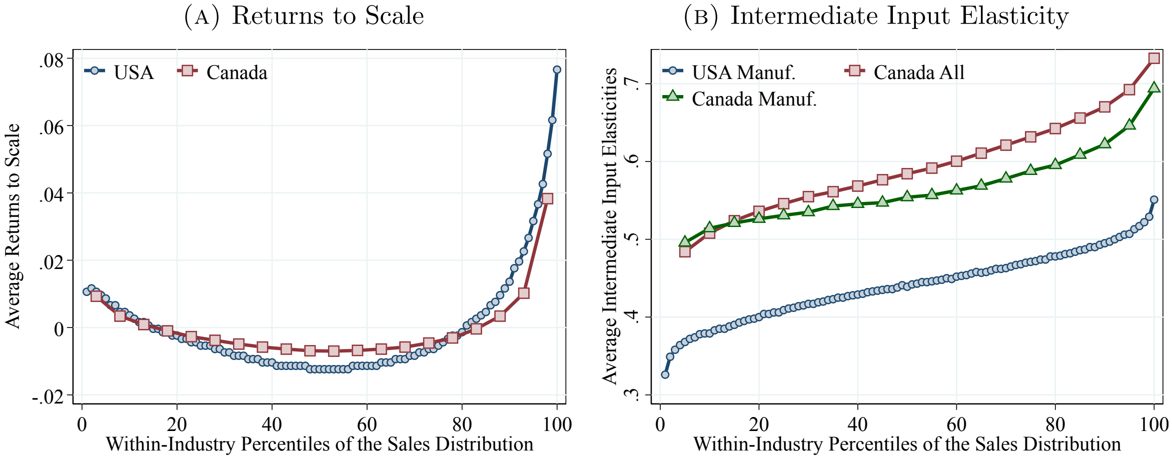
(b) Intermediate Inputs Share

Notes: Figure 2 shows the average estimated factor elasticities within 5%-quantiles of the revenue distribution; Figure 2b shows the intermediate inputs revenue share. In both cases we demean the within-quantile average by NAICS2 industry averages.

8 p.p. higher relative to the average RTS of firms in the bottom half of the revenue 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 due to within-industry differences.

Output elasticities by firm size. As we discussed in Section 2, we measure RTS as the sum of output elasticities with respect to inputs. Therefore, the significant increase in RTS is driven by either of these inputs or a combination of them. Our analysis shows, however, that it is the intermediate input elasticity that entirely accounts for the positive relation between RTS and firm revenue. As shown in Figure 2a, the intermediate input elasticity monotonically increases from -0.09 (relative to the industry average) for the bottom 5% of firms by revenue, to approximately zero for firms around the median, and to 0.09 for the top 5% of firms. This 9 p.p. gap in intermediate input elasticities between the top 5% and median firms thus explains all of the 8 p.p. gap in RTS in this range. Furthermore, Figure 2b shows that the corresponding variation in the intermediate input revenue share is very similar—with larger firms spending more on intermediate inputs relative to small firms as a portion of their revenue—as expected since our estimation treats them as flexible input. This finding also highlights that estimating gross output production functions is essential: the sum of capital and labor elasticities is declining across the firm revenue distribution, indicating that the use of value added production functions can lead to

Figure 3 – Returns To Scale and Elasticities of Intermediate Inputs in Manufacturing



Notes: Figure 3a shows the average RTS within percentiles of the sales growth distribution demeaned by industry averages for the Canadian and US manufacturing sectors. Canadian results shown within 5%-quantiles of the revenue distribution. Figure 3b 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.

misleading results.[16]

Total factor productivity by firm size. Intuitively, one would expect that firms at the top of the firm size distribution are also the most productive. To see if this is the case, we rank firms into 100 bins according to their RTS within each industry. Then, we measure the relative TFP of a firm as its TFP rank within these narrow bins. Figure 4 shows the average TFP percentile across the firm revenue distribution for the pooled sample. Similar to previous literature (see Leung *et al.* (2008) or Baldwin *et al.* (2002)), we find that relative TFP increases by firm size until the top decile of the revenue distribution. For the top 10% of the largest firms, however, relative TFP flattens out. This contrasts with the previously documented pattern of RTS by firm size, which increases in a convex fashion at the top of the revenue distribution (Figure 1). Therefore, we conclude that the largest firms tend to feature the highest RTS, and not necessarily the highest TFP as commonly assumed.

4.2.1 Robustness Checks

US-Manufacturing. Our results are not driven by particular features of the Canadian economy but are also present in other countries and industries. In particular,

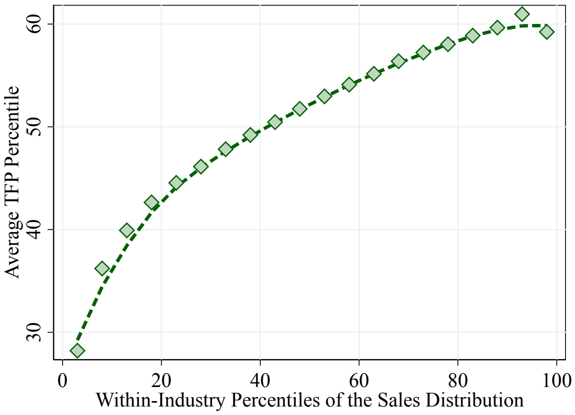
[16] We show capital and labor elasticities in Appendix Figure OA.2. In Canada, both are, on average, declining in firm size. In US manufacturing, the labor elasticity is also declining, while the capital elasticity is u-shaped in firm revenue.

our main results hold within the US manufacturing sector as well. First, Figure 3 displays the firm-level RTS relative to its industry average, showing that RTS is u-shaped with respect to firm revenue, with a larger rise at the top, increasing about 9 p.p from the middle 50th to the top 1% of the revenue distribution. For comparison, we also add a corresponding series for the Canadian manufacturing sector to this figure, which shows a similar pattern. Second, most of the increase is again due to a significant rise in the output elasticity of intermediate inputs, that rises from 0.4 at the bottom of the firm-size distribution to around 0.55 for the largest firms within US manufacturing.

Cobb-Douglas specification. One additional concern is that our results are driven by the particular method we use in our estimation. To address this concern, we first we re-estimate the production function by imposing homogeneous relative factor elasticities while still allowing for RTS differences. The basic idea is to isolate heterogeneity in RTS regardless of different relative output elasticities. In particular, we estimate the production function by restricting to homogeneous relative output elasticities while allowing for heterogeneity in RTS, that is, we assume the production function is specified as $Y_{jt} = e^{\nu_{jt}} \cdot \left(M_{jt}^{\gamma_M} K_{jt}^{\gamma_K} L_{jt}^{\gamma_L}\right)^{\eta_{jt}}$. Then, the returns-to-scale parameter η_{jt} is estimated from the first stage of the estimation procedure described in Section 2. Similarly to the baseline results, the Cobb-Douglas series in Figure 5 show that RTS increases with firm size by about 10 p.p., with a stronger increase in the bottom half and a weaker increase in the top half of the revenue distribution relative to our baseline results.

Clustering specification. In a second robustness exercise, we re-estimate firms' production technology within clusters of firms with similar characteristics. The motivation for this exercise is that we ideally want to estimate a separate production function for each firm, but this is econometrically not feasible. Clustering similar firms is feasible way of approaching this ideal. In practice, we cluster firms using information 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. We further impose that a firm belongs to only one cluster throughout its lifecycle. Finally, we then estimate the nonparametric production function by each cluster. After esti-

Figure 4 – Firm Size and Productivity



Notes: Figure 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.

mation, we sort firms by revenue within an industry, as we do in our baseline results. The Cluster series in Figure 5 shows that while these estimates are less smooth along the revenue distribution, the main patterns are as in the baseline results: RTS are similar within the bottom half of firms, and then increase by close to 10 p.p. from the median to the top 5% of firms within an industry.

