Customer Accumulation, Firm Size, and Aggregate Productivity*

Mahmood Haddara§

January 19, 2026

Click here for most recent version

Abstract

Firms differ in how well they can produce as well as how many customers they can sell to. In this paper, I study the possibility of mismatch between these two objects: productive firms with few customers and unproductive firms with many. I argue that this is an important channel for understanding firm size dispersion and a key determinant of aggregate productivity. To do so, I develop a dynamic macroeconomic model where firms endogenously differ in both productivity and customer access. Customer accumulation in the model is gradual and uncertain, naturally allowing for productivity-customer mismatch. I show that the firm's static optimality conditions distinguish customer access from productivity, allowing me to separately identify the two objects in firm-level data. Applying this method to administrative data covering the universe of Canadian firms, I uncover a surprising degree of mismatch. This leads to a strong policy implication: A tax on high-revenue firms — clearly detrimental in standard models — can increase aggregate productivity and welfare. This is because the tax shifts production away from unproductive large firms, reducing misallocation by enough that the net effect is positive. I also find substantial welfare gains from subsidizing firm investment in customer accumulation, highlighting an understudied margin for policy intervention.

Keywords: Firm dynamics, firm heterogeneity, customer accumulation, markups, productivity.

JEL codes: E22, D24, D25, M13.

*I am indebted to Gueorgui Kambourov, Burhan Kuruscu, Murat Celik, and Daniel Trefler for their invaluable advice and mentorship. I benefited from the insightful comments of Abdelrahman Amer, Laurent Cavenaile, Fatih Guvenen, Lu Han, Ismail Landolsi, Peter Morrow, Sergio Ocampo, Mark Rempel, Diego Restuccia, Baxter Robinson, Kurt See, Joseph Steinberg, and Jeffrey Sun. I thank participants at University of Toronto, Canadian Economics Association, and Bank of Canada presentations for their helpful questions and feedback. I gratefully acknowledge funding from the SSHRC Doctoral Fellowship and CRDCN Emerging Scholars Grant. Data analysis was conducted at the University of Toronto Research Data Centre, a part of the Canadian Research Data Centre Network (CRDCN). This service is provided through the support of the Canada Foundation for Innovation, the Canadian Institutes of Health Research, the Social Sciences and Humanities Research Council, and Statistics Canada, and through the support of the University of Toronto. All views expressed in this work are my own.

§Department of Economics, University of Toronto. Email: m.haddara@mail.utoronto.ca.

1 Introduction

Firm size varies enormously within modern economies. This variation plays an important role in shaping aggregate productivity and determining the effects of economic policy.[1] In this paper, I study two central drivers of firm size dispersion: productivity and customer accumulation.[2] I argue that these two dimensions often diverge, with productive firms having few customers and vice versa. Moreover, I show that this mismatch is crucial for understanding firm heterogeneity and has strong policy implications.

To make this argument, I develop a macroeconomic model where firms endogenously differ in two dimensions: productivity, capturing attributes such as efficiency or quality, and customer access, defined as the share of customers that a firm can sell to. Firms can invest in improving both objects, but the benefits of these investments are uncertain and take time to materialize. As such, more productive firms do not always have more customers. I propose a novel method to separately identify productivity and customer access and use it to calibrate the model to administrative data covering the universe of Canadian firms. The calibrated model features a negative correlation between productivity and customer access, indicating substantial mismatch. This allows the model to capture empirical patterns that are challenging to replicate in alternative frameworks, such as those where firms differ only in productivity or where more productive firms always have more customers.

Productivity-customer mismatch also leads to a strong policy result: A tax on high-revenue firms can improve welfare, in stark contrast to standard models. This is because the model features a significant number of unproductive large firms that are propped up by high customer access. Shifting production away from these firms improves allocative efficiency, an effect quantitatively strong enough to overturn the usual distortionary effects of such a tax. I also find that firm investment in customer access should be subsidized, with the optimal flat subsidy raising consumption-equivalent welfare by 4%. This is because consumers capture part of the surplus generated by firm investment, resulting in firms undervaluing customer accumulation. Finally, I contribute to the ongoing discussion on the importance of customer accumulation for firm size dispersion. Using my rich framework where productivity and customers endogenously interact, I ask the following question: How much variation in firm size would there be if all firms could sell to all customers? In my model, the variance of log revenue would fall by 63%, indicating that the majority of firm size dispersion is attributable to the customer margin.

The model I develop combines elements from the literature on firm dynamics, innovation, and customer

[1] A long line of literature has emphasized these points. See, for example, Hopenhayn and Rogerson (1993), Restuccia and Rogerson (2008), Hsieh and Klenow (2009), Moll (2014), and Edmond, Midrigan, and Xu (2023).

[2] Both of these dimensions have played a prominent role in the literature studying firm heterogeneity. Seminal contributions regarding productivity include Lucas (1978), Jovanovic (1982), Hopenhayn (1992), and Melitz (2003). See Arkolakis (2010), Dinlersoz and Yorukoglu (2012), and Gourio and Rudanko (2014) for influential models of customer accumulation.

accumulation. A continuum of identical households supply labor, the economy's only input, and consume a basket of imperfectly substitutable varieties. A key feature of my model is that this basket is not, in general, equal to the entire set of varieties in the economy. Instead, each household can only purchase from a subset of all firms, as in Arkolakis (2010) and similar models of customer acquisition. The share of households that can purchase from a firm is defined as the *customer access* of that firm, an object endogenously determined by the firm's dynamic investment decisions.[3] Since firm investment in customer access expands the set of household varieties, it carries a standard "love-of-variety" externality (Dixit & Stiglitz, 1977; Krugman, 1980; Romer, 1987).

There is free entry and firms engage in monopolistic competition. A firm's optimal revenue is increasing in both its productivity and its customer access. The former operates through the intensive margin, reducing marginal cost and thereby allowing firms to charge a lower price and sell more to each customer. The latter operates through the extensive margin, allowing firms to sell to additional customers. Due to a Kimball (1995) aggregator in household preferences, firms face variable demand elasticities and therefore charge variable markups.[4] This generates cross-sectional misallocation of production — more productive firms face a lower elasticity and therefore underproduce relative to less productive firms. However, unlike standard models of markup misallocation (e.g., Edmond, Midrigan, and Xu, 2023), more productive firms do not necessarily have higher revenue. Consequently, misallocation is on the basis of productivity rather than size, such that shrinking unproductive large firms (and expanding productive small firms) raises aggregate productivity. By contrast, the optimal policy in Edmond, Midrigan, and Xu (2023) is to expand all large firms and shrink all small firms.

By breaking the perfect link between productivity, size, and markups, the Kimball aggregator also provides the key to identification. With Kimball demand, firms that are large due to productivity are fundamentally different from those that are large due to customer access. The former operate further *along* the demand curve, facing a lower elasticity and charging a higher markup. The latter face a demand curve that is *shifted* further out, which does not affect the elasticity they face.[5] Consequently, the model admits a closed-form expression that allows a firm's customer access and productivity to be inferred from its observed

[3] There are several reasons why a customer may not be able to buy from a particular firm. The obvious limitation is simply information: The customer may not know that the firm exists. Customer access can also be thought of more generally, however, including dimensions such as geographic expansion or retailer agreements. My focus is on the implications of variation in customer access, not its causes. As such, I do not take a particular stance on which of these mechanisms is at play.

[4] The Kimball aggregator is part of the more general class of homothetic direct implicit additivity (HDIA) aggregators studied by Matsuyama (2023).

[5] See Afrouzi, Drenik, and Kim (2023) for empirical evidence supporting this distinction.

revenue and markup.[6] With the more commonly used CES aggregator, customer access and productivity are isomorphic from the perspective of revenue and markups, preventing the use of such an identification strategy.

Firm productivity and customer access each take on discrete points on their respective "ladders," as in standard "quality ladder" models (Aghion & Howitt, 1992). Entering firms draw productivity from an exogenous distribution, whereas their customer access is endogenously determined by an initial investment. Incumbent firms undertake investments to endogenously improve each variable, both of which are subject to stochastic negative shocks. The firm's dynamic problem is complex, because the marginal benefit of increasing one variable depends on the current and future values of the other. Firms must hire overhead labor to operate, resulting in endogenous exit whenever the value of operating falls below the fixed cost.

I calibrate the model using the National Accounts Longitudinal Microdata File (NALMF), an administrative dataset covering the universe of Canadian firms from 2001-2019. The universal coverage of the NALMF is a significant asset relative to more commonly used datasets (e.g., Compustat) that disproportionately represent large and/or old firms. The inclusion of small and young firms is particularly useful when studying customer accumulation, since the customer margin is likely to be especially relevant for such firms. The NALMF is also advantageous because it includes detailed firm-level income statements. This information is often missing from datasets with comparable coverage, such as the US Business Dynamics Statistics (BDS).

The calibrated model features a negative correlation between productivity and customer access, indicating substantial mismatch. Two features of the model are crucial for generating this result. First, customer accumulation is gradual, implying that high-productivity firms take time to accumulate customers. This is particularly relevant for young firms, which empirically have fewer average customers but higher average productivity (Foster, Haltiwanger, & Syverson, 2016; Argente, Fitzgerald, Moreira, & Priolo, 2025). Since young firms represent a large share of the economy and often exit before accumulating many customers, this force is quantitatively strong. Second, firm selection is on the basis of overall profitability rather than productivity alone (Foster, Haltiwanger, & Syverson, 2008). As such, unproductive large firms can survive by relying on high accumulated customer access, whereas firms with low customer access can exit despite having high productivity.

The substantial mismatch in the calibrated model allows it to capture three elusive patterns in the data: i) On average, entrants are smaller than incumbents but have higher productivity (Eslava, Haltiwanger, Kugler, & Kugler, 2004; Foster, Haltiwanger, & Syverson, 2008; Chang & Wu, 2014). ii) Sales per customer

[6] To address concerns with the accuracy of markup estimates, I implement this method using indirect inference (Gourieroux, Monfort, & Renault, 1993; Smith Jr., 1993). In short, rather than calibrate my model to match potentially erroneous markup estimates, I target moments related to the sales-cost ratio of firms, which is readily observable and has a close theoretical link to the markup. Although their overall identification strategy is quite different, this is similar to how Ignaszak and Sedlacek (2023) avoid concerns with markup estimation in their calibration.

is negatively correlated with total sales (Bernard, Dhyne, Magerman, Manova, & Moxnes, 2022). iii) The mean and variance of revenue steeply increase with firm age, yet the mean of the sales-cost ratio barely changes and its variance decreases.[7] These patterns are challenging to capture without mismatch — if more productive firms always had more customers, we would expect small firms to have low productivity, sales per customer to be positively correlated with total sales, and revenue to co-move closely with sales-cost ratios.[8]

I then turn to policy analysis. Motivated by the mismatch result, I revisit the efficiency implications of taxing large firms. In models where size is synonymous with productivity, policies that cause large firms to contract are typically detrimental. After all, large firms are the most productive firms in these models by construction, so shifting production away from them lowers aggregate productivity. The effect is not so clear in my model, because some large firms are unproductive, relying instead on high customer access. To explore this question concretely, I consider a 3% tax on all firm revenue above $20 million, in the spirit of Canada's recently abolished Digital Services Tax.[9] For comparison, I also simulate the tax in a fully recalibrated version of the model where all firms can sell to all customers, implying a one-to-one mapping between size and productivity.

The tax results in a consumption-equivalent welfare change of +0.1% in the full model, compared to −1.1% in the productivity-only model. In other words, mismatch completely overturns conventional wisdom regarding this policy. Moreover, the policy can be improved with an even higher tax rate. The welfare-maximizing rate in the full model is 10%, yielding a 0.9% welfare gain compared to a 2.3% loss in the productivity-only model.[10] A decomposition of TFP reveals a key difference between the two models in the policy's effect on allocative efficiency. Whereas this effect is unambiguously negative in the productivity-only model, there is a positive component in the full model due to production shifting away from unproductive large firms. This channel is quantitatively strong enough that the overall effect on aggregate productivity and welfare becomes positive.

I then further demonstrate the usefulness of the model through two additional exercises. First, I consider policies relating to firm investment in customer access. Theoretically, it is unclear whether this investment should be taxed or subsidized. On the one hand, firms do not receive the full benefit of their investment due to positive consumer surplus and the usual "love-of-variety" externality. On the other hand, there is a

[7] I document these facts in my data. Although the mean and variance of revenue increasing with firm age is well known, I am not aware of prior literature discussing these facts about the sales-cost ratio.

[8] One may be concerned that weakening the link between productivity and size would contradict the negative empirical relationship between a firm's revenue and its labor share (Autor, Dorn, Katz, Patterson, & Van Reenen, 2020; Gouin-Bonenfant, 2022). The model retains this relationship due to the presence of overhead labor. Wages paid to overhead workers mechanically constitute a smaller share of revenues for large firms, lowering their labor share. This mechanism has been previously discussed by Bartelsman, Haltiwanger, and Scarpetta (2013) and Autor, Dorn, Katz, Patterson, and Van Reenen (2017).

[9] This exercise is not intended to be a comprehensive evaluation of the Digital Services Tax in particular, which targets specific firms and is motivated by profit-shifting and other concerns. It simply provides a convenient template for a policy targeting high-revenue firms.

[10] I hold the $20 million threshold fixed in this exercise, maximizing only over the tax rate.

negative "business-stealing" externality: part of the marginal benefit that firms consider in their decision is a transfer from other businesses, rather than new surplus. I find that the positive externalities dominate, and a 36% subsidy raises consumption-equivalent welfare by 4%.

The second exercise shifts from policy analysis to contribute to an ongoing discussion on the determinants of firm variation. Recent studies have argued empirically that the vast majority of firm size dispersion is attributable to customer accumulation (Einav, Klenow, Levin, & Murciano-Goroff, 2021; Afrouzi, Drenik, & Kim, 2023). Yet empirical analysis is silent on the endogenous interaction between customers and productivity. If the most productive firms always endogenously acquire the most customers (as in Arkolakis, 2010), then the variation that empirical studies attribute to the customer margin is in fact fundamentally driven by productivity. On the other hand, if productivity and customers are completely orthogonal, these studies suggest a surprisingly minimal role for productivity in driving firm heterogeneity.

To investigate this, I use my model to simulate a simple counterfactual: How much firm size dispersion would there be if all firms could sell to all customers? I find that this economy has 63% lower variance of log revenue, indicating that almost two-thirds of firm size dispersion is attributable to the customer margin. Although this is lower than the 80-90% suggested by empirical studies, it affirms a central role for the customer margin even after endogenous interactions are accounted for. Moreover, the exercise reveals additional insights about the relationship between customers and productivity. I find that a substantial portion of the overall customer effect comes from its indirect, endogenous impact on the productivity distribution. Large variation in customer access generates large variation in incentives to innovate. As such, underlying productivity dispersion would be much lower if the customer channel were shut down, i.e., if all firms could sell to all customers.

My paper closely relates to the extensive literature on firm heterogeneity (Lucas, 1978; Hopenhayn, 1992; Melitz, 2003; Luttmer, 2007; Atkeson & Burstein, 2008; Edmond, Midrigan, & Xu, 2023; Berlingieri, Jaimovich, Terry, & Vincent, 2025). While this literature has typically modeled firm size as being directly tied to productivity, I allow for unproductive large firms by incorporating customer access. I show that this implies a distinct form of inefficiency compared to models that break the productivity-size link through financial frictions (see, e.g., Buera, Kaboski, and Shin, 2011; Moll, 2014). In such models, there is misallocation because the social planner would reallocate capital from unproductive firms to more productive ones. If capital cannot be reallocated, however, then the decentralized allocation of production is "constrained" efficient.[11] In my model, there is misallocation even if customers cannot be reallocated. The reason is that, taking customer stocks as given, variable demand elasticities induce more productive firms to underproduce

[11] This is also true if capital can be reallocated but the social planner is subject to the same frictions as firms, as in adjustment cost models.

relative to less productive ones. As such, there is a particular motive to reallocate *production* from unproductive high-customer firms to productive low-customer firms, even if their respective customer stocks are unchanged. This novel margin directly leads to my surprising tax result.

My paper also builds on previous models of endogenous customer accumulation, beginning with Phelps and Winter (1970) and including Ravn, Schmitt-Grohe, and Uribe (2006), Arkolakis (2010), Nakamura and Steinsson (2011), Dinlersoz and Yorukoglu (2012), Drozd and Nosal (2012), Gourio and Rudanko (2014), Arkolakis (2016), Sedlacek and Sterk (2017), Roldan-Blanco and Gilbukh (2021), Afrouzi, Drenik, and Kim (2023), Cavenaile, Celik, Perla, and Roldan-Blanco (2023), Ignaszak and Sedlacek (2023), Steinberg (2023), Bornstein (2025), and Hubmer and Nord (2025). This literature has generally either abstracted from productivity heterogeneity or considered settings where more productive firms always have more customers. I uncover a high degree of mismatch between productivity and customer access and demonstrate surprising policy implications of this finding.

Of particular relevance to my paper is recent work by Afrouzi, Drenik, and Kim (2023), Ignaszak and Sedlacek (2023), and Bernard, Dhyne, Magerman, Manova, and Moxnes (2022). Afrouzi, Drenik, and Kim (2023) study a notion of "customer misallocation," but do so in a model where firms can purchase as many customers as desired instantly within each period. By contrast, my model captures the gradual nature of customer accumulation stressed by empirical literature (Foster, Haltiwanger, & Syverson, 2016; Argente, Fitzgerald, Moreira, & Priolo, 2025). The result is that productive firms in my model can have few customers, whereas productivity and customers co-move strongly in their setup. Our different notions of customer misallocation reflect the distinct objectives of our two models: Afrouzi, Drenik, and Kim (2023) aim to revisit cross-sectional misallocation caused by markups (as in Edmond, Midrigan, and Xu, 2023), whereas I analyze dynamic misallocation due to productivity-customer mismatch. This mismatch directly leads to my novel policy results, which are absent in their framework. I also contribute to the literature by applying a new identification strategy to administrative data covering the universe of firms. Conversely, Afrouzi, Drenik, and Kim (2023) rely on Compustat for customer acquisition expenditures, which excludes many of the small and young firms for whom customer accumulation is arguably most relevant.

Ignaszak and Sedlacek (2023) show theoretically that firms with higher expected customer growth will invest more in innovation, and quantify the importance of this mechanism for aggregate growth. They do so with a model that shares some similarities to mine, such as endogenous investment in both productivity and customers. As with Afrouzi, Drenik, and Kim (2023), part of my contribution is to introduce a new identification strategy and apply it to a richer and more comprehensive dataset. More importantly, however, I study a very different question. Whereas Ignaszak and Sedlacek (2023) emphasize the aggregate growth effects of productivity-customer complementarity, I focus on the policy implications of productivity-customer

mismatch. As a result, my model includes additional features that are crucial for studying my particular question. For example, they restrict attention to CES demand, whereas I allow for variable elasticity through a Kimball aggregator. This generates a distinct form of misallocation, directly leading to my surprising revenue tax result. Moreover, I consider a substantially more general dynamic process for productivity, reflecting my focus on firm-level heterogeneity rather than long-run aggregate growth.

