Market Concentration and Aggregate Productivity: The Role of Demand

Jeremy Pearce[*] Liangjie Wu[†] [‡§]

July 2026

Abstract

This paper focuses on large firms that build customer capital through investments in demand. Grounded in facts we develop on firms' expenditures and customer capital, we build a model with oligopolistic firms with heterogeneous productivity that invest in demand and study how this affects concentration and aggregate productivity. Estimating the joint dynamics of productivity and customer capital, we find TFP losses sustained by dynamic investments in demand, and decompose these dynamics through the lens of the model. "Good" concentration is productivity-driven; "bad" demand-driven concentration can persist due to firms' incentives to protect their accumulated customer capital. This creates a legacy effect where unproductive firms retain large market share. Aggregate productivity is 5% lower due to this effect. Standard size-based policies can backfire through entrenching incumbents.

Key Words: Firm Dynamics, Productivity, Demand, Customer Capital, Market Concentration, Competition, Misallocation.

JEL Code: O31, O32, O34, O41, D22, D43, L11, L13, L22

[*]Federal Reserve Bank of New York, jeremy.pearce@ny.frb.org;

[†]The University of Edinburgh, liangjiewu.robin@gmail.com;

[‡]Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

[§]We thank Amir Sufi for making the sales-and-marketing expenditure data available. We also thank Tincho Almuzara, Mary Amiti, and Simone Lenzu for helpful comments, and seminar participants at BSE (Growth and Productivity), EIEF, New York Fed, NY Macro, Richmond Fed, Rutgers, SED, and St. Louis Fed. The views expressed here are the authors' and do not necessarily reflect those of the Federal Reserve Bank of New York or the Federal Reserve System.

1 Introduction

Competition in the typical market is competition among a few large firms, each large enough to move market aggregates. These firms build and hold their positions on two margins, improving productivity and investing in demand, and both margins are persistent. Productivity sits at the center of the modern framework of firm dynamics, where size and productivity tend to move together through the basics of market competition (Atkeson and Burstein, 2008; Edmond et al., 2023). The dynamics of demand investment have not played a major role in this literature, even as concentration and market power continue to draw attention (De Loecker et al., 2020; Olmstead-Rumsey, 2022; Akcigit and Ates, 2023; De Ridder, 2024). Further, firms spend more building demand than improving productivity: marketing and sales average 4.2% of revenue, against 2.4% for R&D.[1] This paper studies strategic demand investment under oligopoly and traces its consequences for market concentration and aggregate productivity.

We capitalize firm-level marketing expenditure into a stock of *customer capital*, measure it alongside productivity for publicly traded U.S. firms, and ask whether the concentration that demand investment produces is efficient or distortionary. We document three facts.

> **Fact 1 (Customer capital tracks firm size more closely than productivity does).** Within industries, customer capital correlates with revenue at 0.73–0.83; labor productivity at 0.50–0.53.
>
> **Fact 2 (Customer capital is more persistent than productivity, and its investment is path dependent).** A simple AR(1) is 0.98 against 0.86 for labor productivity, and marketing investment growth is positively autocorrelated.[2]
>
> **Fact 3 (Markets are granular).** A typical market has 4–7 effective competitors, across both publicly traded and consumer-packaged goods firms.[3]

A long literature interprets the gap between large firms and productive firms as allocative inefficiency, the output lost when inputs do not flow to their most productive

[1] We follow He et al. (2024), who separate sales and marketing expenditure from SG&A in Capital IQ filings and impute missing values with a language-model extraction validated against external data.

[2] David and Venkateswaran (2019) report negative growth autocorrelation for physical capital. The positive sign for customer-capital investment motivates the path dependency discussed in Section 3.

[3] We use the sales-weighted inverse Herfindahl index across various market definitions.

use (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). Customer capital can enrich our understanding of this reading: as a measured factor behind size it explains part of the gap, while its path-dependent accumulation (Fact 2) opens a dynamic channel for more misallocation. Whether the overall investment in marketing is productive turns on the incentives of the handful of strategic rivals who direct the flow of demand (Fact 3).

We build a model that connects these facts to equilibrium dynamics. Granular firms compete in quantities (Cournot) under CES demand, differ in productivity and customer capital, and invest in marketing for two reasons: to complement their productivity (the *productivity incentive*) and to capture the market power that rising markups attach to size (the *size incentive*). In a world without strategic incentives (e.g., constant markups), the size incentive vanishes and marketing loses its strategic component.