Intangible capital specification. Finally, we include a measure of intangibles in our measure of the capital stock and re-estimate firms' production functions. In theory, including intangibles in firms' capital stock changes 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 leads to underestimation of the capital elasticity (and of RTS) and overestimation of TFP for larger firms. Indeed, the Intangible Capital series in Figure 5 shows that the positive relation between firm size and RTS becomes even stronger when including intangible capital.[17]

Ranking firms by employment or value added. Appendix Figure OA.3 shows our main findings when ranking firms, within industry, by employment or value added instead of by revenue. The patterns for RTS are very similar. However, the output elasticities show different patterns: firms with high employment or high value added tend to have higher labor elasticities, whereas the intermediate input elasticity shows

[17] The positive relation between firm size and conditional TFP percentile remains unchanged, perhaps due to the fact firms that have the same RTS also have similar intangible intensities.

only a small uptick for the largest firms. This is somewhat mechanical, as we except high employment or high value added firms to be labor-intensive by construction of the ranking. For this reason, we prefer to rank firms by firm revenue—a factor-neutral approach—in our main approach. It is reassuring that the pattern for RTS by firm size is very stable across the three ranking methods.

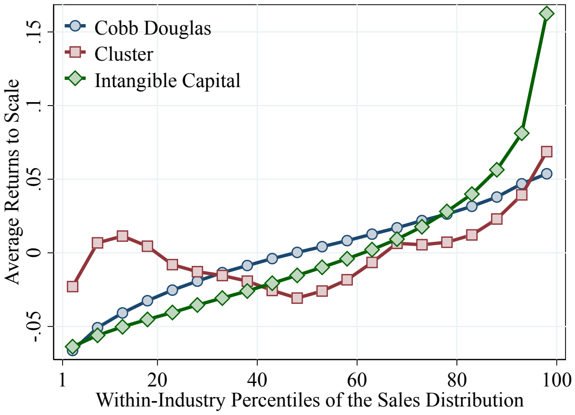
Markups. Our RTS estimates are based on revenue elasticities of the three inputs. One may be concerned that the positive relation between RTS and firm size could be driven by markup variation (e.g. De Loecker *et al.* (2020)). Ideally, we would want to separately estimate physical output elasticities and markup. However, doing so would require firm-level information on output prices and physical quantities, which we do not have in the Canadian or US datasets. Even with such output price and physical quantity information, it is challenging to estimate the physical elasticity for multi-product firms, in the absence of information on product-specific inputs. It is important to note that if larger firms charge higher markups—as implied by models with oligopolistic competition (Atkeson and Burstein, 2008) or with monopolistic competition and log-concave demand systems (Edmond *et al.* (2023))—then physical RTS would be increasing even more with firm size compared to measured revenue RTS.[18] 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 in the Appendix shows that, on average, markups monotonically increase with firm revenue, consistent with De Loecker *et al.* (2020). These theoretical and empirical considerations reinforce our interpretation of measured RTS differences along the firm-size distribution as representing differences in production technologies.

4.3 Identification: Output Elasticities and Input Shares

Typically, estimates of output elasticities of inputs 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 first order condi-

[18] It can easily be shown that the physical output elasticity is the product of revenue elasticity and markup.

Figure 5 – 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.

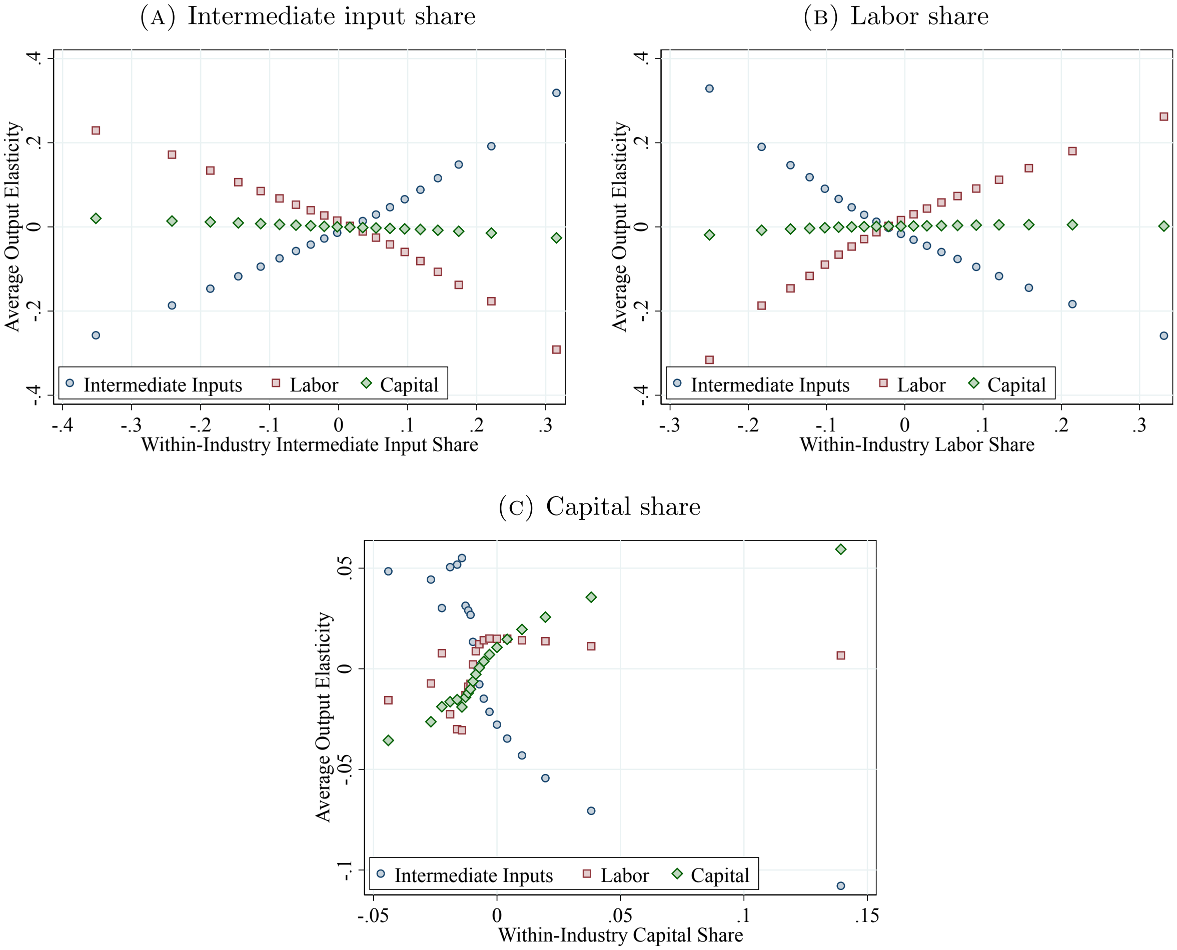
tions (FOCs) of profit-maximizing firms. Nevertheless, output elasticities tend to be positively correlated with the respective factor shares, as we discuss here.