Finally, Bernard, Dhyne, Magerman, Manova, and Moxnes (2022) argue that a negative correlation between customers and productivity (i.e., mismatch) is needed to explain the joint behavior of sales and sales per customer. Whereas they impose this negative correlation exogenously in their model, I propose a theory that generates it endogenously. I also highlight additional empirical patterns that can be explained by this mismatch and demonstrate novel policy implications.

My work also connects to empirical studies on the role of demand in driving firm heterogeneity, with customer accumulation often particularly emphasized (Foster, Haltiwanger, & Syverson, 2008, 2016; Hottman, Redding, & Weinstein, 2016; Pozzi & Schivardi, 2016; Einav, Klenow, Levin, & Murciano-Goroff, 2021; Eslava, Haltiwanger, & Urdaneta, 2024). Whereas this literature has exploited specialized datasets covering particular subsets of the economy, I calibrate a structural model to administrative data covering the universe of firms. To do so, I develop a novel identification strategy that makes use of the model's firm-optimality conditions. The structure of the model also allows me to account for the endogenous interaction between productivity and customer access, a channel absent in empirical analysis.

2 Model

2.1 Environment

Time is discrete and indexed by t. The economy features a continuum of infinitely lived households indexed by $i \in [0, 1]$ and an infinite mass of atomistic firms indexed by $j \in \Omega_t$. The set Ω_t is determined in equilibrium by endogenous entry and exit. Firms produce a unique differentiated variety. Since each firm corresponds to exactly one variety, I use the terms "variety" and "firm" interchangeably.

2.2 Households

Household i has lifetime utility given by:

$$\sum_{t=0}^{\infty} \beta^t \left[\frac{C_{it}^{1-\gamma}}{1-\gamma} - \xi \frac{L_{it}^{1+\nu}}{1+\nu} \right], \tag{1}$$

where β, γ, and ν are exogenous parameters, C_{it} is consumption, and L_{it} is labor supplied. Household consumption C_{it} is a composite aggregate of varieties c_{ijt}, implicitly defined by:

$$\int_{\omega_{it}} \Upsilon\left(\frac{c_{ijt}}{C_{it}}\right) dj = 1. \tag{2}$$

The function $\Upsilon(\cdot)$ is assumed to be twice differentiable and satisfy $\Upsilon(1), \Upsilon'(\cdot) > 0, \Upsilon''(\cdot) < 0$, as in Kimball (1995). These assumptions imply constant returns to scale and a general form of imperfect substitutability between varieties j. In the case where $\Upsilon(x) = x^{1-\sigma^{-1}}$, this reduces to the CES aggregator.

The integral is taken over the set ω_{it}, a household specific subset of Ω_t, the set of all varieties. This subset is given by a random sample from Ω_t, where the probability that firm j belongs to ω_{it} is θ_{jt}, the *customer access* of j in period t. The assumption that this probability is a firm-level object, i.e., θ_{jt} rather than θ_{ijt}, greatly aids tractability. In particular, it ensures that the distribution of prices within each choice set ω_{it} is the same, a property which I refer to as *symmetry* of household choice sets. This in turn implies that each household has the same demand function.[12] As such, the only relevant consideration for a firm is *how many* customers they can sell to, not who those customers are. Finally, note that this setup nests standard models, where all households can buy all varieties, in the case where $\theta_{jt} = 1$ for all j, t and therefore $\omega_{it} = \Omega_t$ for all i, t.

All firms in the economy are owned equally by all households, such that each household receives an equal share of aggregate profits Π_t in each period. In addition, households are paid a wage w_t for each unit of labor supplied. Households consume their income in every period (i.e., they are hand to mouth), such that:

$$\int_{\omega_{it}} p_{jt} c_{ijt} dj = w_t L_{it} + \Pi_t. \tag{3}$$

To summarize, the household problem in each period t is given by:

$$\max_{\{c_{ijt}\}, L_{it}} \frac{C_{it}^{1-\gamma}}{1-\gamma} - \xi \frac{L_{it}^{1+\nu}}{1+\nu} \tag{4}$$

s.t.

$$\int_{\omega_{it}} \Upsilon\left(\frac{c_{ijt}}{C_{it}}\right) dj = 1$$

$$\int_{\omega_{it}} p_{jt} c_{ijt} dj = w_t L_{it} + \Pi_t.$$

[12] This result depends on the fact that there are infinitely many firms, so that the law of large numbers holds. See Appendix A.1 for more details.

2.3 Firms

Each firm produces its unique variety j with production function

$$q_{jt} = z_{jt}\ell_{jt}^{\alpha}, \tag{5}$$

where ℓ_{jt} is the quantity of labor hired and z_{jt} is productivity. As in standard models, productivity should be interpreted broadly. It includes quality, taste, and all demand-side factors aside from customer access that affect a firm's revenues.[13]

In addition to productivity z_{jt}, firms also differ in customer access θ_{jt}. As discussed above, θ_{jt} is the fraction of households i that have firm j in their choice set at time t:

$$\theta_{jt} \equiv \int_0^1 \mathbf{1}\{j \in \omega_{it}\}di. \tag{6}$$

All firms are price-takers in the input market and hire labor for the common price w_t. It will be convenient to define the firm's *static profit* (i.e., profit before investment and fixed costs) as:

$$\pi(\theta, z) \equiv \max_q \; p(\theta,q)q - w\underbrace{\left(\frac{q}{z}\right)^{1/\alpha}}_{=\ell}, \tag{7}$$

where $p(\theta, q)$ is the inverse demand curve arising from household optimization.

2.4 Dynamic Processes

I consider a discrete "ladder" for the state variables $\{\theta_1, .., \theta_{N_\theta}\} \times \{z_1, .., z_{N_z}\}$.

In each period, firms first draw exogenous negative shocks. If the firm was at position n_s, $s \in \{\theta, z\}$ in period $t-1$, its position after drawing the period t negative shock for variable s will be $n_s' \in \{1, 2, ..., n_s\}$. In words, it will either be pushed down the ladder or remain in its previous position. The value n_s' is drawn from an exogenous distribution $G_{s-}(n)$.

If $n_s' = n_s$, i.e., the firm avoided a negative shock to variable s, it is eligible for successful investment with endogenous probability λ_s. Success probability λ_s in period t requires that the firm hire x_s units of labor in period $t-1$, given by:

$$x_s(\theta, z, \lambda_s) = h_s(\lambda_s)\theta^{\eta_{\theta,s}} z^{\eta_{z,s}}, \tag{8}$$

[13] The justification for z including demand-side attributes is discussed in Section 2.7.1.

where $h'(\lambda) > 0, h''(\lambda) > 0$, $h_s(0) = 0$, $\lim_{\lambda \to 1} h(\lambda) = \infty$ and $\eta_{ss'} \in (-\infty, \infty)$ are exogenous parameters.

If investment is successful, firms draw their new position $n_{s,t} \in \{n_{s,t-1} + 1, n_{s,t-1} + 2, ..., N_s\}$ from an exogenous distribution $G_{s+}(n)$.

2.5 Dynamic Problem

At the beginning of each period, firms draw their state (θ, z) and decide to either exit (receiving value 0) or hire l_F units of overhead labor.

Operating firms receive static profit $\pi(\theta, z)$ according to their state (θ, z), as given in Equation 7. They choose investment success probabilities $\lambda_\theta, \lambda_z$, which determine their probability distribution over states (θ', z') in the next period. At the end of the period, firms may exogenously exit with probability $1 - \psi$.

The value of operating in state θ, z is therefore given by:

$$V(\theta, z) = \max_{\lambda_\theta, \lambda_z} \left\{ \pi(\theta, z) - w \sum_{s=\theta,z} x_s(\theta, z, \lambda_s) + \beta\psi E\Big[\max\{V(\theta', z') - w l_F, 0\} \mid \lambda_\theta, \lambda_z, \theta, z \Big] \right\}. \tag{9}$$

Firms endogenously exit whenever this value falls below the fixed cost $w l_F$.

2.6 Entry

In each period, an unbounded mass of entrepreneurs may pay an entry cost of κ units of labor in order to start a new firm. Each new firm begins with the lowest possible customer access θ_{min} and randomly drawn productivity. The position of each entrant on the productivity ladder is a Poisson variable with mean ζ. After drawing z, entrants can make an initial customer investment, before production begins, using the same technology as incumbents. Specifically, to increase from the minimum customer level with probability λ_E, they must hire $x_E(z, \lambda_E) = h_\theta(\lambda_E)\theta_{min}^{\eta_{\theta,\theta}} z^{\eta_{z,\theta}}$ units of labor. Conditional on success, they draw their starting position on the θ ladder from the distribution $G_{\theta+}(n)$. If their investment does not succeed, they remain at the lowest customer level.[14]

The value of entering with productivity z is therefore given by:

$$V_E(z) = \max_{\lambda_E} - w x_E(z, \lambda_E) + E_\theta[\max\{V(\theta, z) - w l_F, 0\}]. \tag{10}$$

Notice that the expectation is only taken with respect to θ, whereas z is known. Moreover, since the initial customer investment occurs within the same period, there is no discount factor $\beta\psi$.

[14] Note that entrants are not subject to negative shocks, either to θ or z.

The free entry condition is given by:

$$w\kappa = E_z[V_E(z)]. \tag{11}$$

2.7 Discussion of Model Assumptions

2.7.1 Household Homogeneity

As mentioned in Section 2.2, an important assumption for tractability is that household choice sets are symmetric, implying each household has the same demand function. This allows me to only keep track of *how many* households each firm can sell to, not which households they are. As such, one should interpret the Kimball demand function as representing an average demand within households that a firm has access to.

I make this assumption because the core question of this paper relates to variation in the *number* of customers that firms can sell to, not the composition of those customers.[15] Moreover, differences in customer composition across firms can be captured in a reduced-form way through productivity z. If two firms have the same number of customers but the customers of one firm have, on average, higher willingness to pay, the model can capture this by assigning the two firms the same θ and one firm a higher z. To see this, note that we could alter Equation 2 to instead implicitly define C_{it} by:

$$\int_{\omega_{it}} \Upsilon\left(\frac{\varphi_{jt} c_{ijt}}{C_{it}}\right) dj = 1, \tag{12}$$

where φ_{jt} is a preference shifter that captures demand-side factors, including differences in customer composition. If we did, it is easy to show there would be a direct isomorphism between z and φ. This also justifies the interpretation of z as including other demand-side factors such as quality.

2.7.2 Investment Cost Functions

Recall that the investment cost functions for $s \in \{\theta, z\}$ were defined as:

$$x_s(\theta, z, \lambda_s) = h_s(\lambda_s)\theta^{\eta_{\theta,s}} z^{\eta_{z,s}},$$

where λ_s is the investment success probability.

It is worth discussing briefly the role of the $\eta_{ss'}$ parameters, which capture the effect of the level of variable s on the cost of investment in s'. It is easy to show that the firm's static profit function, shown in

[15] See Sedlacek and Sterk (2017) and Cavenaile, Celik, Perla, and Roldan-Blanco (2023) for explicit treatments of customer composition.

Equation 16, is supermodular. In other words, the cross-partial derivative is positive: $\frac{\partial^2 \pi(\theta,z)}{\partial \theta \partial z} > 0$. This implies that if we set $\eta_{ss'} = 0$ for all s, s', investment in one variable would always be increasing in the level of the other variable. Firms with higher z would invest more in θ, and vice versa. While this may be true, it is not clear that it should be imposed ex-ante. If the cost of investing in s is higher when s' is higher, this may overturn the complementarity from the profit function. Indeed, a long line of innovation literature, motivated by empirical evidence from Akcigit and Kerr (2018), assumes that innovation costs are higher for large firms. As such, I allow $\eta_{ss'}$ parameters to potentially be nonzero, with the data informing their values through the calibration process.

2.7.3 Dynamic Processes

A crucial assumption of the model is that customer accumulation is associated with uncertainty and "time to build." Firms make an investment in period t with stochastic return in period $t+1$, rather than instantly and deterministically acquiring customers as in many alternative frameworks (Arkolakis, 2010; Afrouzi, Drenik, & Kim, 2023; Steinberg, 2023). This is the key assumption, in addition to the nonzero cost elasticities $\eta_{ss'}$, that allows for productivity-customer mismatch. If the cost elasticities $\eta_{ss'}$ were all zero, then, as described in the previous subsection, firms with higher productivity would always have greater incentives to invest in customers. If, in addition, customer acquisition was instant and deterministic, these greater incentives would always translate into more customers.

This assumption is intuitive — after all, even the most productive startups often take years to develop large customer bases. More importantly, however, it is strongly supported by the data. Foster, Haltiwanger, and Syverson (2016) show that new manufacturing establishments face persistently lower demand, even with high productivity, that takes years to gradually overcome. Similarly, Argente, Fitzgerald, Moreira, and Priolo (2025) provide evidence that retail firms accumulate customers gradually and are subject to customer accumulation frictions. Finally, as I show in Section 3.5, the mismatch generated by this assumption is critical in allowing the model to reconcile several seemingly contradictory empirical patterns.

Finally, it is worth noting that the dynamic processes described in Section 2.4 allow for successful investments to result in multiple "steps" up the relevant ladder. This is in contrast to standard quality ladder models, which typically consider innovations to occur one step at a time. In these models, firms control the probability that an innovation occurs, and if successful improve their quality by a fixed increment. Although this assumption is not overly restrictive in the context of long run aggregate growth, it is necessary to relax it in my setting to capture the rich variation in firm-level growth.[16] The approach I use does so with min-

[16] Of course, the "one step at a time" restriction is less problematic in continuous time models. I use discrete time because it more cleanly maps to the data on firm life-cycle profiles, which — as we will see — play a central role in my calibration.

imal cost to tractability and is similar to that used by Lim, Trefler, and Yu (2022) and Akcigit, Ates, and Impullitti (2024). Specifically, firms still only control the success probability but, in the event of success, can potentially take multiple steps forward. The exact number of steps they take forward, conditional on success, is drawn from an exogenous distribution.[17]

2.8 Equilibrium

As discussed above, household choice sets are *symmetric* in the sense that all households face the same distribution of prices. As a result, we can write $P_{it} = P_t$, where P_{it} is the price index faced by household i in period t.[18] Consequently, each household solves identical optimization problems and we can write $L_{it} = L_t$, $C_{it} = C_t$, and $c_{ijt} = c_{jt}$ for each i.

Normalizing $P_t = 1$, the first order conditions of the household problem give rise to the labor supply equation:

$$\xi L_t^{\nu} = C_t^{-\gamma} w_t \tag{13}$$

and inverse demand equation:

$$p_{jt} = D_t \Upsilon' \left(\frac{c_{jt}}{C_t} \right), \tag{14}$$

where D_t is given by:

$$D_t = \left(\int_{\omega_{it}} \Upsilon' \left(\frac{c_{jt}}{C_t} \right) \frac{c_{jt}}{C_t} dj \right)^{-1}. \tag{15}$$

Due to the symmetry of the choice sets ω_i, the integral above is the same for all i.

We can then write the static profit of firm j as:

$$\pi(\theta, z) \equiv \max_{q} \; D \Upsilon' \left(\frac{q}{\theta C} \right) q - w \underbrace{\left(\frac{q}{z} \right)^{1/\alpha}}_{=\ell}, \tag{16}$$

where $q_{jt} = c_{jt}\theta_{jt}$ follows from the fact that firm j sells the same amount c_{jt} to each of its customers.

I consider a stationary equilibrium such that aggregates (w, C, D) and the distribution of firms over states $\rho(\theta, z)$ are time invariant.

[17] It should be emphasized that firm investment affects the entire probability distribution across possible outcomes. It is only the distribution conditional on success that is exogenously fixed. Another way of saying this is that firms choose from a family of distributions over future outcomes, with the overarching family determined by the exogenous function G_{s+}.

[18] P_i is the cost of one unit of consumption, defined formally as $P_i \equiv \min \int_{\omega_i} p_j c_{ij} dj \; s.t. \; C_i = 1$. When Υ is the CES aggregator, this takes the usual form $P_i = \left(\int_{\omega_i} p_j^{1-\sigma} dj \right)^{\frac{1}{1-\sigma}}$.

Definition 1. A *stationary symmetric equilibrium* is a set of aggregates $\{C, D, L, \Pi, w\}$, policy functions $\{\lambda_E^*(z), d^*(\theta,z), \lambda_\theta^*(\theta,z), \lambda_z^*(\theta,z), q^*(\theta,z), \ell^*(\theta,z)\}$, value functions $\{V_E(z), V(\theta,z)\}$, distribution over firm types $\rho(\theta,z)$, and masses of operating firms and entrants $\{|\Omega|, \mathcal{E}\}$ that satisfies:

1. Household optimization: L and C solve each household's problem given aggregates w and Π;

2. Static firm optimization: $q^*(\theta,z)$ satisfies the static maximization of the firm given w, C, D, and $\ell^*(\theta,z) = \left[\frac{q^*(\theta,z)}{z}\right]^{1/\alpha}$;

3. Dynamic firm optimization: $V_E(z), V(\theta,z)$ solve the entrant and incumbent Bellman equations, $\lambda_E^*(z)$, $\lambda_\theta^*(\theta,z)$, and $\lambda_z^*(\theta,z)$ are the associated policy functions, and $d^*(\theta,z) = \mathbf{1}\{V(\theta,z) - wl_F \geq 0\}$ is the optimal decision function for operating;

4. Labor market clearing:

$$L = \sum_\theta \sum_z \left[\ell^*(\theta,z) + x_z(\theta,z,\lambda_z^*(\theta,z)) + x_\theta(\theta,z,\lambda_\theta^*(\theta,z)) + l_F\right]\rho(\theta,z)\,|\Omega| + \mathcal{E}\left[\kappa + \sum_z \mathbb{P}_E(z)\, d_E^*(z)\, x_E(z,\lambda_E^*(z))\right],$$

   where $\mathbb{P}_E(z)$ is the probability mass function for the productivity of entrants;

5. Goods market clearing: $\sum_\theta \sum_z \Upsilon\left(\frac{q^*(\theta,z)}{\theta C}\right)\theta\rho(\theta,z)|\Omega| = 1$ and the demand index D is given by $D = \left(\sum_\theta \sum_z \Upsilon'\left(\frac{q^*(\theta,z)}{C\theta}\right)\frac{q^*(\theta,z)}{C\theta}\theta\rho(\theta,z)|\Omega|\right)^{-1}$;

6. Free Entry: $w\kappa = \sum_z \mathbb{P}_E(z) V_E(z)$;

7. Stationary distribution:

$$\mathcal{E} = |\Omega|\left[1 - \sum_\theta \sum_z \sum_{\theta'} \sum_{z'} Pr\left(\theta', z' | \theta, z, \lambda_z^*(\theta,z), \lambda_\theta^*(\theta,z)\right)\psi d^*(\theta', z')\rho(\theta,z)\right],$$

   and

$$\rho(\theta', z') = \frac{\mathcal{E}}{|\Omega|}\,\mathbb{P}_E(z')\,\Pr(\theta' \mid z', \lambda_E^*(z'))\,d^*(\theta', z') + \sum_\theta \sum_z \Pr(\theta', z' \mid \theta, z, \lambda_z^*(\theta,z), \lambda_\theta^*(\theta,z))\,\psi\, d^*(\theta', z')\,\rho(\theta,z),$$

   where $\mathbb{P}_E(z)$ is the probability mass function for the productivity of entrants.