We focus on two results in the oligopolistic model. In classic models of competition and innovation, rents are lowest when firms are neck and neck, so investment concentrates there (Aghion et al., 2005). Competition with variable markups changes this structure: when markups rise with share, the return to widening a gap is largest for the firm already ahead, and the marketing incentive peaks when firms already have an advantage, so the largest investments go to defending existing leadership. The strategic feedback raises the persistence of customer capital above its depreciation rate. Firms with identical productivity therefore can develop persistent size dispersion, and marketing alone can sustain concentration. We term this a *legacy effect*: past marketing raises current marketing with productivity held fixed, so firms can hold high market shares by defending accumulated customer capital.

We then use this framework to study concentration. Concentration driven by productivity alone tends to be efficient, as an increase in productivity at the frontier dominates the markup loss it brings. With a second state imperfectly co-moving with productivity, concentration can be good or bad: productivity-driven concentration raises measured TFP and demand-driven concentration lowers it, at identical shares and markups in the two cases. The concentration-productivity relationship therefore has no fixed sign; its slope takes the sign of the demand–productivity alignment, which the strategic feedback holds positive on average while defending misaligned positions when productivity reverses.

We microfound demand with rational inattention and, because the granular dynamic

game is otherwise intractable, solve it globally with a new computational method: the value function is a linear-quadratic block plus probit corrections that preserve the nonlinearity in productivity and measure how far the equilibrium departs from a linear-quadratic benchmark.

We then measure this misallocation in the data. Following Hsieh and Klenow (2009) with customer capital in place of capital, the cross-sectional dispersion of demand relative to productivity implies a static TFP loss of 61%; tracing it through the joint dynamics of productivity and customer capital, about 70% is sustained by the accumulation process rather than by permanent differences across firms. Persistence alone cannot say whether it reflects efficient sorting of demand toward productive firms or distortionary entrenchment of incumbents, and we thus use the model to guide this analysis.[4]

We then use the quantified framework to study aggregate productivity and welfare. Marketing moves aggregate productivity through two channels: it aligns demand with productive firms, which raises the productivity frontier, and it sustains the concentration and markups that hold output below that frontier. The second channel dominates. Shutting marketing down removes the alignment gain, lowering frontier productivity by 4.9%, but welfare rises by 5.2% once marketing labor is counted. A planner who reallocates customer capital freely raises productivity by 23%; a planner who controls only marketing raises it by 6% relative to the no-marketing benchmark; and the feasible first best raises it by 4.7% over the equilibrium, so aggregate productivity is about 5% below its efficient level.

These dynamics change how we evaluate size-based policy. Because private returns to marketing rise with the accumulated stock while social returns fall with it, instruments that reward size, output subsidies among them (Edmond et al., 2023), can push marketing in the wrong direction and entrench incumbents. For a small subsidy, output subsidies still increase welfare. At large subsidies, output subsidies entrench incumbents as they continue to invest in building demand even with low productivity.

The remainder of this section reviews the literature, while the rest of the paper is structured as follows. Section 2 documents the marketing facts using firm-level data. Section 3 introduces the theoretical model with oligopolistic firms with heterogeneous

[4]NielsenIQ scanner data corroborate the construct: the Hottman et al. (2016) residual-demand measure behaves like the marketing-based stock, and its entry and exit flows give an independent check on the model's decay and marketing-cost parameters (Appendix C).

productivity and customer capital. Section 4 measures the misallocation between customer capital and productivity and calibrates the dynamic game to the estimated dynamics. Section 5 discusses productivity shocks and policy implications for how the quantified model changes our understanding of size-based policies, marketing, and antitrust; Appendix C corroborates the customer-capital construct in NielsenIQ scanner data.

Related Literature. A growing literature is studying how firms build their customer base, treating demand as a slow-moving state that firms accumulate through investment (Dinlersoz and Yorukoglu, 2012; Drozd and Nosal, 2012; Gourio and Rudanko, 2014; Ignaszak and Sedlácek, 2022; He et al., 2024). Demand persists through consumer inertia (Bornstein, 2019), through brand preferences that hold across decades and locations (Bronnenberg et al., 2009, 2012), and through relationship capital in buyer–seller settings (Heise, 2024); young plants, entrants, and exporters grow by accumulating demand (Foster et al., 2008, 2016; Fitzgerald et al., 2024; Argente et al., 2025); and in product-level data residual demand carries the largest share of the variation in firm size (Hottman et al., 2016; Eslava et al., 2024), with product creation central to sustaining it (Einav et al., 2021; Argente et al., 2024). Customer capital is an essential intangible asset (Haskel and Westlake, 2017; Crouzet and Eberly, 2019; Bronnenberg et al., 2022; Pearce and Wu, 2024). Kaplan and Zoch (2020) note the rising role of labor in expansionary activities. The strategic margin of the accumulation is studied less (Cavenaile and Roldan-Blanco, 2021; Afrouzi et al., 2023; Cavenaile et al., 2023; Bornstein and Peter, 2025; Cavenaile et al., 2025; Hubmer and Nord, 2025). Afrouzi et al. (2023) also connect customer acquisition to market power and misallocation; we differ in this paper by focusing on an oligopolistic setting, where firms defend the stock they have accumulated. The closest antecedent in strategic competition for customers is Cavenaile et al. (2025), where advertising shifts demand across oligopolists and firms strategically invest in marketing. The key difference is that the investments are period-by-period and do not treat customer capital as a state variable, whereas we focus on customer capital as a dynamic state variable.