Figure 6 shows (demeaned) output elasticities for all three inputs, when ordering firms by their intermediate input shares (Figure 6a), by their labor shares (Figure 6b), and by their capital shares (Figure 6c).[19] For all three inputs, the respective output elasticity is positively correlated with the corresponding 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. For intermediate inputs, the positive correlation is very strong, which is expected since we treat them as flexible input and we use the firm's FOC to estimate the intermediate input elasticity (see equation (1)). On the other hand, for labor and capital, our estimation does not rely on firms FOCs. Yet, we find that their input shares are also strongly positively correlated with their corresponding output elasticities. These results suggest that heterogeneity in output elasticities, therefore in RTS, reflects differences in input shares.[20]

[19] In the Appendix, we plot the relationship between input shares and firm revenue in Figure OA.4. We find that the intermediate input share increases with firm revenue for the entire Canadian private sector, within manufacturing, and within manufacturing in the US. Instead, labor and capital shares decline by firm revenue.

[20] These findings suggest that high-RTS firms should have relatively low profit shares. This is indeed the case: Figure OA.5 in the Appendix shows that, on average across firms, the EBITDA-

Figure 6 – Output Elasticities and Factor Shares of Revenue



(a) Intermediate input share

(b) Labor share

(c) Capital share

Notes: Figure 6 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 2-digit NAICS industry.

4.4 Firm Lifecycle Profile

Intuitively, heterogeneity in returns to scale has important implications not only for the firm size distribution, but for the evolution of each firm over the life-cycle. In particular, one would expect that firms with high RTS tend to grow faster and reach a higher optimal size relative to firms with the similar TFP but lower RTS. To examine these life-cycle patterns, we construct a balanced panel of firms and group them based on their production function characteristics when they entered the economy. Specifically, we construct a sample of firms which i) are born between 2002 and 2005,

revenue ratio correlates negatively with RTS.

Table V – Probit Regressions of Firm Exits

	(1)	(2)
TFP Percentile	-0.001***	0.005***
	(0.000)	(0.000)
RTS	-0.600***	-4.527***
	(0.020)	(0.088)
N	4.1M	3,.4M
Constant	Y	Y
Industry FE	Y	Y
First-difference		Y
Pseudo R2	0.010	0.018

Note: Robust standard error reported. We first-difference both regressors in Column (2).

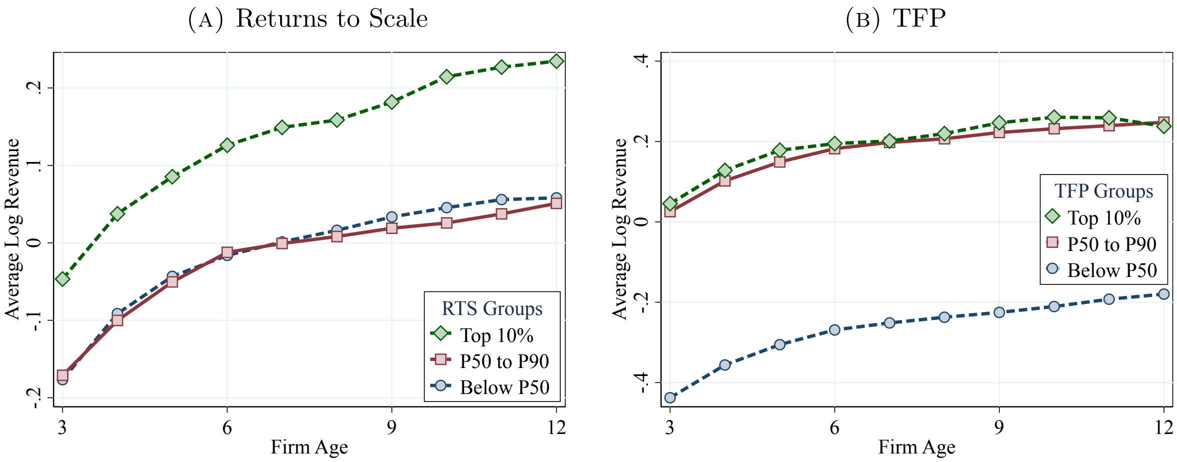
and ii) are observed for at least 12 consecutive years in the sample. We group firms based on their initial RTS demeaned at the industry level and initial TFP. Figure 7 shows the average log revenue, also demeaned at the industry level, against firm age for different firm groups. We find that firms with high initial RTS and TFP both start with higher revenue relative to other firms of the same age within the same industry. More interestingly, firms with higher initial RTS grow significantly faster than those with low initial RTS. Firms in the top 10% of the initial RTS distribution grow about 30 log points over the next 10 years, whereas those in the bottom 90% grow, on average, only about 20 log points. This finding corroborates our interpretation that firms with high measured RTS operate technologies that are more scalable, allowing them to grow significantly more over their lifecycle.

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.[21] In fact, higher initial TFP is associated with slightly smaller subsequent firm growth rates. This pattern is consistent with TFP as a mean-reverting process and might explain why some firms do not expect to growth, as documented by Hurst and Pugsley (2011). Our results indicate that these firms, although highly productive, might be characterized by low RTS, and hence small optimal size.

The previous results focus on the life-cycle patterns of surviving firms. We also study how RTS and TFP heterogeneity affect firm exit. Specifically, we estimate a

[21] By construction, the three TFP groups in Figure 8b all have identical average RTS.

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


(a) Returns to Scale

(b) TFP

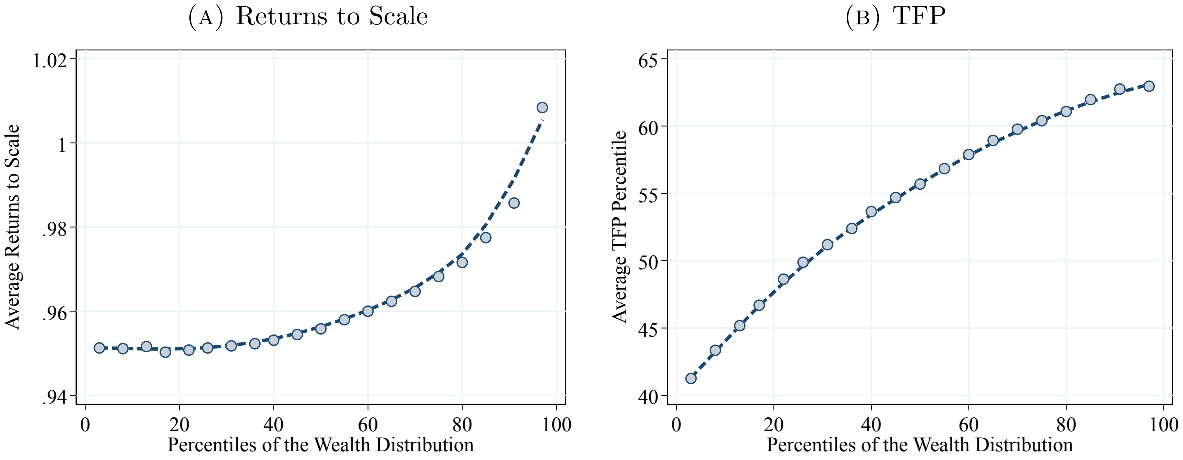
Notes: Figures 8a and 8b compares the life-cycle profile of revenue between firms with different initial RTS and TFP. It is constructed using a balance panel of firms which 1) are born between 2002 and 2005, and 2) survive for at least 12 years. We demean firms' initial RTS at the NAICS2 industry level and construct the TFP as described before. We bin firms into three groups based on their demeaned initial RTS in Figure 8a and three groups based on initial TFP in Figure 8b. Firm log revenue is also demeaned at the NAICS2 industry level.

probit regression of firm exit on TFP percentile and RTS. The results are reported in Table V. In Column (1), we use the levels of standardized production parameters, and in Column (2) we use the first-differenced parameters, controlling for NAICS2 industry fixed effects in both columns. We see that in both specifications, a higher RTS is associated with a lower probability of firm exit. The effect of TFP on firm exit is significantly smaller, with opposite signs in levels and first-differences. We conclude that from an ex ante perspective, RTS rather than TFP heterogeneity predicts differences in firm growth over the life cycle.