3 Calibration

3.1 Data

I calibrate the model using the National Accounts Longitudinal Microdata File (NALMF), which is an administrative dataset covering the universe of Canadian firms from 2001-2019. The cleaned dataset contains 1,755,000 unique firms and 13,123,000 firm-year observations.[19] Table 1 lists basic summary statistics.

Table 1: Summary Statistics

Variable	Mean	Std. Dev.
Revenue	\$1,967,504	\$7,068,270
Expenses	\$1,864,813	\$6,881,013
Employees	8.57	37.31
Firm age (years)	9.36	8.07

Notes: This table presents summary statistics from the cleaned NALMF dataset, covering years 2001-2019. Revenue and expenses are in 2019 Canadian dollars.

3.2 Functional Forms

3.2.1 Kimball Aggregator

Following recent literature (Afrouzi, Drenik, & Kim, 2023; Edmond, Midrigan, & Xu, 2023; Boar & Midrigan, 2024), I quantify the model using the Klenow and Willis (2016) specification of the Kimball aggregator:

$$\Upsilon(x) = 1 + (\sigma - 1)\exp\left(\frac{1}{\epsilon}\right)\epsilon^{\frac{\sigma}{\epsilon}-1}\left[\Gamma\left(\frac{\sigma}{\epsilon}, \frac{1}{\epsilon}\right) - \Gamma\left(\frac{\sigma}{\epsilon}, \frac{x^{\frac{\epsilon}{\sigma}}}{\epsilon}\right)\right], \tag{17}$$

where $\Gamma(s, r) \equiv \int_r^\infty t^{s-1}e^{-t}dt$ is the *upper incomplete Gamma function*, $\sigma > 1$, and $\epsilon > 0$. In the limit as $\epsilon \to 0$, this converges to the usual CES aggregator. With $\epsilon > 0$, which I assume, the elasticity of demand is higher at higher prices. This property, known as Marshall's Second Law, is a standard assumption in models of endogenous markups.

With this specification, we have:

$$\Upsilon'(x) = \frac{\sigma - 1}{\sigma}\exp\left(\frac{1 - x^{\frac{\epsilon}{\sigma}}}{\epsilon}\right). \tag{18}$$

[19] These numbers are rounded due to confidentiality restrictions.

3.2.2 Investment Cost Function

Recall that the investment cost functions for $s \in \{\theta, z\}$ were defined as:

$$x_s(\theta, z, \lambda_s) = h_s(\lambda_s)\theta^{\eta_{\theta,s}} z^{\eta_{z,s}}.$$

I consider the functional form for $h_s(\cdot)$ given by:

$$h_s(\lambda) = \chi_s\Big[-\log(1-\lambda) - \lambda\Big], \tag{19}$$

where χ_s is an exogenous parameter to be calibrated. This functional form is convenient in that it depends only on a single parameter and satisfies the assumptions $h'(\lambda) > 0, h''(\lambda) > 0$, $h_s(0) = 0$, $\lim_{\lambda\to 1} h(\lambda) = \infty$. Moreover, the fact that $h'(0) = 0$ is helpful in avoiding corner solutions.

3.2.3 Step Size Distributions

Recall that the exogenous functions G_{s-}, G_{s+} governed the distributions across negative and positive step sizes respectively.

I consider the same functional form for G_{s+} as Akcigit, Ates, and Impullitti (2024). In particular, I define a benchmark distribution for firms that are at the lowest point, i.e., $n_s = 1$. For these firms, the probability of moving to a point $n \in \{2, ..., N_s\}$, conditional on successful investment, is given by:

$$\bar{G}_{s+}(n) = \frac{n^{-\phi_{s+}}}{\sum_{r=2}^{N_s} r^{-\phi_{s+}}}. \tag{20}$$

This depends only on the parameter $\phi_{s+} > 0$, which governs the rate at which probability decays for higher step sizes. As $\phi_{s+} \to \infty$, the distribution reduces to the single step model.

For each $n_s > 1$, firms face a modified version of this benchmark function, such that all mass below $n_s + 1$ is shifted to the point $n_s + 1$. Specifically:

$$Pr(n_s' = n |\ n_s' > n_s) = \begin{Bmatrix} \bar{G}_{s+}(n) & \text{if } n > n_s + 1 \\ \sum_{r=2}^{n} \bar{G}_{s+}(r) & \text{for } n = n_s + 1 \end{Bmatrix}. \tag{21}$$

Notice that it is possible for any firm to jump to the highest level N_s, allowing for a small number of extremely high-growth firms. Yet the higher the firm currently is on the ladder, the more likely it becomes that they take only one step forward.

I consider an analogous process for negative shocks. In this case, the benchmark distribution is defined

for firms at the top of the ladder, i.e., $n_s = N_s$. For these firms, the probability of moving to position n is given by:

$$\bar{G}_{s-}(n) = \frac{(N_s + 1 - n)^{-\phi_{s-}}}{\sum_{r=1}^{N_s}(N_s + 1 - r)^{-\phi_{s-}}},$$

where again, $\phi_{s-} > 0$. In this case, $\phi_{s-} \to \infty$ corresponds to a model where s never decreases.

For firms with $n_s < N_s$, the distribution is given by:

$$Pr(n'_s = n) = \begin{Bmatrix} \bar{G}_{s-}(n + N_s - n_s) & \text{if } 1 < n \leq n_s \\ 1 - \sum_{r=2}^{n_s} \bar{G}_{s-}(r + N_s - n_s) & \text{if } n = 1 \end{Bmatrix}. \tag{22}$$

This setup implies that firms are less likely to experience a drastic negative shock the higher they are on the ladder.

3.3 Calibration Strategy

I first state a key result that will inform my calibration strategy.

Proposition 1. *Let $\mathcal{R}(\theta, z)$ and $\mu(\theta, z)$ denote the optimal revenue and markup for a firm in state (θ, z). Then:*

$$\mathcal{R}(\theta, z) \propto \theta \exp\left(\frac{\sigma}{\epsilon}\mu(\theta, z)^{-1}\right)\left(\frac{\mu(\theta, z) - 1}{\mu(\theta, z)}\right)^{\frac{\sigma}{\epsilon}}.$$

It is important to emphasize that this result depends only on the static firm optimality conditions, not any of the dynamic assumptions of the model. In other words, this will be true for any given (θ, z), regardless of how θ and z are endogenously determined. The key insight is that z enters the equation above only through the firm's markup μ. Consequently, revenue and markups together can identify customer access θ separately from productivity z.

The intuition for this result can be seen by dividing both sides of the equation by θ, yielding

$$\frac{\mathcal{R}(\theta, z)}{\theta} \propto \exp\left(\frac{\sigma}{\epsilon}\mu(\theta, z)^{-1}\right)\left(\frac{\mu(\theta, z) - 1}{\mu(\theta, z)}\right)^{\frac{\sigma}{\epsilon}}.$$

The left-hand side is sales per customer, whereas the right-hand side is a function only of the markup. In other words, there is a direct link in this model between the sales per customer of the firm and its markup, as supported empirically by Afrouzi, Drenik, and Kim (2023). As a result, the firm's sales per customer can be inferred from its markup. Since total sales are observable, knowing a firm's sales per customer immediately allows us to infer its number of customers.

If markups were observable at the firm-level, this result could in principle be used to infer customer access and productivity at the firm-level. In practice, however, estimating a firm's markup from revenue data alone is controversial (Bond, Hashemi, Kaplan, & Zoch, 2021). As such, I instead make use of the fact that there is a link between a firm's observed sales-cost ratio and its markup.[20] Specifically, the sales-cost ratio is given by:

$$\frac{\text{Total Revenue}}{\text{Total Expenses}} = \frac{pq}{w\ell + wl_F + wx_z + wx_\theta} = \alpha^{-1}\mu\left[\frac{\ell}{\ell + l_F + x_z + x_\theta}\right]. \tag{23}$$

This motivates the use of Simulated Method of Moments (SMM), with moments related to the joint distribution of revenue and the sales-cost ratio informing parameters that govern the (θ, z) distribution. The idea is to use an observable object (sales-cost ratio) that is informative about the unobservable object (the markup), rather than use an erroneous estimate of the unobservable object. This strategy is in line with the indirect inference approach, which has been widely studied and implemented (Gourieroux, Monfort, & Renault, 1993; Smith Jr., 1993; Lentz & Mortensen, 2008; Guvenen & Smith, 2014).

I set the exogenous exit rate $1 - \psi$, annual discount factor β, utility curvature γ, and Frisch elasticity $1/\nu$ to standard values in the literature.[21] I choose the disutility of labor ξ and entry cost κ to normalize $C = 1$ and $\frac{w}{CD} = 1$ in the baseline equilibrium.

The remaining 15 parameters are calibrated using Simulated Method of Moments (SMM). I target the 82 moments listed in Table 2.

Table 2: Targeted Moments

Exit rate
Aggregate profit share
Mean log revenue, ages 1 to 20
Mean log sales-cost ratio, ages 1 to 20
Variance of log revenue, ages 1 to 20
Variance of log sales-cost ratio, ages 1 to 20

Notes: This table shows the 82 moments that I target in the Simulated Method of Moments (SMM) estimation. Log revenue and log sales-cost ratio are residualized with 4 digit NAICS-year fixed effects.

[20] Although their overall estimation process is quite different, the strategy of targeting sales-cost ratios rather than markups is also used by Ignaszak and Sedlacek (2023).

[21] It should be noted that the two traditional channels that γ governs — risk aversion and intertemporal substitution — are both irrelevant in this model. There is no household risk and households do not make a consumption-savings decision. The only reason the value of γ matters is because it affects the labor supply equation through the marginal utility of consumption. This is why I describe it as a utility curvature parameter, rather than relative risk aversion or (inverse) elasticity of intertemporal substitution.

Table 3: Model Parameters

Parameter	Description	Value
Panel A: Externally Set		
$1/\nu$	Frisch elasticity	1.0000
γ	Utility curvature	2.0000
β	Discount factor	0.9600
$1-\psi$	Exogenous exit rate	0.0200
Panel B: Internally Calibrated		
χ_z	Productivity investment (scale)	3.7631
χ_θ	Customer investment (scale)	0.0145
η_{zz}	Productivity investment (own-elasticity)	2.5394
$\eta_{\theta z}$	Productivity investment (cross-elasticity)	0.0951
$\eta_{\theta\theta}$	Customer investment (own-elasticity)	-0.1984
$\eta_{z\theta}$	Customer investment (cross-elasticity)	1.0716
$\phi_{\theta+}$	Customer step distribution (increasing)	2.8285
$\phi_{\theta-}$	Customer step distribution (decreasing)	3.1039
ϕ_{z+}	Productivity step distribution (increasing)	7.1419
ϕ_{z-}	Productivity step distribution (decreasing)	3.0282
l_f	Overhead labor	0.0001
ζ	Mean entrant productivity	17.8970
α	Production function curvature	0.7711
σ	Kimball (elasticity)	4.5631
ϵ	Kimball (superelasticity)	5.9054
Panel C: Normalizations		
κ	Entry cost	2.12×10^{-6}
ξ	Labor disutility	2.4789

Notes: This table presents the estimated values for all model parameters. Panel A shows externally set parameters according to standard values in the literature. Panel B shows internally calibrated parameters, which are estimated to match the 82 moments in Table 2. Parameters in Panel C are chosen to normalize the values of C and $\frac{w}{CD}$ in the baseline equilibrium.

3.4 Results

Table 3 displays the full list of parameters. A few comments are in order. Firstly, the calibrated values of standard parameters α, ϵ, and σ are consistent with previous literature. The value $\sigma = 4.5631$ is on the low end of the typical range, but is close to the value $\sigma = 5$ considered by Klenow and Willis (2016). It is also worth noting that σ is positively related to dispersion in sales per customer. Models with representative consumers require higher values of σ to match observed size dispersion than models — such as this one — that also have an extensive margin of demand.

Secondly, the calibrated value for ζ implies that entrants have an average productivity around 33% lower than the mean productivity of incumbent firms. Most of these entrants, however, do not ever operate. This is because, in the calibrated model, firms at the bottom of the θ ladder will always exit, regardless of how high their z is. As such, only entrants that are able to reach a significant number of customers through

their initial investment will operate. Combined with the low entry cost κ, one should think of entry as simply acquiring an idea. Ideas are typically of low quality, and only materialize into an operating firm when the idea is good enough and the firm using the idea is able to reach enough customers. Conditional on an entrant operating, however, its productivity is actually higher on average than incumbents, as will be discussed further shortly.

Thirdly, the cost elasticity (η) parameters are all positive, with the exception of $\eta_{\theta\theta}$. The fact that $\eta_{\theta\theta} < 0$ means that high levels of customer access make it easier for firms to acquire even more customers. This can be thought of as capturing word of mouth or peer effects, as empirically documented by Argente and Mendez (2025) and explicitly assumed in recent models such as Hubmer and Nord (2025). On the other hand, $\eta_{z\theta}$ is positive, indicating that it is more costly for productive firms to reach customers. Recall that part of what z captures is consumer tastes and demand for a firm's product. As such, this could reflect a tradeoff between valuation and scope — products that are highly valued by a niche set of customers may find it more difficult to expand to a broader audience.

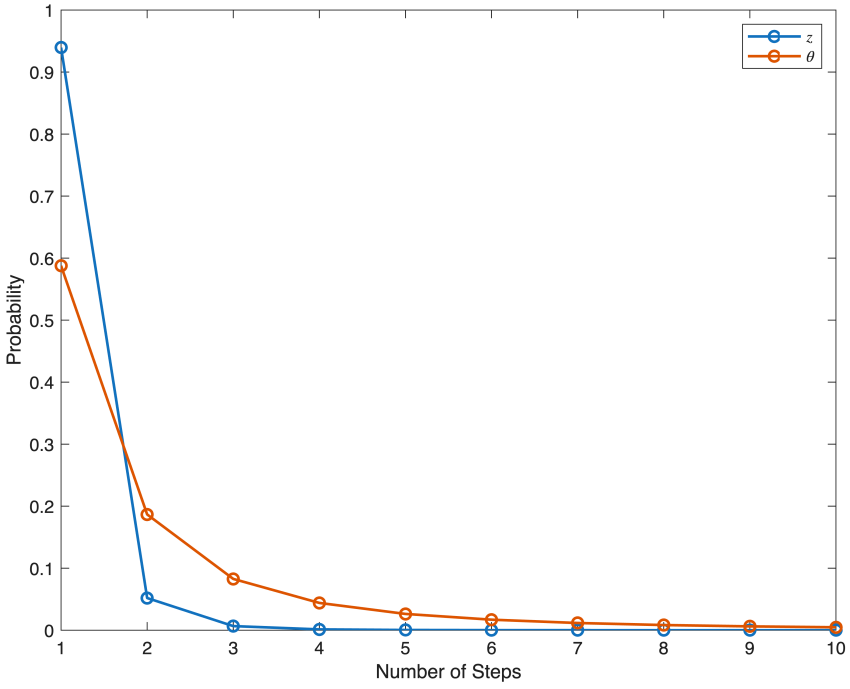
The value of $\eta_{\theta z}$ is close to zero, suggesting a minimal role for customer levels in influencing innovation costs. Interestingly, however, this is paired with a substantially larger value for η_{zz} than is typically estimated in productivity-only models. Those with CES demand have estimated values in the range of 1.2-1.3 (Akcigit & Kerr, 2018; Ignaszak & Sedlacek, 2023), whereas Yao (2022) estimates a slightly higher value of 1.6 with Kimball demand. These estimates are typically based on the firm-level relationship between R&D or growth rates and firm size. Since firm size includes both productivity and customer access, my results suggest that these models are capturing an average of the low $\eta_{\theta z}$ and high η_{zz}. This would mean that it is not firm size *per se* that raises innovation costs, but firm productivity. Consequently, there may be large firms with low innovation costs, if their large size is due to high customer access rather than productivity.

Finally, the step size (ϕ) parameters indicate similar negative shock processes for productivity and customer access, but a slower positive process for productivity. Figure 1 compares the distribution across positive step sizes for a firm at the bottom of the relevant ladder.[22] Conditional on a successful innovation, firms overwhelmingly take only one productivity step forward. This justifies the choice of many quality ladder models to restrict innovations to occur one step at a time, even within discrete time models. For customer access, however, the probability of taking multiple steps is much higher. For a firm at the bottom of the ladder, the probability of taking even five steps forward is as high as 2.6%. This process suggests rich heterogeneity in customer growth rates, including the potential for explosive growth early in the firm's life.

Figure 2 shows the negative step distributions, which are very similar for both θ and z. The probability of drawing a negative shock in a given period is around $1 - 0.85 = 0.15$ for both variables. Conditional

[22] Recall that as firms progress up the ladder, the probability of taking only one step forward increases.

Figure 1: Positive Step Distribution



Notes: This figure shows the calibrated positive step size distribution for a firm at the bottom of the relevant ladder. The y-axis shows the probability of taking different step sizes, conditional on successful investment. Note that the distribution is defined over the entire ladder, but only 10 steps are shown for visual clarity.