Our approach builds on the firm-dynamics tradition where we focus on granular firms (Jovanovic, 1982; Hopenhayn, 1992; Klette and Kortum, 2004; Gabaix, 2011; Baqaee and Farhi, 2019), and on the misallocation literature, which reads the gap between large firms and productive firms as lost output (Restuccia and Rogerson, 2008; Hsieh and

Klenow, 2009; David and Venkateswaran, 2019). How firms invest to expand, and how market structure shapes that investment, is the subject of a second branch (Aghion et al., 2005; Liu et al., 2022; Peters, 2020; Akcigit and Ates, 2023), built mostly to answer growth questions. We relate these two literatures through studying investment in an oligopolistic setting. Investment in our economy can be related to margins that do not enhance productivity, which turns the investment machinery common for R&D in the growth literature into a misallocation question. Our framework also stays tractable at empirically relevant firm counts, so concentration is a quantitative equilibrium object rather than a feature fixed by the market structure.

A third literature studies the efficiency cost of market power (Atkeson and Burstein, 2008; De Loecker and Eeckhout, 2018; De Loecker et al., 2020; Olmstead-Rumsey, 2022; De Ridder, 2024); the rise of superstar firms and the labor share decline (Autor et al., 2020; Kehrig and Vincent, 2021) are consistent with a significant role for demand, which our facts show tracks firm size more closely than productivity does. When size reflects productivity alone, correcting the static markup distortion and raising efficiency go hand in hand (Edmond et al., 2023); with a second, demand-driven source of size the two separate. We extend the insight of Atkeson and Burstein (2008) that variable markups generate misallocation into a dynamic setting, where a size-based instrument faces a trade-off between compressing markups and degrading the dynamic sorting of customers toward productive firms: larger is not necessarily better and this message is central to our discussion and quantitative results.

2 Measurement and Facts of Customer Capital

We construct customer capital from firm-level marketing data and document three facts: it tracks firm size more closely than productivity does, it is more persistent than productivity and its investment is path dependent, and markets are granular. These facts motivate the dynamic model of Section 3; Section 4 then measures the misallocation between size and productivity that they imply.

2.1 Data and Measurement

We draw on Capital IQ matched to Compustat for firm-level marketing expenditures, revenue, and employment of publicly traded U.S. firms. Capital IQ reports sales and marketing expenditures, following the classification in He et al. (2024). The sample covers 1997 to 2022 but in order to use a developed measure we start in 2007. All dollar variables are deflated to 2017 real dollars.

We construct a stock of customer capital via the perpetual inventory method:

$$B_{it} = (1 - \delta)\, B_{i,t-1} + M_{it}, \tag{1}$$

where M_{it} is real marketing expenditure and $\delta = 0.15$ is the annual depreciation rate.[5] We initialize the stock in 1997, the first year of Capital IQ marketing coverage. We measure labor productivity as log real revenue per worker.[6] The stock B_{it} is the empirical counterpart of the demand state variable b_{it} in the model of Section 3.[7]

Customer Capital and Firm Size. Table 1 reports the moments behind these facts for two samples: firms with observed Capital IQ marketing data, and an expanded sample that imputes marketing for firms reporting SG&A but not marketing, using the industry $\times$ year median marketing/SG&A ratio.[8] All the main qualitative results of this paper are robust to this classification choice. Customer capital correlates with revenue at 0.73–0.83, against 0.50–0.53 for revenue per worker, and its industry-demeaned AR(1) is 0.98, against 0.86 for productivity (within firms, 0.81 versus 0.49). Marketing investment is itself path dependent: its growth is positively autocorrelated (Appendix B.8).

Market Structure. Competition for customer capital takes place among a handful of large firms. We measure the sales-weighted inverse Herfindahl index of market shares,

[5] We follow He et al. (2024) for the construction of marketing expenditures and use the depreciation rate from purchase price allocation useful-life estimates. Our results are robust to alternative depreciation rates $\delta \in \{0.05, 0.10, 0.20, 0.25, 0.30\}$.

[6] The misallocation and calibration exercises of Section 4 instead use a revenue-productivity residual that nets customer capital out of revenue.