4.5 Implications for Wealth and Wage Inequality

We conclude this section by looking at the relation between wealth, wage, and returns to scale. First, we look at how the production function parameters vary by owner's equity wealth. The share of self-employed and business owners in Canada equals 11.7% of individual tax filers. The average number of firms owned by a business owner is 1.96, with a standard deviation of 2.34. We calculate the equity wealth of each individual by aggregating the value of the firms they own, weighted by the ownership shares. Then, for each individual, we calculate their RTS and TFP percentile by taking a equity-value weighted average of the firms they own. Figure 8 show the results. Similarly to our results by firm size, we find that high-wealth individuals tend

Figure 8 – Returns to Scale and Productivity by owner's wealth


(a) Returns to Scale

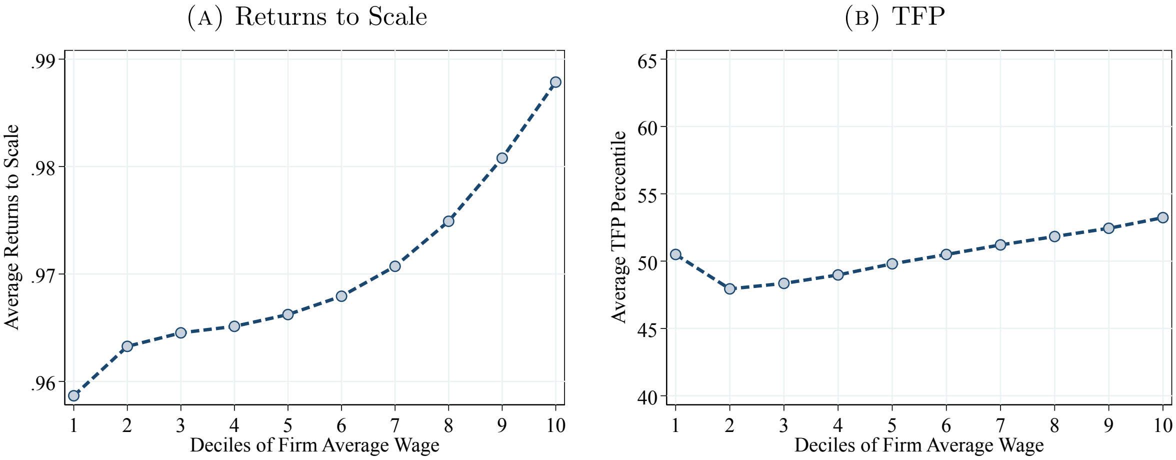
(b) TFP

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

to own firms that on average have higher 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. Given that firms owned by rich households tend to have high RTS and high TFP, it is hard to disentangle the strength of each for wealth accumulation. We use the model of the next section to shed some light on this issue.

It is well known that large firms tend to pay higher wages to similar workers than smaller firms (Bloom *et al.*, 2018; Brown and Medoff, 1989). Given our results, however, it is unclear if this relation derives from large firms having higher TFP or higher RTS. To distinguish between these two explanations, we rank firms in our sample by their average wage and compute average RTS and TFP within firm wage deciles. We find that firms that pay on average more to their workers tend to have higher RTS—and therefore tend to be larger—than firms with low average wage, as shown in Panel A of Figure 9. In contrast, there is much less variation in TFP across the firm wage distribution as shown in panel B of Figure 9. We conclude that the wage-firm size relation is mostly driven by heterogeneity in RTS.

Figure 9 – Returns to Scale and Productivity and Firm Average Wage


(a) Returns to Scale

(b) TFP

Notes: Figure 9 shows the average RTS and TFP by deciles of firms' average wage.

5 Misallocation with RTS Heterogeneity

So far, we have shown sizable heterogeneity in RTS across incorporated firms in Canada and manufacturing firms in the US. More importantly, larger firms are likely to have more scalable production technologies. In this section, we show why this heterogeneity matters in an application to efficiency costs of misallocation, indicating the potential role of RTS heterogeneity for an array of quantitative questions, such as optimal taxation of capital, firm recruiting intensity (Gavazza *et al.*, 2018), or firm cyclicality (Clymo and Rozsypal, 2023). In particular, we embed firm heterogeneity in RTS (η) and TFP (z) in a 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—to the standard setting with heterogeneity in z only—the z-economy. To build intuition, we begin by deriving an analytical result in a static endowment economy. We 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. Firms $i \in [0, \chi]$, where $\chi \in (0, 1)$, are constrained, facing an input price wedge $\tau \geq 0$. Their production technologies are iso-elastic: each firm is characterized by a pair of parameters (η, z) with decreasing RTS $\eta \in (0,1)$ and TFP z, and their output is given by $y = f(x; z, \eta) = z \cdot x^{\eta}$. The residual fraction of firms $i \in (\chi, 1]$ is unconstrained and has constant RTS.[22] The following proposition characterizes misallocation in terms of the share of the economy that is constrained as well as the returns to scale 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*

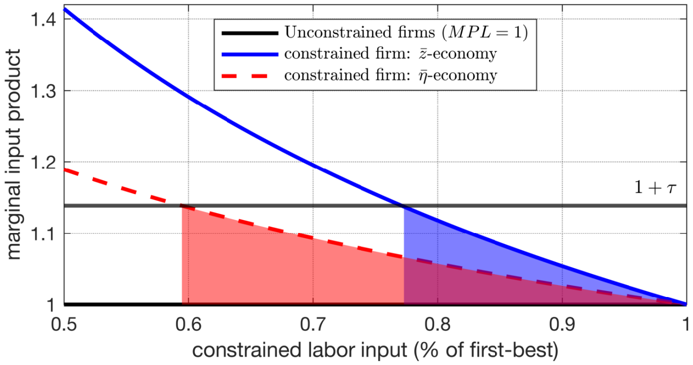
$$\Delta \ln Y(\tau) = \underbrace{\frac{\tau^2}{2}}_{\text{size of friction}} \cdot \underbrace{\int_0^{\chi} w_i \cdot di}_{\text{output share of constrained firms}} \cdot \underbrace{\int_0^{\chi} \frac{w_i}{\int_0^{\chi} w_j dj} \cdot \frac{\eta_i}{1-\eta_i} \cdot di}_{\text{avg. } \frac{RTS}{1-RTS}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

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

The proposition states that misallocation is proportional to the size of the friction, the output share of constrained firms, as well as increasing and convex in the (weighted-average) RTS of constrained firms. Therefore, for a given friction, misallocation is more severe when constrained firms exhibit higher RTS. In addition, due to the convexity of misallocation in RTS, higher dispersion of RTS also translates into more misallocation. Intuitively, a given input price wedge translates into a larger quantity adjustment when RTS are high—when marginal products decline more slowly—as constrained firms reduce their inputs more. Therefore, misallocation is larger. Firm TFP, on the other hand, matters only indirectly via the output share of constrained firms. We illustrate this in Figure 10, which depicts the marginal input product of firms that would be "large" in the first-best, and which matter the most for misallocation: the solid blue line represents the conventional setting where large firms have high TFP ($\bar{z}$), the dashed red line represents an economy where these large firms have high RTS ($\bar{\eta}$). For a given wedge τ, misallocation—the area under the curve—is larger in the $\bar{\eta}$-economy.