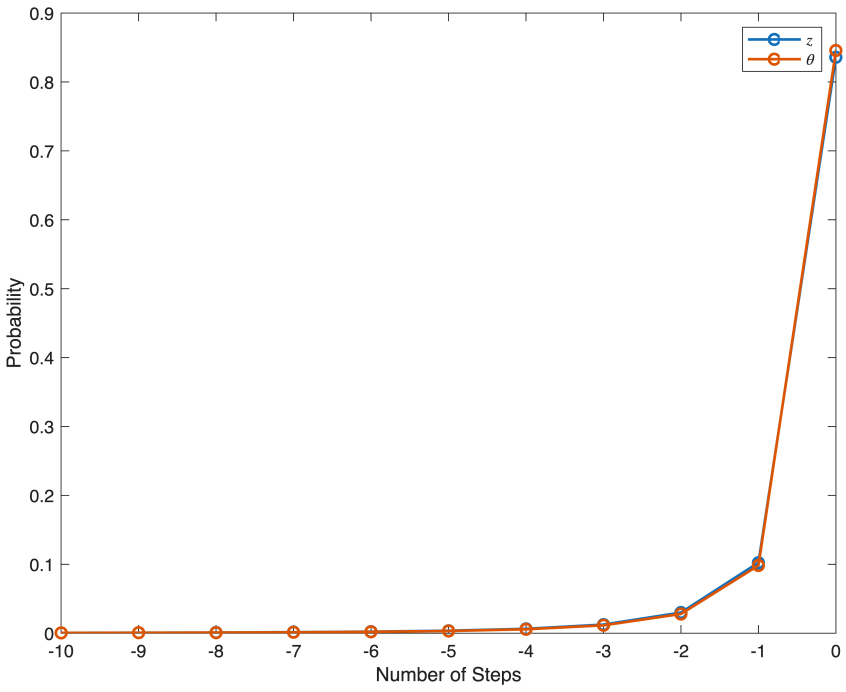
on drawing a negative shock, the size could be as large as six steps backwards before the probability falls below 1%. This suggests relatively volatile processes in both variables, in line with the high levels of churn observed in the data.

3.5 Model Properties

I now proceed to discussing the model's empirical performance. Does the model fit the data well? Moreover, what additional explanatory power does productivity-customer mismatch provide?

I begin by evaluating the model's performance on targeted moments. Figure 3 compares the model to the data with respect to the four targeted life cycle profiles. The model does quite well, particularly considering how few parameters there are compared to the number of moments. It is reassuring that the model matches the variance of log revenue, a key measure of firm size dispersion, especially closely. After all, understanding firm size dispersion more deeply is the fundamental objective of this paper. It is also worth noting that variation in the sales-cost ratio, both between and within ages, is an order of magnitude lower than variation in revenue. Although the model struggles to replicate the exact shape of the two sales-cost ratio profiles, the

Figure 2: Negative Step Distribution



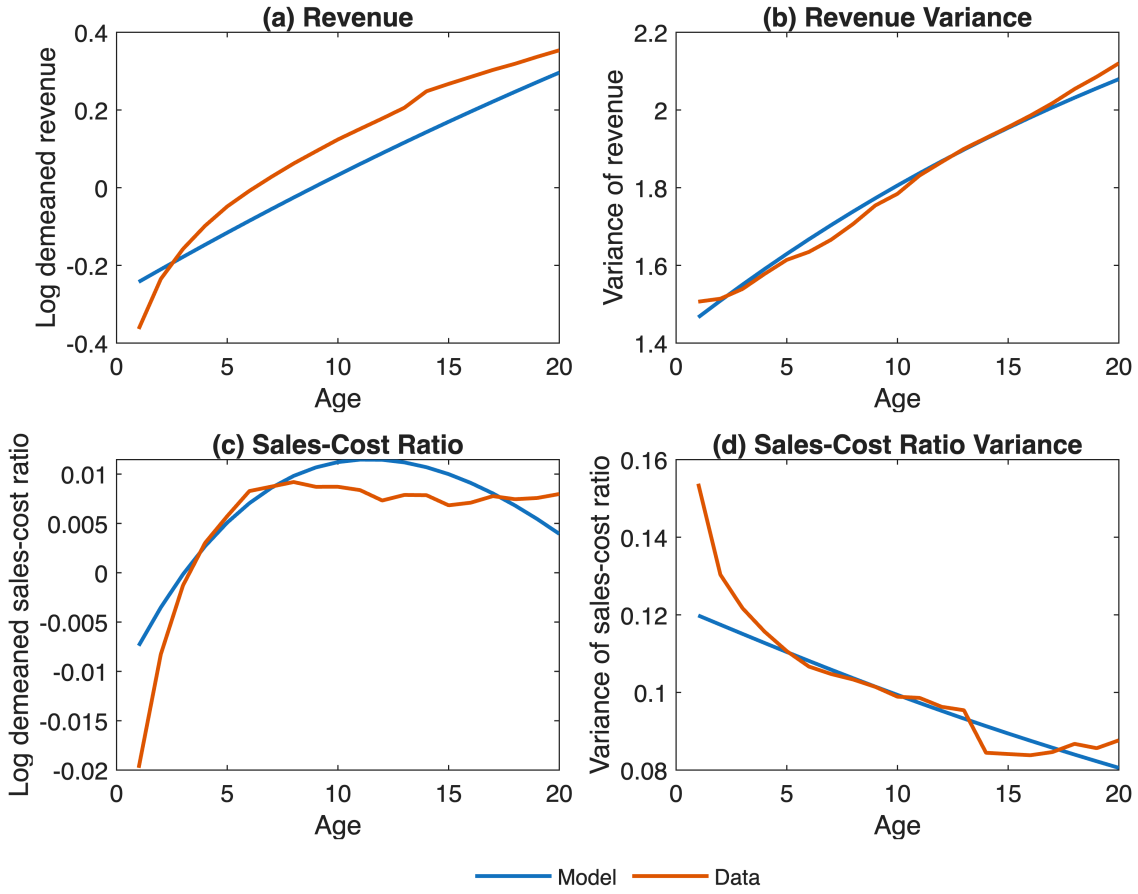
Notes: This figure shows the calibrated negative step size distribution for a firm at the top of the respective ladder. At the start of each period, each firm will draw a value on the x-axis with associated probability on the y-axis. If they draw a value of 0, they remain in their current position and are eligible for successful investment. Note that the distribution is defined over the entire ladder, but only 10 steps are shown for visual clarity.

fact that it matches this striking feature about their levels is encouraging.

In general, the markedly different behavior of the sales-cost ratio compared to revenue is not a trivial pattern to match. While the mean of revenue increases steeply over the life cycle, the sales-cost ratio varies only marginally. Even more strikingly, the variances display diametrically opposite behavior: increasing over the life cycle for revenue and decreasing for the sales-cost ratio. The model is able to match these differences because there are two dimensions of firm heterogeneity which need not move together, and in fact often diverge. Compare this, however, to other frameworks. In models with only productivity (such as Hopenhayn, 1992) or with no productivity-customer mismatch (such as Arkolakis, 2010), revenue and the sales-cost ratio must go hand in hand. As such, it would be difficult to reconcile the opposing behavior found in the data. This clearly illustrates the value of a model with productivity-customer mismatch in better capturing the rich firm heterogeneity found in the data.

Table 4 shows the model's performance with respect to aggregate moments. It matches the overall variance of revenue and the sales-cost ratio quite well. Although I target the within-age variance from ages

Figure 3: Life-cycle moments (Model vs Data)



Notes: This figure shows the model's fit relative to the four life-cycle profiles in Table 2. Log revenue and log sales-cost ratio are residualized with 4 digit NAICS-year fixed effects. Subfigures (a) and (c) show the mean of the residualized objects in each age group. Subfigures (b) and (d) show the variance of the residualized objects in each age group.

1 to 20, I do not target any information about the distribution of firms across ages. As such, it is reassuring that the model is still able to match the overall variance across all firms. Compared to the data, the model generates a moderately higher covariance between revenue and the sales-cost ratio. If anything, however, this would suggest that the model is overestimating the covariance between size and productivity. This is because the sales-cost ratio is closely related to the markup, which is an increasing function of productivity. As such, matching this moment more closely would likely further reduce the covariance between size and productivity and strengthen my results.

The model also tightly replicates the negative relationship between labor share and size. This mechanically corresponds to a positive relationship between revenue productivity and size, as in the data (see,

e.g., Bartelsman, Haltiwanger, and Scarpetta, 2013).[23] The model is able to match these patterns despite the fact that large firms are not necessarily productive. Instead, these relationships are primarily driven by the presence of overhead labor. Since wages paid to overhead workers wl_F mechanically constitute a smaller share of revenue for large firms, there would be a positive relationship between measured revenue productivity and size even if all firms had the same underlying productivity z. This mechanism has been previously discussed by Bartelsman, Haltiwanger, and Scarpetta (2013) and Autor, Dorn, Katz, Patterson, and Van Reenen (2017).

Table 4: Aggregate Moments (Model vs Data)

Moment	**Model**	**Data**
Panel A: Targeted Moments		
Exit Rate	0.097	0.097
Profit Share	0.128	0.116
Panel B: Untargeted Moments		
Variance of log revenue	1.814	1.898
Variance of log sales-cost ratio	0.103	0.117
Covariance of log revenue and log sales-cost ratio	0.223	0.098
Regression of labor share on size	-0.123	-0.191

Notes: This table shows the model's fit relative to targeted and untargeted aggregate moments. The data value for the profit share is taken from Gouin-Bonenfant (2022).

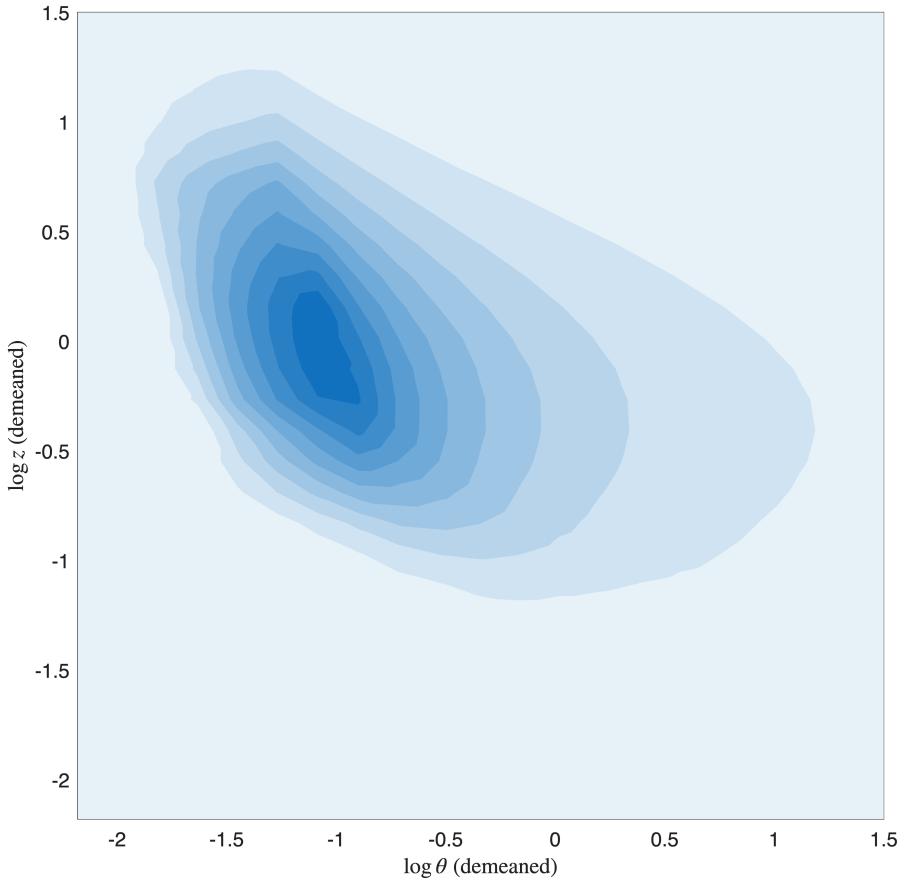
I now discuss properties of the fundamental state variables θ and z and how they further enhance our understanding of the data. Figure 4 shows the stationary distribution over $\log(\theta)$ and $\log(z)$. The most striking feature of the distribution is the strong negative relationship between the two variables, with $corr(\log\theta, \log z) = -0.30$. Since this may seem surprising, I explain this result in two steps. The first key step to this result is that the model allows for mismatch between productivity and customers. As discussed in Section 2.7.3, alternative models that allow for customers to be acquired instantly and deterministically effectively rule out mismatch, inevitably giving rise to strong positive correlations. More productive firms have a greater incentive to invest in customers, and this incentive can always be immediately realized, so more productive firms always have more customers.

The next step, given that productivity and customers do not necessarily go hand in hand, is selection. Since θ and z are both firm state variables, selection takes place on both margins. The productivity threshold below which a firm will exit varies according to its level of customer access. In particular, a firm with high

[23] Let revenue productivity be defined as the ratio of revenue to employment, and let the labor share be defined as the ratio of total labor compensation to revenue. Then, since wages do not differ across firms in the model, we have $\log \text{Labor Share} = -\log \text{Revenue Productivity} + \text{Constant}$.

customer access can continue to operate at a productivity level low enough to force out any low-customer firm. This immediately leads to the negative correlation between θ and z: Conditional on higher θ, expected z is lower because the minimum z required to operate is lower.[24]

Figure 4: Joint Distribution of Productivity and Customer Access

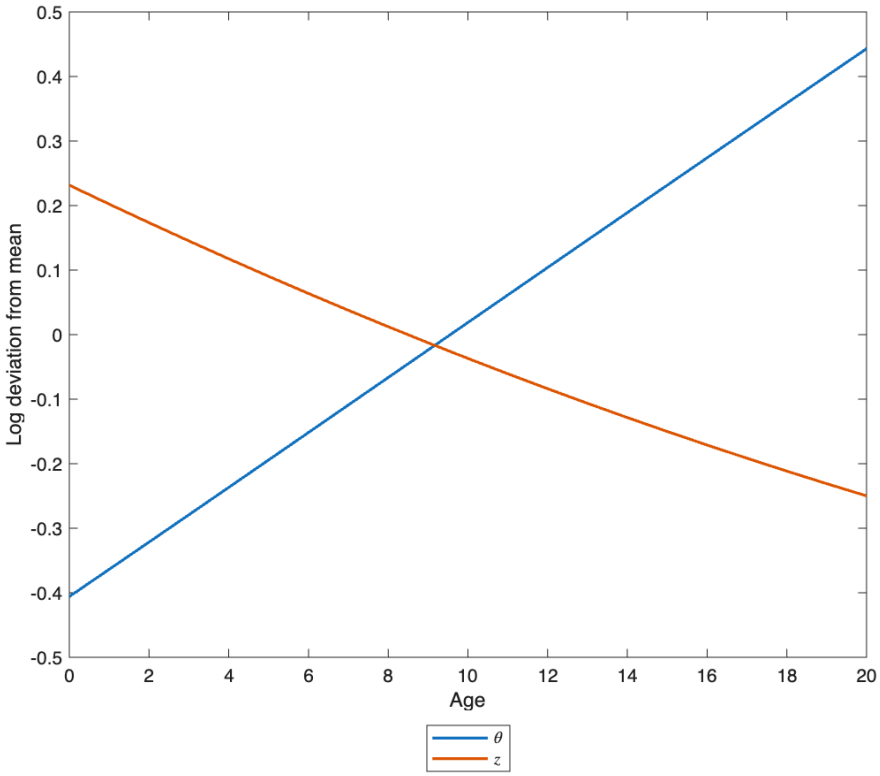


Notes: This contour graph shows the stationary joint distribution of firms over demeaned $\log\theta$ and demeaned $\log z$. Darker shades indicate a higher density of firms. For visual clarity, I depict here a subset of the state space that contains 87% of all firms. Figure 6 shows the full support of the marginal distributions over $\log\theta$ and $\log z$. Similarly, although the underlying state space is discrete, I present a continuous distribution here for visual clarity. See Appendix B.2 for the raw underlying matrix.

This feature of the joint distribution implies a negative correlation between a firm's customers and its sales per customer, a fact documented empirically by Bernard, Dhyne, Magerman, Manova, and Moxnes (2022). More broadly, negative correlations between productivity and demand-side factors are widely reported in

[24]Of course, the minimum z being lower does not necessarily imply that the conditional mean will be lower. This is a quantitative question, and depends on the nature of the conditional distribution. It turns out to be true that the conditional mean is lower in the calibrated model.

Figure 5: Customer and Productivity Life Cycle Profiles



Notes: This graph shows the mean of $\log\theta - \overline{\log\theta}$ and $\log z - \overline{\log z}$ by age, where bars denote economy-wide means.

empirical studies (Roberts, Yi Xu, Fan, & Zhang, 2018; Blum, Claro, Horstmann, & Rivers, 2024; Aghion, Bergeaud, Boppart, Klenow, & Li, 2025). Alternative models of customer acquisition, with strong positive correlations between customers and productivity, are inconsistent with these findings.[25]

Next, Figure 5 shows the life cycle profiles of productivity and customer access. Despite entrants being significantly smaller than incumbents, as in the data, they are actually more productive on average. This is in line with empirical evidence documented by Foster, Haltiwanger, and Syverson (2008) and others. Again, it is clear that the mismatch plays a central role in reconciling these two patterns. If productivity and customers always go together, then small firms must be unproductive. As such, entrants cannot simultaneously be small and productive. The model with mismatch is able to naturally reconcile both of these empirical findings. The entirety of the positive life cycle trend in revenue is attributable to customer access, as hypothesized by Foster, Haltiwanger, and Syverson (2016). In fact, average productivity is lower for older firms. This finding is easily explained by the selection margin described above. Since older firms have accumulated more customers on average, the threshold z for continuing to operate is lower, and this leads to a lower average z.

[25] As emphasized previously, the key feature that prevents other models from matching these facts is instant customer acquisition. Ignaszak and Sedlacek (2023) model gradual customer accumulation and similarly find a negative correlation between productivity and customers.

To summarize, the model jointly rationalizes the following collection of empirical facts. First, the dramatically different behavior of revenue and the sales-cost ratio over the life cycle. Second, the positive relationship between revenue productivity and size. Third, the negative correlation between sales per customer and total sales. Fourth, the smaller size but higher productivity of entering firms compared to incumbents. These four facts cannot all be explained by productivity-only models or customer accumulation models where more productive firms always have customers. As such, productivity-customer mismatch is an important and useful mechanism for understanding firm-level variation in the data.

In addition to capturing empirical patterns, the model also provides novel insights as to their explanation. I conclude this section by discussing one such pattern. Figure 6 shows the respective marginal distributions of $\log \theta$ and $\log z$. As is clear from the figure, customer access displays much higher variation than productivity. Intuitively, this is closely tied to the empirically higher variation in revenue compared to the sales-cost ratio. More importantly, the right tail of the $\log \theta$ distribution is quite long and much longer than that of $\log z$. This contributes to our understanding of the phenomenon of "superstar" firms, which are at the far right tail of the size distribution and therefore often assumed to be at the far right tail of the productivity distribution. My model suggests an alternative interpretation: The astronomical size of these superstar firms comes from extremely high values of customer access. The largest firms in my model are of course highly productive, but their productivity is nowhere near as high, relative to the mean, as their customer access is.