[22] Alternatively, unconstrained firms exhibit decreasing RTS but there is free entry in that sector, such that there are constant RTS at the sectoral level.

Figure 10 – Efficiency costs in endowment economy



5.2 Quantitative Dynamic Model

We now consider a dynamic economy in the tradition of Quadrini (2000); Cagetti and De Nardi (2006), where 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 data, and compare it with the misallocation in an economy where firms differ only in TFP.

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

$$\sum_{t \geq 0} \beta^t \ln(c).$$

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 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, R 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, and follows a joint first-order Markov process. Asset markets are incomplete, and agents can invest their wealth $a \geq 0$ into an annuity, paying an interest rate r. Upon death, individuals are replaced by the same 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 needs to be financed by the entrepreneur's own wealth. Thus, static profit maximization entails a net profit of

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

implying input choices $k(a,z,\eta), \ell(a,z,\eta)$ and output $y(a,z,\eta)$.[23] Thus, the dynamic agent problem can be written in recursive form as

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

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

5.2.2 Equilibrium

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 $R = r + \delta - p \cdot (1+r)$.

[23] 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, e.g., with a collateral constraint on k only). Relatedly, we also do not include intermediate inputs in our production function here. Including them would tend to magnify the level of efficiency costs of financial frictions, but not change the comparison between the (η, z)- and the z-economy.

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

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

5.2.3 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. Before going there, we briefly discuss fixed common parameters. We set the death probability to $\frac{1}{80}$, corresponding to an expected life expectancy of 80 years.[24] We use a Cobb-Douglas production function (f) with capital share $\alpha = 0.4$ and depreciation rate $\delta = 0.05$. We model the process for labor efficiency units h as a log-normal AR(1) with autocorrelation of 0.9, cross-sectional standard deviation of 1.3, and normalized mean $\mu_h = -\frac{\sigma_h^2}{2}$, which we estimate using Canadian data on individual post-tax earnings. We calibrate both economies at $\lambda = 0.3$, meaning that 30% of input expenditures need to be financed with the owner's wealth, and then vary λ in counterfactuals.[25]

(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 right column of Table VI. We now provide intuition on how the parameters are identified: the effective discount factor β affects the aggregate capital-output ratio. The (common) RTS parameter η is closely related to the population fraction of entrepreneurs, as it impacts their income share. We model the z-process as log-normal AR(1) with normalized mean $\mu_z = -\frac{\sigma_z^2}{2}$. Its auto-correlation (ρ_z) affects the transition into (and out of) entrepreneurship. Finally, the firm size distribution is closely related to the cross-sectional dispersion of $z(\sigma_z)$. We also stretch out the top 1% of the z- distribution with a Pareto tail, where ξ_z

[24] 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 3.4% in the (η, z)-model and 2.2% in the z-model, in the ballpark of the value for Canada of 4.9%.

[25] Defining the debt d of entrepreneurs as $d = \max\{0, k - a\}$, the aggregate debt-to-capital ratio equals 81% in the (η, z)-model and 71% in the z-model, in the range of the value for Canada of 70%.

Table VI – 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	0.799	0.804
Top 1% revenue share	0.522	0.522	0.515
Top 0.1% revenue share	0.282	0.282	0.284
RTS: Top 5% vs Bottom 50% (by revenue)	0.083	0.076	0*
Capital-output ratio	2.970	2.971	2.970
B. Internally calibrated parameters			
Mean RTS	μ_η	0.794	0.683
Standard deviation RTS	σ_η	0.049	
Standard Deviation TFP	σ_z	0.571	0.910
Persistence TFP	ρ_z	0.952	0.970
Pareto tail TFP	ξ_z		2.880
Correlation (z, η)	$\sigma_{z,\eta}$	-0.219	
Discount factor	β	0.889	0.902

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

denotes the tail coefficient. This allows the model to match the right tail of the firm size distribution as well. Overall the fit of the model to the targeted moments is excellent.

(η, z)-Economy: In a nutshell, we repeat the calibration of the z-economy, and in addition discipline the heterogeneity in η by matching the observed dispersion of RTS along the revenue distribution (middle column of Table VI). Concretely, we model η as a truncated normal AR(1) in the interval $(0, 1)$ with parameters $(\mu_\eta, \sigma_\eta, \rho_\eta)$. We ex ante fix the auto-correlation to a high value of $\rho_\eta = 0.98$, which is the persistence of RTS in our empirical analysis. Its mean μ_η again controls the fraction of entrepreneurs. Its cross-sectional standard deviation σ_η is closely linked to the difference in the average RTS of the top 1% vs. the bottom 50% of firms ordered by revenue. We also allow z and η to be correlated by setting log TFP $\ln z = \tilde{z} + \sigma_{\eta,z} \cdot \frac{\sigma_z}{\sigma_\eta} \cdot (\eta - \mu_\eta)$, where $\tilde{z}$ follows a normal AR(1) process with parameters $(\sigma_z, \rho_z, \mu_z = -\frac{\sigma_z^2}{2})$. Intuitively, both $\sigma_{\eta,z}$ and σ_z affect moments of the firm size distribution: e.g., if the empirical dispersion in RTS was small, then a high residual TFP dispersion σ_z would be required to match the high revenue concentration of firms. Conversely, if the observed dispersion in RTS was very high, then the calibration would infer a more negative value for the correlation parameter $\sigma_{\eta,z}$. We calibrate the TFP parameters residually in this way, instead of directly feeding the estimated joint distribution of η and z into the model, because when firms operate different production functions that differ in η, inferred relative TFP depends on the unit choice. This is the case in our model setting, as well as in some of our empirical approaches (when we cluster firms, such that firms within an industry do not all share the same production function).[26] In sum, we calibrate six parameters to match seven moments. This model version also provides an excellent fit to the data. It is noteworthy that 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 suffices.

5.2.4 Quantitative findings

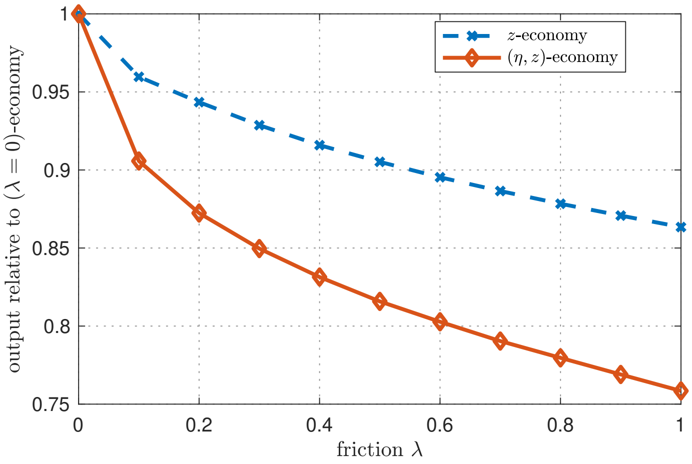
The two model economies are observationally equivalent in terms of the fraction of entrepreneurs, the persistence of entrepreneurship, the firm size distribution, and the ratio of wealth (capital) to output. We now evaluate the efficiency losses associated with the same financial friction in both economies. Figure 11 below compares the output losses induced by raising the financial friction from the unconstrained case of $\lambda = 0$ all the way up to $\lambda = 1$, across stationary equilibria of the two model versions. We see that, e.g., if entrepreneurs need 30 cents of their own wealth to finance a dollar of input expenditure, the (η, z)-economy with heterogeneity in both TFP and RTS—disciplined by our empirical estimates—features an output loss of 16.3 log points, relative to the frictionless economy. However, in the conventional z-economy that imposes homogeneous RTS, the output loss is significantly lower, at 7.4 log points.