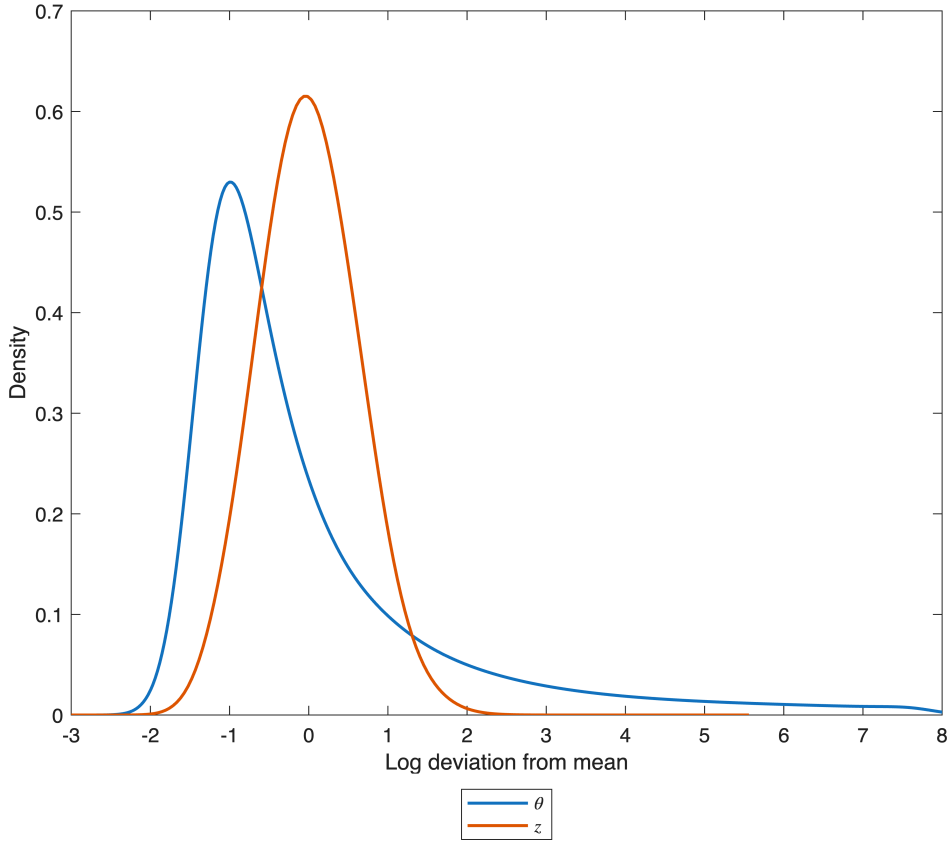
4 Quantitative Exercises

I consider three main exercises in this section. First, I revisit the conventional wisdom that taxing large firms lowers aggregate productivity. Second, I analyze policies that directly influence the customer investment decisions of firms. Third, I add to the ongoing discussion regarding the importance of customer accumulation for firm heterogeneity.

4.1 Taxing Large Firms

The rise of "superstar firms" in recent decades, combined with the shrinking middle class and growing wealth inequality, has fueled calls for greater taxation of large companies. Through the lens of a productivity-only model, these large companies are necessarily the most productive firms in the economy. Consequently, taxing them seems clearly ill-advised. In this section, I show that this conventional wisdom does not carry over to my model. In fact, a tax on high-revenue firms can actually *increase* welfare. Although I explain the drivers of this result in detail shortly, the intuition is straightforward: In the model, there are a significant number of unproductive large firms that are propped up by high customer access. Taxes on high revenue

Figure 6: Marginal Distributions



Notes: This graph shows the respective marginal distributions of productivity z and customer access θ. Although the underlying state space in this model is discrete, for visual clarity I present continuous Kernel densities in this figure. See Appendix B.2 for the raw underlying histograms.

shift production away from these firms. This channel is strong enough that even though some productive firms contract as a result of the policy, its overall effect is positive.

This intuition applies to a large class of policies that reallocate production from large to small firms. To illustrate it concretely, however, I consider a specific family of revenue taxes in the spirit of Canada's recently rescinded Digital Services Tax (DST).[26] This family takes the following form: All revenue above \$20 million is taxed at a rate of $\tau_{\mathcal{R}}$, whereas revenue below \$20 million is untaxed. In the case of Canada's DST, $\tau_{\mathcal{R}} = 0.03$.[27] I also present results for the value of $\tau_{\mathcal{R}}$ that maximizes the consumption equivalent welfare effect, taking as given the threshold of \$20 million.

[26] I use the structure of Canada's DST as a convenient template for a revenue tax targeting large firms. Unlike the actual DST, which applied only to certain digital services, my analysis applies the tax uniformly across all firms. The goal is to study the effects of targeting large firms, not the particular considerations of digital services.

[27] Note that because the policy and my data are Canadian, the threshold of \$20 million is in Canadian dollars. At the time of writing, this corresponds to around 14 million USD.

I consider a revenue tax rather than the more commonly studied profit tax for several reasons. Firstly, it most cleanly targets the key mechanism at play, which is the cross-sectional misallocation of production. A profit tax, in addition to reallocating cross-sectional production, also has a strong dynamic effect on investment decisions. This introduces additional confounders to the overall effect. Secondly, a revenue tax offers a direct comparison to the size-dependent subsidies studied in productivity-only settings such as Edmond, Midrigan, and Xu (2023). Finally, although practical implementation is not the primary concern of this exercise, it is worth noting that there are growing concerns over the practical efficacy of profit taxes. This has led to a number of recent tax policies targeting revenue.[28]

In order to provide a quantitative benchmark for my results, I also implement the revenue tax in a fully recalibrated productivity-only version of the model. The productivity-only model is identical to the full model, except that $\theta = 1$ for all firms. This recalibrated model allows me to quantify the intuition that such a policy would be detrimental in a more conventional framework.

To evaluate all policies, I use the standard definition of consumption-equivalent welfare (CEW) effects: The CEW of a policy is the percentage change in consumption that would be required to generate the same welfare change in the absence of the policy. Specifically, recall that the lifetime utility of households is given by:

$$\sum_{t=0}^{\infty} \beta^t \left[\frac{C_{it}^{1-\gamma}}{1-\gamma} - \xi \frac{L_{it}^{1+\nu}}{1+\nu} \right].$$

In a symmetric stationary equilibrium with $C_{it} = C$ and $L_{it} = L$ for all i, t, we can write lifetime utility as:

$$U(C, L) = \frac{1}{1-\beta} \left[\frac{C^{1-\gamma}}{1-\gamma} - \xi \frac{L^{1+\nu}}{1+\nu} \right]. \tag{24}$$

Let C_{old}, L_{old} denote the level of consumption and labor in the baseline stationary equilibrium. Let C_{new}, L_{new} denote the level of consumption and labor in the stationary equilibrium with some new policy. The CEW is then the amount Ξ such that:

$$U(C_{old} + \Xi, L_{old}) = U(C_{new}, L_{new}). \tag{25}$$

For ease of comparison, I report values of Ξ as percentages of baseline consumption. Table 5 presents the results. The 3% tax generates substantial welfare losses in the productivity-only model, yet modest gains in the full model. When maximizing the welfare effect in the full model, the result is even more striking.

[28] The difficulty of attributing profits from digital services has led to many countries (e.g., the UK, Spain, Italy, and Austria) implementing taxes on digital *revenue*. These taxes typically target large firms, as in the policy I consider. Another concern is multinational profit shifting, as studied recently in Dyrda, Hong, and Steinberg (2024).

The optimal tax in the full model is 10%, a policy which dramatically reduces welfare in a productivity only model. With heterogeneous customer access, however, this tax actually yields substantial welfare *gains*.

Table 5: Taxing Large Firms

	DST (3%)		Optimal (10%)	
	Full Model	Productivity Only	Full Model	Productivity Only
CEW	+0.1%	−1.1%	+0.9%	−2.4%
ΔTFP	+0.5%	−1.2%	+2.4%	−2.8%
ΔY	−0.4%	−0.8%	−0.9%	−1.7%

Notes: This table presents the effects of two size-dependent revenue taxes in the full model and a "productivity-only" version of the model where $\theta = 1$ for all firms. Both taxes apply to all revenue above $20 million. The 3% rate is chosen based on Canada's DST, whereas 10% is the rate that maximizes CEW, the consumption-equivalent welfare effect, in the full model. Y is aggregate output, L_p is total production labor, $TFP \equiv \frac{Y}{L_p^{\alpha}}$, and L is total labor.

Why are the results so different in the two models? To understand this, it will be useful to unpack the effect on TFP. Consider the following accounting decomposition of TFP:

$$TFP = |\Omega| \times \overline{Z} \times Q, \tag{26}$$

where $|\Omega|$ is the mass of operating firms, $\overline{Z}$ represents underlying productivity given the stationary distribution $\rho(\theta, z)$, and $Q \leq 1$ captures allocative efficiency. If $Q = 1$, the cross-sectional allocation of production is efficient.[29] In general, $Q < 1$ due to the well-known distortions caused by variable markups (see, e.g., Edmond, Midrigan, and Xu, 2023).

Motivated by this decomposition, I present in Table 6 the percentage changes in each of these three components. The table paints a clear picture as to the underlying tradeoff of the policy: On the one hand, the tax dampens aggregate labor demand, reducing the wage and encouraging the entry of new firms. This raises TFP through the mass of operating firms $|\Omega|$. On the other hand, the tax disincentivizes investment and distorts the allocation of production, lowering $\overline{Z}$ and Q. This fundamental tradeoff is the same in both models. In the full model, however, the reduction in $\overline{Z}$ is less severe and the reduction in Q almost vanishes. As a result, the net effect becomes positive.

To understand the different effects on $\overline{Z}$, note that revenue and productivity are synonymous in the productivity-only model. As such, taxing high revenue directly reduces the marginal benefit of innovation

[29] For a formal definition of efficiency, see Appendix A.2.

Table 6: Taxing Large Firms (Decomposition)

	DST (3%)		Optimal (10%)	
	Full Model	Productivity Only	Full Model	Productivity Only
$\Delta\lvert\Omega\rvert$	+3.3%	+3.2%	+16.5%	+70.0%
$\Delta\overline{Z}$	−2.8%	−4.0%	−12.0%	−42.3%
ΔQ	−0.0%	−0.2%	−0.2%	−0.9%

Notes: This table shows the percentage change in the three components that constitute TFP in this model based on Equation 26. Note that Equation 26 implies an exact decomposition in logarithms, whereas the changes presented in this table are percentages. Therefore, the percentage changes in $|\Omega|$, $\overline{Z}$, and Q do not add up to the total percentage change in TFP, particularly when the percentage changes are large.

and thereby $\overline{Z}$. Conversely, the full model features two margins for improving $\overline{Z}$ — customer and productivity investment — neither of which solely determine firm-level revenue. As such, the disincentive caused by a revenue tax is diluted in this model and the overall effect is less harmful.

As for the effect on Q, it is helpful to recall the nature of misallocation in this class of models. In both this and standard models, market power induces firms to produce less than the competitive equilibrium. Misallocation occurs because high-markup firms face a lower elasticity than low-markup firms, so the markup distortion is stronger for them. Policies that reallocate production from low-markup to high-markup firms will improve allocative efficiency, and vice versa.

In a productivity-only model, "high markup" is synonymous with high revenue. The tax reallocates production from high-revenue firms to low-revenue firms, so it clearly worsens allocative efficiency. The key difference in the model with customer access is that high-markup firms are not necessarily high-revenue firms. Instead, as emphasized in Section 4 and supported empirically by Afrouzi, Drenik, and Kim (2023), markups are determined by revenue *per customer*. The following result relates this property directly to firm fundamentals θ and z.

Lemma 1. *Let $\mu(\theta, z)$ denote the optimal markup for a firm with customer access θ and productivity z. Then:*

1. $\frac{\partial \mu(\theta,z)}{\partial z} > 0$
2. $\frac{\partial \mu(\theta,z)}{\partial \theta} \leq 0$, *with equality if and only if* $\alpha = 1$

The reasoning behind this result is as follows. In standard models with variable markups, lower marginal costs typically lead to higher markups. The same property holds in this model, except that the relevant object is marginal cost *per customer*. When there are decreasing returns, the cost of producing a given

amount *per customer* rises with θ, so the firm charges a lower markup. On the other hand, higher z still leads to a lower marginal cost per customer and therefore a higher markup.

The upshot is that shifting production from high revenue to low revenue firms need not worsen misallocation. The precise effect depends on how the reallocation occurs with respect to θ and z. Table 7 investigates this in more detail. For simplicity, I split the economy into four groups based on whether their θ and z are high (above average) or low (below average). I then compute the production shares of these groups before and after the tax. The table shows that the policy reduces the production share of high-θ, low-z firms and reallocates it to the other three groups. While the expansion of low-θ, low-z firms has a negative effect on allocative efficiency, the expansion of high-z firms has a positive effect. These combine to yield a much smaller reduction in Q relative to the productivity-only model.

Table 7: Production Shares (No Tax $\rightarrow$ 10% Tax)

	Low z	High z
Low θ	9.8% $\rightarrow$ 10.5%	4.1% $\rightarrow$ 5.5%
High θ	33.7% $\rightarrow$ 26.3%	52.4% $\rightarrow$ 57.7%

Notes: This table splits firms into four subgroups based on whether their θ and z are high (above average) or low (below average). It shows the production share of each group before and after a 10% tax is imposed on all revenue above $20 million. The format is [share before] $\rightarrow$ [share after].

The final question to answer is why low-z, high-θ firms contract, rather than the other groups. To understand this, let us briefly consider the mechanics of this policy. There are three main effects. Firstly, all firms whose optimal revenue (in the absence of a tax) was above the tax threshold contract, since their marginal revenue decreases. Secondly, firms whose optimal revenue was slightly above the tax threshold reduce their output and "bunch" at the threshold. Finally, aggregate labor demand falls, reducing the wage. This exerts an upward force on the production of all firms.

The key insight is that the first effect, a reduction in marginal revenue, has a stronger impact on low-z firms. This is because low-z firms sell less to each customer, facing a higher elasticity and consequently reacting more strongly to negative shocks. As such, when all three forces are quantitatively combined, the only group to contract in relative terms is the low-z, high-θ group. Many of these firms have high enough θ to be exposed to the tax, and low enough z that the resulting contraction is relatively strong.

4.2 Customer Investment Policies

I now consider policies that directly relate to the customer margin. In particular, I ask whether policymakers can improve upon the decentralized equilibrium by taxing or subsidizing customer investment.[30] I consider a proportional tax τ_θ levied on customer investment, such that the new cost function for firms becomes:

$$(1+\tau_\theta)wx_\theta(\theta,z,\lambda_\theta). \tag{27}$$

Similarly, the cost function for entrants becomes:

$$(1+\tau_\theta)wx_E(z,\lambda_E). \tag{28}$$

If $\tau_\theta < 0$, this operates as a subsidy. The government budget is balanced through a lump sum transfer T, equally distributed among all households.

I define the optimal policy to be the value τ_θ that maximizes consumption-equivalent welfare. As Table 8 shows, the resulting optimum is a 36% subsidy on customer investment. This policy dramatically increases aggregate productivity and consequently output and welfare. The rise in output is dampened (relative to the increase in TFP) by a reallocation of labor from production to customer investment, and the rise in welfare is dampened (relative to the increase in output) by an overall increase in labor supply.

To analyze this policy, it will be convenient to consider an alternate decomposition of TFP:

$$TFP = |\Omega| \times E[\theta z] \times \iota,$$

where ι is an accounting residual that (as we will see) plays a negligible role in the effect of this policy. We can then write:

$$\Delta \log(TFP) = \Delta \log(|\Omega|) + \Delta \log(E[\theta z]) + \Delta \log(\iota). \tag{29}$$

We can also further decompose the $E[\theta z]$ term as:

$$E[\theta z] = E[\theta]E[z] + Cov(\theta, z). \tag{30}$$

[30] One may wonder how practical it is to implement such policies. Implementation is not the primary concern of this exercise, but it does not, in any case, appear to be a major obstacle. Many countries have R&D subsidies or tax credits, which require firms to classify certain expenses as meeting the definition of R&D. In Canada, this definition is that "the work must be conducted for the advancement of scientific knowledge or for the purpose of achieving a technological advancement" and "the work must be a systematic investigation or search that is carried out in a field of science or technology by means of experiment or analysis" (Canada Revenue Agency, 2024). One can easily imagine a similar definition for activities that expand the number of customers that can purchase from a firm.

Table 8: Optimal Customer Investment Policy

τ_θ	-0.36
CEW	$+4.0\%$
ΔY	$+5.8\%$
ΔL_p	-6.1%
ΔTFP	$+11.0\%$
ΔL	$+1.9\%$

Notes: This table shows the aggregate effects of the policy τ_θ that maximizes CEW, the consumption-equivalent welfare effect of the policy. Y is aggregate output, L_p is total production labor, $TFP \equiv \frac{Y}{L_p^\alpha}$, and L is total labor.

Table 9: Optimal Customer Investment Policy (Decomposition)

Object	Percentage of Total Effect	
$\lvert\Omega\rvert$		-68.9%
ι		-5.7%
$E[\theta z]$		$+174.6\%$
$E[\theta]$	$+88.5\%$	
$E[z]$	$+43.1\%$	
$Cov(\theta, z)$	-31.6%	

Notes: This table decomposes the changes in $\Delta \log TFP$ and $\Delta \log E[\theta z]$ based on Equations 29 and 30 respectively. Total effects are decomposed into percentages using a Shapley decomposition. The exact linear decomposition of log TFP given in Equation 29 means that the Shapley contribution of object $X \in \{|\Omega|, E[\theta z], \iota\}$ is simply $\frac{\Delta \log(X)}{\Delta \log(TFP)} \times 100\%$. Equation 30 is nonlinear, so the more general Shapley method is used to compute the contribution.

Table 9 sheds further light on the effects of the policy through the lens of this decomposition. The positive effects of the policy are entirely due to an increase in $E[\theta z]$. This increase is driven primarily by $E[\theta]$, with $E[z]$ playing an important secondary role. This is intuitive: Subsidizing customer investment primarily improves welfare by increasing the average level of customer access. A strong amplifying effect is generated by the endogenous response of productivity, due to the complementarity between customer access and productivity. This is accompanied by a large decrease in the mass of operating firms $|\Omega|$, due to higher wages, and a modest reduction in allocative efficiency. Neither of these are strong enough, though, to overturn the positive effect.