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

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

which depends on the level of the input ℓ; therefore, it depends on the unit of account. In particular, relative TFP of the higher-RTS firm is inversely proportional to the unit of account. Therefore, depending on the choice of unit (e.g., hours vs. full-time equivalents), for the same data, one can find any relationship (both in sign and magnitude) between TFP of these two firms. Therefore, when firms operate different production functions that differ in RTS, relative TFP does not have the usual cardinal interpretation. For a similar argument about unit dependance in the context of house price elasticities, see Greaney (2019).

Figure 11 – Output losses from financial friction in dynamic model



Therefore, incorporating realistic heterogeneity in RTS, while otherwise matching the same observables, amplifies the output losses due to the financial friction by 120%.

To understand this finding, we additively decompose the total log output loss into three terms: (i) static misallocation of production factors holding fixed occupational choice, (ii) misallocation of talent across occupations, and (iii) under-accumulation of capital. Panel A of Table VII shows that static misallocation of production factors across firms contributes 9.5 log points in the (η, z)-economy, more than half of the total GDP loss, and almost twice as much as in the conventional z-economy. This is the channel highlighted in our analytical discussion in Section 5.1, and our quantitative findings are in line with Proposition 1. Under-accumulation of capital contributes the majority of the remaining output loss. Misallocation of talent across occupations is also a little bit larger in the (η, z)-economy, but rather small in both economies: the λ-friction mainly misallocates production factors across firms, but does not distort the decision to enter entrepreneurship vs. employment much. We deliberately used a rather conservative, simple and transparent calibration strategy, avoiding, e.g., fixed costs of entry and exit, which would tend to magnify the overall importance of that channel.

Our findings are robust to alternative approaches of making the financial friction comparable in both economies. In our benchmark scenario we raise λ from 0 to 0.3 in both economies. Panel B of Table VII shows that our findings are even stronger when instead equating observable moments such as the aggregate debt-to-capital ratio or

TABLE VII – DYNAMIC MODEL: DECOMPOSITION OF OUTPUT LOSSES AND ROBUSTNESS

	(η, z)*-economy*	z*-economy*
A. Decomposition of output losses going from $\lambda = 0$ to $\lambda = 0.3$		
Total log GDP loss	16.3	7.4
... due to misallocation of production factors	9.5	5.0
... due to misallocation of talent	0.6	0.5
... due to K accumulation	6.2	1.9
B. Total log GDP loss: robustness to alternative comparisons		
Equating λ	16.3	7.4
Equating aggregate debt/capital ratio	24.9	7.4
Equating dispersion in log marginal products	20.1	7.4

Notes: Panel A additively decomposes the total (steady state) log GDP loss going from $\lambda = 0$ to $\lambda = 0.3$ into: (i) misallocation of production factors (starting from the $\lambda = 0.3$ steady state, fixing K, L, and occupational status, allowing for efficient reallocation of K, L across firms); (ii) misallocation of talent (in addition allowing for efficient change of occupational status), and (iii) dynamic under-accumulation of capital. Panel B reports the total 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, resp. marginal input product dispersion of the z-economy with $\lambda = 0.3$.

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 in the latter case. We then raise λ from 0 to values of 0.803 (resp. 0.488) in the (η, z)-economy, in these two scenarios, to replicate the debt ratio (resp. marginal product dispersion). As a result, the (η, z)-economy generates output losses which are $172 - 236\%$ larger than the ones in the conventional z-economy.

6 Conclusion

In this paper, we have documented significant heterogeneity in firms' scalability (their RTS), even within narrowly defined industries. RTS heterogeneity is large, highly persistent, and systematically linked to firm size: larger firms tend to be those that have higher RTS. Our empirical analysis also showed that differences in the output elasticity of intermediate inputs drive the positive relation between firm size and RTS, whereas labor and capital elasticities are jointly decreasing in firm size.

The documented RTS heterogeneity has important implications for our understanding of the drivers of firm growth, the determinants of the firm size distribution, and the distributional impact of financial constraints and taxes. To demonstrate this, we developed a quantitative model with firm heterogeneity in TFP—as in typical models of entrepreneurship and firm dynamics—and in RTS. We used this model to evaluate the efficiency and output costs of financial frictions. When large firms are characterized by high RTS—as we documented empirically—as opposed to high TFP—the conventional view—efficiency costs are magnified. We provide intuition for this result in a static setting, and then quantify the mechanism in the dynamic model. We show that the same friction generates more than twice the efficiency and output costs in an economy with both RTS and TFP heterogeneity, as observed in our data, compared to the conventional calibration that loads all observed firm heterogeneity on TFP dispersion. These findings suggest that allowing for realistic RTS heterogeneity also has important implications for optimal wealth and capital income taxation.

References

Ackerberg, D. A., Caves, K. and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, **83** (6), 2411–2451. 2.2

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

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

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

Basu, S. and Fernald, J. G. (1997). Returns to scale in us production: Estimates and implications. *Journal of political economy*, **105** (2), 249–283. 4.1

Bloom, N., Guvenen, F., Smith, B. S., Song, J. and von Wachter, T. (2018). 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. 4.5

Brown, C. and Medoff, J. (1989). The employer size-wage effect. *Journal of political Economy*, **97** (5), 1027–1059. 4.5

Cagetti, M. and De Nardi, M. (2006). Entrepreneurship, frictions, and wealth. *Journal of political Economy*, **114** (5), 835–870. 1, 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. 12, 4.1

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

— 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

Demirer, M. (2020). Production function estimation with factor-augmenting technology: An application to markups. *Job Market Paper*. 2, 15

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

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, 15

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)*. 2, 4.1

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*. 26

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

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

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. 5

Hurst, E. and Pugsley, B. W. (2011). *What do small businesses do?* Tech. rep., National Bureau of Economic Research. 4, 4.4

Leung, D., Meh, C. and Terajima, Y. (2008). *Firm size and productivity*. Tech. rep., Bank of Canada. 4.2

Quadrini, V. (2000). Entrepreneurship, saving, and social mobility. *Review of Economic Dynamics*, **3** (1), 1–40. 1, 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. 4.1

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

Online Appendix for “Scalable vs. Productive Technologies”

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

A Data Appendix

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

A.1 Variable Construction

Revenue We use the revenue measure which 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, 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 tangible capital of firm i. The initial year capital stock K is calculated as the book value of tangible capital net of accumulated tangible capital amortization. Tangible investment include investment of building and land, computer, 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.