Having clarified the mechanics of how the policy improves TFP, it is worth discussing the inefficiency involved. Why were $E[\theta]$ and $E[z]$ suboptimal to begin with? Two familiar forces are at play. The first is a "love-of-variety" externality, since customer investment expands the household choice set ω and thereby

provides them with more varieties to choose from. Firms do not consider this positive externality when making their investment decisions, and therefore may under-invest relative to the social planner. The second is a "business-stealing" externality. When a firm enters the choice set of an additional household, the other firms in that choice set will contract. The social planner considers the surplus that the new firm generates net of this reduction, whereas private firms consider the gross surplus. In the calibrated model, the love-of-variety effect dominates and aggregate investment in θ is too low.

A natural question to ask is why my results differ from recent macroeconomic studies of advertising, which typically have found that expenditure is excessive and should be taxed (Cavenaile & Roldan-Blanco, 2021; Cavenaile, Celik, Roldan-Blanco, & Tian, 2025; Hubmer & Nord, 2025). Advertising and customer investment are not the same, but it still may be surprising that my result is so different. There are several factors to consider. The first is that I model customer investment as expanding the choice set of consumers, whereas the studies I cite model it as a preference shifter. Consequently the love-of-variety externality, which drives my result, is absent in their models.

Secondly, these studies typically assume that advertising expenditures are entirely relative. A firm that spends \$1 when aggregate expenditure is \$100 will receive the same benefit as a firm that spends \$1000 when aggregate expenditure is \$100,000. Conversely, I model customer investment as capturing a real, technological cost. This does not mean that there are no congestion effects in my model. Just as higher demand for advertising services raises the price of ad space, or increased geographic expansion raises the price of real estate, higher customer investment raises the wage in my model and therefore increases the cost of investment. The difference lies in the fact that, aside from congestion effects, there is a real intrinsic cost to this investment. If there is no intrinsic cost to advertising, and it is entirely relative, then the conclusion that expenditure is excessive is hardly surprising.[31]

4.3 How Important is Customer Accumulation?

In my final exercise, I shift away from policy analysis to instead ask a positive question: How important is the customer margin for firm size dispersion? Recent empirical literature has argued forcefully for the importance of customers, but their analysis is silent on the endogenous relationship between customers and productivity. In this section, I use the rich structure of my model to provide an equilibrium answer to this question.

[31] Cavenaile, Celik, Roldan-Blanco, and Tian (2025) show in an extension that taxing advertising is still optimal in their model when there are intrinsic costs to advertising. The reason is that the primary driver of their result is not relative expenditure, but aggregate substitution between advertising and R&D. Although I abstract from aggregate growth, and a full discussion of this mechanism is beyond the scope of this paper, it is worth noting that it does not necessarily hold in my setting. As described in Appendix C.1, using Kimball demand (as opposed to CES, which Cavenaile, Celik, Roldan-Blanco, and Tian (2025) assume) greatly strengthens the complementarity between productivity and customer access. Indeed, as Table 9 shows, almost half of the positive effect of the customer investment subsidy comes from its indirect *positive* effect on innovation.

I do this by simulating a simple counterfactual. Suppose that all firms had full access, i.e., $\theta_j = 1$ for all j, as is implicitly assumed in productivity-only models. How would the economy change? Specifically, how much dispersion in firm size would remain in this counterfactual economy?

Table 10 shows the results. Overall size dispersion in the economy, measured by the variance of log revenue, declines by 63.3%. Clearly, however, this is driven not just by the direct effect of shutting down variation in customer access, but also the indirect effect it has on the endogenous distribution of productivity. Recall that the marginal benefit of investing in productivity is a function of the firm's current customer level. Therefore less dispersion in customer access implies less dispersion in incentives to innovation and therefore less dispersion in productivity in equilibrium. This effect is quantitatively strong, bringing down the variance of log productivity by an order of magnitude.

Table 10: Counterfactual: Full Access

	Baseline	Full Access
Variance of log revenue	1.8142	0.6655
Variance of log customer access	2.7801	0.0000
Variance of log productivity	0.3157	0.0349

Notes: This table presents measures of firm size dispersion in the baseline calibrated model and the counterfactual equilibrium where all firms can sell to all customers (i.e., $\theta_j = 1$ for all j).

To further unpack this result, I conduct the following decomposition. First, note that we can mechanically write $\log(pq) = \log(\frac{pq}{\theta}) + \log(\theta)$, i.e., log revenue is the sum of log revenue per customer and log customers. Secondly, revenue per customer is fundamentally a function of the firm's state (θ, z) and the aggregate state of the economy $\mathcal{A} \equiv (w, C, D)$. Therefore we can write $\log(\frac{pq}{\theta}) = f(\theta, z; \mathcal{A})$.

It follows that

$$Var(\log(pq)) = Var(f(\theta, z; \mathcal{A})) + Var(\log(\theta)) + 2Cov(f(\theta, z; \mathcal{A}), \log(\theta)). \tag{31}$$

With this equation in hand, I can break down the Full Access experiment into three separate effects. First, I set the second and third terms to zero, i.e., $Var(\log(\theta)) = Cov(f(\theta, z; \mathcal{A}), \log(\theta)) = 0$, but keep the first term unchanged. This captures the direct "Accounting Effect" of θ becoming a constant, but ignores how sales per customer will respond to this change. I call this an "accounting" effect because it does not require a model — given firm-level data on customers, it could be estimated empirically, as in Einav, Klenow, Levin, and Murciano-Goroff (2021) and Afrouzi, Drenik, and Kim (2023). Second, I update the first term by allowing

the joint distribution of (θ, z) to adjust. This "Distribution Effect" includes θ becoming 1 for all firms and z endogenously responding through investment, entry, and exit. Third, I further update the first term by allowing the aggregates $\mathcal{A}$ to adjust, capturing the "General Equilibrium (GE) Effect."

Table 11: Full Access (Decomposition)

Effect	ΔVar(log Revenue)
1) Accounting Effect	-83.65%
2) Distribution Effect	-16.34%
3) GE Effect	$+36.67\%$
Full Effect	-63.32%

Notes: This table decomposes the total change in variance of log revenue into three intermediate steps. The steps are implemented in sequence, such that the sum of the three intermediate effects is equal to the full effect.

Table 11 shows the results of the decomposition. First, it is worth noting that the so-called "Accounting Effect" is in line with empirical studies such as Einav, Klenow, Levin, and Murciano-Goroff (2021) and Afrouzi, Drenik, and Kim (2023), an encouraging validation of the model.[32] This effect is driven by the fact that productivity dispersion is much lower than customer access dispersion, as shown in Table 10 and previously illustrated in Figure 6.

The Distribution Effect is large and negative, as would be expected. Revenue per customer is a function of both θ and z, so eliminating the variance of the former entirely and reducing the latter by an order of magnitude (see Table 10) unsurprisingly has a large effect on the variance of log revenue, over and above the Accounting Effect. The GE effect is perhaps more interesting, generating a large *increase* in revenue dispersion.

To understand the GE effect, let us consider more closely the mapping between productivity z and revenue per customer. First, consider again the benchmark of CES demand and constant returns to scale. In this model, these assumptions imply that revenue per customer is log linear in $\log z$. Specifically, we have:

$$f(\theta, z; \mathcal{A}) = (\sigma - 1)\log(z) + f_{CES}(\mathcal{A}). \tag{32}$$

The explanation is straightforward: Increasing $\log z$ by a fixed increment always leads to a fixed decrease in log marginal cost (due to constant returns to scale), which leads to a fixed decrease in log prices (due

[32] Afrouzi, Drenik, and Kim (2023) report a variance of log sales per customer that is 11.4% of the total variance of log revenue. This corresponds to an "Accounting Effect" of 88.6%, which is close to the 83.7% I find. Einav, Klenow, Levin, and Murciano-Goroff (2021) conduct a slightly different exercise, which although not directly comparable attributes around 80% of variation to log customers.

to constant markups), which leads to a fixed increase in log quantities (due to constant elasticity). The log linearity implies that the variance of log revenue in the $\theta = 1$ case would simply be given by $(\sigma - 1)^2 Var(\log(z))$. The GE effect is zero, and an order of magnitude decrease in the variance of $\log z$ would necessarily result in an order of magnitude decrease in the variance of log revenue.

In my case, decreasing returns to scale ($\alpha < 1$) and declining elasticity ($\epsilon > 0$) introduce diminishing returns to productivity in the revenue function. As a result, dispersion in revenue is not only determined by dispersion in productivity, but also the shape of the revenue function in equilibrium. This shape depends on aggregate objects because wages influence the strength of decreasing returns to scale, and aggregate consumption influences the strength of declining elasticity.[33] As a result, the GE effect is nonzero.

Figure 7 shows the GE effect in action. The figure plots revenue as a function of z, fixing $\theta = 1$, with the baseline and Full Access aggregates. As the plot shows, the Full Access experiment constitutes a positive aggregate shock, increasing revenue for all firms. Diminishing returns to productivity only begin to set in at extremely high levels of z, which constitute very few firms. For the vast majority of firms, the revenue function is approximately linear, amplifying the underlying dispersion in productivity relative to the baseline economy.

When all three effects are taken together, we find that the overall share of variance attributed to the customer margin is smaller than the empirical "Accounting Effect" would suggest. The fact that firms cannot, in general, sell to all firms amplifies firm size dispersion through its endogenous effect on productivity, but this is more than overturned by its aggregate effect, which dampens firm size dispersion.

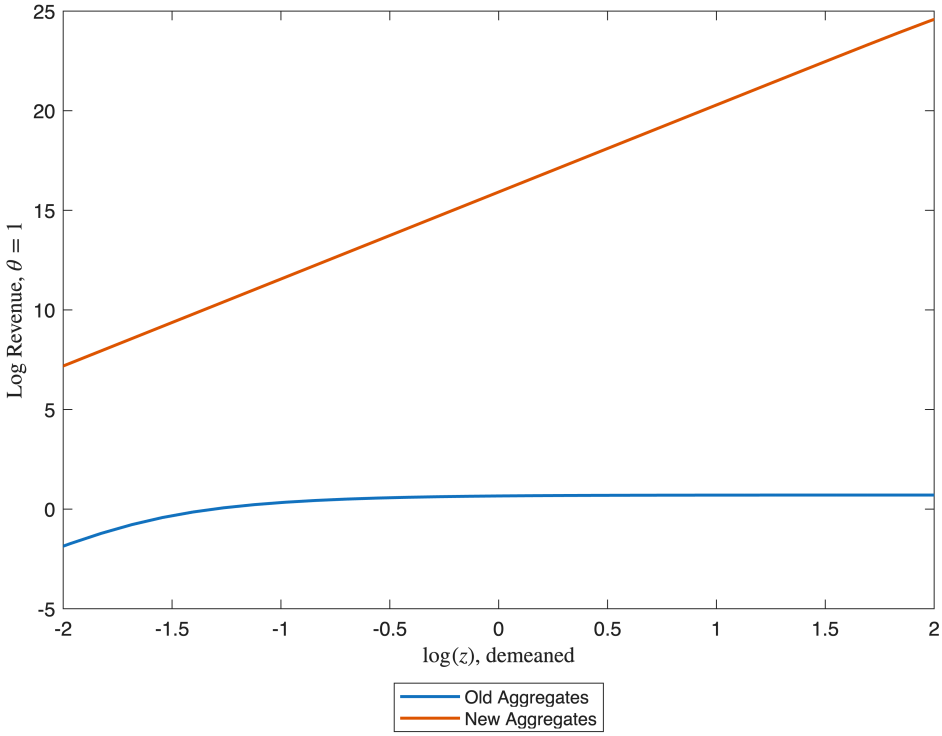
Overall, this exercise reveals insights about the interaction between customer access and firm size dispersion that are missed by a purely empirical approach. Customer access is important not only because its variance is high relative to the variance of sales, but because it interacts strongly with incentives to innovate and the mapping from productivity to revenue.

5 Conclusion

For decades, economists have been fascinated by the vast dispersion observed in firm-level data. Fields such as international trade, business cycles, and endogenous growth have all been greatly influenced by heterogeneous firm models. In this paper, I argue that there is substantial mismatch between two drivers of firm heterogeneity: productivity and customer accumulation. I show that this mismatch is important for understanding firm-level data and has surprising policy implications.

[33] Recall that with Kimball demand, elasticity faced by firm j is a function of *relative* quantity $\frac{c_j}{C}$. The higher aggregate consumption C is, the higher the elasticity will be at a given quantity. This in turn means that the revenue function will be steeper at every given level of productivity.

Figure 7: Revenue and Productivity (Baseline vs Full Access)



Notes: This figure shows log revenue as a function of log productivity z for a firm with $\theta = 1$, i.e., one that can sell to all customers. The blue line corresponds to this revenue function in the baseline equilibrium, whereas the red line corresponds to the revenue function in the Full Access equilibrium, i.e., the equilibrium where all firms have $\theta = 1$.

To do so, I construct a dynamic macroeconomic model with heterogeneity in productivity and customer access. I apply a novel method to calibrate the model to rich administrative data covering the universe of Canadian firms, and use the calibrated model to make four main contributions. First, I show that the economy features substantial mismatch, and that this channel allows the model to jointly match several elusive empirical patterns. Second, I show that this mismatch entirely overturns conventional wisdom regarding taxing large firms. In fact, I find significant welfare gains from a 10% tax on revenue above $20 million. The tax is beneficial because it shifts production away from unproductive large firms, which do not exist in models that equate size with productivity. Third, I show that a subsidy on customer investment can significantly increase welfare, drawing attention to a neglected policy lever. Fourth, I find that a counterfactual economy where all firms could sell to all customers would have 63% lower variance of log revenue. This exercise confirms a critical role for customer accumulation in driving firm size dispersion, and emphasizes the endogenous interaction between customer access and firm-level productivity.

My results suggest several promising channels for future work. Although the assumption of atomistic firms greatly aids tractability, it would be interesting to explore the effects of customer accumulation in a setting where large "superstar" firms have the ability to influence aggregate conditions. How does customer accumulation interact with strategic considerations for oligopolistic firms? Interestingly, some of these large firms (e.g., Meta, Google), directly facilitate customer accumulation for smaller firms through digital advertising. How does this influence — and potentially distort — equilibrium outcomes?

My results also raise interesting questions about the interaction between customer accumulation and aggregate growth. How should a policymaker with limited resources divide expenditure between R&D and customer investment subsidies? Does the low customer access of new firms dampen creative destruction, and has this changed with the rise of digital advertising? These and other questions present fruitful opportunities for further research.

References

Afrouzi, H., Drenik, A., & Kim, R. (2023). Concentration, Market Power, and Misallocation: The Role of Endogenous Customer Acquisition [Working Paper]. (Cit. on pp. 3, 6, 7, 13, 16, 18, 32, 37, 38, 49).

Aghion, P., Bergeaud, A., Boppart, T., Klenow, P. J., & Li, H. (2025). Good Rents Versus Bad Rents: R&D Misallocation and Growth [Working Paper 34190, National Bureau of Economic Research]. (Cit. on p. 27).

Aghion, P., & Howitt, P. (1992). A Model of Growth Through Creative Destruction. *Econometrica*, *60*(2), 323–351 (cit. on p. 4).

Akcigit, U., Ates, S. T., & Impullitti, G. (2024). Innovation and Trade Policy in a Globalized World [Working Paper] (cit. on pp. 14, 17).

Akcigit, U., & Kerr, W. R. (2018). Growth Through Heterogeneous Innovations. *Journal of Political Economy*, *126*(4), 1374–1443 (cit. on pp. 13, 21).

Argente, D., Fitzgerald, D., Moreira, S., & Priolo, A. (2025). How Do Entrants Build Market Share? The Role of Demand Frictions. *American Economic Review: Insights*, *7*(2), 160–176 (cit. on pp. 4, 7, 13).

Argente, D., & Mendez, E. (2025). Cross-Border Product Adoption: Individual Imports, Migrant Networks, and Domestic Retailers [Working Paper] (cit. on p. 21).

Arkolakis, C. (2010). Market Penetration Costs and the New Consumers Margin in International Trade. *Journal of Political Economy*, *118*(6), 1151–1199 (cit. on pp. 2, 3, 6, 7, 13, 23).

Arkolakis, C. (2016). A Unified Theory of Firm Selection and Growth. *The Quarterly Journal of Economics*, *131*(1), 89–156 (cit. on p. 7).

Atkeson, A., & Burstein, A. (2008). Pricing-to-Market, Trade Costs, and International Relative Prices. *American Economic Review*, *98*(5), 1998–2031 (cit. on p. 6).

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & Van Reenen, J. (2017). The Fall of the Labor Share and the Rise of Superstar Firms [Working Paper 23396, National Bureau of Economic Research]. (Cit. on pp. 5, 25).

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & Van Reenen, J. (2020). The Fall of the Labor Share and the Rise of Superstar Firms. *The Quarterly Journal of Economics*, *135*(2), 645–709 (cit. on p. 5).

Bartelsman, E., Haltiwanger, J., & Scarpetta, S. (2013). Cross-Country Differences in Productivity: The Role of Allocation and Selection. *American Economic Review*, *103*(1), 305–334 (cit. on pp. 5, 25).

Berlingieri, G., Jaimovich, N., Terry, S. J., & Vincent, N. (2025). The Empirical Distribution of Firm Dynamics and Its Macro Implications [Working Paper]. (Cit. on p. 6).

Bernard, A. B., Dhyne, E., Magerman, G., Manova, K., & Moxnes, A. (2022). The Origins of Firm Heterogeneity: A Production Network Approach. *Journal of Political Economy*, *130*(7), 1765–1804 (cit. on pp. 5, 7, 8, 26).

Blum, B. S., Claro, S., Horstmann, I., & Rivers, D. A. (2024). The ABCs of Firm Heterogeneity When Firms Sort Into Markets: The Case of Exporters. *Journal of Political Economy*, *132*(4), 1162–1208 (cit. on p. 27).

Boar, C., & Midrigan, V. (2024). Markups and Inequality. *The Review of Economic Studies*, *92*(5), 2828–2860 (cit. on p. 16).

Bond, S., Hashemi, A., Kaplan, G., & Zoch, P. (2021). Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation From Production Data. *Journal of Monetary Economics*, *121*, 1–14 (cit. on p. 19).

Bornstein, G. (2025). Entry and Profits in an Aging Economy: The Role of Consumer Inertia [Working Paper] (cit. on p. 7).

Buera, F. J., Kaboski, J. P., & Shin, Y. (2011). Finance and Development: A Tale of Two Sectors. *American Economic Review*, *101*(5), 1964–2002 (cit. on p. 6).