[1] Queen’s University; mons.chan@queensu.ca

[2] University of Toronto; g.hong@mail.utoronto.ca

[3] University of Pennsylvania; jhubmer@sas.upenn.edu

[4] St. Louis Fed, University of Toronto; serdar.ozkan@gmail.com

[5] The Wharton School-University of Pennsylvania; ssalgado@wharton.upenn.edu

Intermediates: We measure intermediate inputs as the total expenses not related to capital and labor. Specifically, it 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 to 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 asset 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 more than one T4 records in a year if she works for more than one firms. For multiple job holders, we keep the job that offers the largest earnings of the year and call it the main job. In addition, we drop workers with annual earnings from the main job lower than 5000 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. Second, 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 the the PIM to construct capital stock for these observations. Third, we drop firm-year observations with missing and non-positive 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 wage bill-to-revenue ratio below the 1st percentile or above the 99th percentile, with wage bill-to-value added ratio below the 1st percentile or above the 99th percentile, with intermediate input-to-revenue ratio above 0.95 or below 0.05, and with capital stock-to-revenue ratio above the 99.9th percentile. This sample selection procedure leaves us with around 4.3 million firm-year observations.

B Proof of Proposition 1

W.l.o.g., set the productivity of the 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:

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

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

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

Thus, we can write the aggregate misallocation loss as

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

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 proposition follows:

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

C Additional Figures and Tables

Table OA.1 – 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: Table shows 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.2 – 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

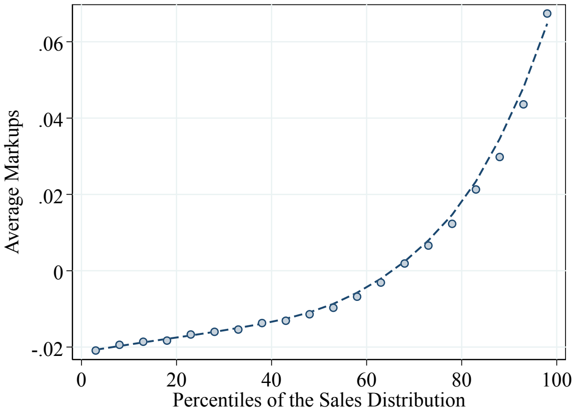
Notes: Table shows moments of the distribution of estimates RTS and output elasticities for the Canadian Manufacturing Sector. The total number of observations is 436,000.

Table OA.3 – 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

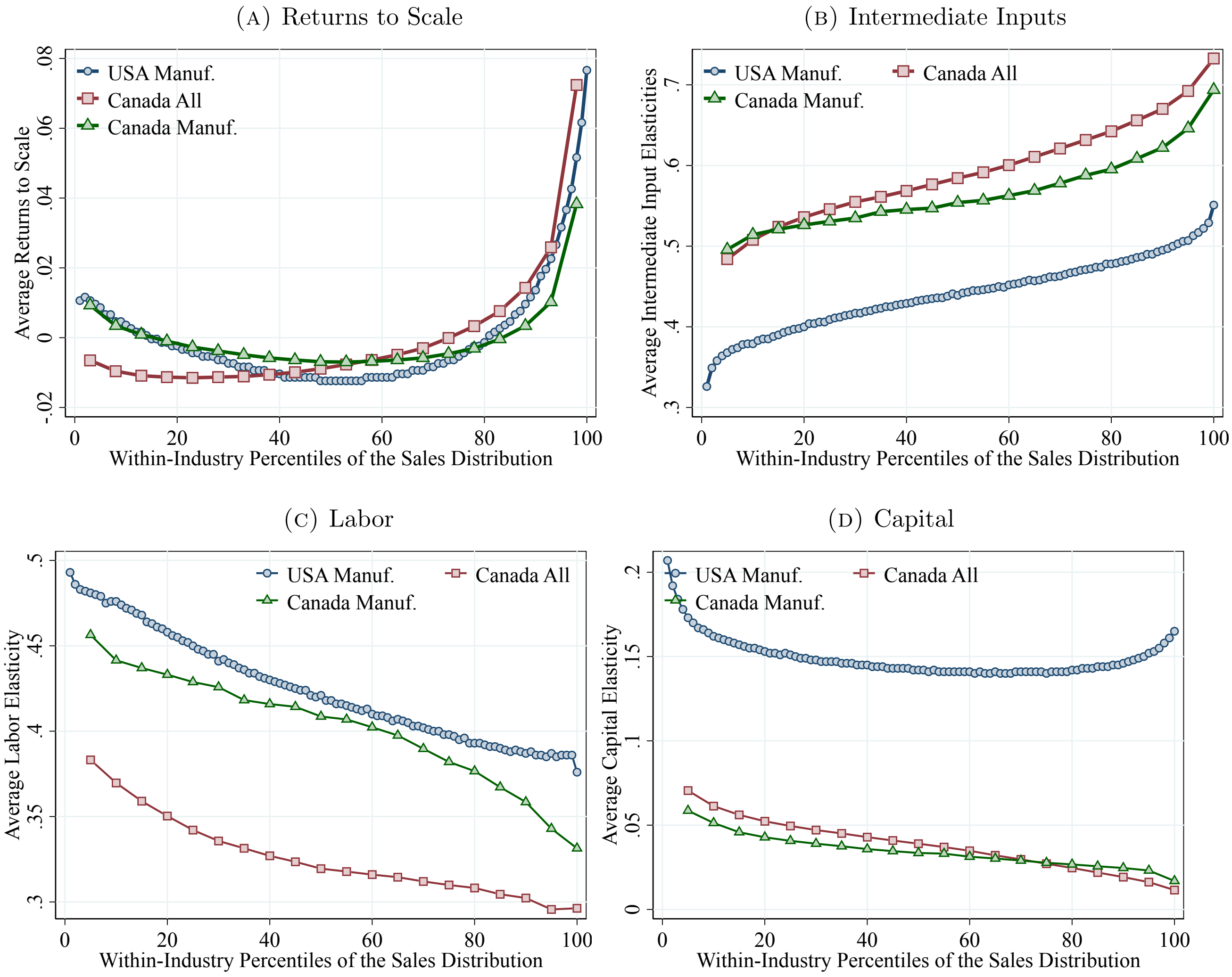
Notes: The numbers of observations are rounded to the nearest hundreds.