Canada Revenue Agency. (2024). What Work is Eligible – Scientific Research and Experimental Development (SR&ED) Tax Incentives [Last Modified: 2024-04-10]. Retrieved September 24, 2025, from https://www.canada.ca/en/revenue-agency/services/scientific-research-experimental-development-tax-incentive-program/sred-eligibility.html (cit. on p. 34).

Cavenaile, L., Celik, M. A., Perla, J., & Roldan-Blanco, P. (2023). A Theory of Dynamic Product Awareness and Targeted Advertising [Working Paper]. (Cit. on pp. 7, 12).

Cavenaile, L., Celik, M. A., Roldan-Blanco, P., & Tian, X. (2025). Style Over Substance? Advertising, Innovation, and Endogenous Market Structure. *Journal of Monetary Economics*, *149*, 103683 (cit. on p. 36).

Cavenaile, L., & Roldan-Blanco, P. (2021). Advertising, Innovation, and Economic Growth. *American Economic Journal: Macroeconomics*, *13*(3), 251–303 (cit. on p. 36).

Chang, S.-J., & Wu, B. (2014). Institutional Barriers and Industry Dynamics. *Strategic Management Journal*, *35*(8), 1103–1123 (cit. on p. 4).

Dinlersoz, E. M., & Yorukoglu, M. (2012). Information and Industry Dynamics. *American Economic Review*, *102*(2), 884–913 (cit. on pp. 2, 7).

Dixit, A. K., & Stiglitz, J. E. (1977). Monopolistic Competition and Optimum Product Diversity. *The American Economic Review*, *67*(3), 297–308 (cit. on p. 3).

Drozd, L. A., & Nosal, J. B. (2012). Understanding International Prices: Customers as Capital. *American Economic Review*, *102*(1), 364–395 (cit. on p. 7).

Dyrda, S., Hong, G., & Steinberg, J. B. (2024). A Macroeconomic Perspective on Taxing Multinational Enterprises. *Journal of International Economics*, *152*, 104022 (cit. on p. 30).

Edmond, C., Midrigan, V., & Xu, D. Y. (2023). How Costly Are Markups? *Journal of Political Economy*, *131*(7), 1619–1675 (cit. on pp. 2, 3, 6, 7, 16, 30, 31, 49).

Einav, L., Klenow, P. J., Levin, J. D., & Murciano-Goroff, R. (2021). Customers and Retail Growth [Working Paper]. (Cit. on pp. 6, 8, 37, 38).

Eslava, M., Haltiwanger, J., Kugler, A., & Kugler, M. (2004). The Effects of Structural Reforms on Productivity and Profitability Enhancing Reallocation: Evidence From Colombia. *Journal of Development Economics*, *75*(2), 333–371 (cit. on p. 4).

Eslava, M., Haltiwanger, J., & Urdaneta, N. (2024). The Size and Life-Cycle Growth of Plants: The Role of Productivity, Demand, and Wedges. *The Review of Economic Studies*, *91*(1), 259–300 (cit. on p. 8).

Foster, L., Haltiwanger, J., & Syverson, C. (2008). Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? *American Economic Review*, *98*(1), 394–425 (cit. on pp. 4, 8, 27).

Foster, L., Haltiwanger, J., & Syverson, C. (2016). The Slow Growth of New Plants: Learning About Demand? *Economica*, *83*(329), 91–129 (cit. on pp. 4, 7, 8, 13, 27).

Gouin-Bonenfant, E. (2022). Productivity Dispersion, Between-Firm Competition, and the Labor Share. *Econometrica*, *90*(6), 2755–2793 (cit. on pp. 5, 25).

Gourieroux, C., Monfort, A., & Renault, E. (1993). Indirect Inference. *Journal of Applied Econometrics*, *8*(S1), S85–S118 (cit. on pp. 4, 19).

Gourio, F., & Rudanko, L. (2014). Customer Capital. *The Review of Economic Studies*, *81*(3), 1102–1136 (cit. on pp. 2, 7).

Guvenen, F., & Smith, A. A. (2014). Inferring Labor Income Risk and Partial Insurance From Economic Choices. *Econometrica*, *82*(6), 2085–2129 (cit. on p. 19).

Hopenhayn, H., & Rogerson, R. (1993). Job Turnover and Policy Evaluation: A General Equilibrium Analysis. *Journal of Political Economy*, *101*(5), 915–938 (cit. on p. 2).

Hopenhayn, H. A. (1992). Entry, Exit, and Firm Dynamics in Long Run Equilibrium. *Econometrica*, *60*(5), 1127–1150 (cit. on pp. 2, 6, 23).

Hottman, C. J., Redding, S. J., & Weinstein, D. E. (2016). Quantifying the Sources of Firm Heterogeneity. *The Quarterly Journal of Economics*, *131*(3), 1291–1364 (cit. on p. 8).

Hsieh, C.-T., & Klenow, P. J. (2009). Misallocation and Manufacturing TFP in China and India. *The Quarterly Journal of Economics*, *124*(4), 1403–1448 (cit. on p. 2).

Hubmer, J., & Nord, L. (2025). Investment in Demand and Dynamic Competition for Customers [Working Paper]. (Cit. on pp. 7, 21, 36).

Ignaszak, M., & Sedlacek, P. (2023). Customer Acquisition, Business Dynamism and Aggregate Growth [Working Paper]. (Cit. on pp. 4, 7, 19, 21, 27, 53).

Jovanovic, B. (1982). Selection and the Evolution of Industry. *Econometrica*, *50*(3), 649–670 (cit. on p. 2).

Kimball, M. S. (1995). The Quantitative Analytics of the Basic Neomonetarist Model. *Journal of Money, Credit and Banking*, *27*(4), 1241–1277 (cit. on pp. 3, 9).

Klenow, P. J., & Willis, J. L. (2016). Real Rigidities and Nominal Price Changes. *Economica*, *83*(331), 443–472 (cit. on pp. 16, 20).

Krugman, P. (1980). Scale Economies, Product Differentiation, and the Pattern of Trade. *American Economic Review*, *70*(5), 950–959 (cit. on p. 3).

Lentz, R., & Mortensen, D. T. (2008). An Empirical Model of Growth Through Product Innovation. *Econometrica*, *76*(6), 1317–1373 (cit. on p. 19).

Lim, K., Trefler, D., & Yu, M. (2022). Looking Backward, Innovating Forward: A Theory of Competitive Cascades [Working Paper]. (Cit. on p. 14).

Lucas, R. E. (1978). On the Size Distribution of Business Firms. *The Bell Journal of Economics*, *9*(2), 508–523 (cit. on pp. 2, 6).

Luttmer, E. G. J. (2007). Selection, Growth, and the Size Distribution of Firms. *The Quarterly Journal of Economics*, *122*(3), 1103–1144 (cit. on p. 6).

Matsuyama, K. (2023). Non-CES Aggregators: A Guided Tour. *Annual Review of Economics*, *15*, 235–265 (cit. on p. 3).

Melitz, M. J. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. *Econometrica*, *71*(6), 1695–1725 (cit. on pp. 2, 6).

Moll, B. (2014). Productivity Losses From Financial Frictions: Can Self-Financing Undo Capital Misallocation? *American Economic Review*, *104*(10), 3186–3221 (cit. on pp. 2, 6).

Nakamura, E., & Steinsson, J. (2011). Price Setting in Forward-Looking Customer Markets. *Journal of Monetary Economics*, *58*(3), 220–233 (cit. on p. 7).

Phelps, E. S., & Winter, S. G. J. (1970). Optimal Price Policy Under Atomistic Competition. In *Microeconomic Foundations of Employment and Inflation Theory* (pp. 309–337). Norton. (Cit. on p. 7).

Pozzi, A., & Schivardi, F. (2016). Demand or Productivity: What Determines Firm Growth? *The RAND Journal of Economics*, *47*(3), 608–630 (cit. on p. 8).

Ravn, M., Schmitt-Grohe, S., & Uribe, M. (2006). Deep Habits. *The Review of Economic Studies*, *73*(1), 195–218 (cit. on p. 7).

Restuccia, D., & Rogerson, R. (2008). Policy Distortions and Aggregate Productivity With Heterogeneous Establishments. *Review of Economic Dynamics*, *11*(4), 707–720 (cit. on p. 2).

Roberts, M. J., Yi Xu, D., Fan, X., & Zhang, S. (2018). The Role of Firm Factors in Demand, Cost, and Export Market Selection for Chinese Footwear Producers. *The Review of Economic Studies*, *85*(4), 2429–2461 (cit. on p. 27).

Roldan-Blanco, P., & Gilbukh, S. (2021). Firm Dynamics and Pricing Under Customer Capital Accumulation. *Journal of Monetary Economics*, *118*, 99–119 (cit. on p. 7).

Romer, P. M. (1987). Growth Based on Increasing Returns Due to Specialization. *The American Economic Review*, *77*(2), 56–62 (cit. on p. 3).

Sedlacek, P., & Sterk, V. (2017). The Growth Potential of Startups Over the Business Cycle. *American Economic Review*, *107*(10), 3182–3210 (cit. on pp. 7, 12).

Smith Jr., A. A. (1993). Estimating Nonlinear Time-Series Models Using Simulated Vector Autoregressions. *Journal of Applied Econometrics*, *8*(S1), S63–S84 (cit. on pp. 4, 19).

Steinberg, J. B. (2023). Export Market Penetration Dynamics. *Journal of International Economics*, *145*, 103807 (cit. on pp. 7, 13).

Yao, J. (2022). Knowledge is (Market) Power [Working Paper]. (Cit. on p. 21).

Appendix for "Customer Accumulation, Firm Size, and Aggregate Productivity"

Appendix A Theory

A.1 Household Choice Sets

In this section I provide additional details and intuition about the construction of household choice sets.

Suppose we were given a set of values for customer access $\{\theta_j\}_{j\in\Omega}$. In principle, there are infinitely many ways we could map firms j to households i that would be consistent with this set. Let us consider a simple example where there are two types of firms: type 1 firms have $\theta = 0.2, z = 1$ and type 2 firms have $\theta = 0.4, z = 2$. There are an infinite number of atomistic firms, with the mass of each type denoted by ρ_1 and ρ_2 respectively.

One way we could map firms to customers, consistent with these values of θ, would be the following: Every household with $i \in [0, 0.2]$ can buy from all firms of the first type, every household with $i \in (0.2, 0.6]$ can buy from all firms of the second type, and every household with $i \in (0.6, 1]$ cannot buy from any firm.

However, we could also say that households $i \in [0, 0.2]$ can buy from *all* firms (regardless of type), households $i \in (0.2, 0.4]$ can buy from all type 2 firms (but none of type 1), and customers with $i \in (0.4, 1]$ cannot buy from any firm.

These two mappings are very different because they imply different choice sets and ultimately different household behavior. In the first mapping, households with $i \in [0, 0.2]$ can only buy from the lower productivity type 1 firms. In the second mapping, these same households can now buy from both types of firms. The aggregate price in their choice set will be different and therefore their demand function will be different.

Mappings such as these would be highly intractable, because solving the problem of a particular firm j would require knowing the entire distribution of prices that *each* household they can sell to faces. For a continuum of households, this is an infinite dimensional problem.

The mapping I use in the main text is symmetric, in the sense that it ensures each household i will have the same distribution of firm types within its choice set. Since each firm of the same type will — in equilibrium — choose the same price, this ensures that each household i faces the same aggregate price level $P_i = P$. Consequently, each household will have the same demand function, and it suffices to know the *mass* of households that a firm can sell to in order to solve its problem.

Applied to the example above, the mapping I use in the main text would be as follows. *Every* household $i \in [0, 1]$ can purchase from a fraction 0.2 of all type 1 firms and a fraction 0.4 of all type 2 firms. This

implies that the mass of customers for each type of firm would be $\int_0^1 \theta \, di = \theta$, consistent with the definition of θ.

Finally, note that a key assumption throughout this discussion has been that there are infinitely many firms of each type. If there were a finite number of firms, this would not hold, especially if the number of firms is small. For example, if there is only one type 1 firm and two type 2 firms, it would be impossible to construct choice sets that satisfy $\theta_1 = 0.2$ and $\theta_2 = 0.4$. As such, the assumption made in the main text that there are an infinite number of atomistic firms is crucial for the symmetry of household choice sets.

A.2 Allocative Efficiency

The main text presented a decomposition of TFP as:

$$TFP = |\Omega| \times \overline{Z} \times Q, \tag{33}$$

where $TFP \equiv \frac{Y}{L_p^\alpha}$ and $Q \leq 1$ was said to represent allocative efficiency. In this subsection, I formally define the notion of efficiency that I use to construct this decomposition. To do so, I define a *static* social planner's problem in the following vein. Taking as given the mass of operating firms $|\Omega|$, the distribution of those firms over (θ, z), and a certain quantity of production labor L_p, the social planner decides how much each firm should produce. Specifically, the planner is choosing how to allocate the production labor L_p across firms. Hence, the difference between output in the decentralized equilibrium and output in this static social planner's problem entirely reflects cross-sectional misallocation.

Formally, this problem can be represented as follows:

$$\max_{q(\theta,z)} \frac{C^{1-\gamma}}{1-\gamma} - \xi \frac{L^{1+\nu}}{1+\nu} \tag{34}$$

s.t.

$$|\Omega| \sum_\theta \sum_z \Upsilon\left(\frac{q(\theta,z)}{\theta C}\right) \theta \rho(\theta, z) = 1$$

$$|\Omega| \sum_\theta \sum_z \left(\frac{q(\theta,z)}{z}\right)^{1/\alpha} \rho(\theta, z) = L_p$$

Let Y_{SP} denote the output arising from the solution to this problem. Recall that $Y = C$ in this model, so that this corresponds to the optimal consumption chosen by the planner. We can then define $\overline{Z}$ as:

$$\overline{Z} \equiv \frac{Y_{SP}}{|\Omega| L_p^{\alpha}}, \tag{35}$$

so that TFP under the efficient allocation is:

$$TFP_{SP} = |\Omega| \overline{Z} \tag{36}$$

Defining $Q \equiv \frac{TFP}{TFP_{SP}}$, the decomposition in the main text follows by construction. Moreover, we can see that the interpretation of Q as capturing allocative efficiency is justified: $Q = 1$ if and only if a production allocation coincides with that of the social planner, and $Q < 1$ whenever there is misallocation. Moreover, $\frac{1}{Q} - 1$ captures precisely the output loss from misallocation. I find this quantity to be 6.4% in the full model and 2.7% in the productivity-only model, in line with previous literature (Afrouzi, Drenik, & Kim, 2023; Edmond, Midrigan, & Xu, 2023).

Appendix B Computation

B.1 Grids

As discussed in Section 2, I model the evolution of θ and z using a two-dimensional "quality ladder" setup. I use 55 grid points for each ladder, implying $55^2 = 3025$ total firm states. Grids are evenly spaced in logarithms.

Denote the upper and lower bounds of the ladders as $\{\overline{\theta}, \underline{\theta}, \overline{z}, \underline{z}\}$. I set $\overline{\theta} = 1$, which comes naturally from the definition of θ as the probability that a given household can purchase from a given firm.

The three remaining values are computational impositions that do not exist in the real world, and therefore must be chosen carefully. There are several considerations. Firstly, $\underline{\theta}$ should be low enough that it reasonably approximates the level of customer access that a new firm would have. Secondly, $\underline{z}$ and $\overline{z}$ should be small and large enough respectively that they are not binding in the calibrated model or any of the counterfactuals. A helpful feature of the Kimball setup in this regard is that, as discussed in Section 4.3, there are decreasing returns to productivity in the revenue function. As such, there will always be a large enough z such that firms invest a minuscule amount in innovation and endogenously stop growing. Similarly, regardless of θ, there will always be a small enough z that revenue becomes minuscule and firms exit.

With these in mind, I set $\underline{\theta} = 5 \times 10^{-5}$, $\underline{z} = 5 \times 10^{-2}$, and $\overline{z} = 100$. The value of $\underline{z}$ is low enough that even firms with $\theta = 1$ will exit. Similarly, the value of $\overline{z}$ is large enough that firms significantly reduce investment well before reaching it. In the calibrated model, the fraction of operating firms with $z = \overline{z}$, across all levels

of θ, is approximately 6×10^{-13}.

One way to assess $\underline{\theta}$ is to convert it to a number of customers. The population of Canada in the time period covered by my data ranged from 31 million (in 2001) to 37 million (in 2019). If we take 35 million as a rough average, then $\theta = \underline{\theta}$ corresponds to 1750 individuals. This seems to be a reasonable approximation of what a new firm could reach without significant advertising. Of course, it is difficult to directly compare model and data in this way, particularly when we consider intermediate good producers for whom the "consumer" is actually other firms. As such, a more direct way to justify this choice is simply that in the calibrated model, even firms with $z = \overline{z}$ will not operate at $\underline{\theta}$. This means that the only entrants that operate are those that successfully increase θ through their initial customer investment. This at least reassures us that the lower bound for θ is not binding.

With these bounds and 55 grid points, the increment between consecutive points on the ladder is (in logarithms) 0.1408 for z and 0.1834 for θ. These grids are dense enough to allow the model to generate realistic growth rates in revenue and sales-cost ratio (SCR), as shown in Table A1

Table A1: Mean Growth Rates

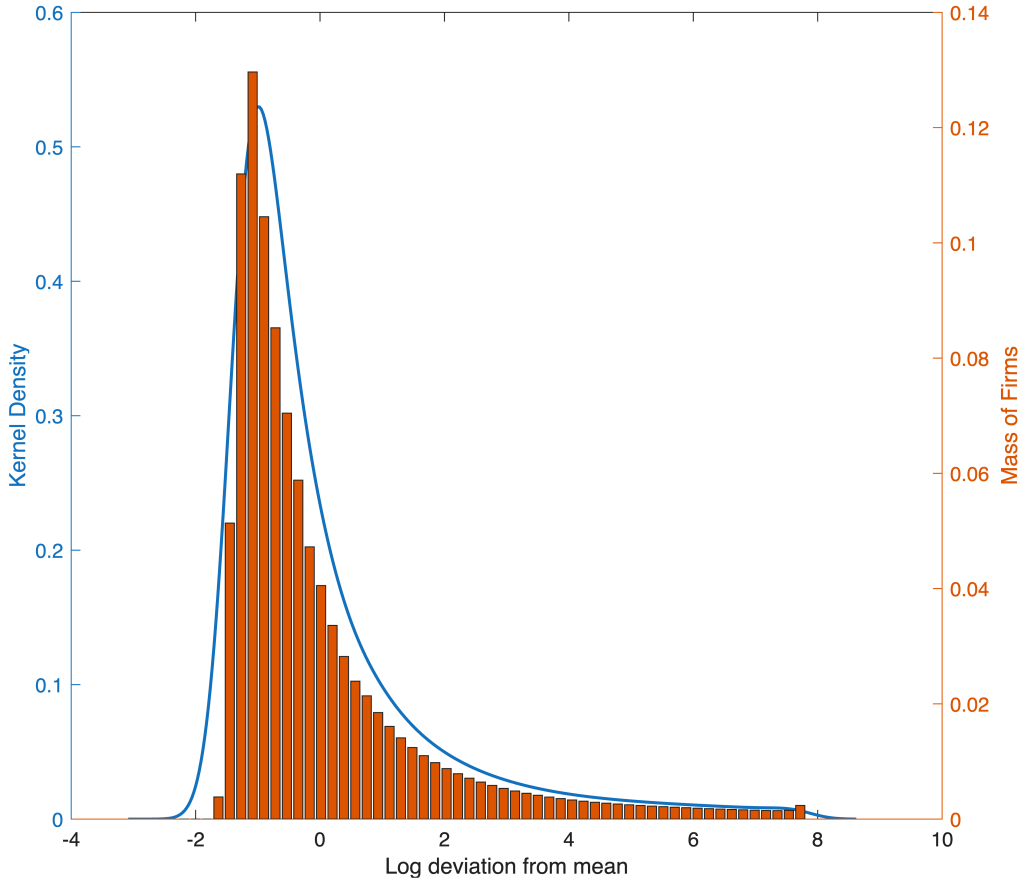
Moment	Model	Data
$E[\Delta \log Revenue \mid \Delta \log Revenue > 0.01]$	0.1778	0.2546
$E[\Delta \log Revenue \mid \Delta \log Revenue < -0.01]$	−0.2068	−0.2979
$E[\Delta \log SCR \mid \Delta \log SCR > 0.01]$	0.0603	0.1390
$E[\Delta \log SCR \mid \Delta \log SCR < -0.01]$	−0.1097	−0.1393

Notes: This table shows mean conditional growth rates of log revenue and log SCR in the model and data. SCR is defined as total revenue divided by total expenses. Both log revenue and log SCR are residualized with 4 digit NAICS-year fixed effects.

B.2 Continuous Illustrations

Although the firm state space in this model is clearly discrete, I present multiple continuous figures in the main text for illustrative ease. Specifically, to compare the respective marginal distributions of θ and z in Figure 6, I use continuous Kernel densities, and to depict the stationary distribution in Figure 4, I smooth and interpolate the underlying matrix. In this section, I present the continuous figures alongside their raw, discrete counterparts to verify that the continuous depictions are accurate.

Figure A1: Marginal Distribution ($\log \theta$)



Notes: This graph compares the continuous Kernel density for $\log \theta$ depicted in Figure 6 to the raw underlying histogram.

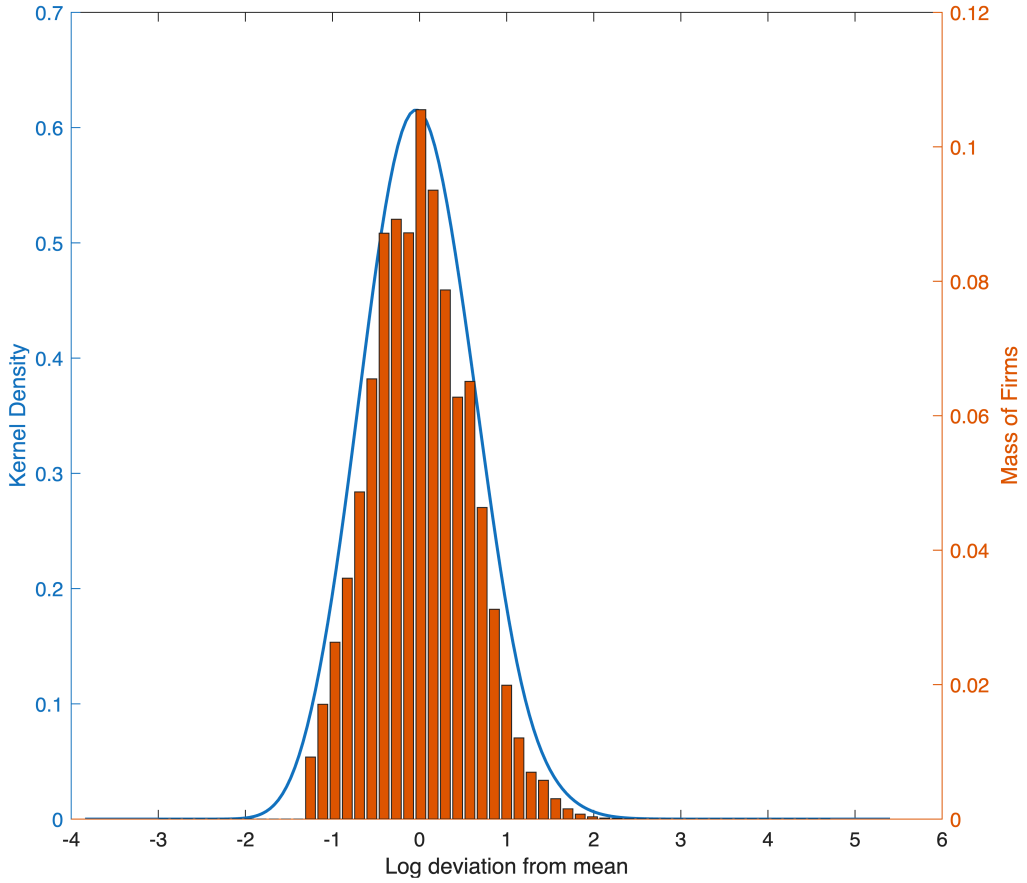
Appendix C Quantitative Analysis

C.1 Productivity-Customer Complementarity

In this section, I provide further detail on the particularly strong complementarity between productivity and customer access in my model. Figure A4 shows demeaned log revenue as a function of log demeaned customer access (θ) and productivity (z). It is instructive to compare this to the widely used benchmark of CES demand and constant returns to scale (CRS) production. In my model, it is easy to show that these assumptions (which correspond to $\epsilon \to 0$ and $\alpha = 1$) lead to a log-linear revenue function in $\log \theta$ and $\log z$. In other words, the contours would be diagonal lines. The curvature shown in Figure A4 indicates much stronger complementarity between productivity and customer access.

The source of this greater complementarity is the fact that Kimball demand features a "choke" quantity

Figure A2: Marginal Distribution ($\log z$)

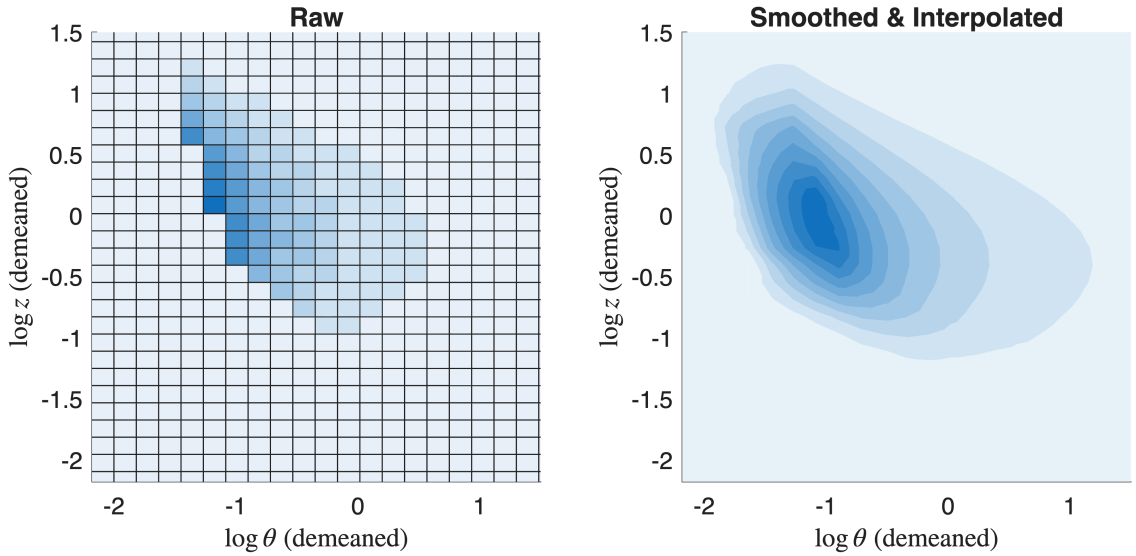


Notes: This graph compares the continuous Kernel density for $\log z$ depicted in Figure 6 to the raw underlying histogram.

above which marginal revenue becomes negative. Specifically, with the functional form I consider, the firm will never produce more than $q^C \equiv \sigma^{\frac{\sigma}{\epsilon}} C$ per customer. This implies that there is a limit to which firms can grow exclusively through productivity. If a firm's customer access θ is fixed, its revenue will be bounded above by $p(q^C)q^C\theta$, even if $z \to \infty$. On the other hand, marginal revenue with CES demand is always positive, such that any quantity can be profitably sold to a single customer given high enough z. As a result, revenue diverges to infinity as $z \to \infty$.

The Kimball "choke quantity" has intuitive appeal. If a company only ever has access to 100 customers, we would expect there to be — in general — a limit to how much they could grow through productivity alone. Abstracting from this limit, with CES demand, would likely under-estimate the quantitative importance of customer access. Indeed, note that in the CES-CRS case, θ and z are isomorphic in the sense that two firms

Figure A3: Joint Distribution



Notes: This graph compares the continuous joint distribution over $\log\theta$ and $\log z$ depicted in Figure 4 to the raw underlying matrix. 10 shades of color are used, with darker shades corresponding to a higher density of firms.

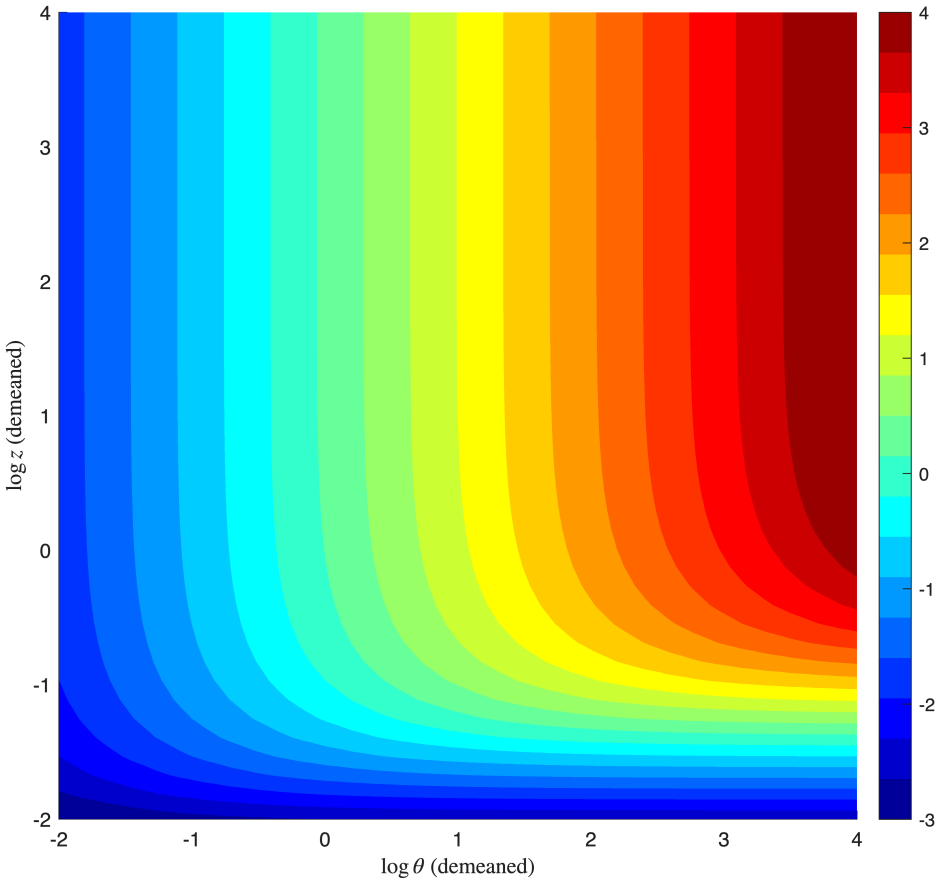
with the same value of $z^{\sigma-1}\theta$ will have the same revenue, regardless of the separate values of θ and z.[34] As such, little information is lost, at least from the perspective of firm size dispersion, if we simply abstract from θ altogether. In this model, with Kimball demand, we can see that is not the case. The curvature of the contour lines implies that the product of $z^{\sigma-1}$ and θ (or the sum of their logarithms) is not a sufficient statistic for firm size. Taking a convex combination of two points with the same revenue yields a higher — rather than equal — revenue. As such, the separate values of θ and z are needed to determine a firm's revenue, and similarly the full joint distribution is required to characterize the economy-wide size distribution. To summarize, relaxing the restriction of CES demand departs from perfect isomorphism between θ and z, amplifying the importance of θ for firm size dispersion.

Another reason why this complementarity is important is because it strengthens the interplay between θ and z in the firm's dynamic investment decision. In the CES-CRS case, the firm's level of θ already plays a strong role in determining its optimal level of investment in z, and vice versa.[35] Intuitively, firms with higher z extract more profit *per customer*, so firms with more customers (i.e., higher θ) gain more from a given increase in z. The deviation from CES-CRS in this model, and resulting curvature of the revenue contour lines, strengthens this interaction further. A firm that is two log points below the mean $\log\theta$ experiences a much smaller gain from increasing $\log z$ than a firm that is two log points above the mean. This is in

[34] A similar isomorphism holds in the decreasing returns case. The only reason I compare to the CRS case is that the expression is analytically simpler. The key assumption driving the isomorphism is CES demand.

[35] The CES-CRS case is analyzed by Ignaszak and Sedlacek (2023), where they show the interaction between customer accumulation and innovation incentives is a strong driver of aggregate growth.

Figure A4: Revenue Function



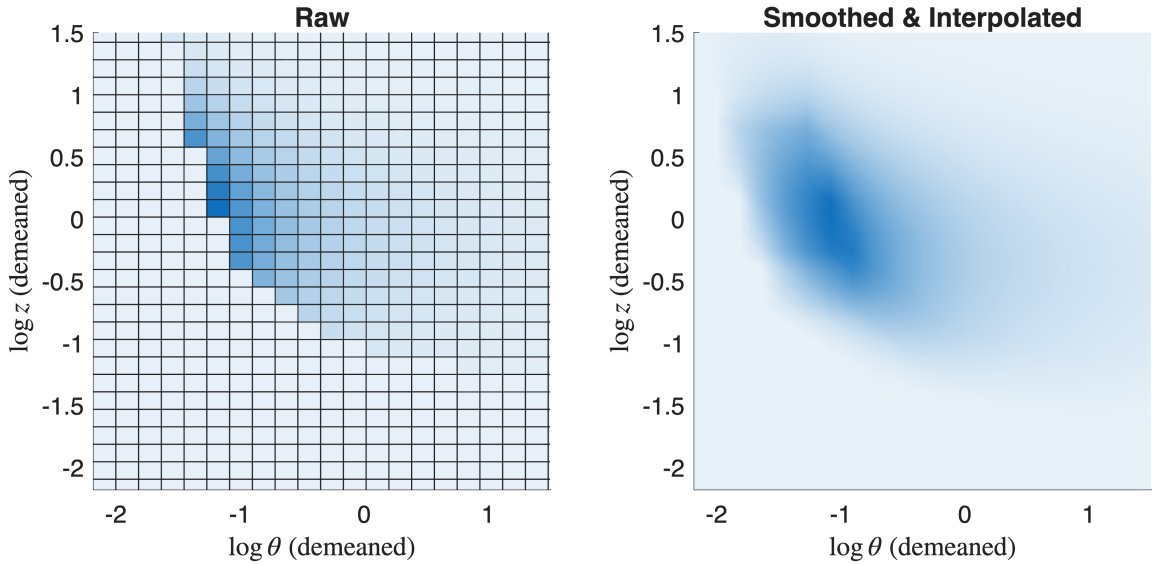
Notes: This contour graph shows log demeaned revenue as a function of log demeaned customer access (θ) and productivity (z). For visual clarity, revenue is split into 20 bins, with each bin assigned a unique color.

contrast to the CES-CRS case, where the proportional gain (specifically, $d \log \text{Revenue}$) from a given increase in $\log z$ is independent of θ. In short, this model features stronger interplay between θ and z than the CES-CRS model. This is important because, as we will see, a significant channel through which customer access influences firm size dispersion is by generating dispersion in incentives to invest in z. Similarly, a significant portion of the gains from subsidizing θ will come from the indirect positive effect on investment in z. Both of these effects are strengthened by the greater complementarity between θ and z.

C.2 Finer Depiction of Joint Distribution

For visual clarity, the joint distribution depicted in Figure 4 used only 10 shades of color. In Figure A5, I present a finer depiction of the joint distribution, both in its raw and smoothed/interpolated form.

Figure A5: Joint Distribution (Finer Depiction)



Notes: This depicts the stationary distribution over $\log\theta$ and $\log z$, both in a smoothed and interpolated "continuous" depiction and the raw underlying matrix. 256 shades are used, with darker shades corresponding to a higher density of firms.

C.3 Constrained Optimal Subsidy

The optimal customer investment policy computed in Section 4 involves large government expenditure, close to 10% of pre-subsidy output. Of course, this large investment is repaid with a large return in output and welfare, but it is still reasonable to wonder how practical such a policy is. Political considerations may prevent such a large investment by the government, particularly when the outcome of the policy is uncertain. In addition, I finance the policy in the model with non-distortionary lump sum taxes, whereas in practice the taxation used to fund the policy would likely be distortionary. With this in mind, Table A2 presents results for a constrained optimal policy, with total expenditure restricted to be less than 1% of pre-subsidy output. The resulting constrained optimum is a 6% subsidy, which still has a significant positive effect on all aggregates.

Table A2: Customer Investment Policies

	Optimal	Constrained
τ_θ	-0.36	-0.06
CEW	$+4.0\%$	$+0.8\%$
ΔY	$+5.8\%$	$+0.9\%$
ΔL_p	-6.1%	-0.8%
ΔTFP	$+11.0\%$	$+1.5\%$
ΔL	$+1.9\%$	$+0.1\%$
$T/Y_{baseline}$	8.8%	1.0%

Notes: CEW refers to the consumption-equivalent welfare effect of the policy. Y is aggregate output, L_p is total production labor, $TFP \equiv \frac{Y}{L_p^\alpha}$, and L is total labor. The optimal policy maximizes the consumption equivalent welfare change. The constrained policy maximizes the consumption equivalent welfare change, subject to the condition that total expenditure cannot exceed 1% of pre-subsidy output.