FIGURE OA.1 – ESTIMATED MARKUPS AND FIRM REVENUE



Notes: Figures 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

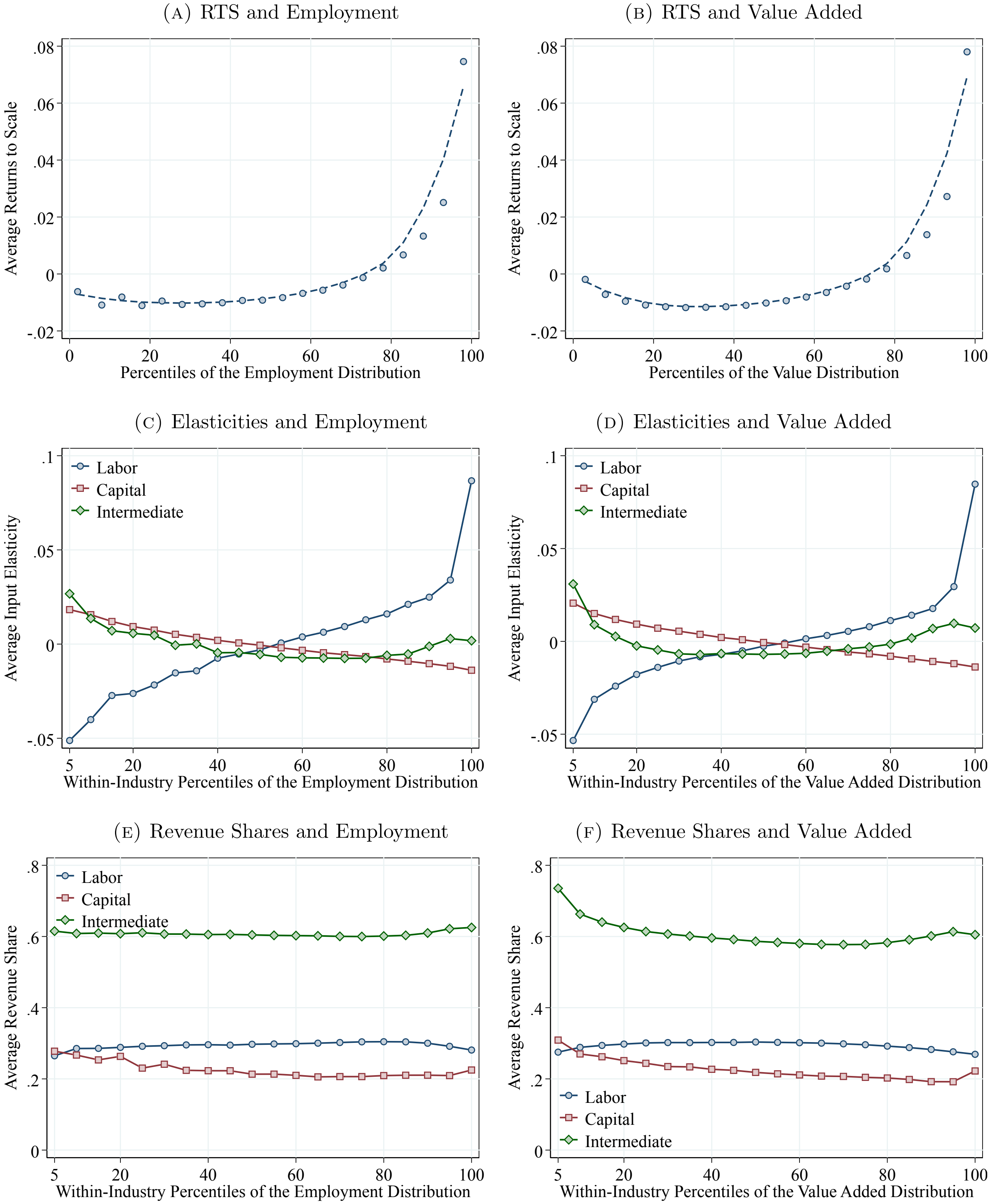
(b) Intermediate Inputs

(c) Labor

(d) Capital

Notes: Figures 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

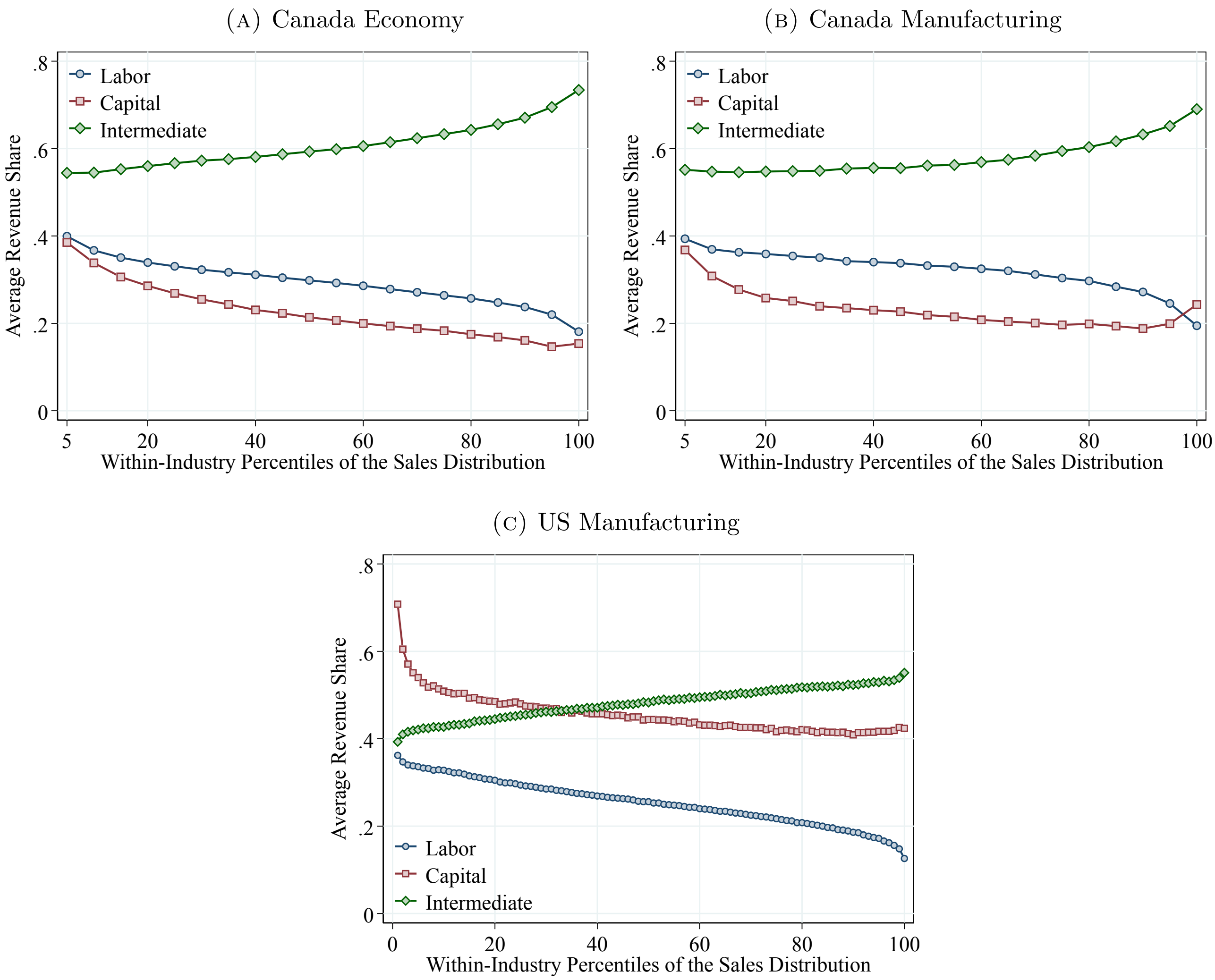
(d) Elasticities and Value Added

(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 column) and within-industry value added ranks (right column). 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



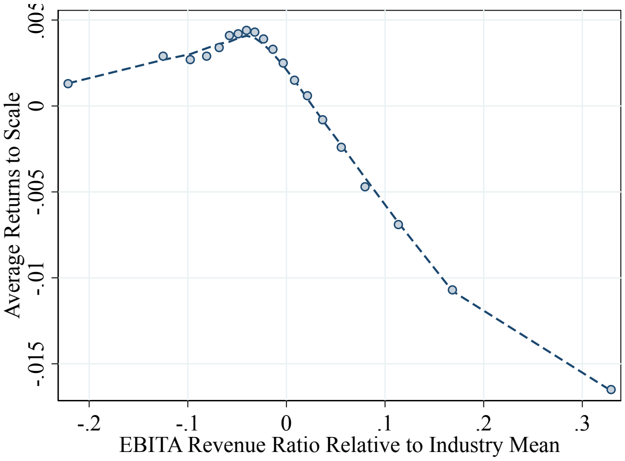
(a) Canada Economy

(b) Canada Manufacturing

(c) US Manufacturing

Notes: Figure OA.4 shows revenue shares across the firm-size distribution for Canada and for the US Manufacturing. 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 presented in 5 percentiles groups.

Figure OA.5 – Profits and Returns to Scale



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