[7] Appendix C adds a structural cross-check in NielsenIQ scanner data, decomposing market shares into demand, productivity, and markups following Hottman et al. (2016) to recover a residual-demand measure of customer capital.

[8] For firms without observed marketing, we compute the NAICS-3 $\times$ year median ratio of marketing to SG&A among firms reporting both, then multiply by the non-reporting firm's SG&A to impute marketing expenditure. Customer capital is then accumulated via the same perpetual inventory method.

Table 1: Customer Capital, Productivity, and Employment

	Customer Capital (b)	Productivity (a)	Employment
Panel A: Marketing Sample ($N = 30{,}816$ firm-years)			
Corr(log revenue, $\cdot$)	0.73	0.53	0.93
AR(1) coefficient	0.98	0.86	0.99
Correlation (x_t, x_{t-1})	0.99	0.88	0.99
AR(1), firm FE	0.81	0.49	0.79
Panel B: Imputed Sample ($N = 56{,}349$ firm-years)			
Corr(log revenue, $\cdot$)	0.83	0.50	0.92
AR(1) coefficient	0.98	0.88	0.99
Correlation (x_t, x_{t-1})	0.99	0.90	0.98
AR(1), firm FE	0.89	0.50	0.78

Notes: Correlations and AR(1) coefficients computed within NAICS-3 $\times$ year cells. Marketing sample: customer capital is log B_{it} from capitalized marketing expenditure ($\delta = 0.15$). Imputed sample adds firms where marketing is imputed from the NAICS-3 $\times$ year median marketing/SG&A ratio. Productivity is log revenue per worker; employment is log headcount. AR(1) from OLS with robust standard errors. Firm FE: within-firm AR(1) with year dummies. Sample: publicly traded U.S. firms, 2007–2022.

and find 4–7 major competitors across market definitions. This is roughly similar to concentration in consumer packaged goods, which contains all bar-codes in a market (Pearce and Wu, 2024). Firms are therefore large enough to internalize their effect on the market, the strategic force the model builds on.

All of these ingredients suggest important elements of the model. The importance of customer capital for firm size in both the level and persistence indicates that firms naturally should have two state variables: customer capital and productivity. Given that this investment is persistent and path dependent, its accumulation cannot be an exogenous force to the firm; and competition for it plays out among a handful of large firms. Section 3 builds an oligopoly economy consistent with these facts. Firms invest in customer capital under two incentives, a *productivity incentive* that sorts demand toward productive firms and a *size incentive* through which variable markups reward large firms regardless of productivity, and Section 4 measures how much misallocation between size and productivity the two incentives generate.

3 Model

Section 2 documents that firm size tracks customer capital rather than productivity, and that customer capital is the more persistent of the two stocks. We study these phenomena by developing a model in which a handful of oligopolistic firms (as in Atkeson and Burstein, 2008) hold two stocks: productivity, which evolves exogenously, and customer capital, which they accumulate through costly marketing. Unlike frameworks in which customer capital is acquired competitively and tracks productivity with a lag (Gourio and Rudanko, 2014), granular firms here internalize that a larger market share carries a higher markup, so marketing responds to size itself and demand acquires its own dynamics. Our contribution is twofold: the model separates the *productivity incentive* from the *size incentive* in demand investment, and a new computational method solves the resulting dynamic game globally at empirically relevant firm counts.

Section 3.1 lays out the environment, and Section 3.2 reduces the aggregate equilibrium to self-contained industry blocks. Section 3.3 characterizes the static pricing equilibrium within an industry and the paper's central asymmetry: profits depend only on the sum of productivity and customer capital, while efficiency depends on the split. Section 3.4 defines the dynamic marketing equilibrium; Section 3.5 works out the two-firm case, in which the economics is analytical, and Section 3.6 shows how we compute the equilibrium at general firm counts. Section 3.7 derives the welfare accounting used in the quantitative analysis. Proofs and derivations for this section are collected in Appendix A.

3.1 Environment

Time is discrete. The economy carries a unit continuum of industries, each holding I firms, and a representative household who consumes across all of them. Calligraphic letters denote economy-wide aggregates ($\mathcal{C}_t$, $\mathcal{L}_t$); plain letters denote industry-level objects, and we drop the industry index k when we focus on a specific industry to understand industry-level mechanics.

Households. The representative household supplies labor $\mathcal{L}_t$ and maximizes

$$\max_{c_{ikt},\mathcal{L}_t} \sum_{t=0}^{\infty} \beta^t \left(\ln \mathcal{C}_t - \mathcal{L}_t\right), \tag{2}$$

subject to the budget constraint $\int_0^1 \sum_{i=1}^{I} p_{ikt} c_{ikt}\, dk = \mathcal{L}_t + \Pi_t$, where aggregate consumption is the Cobb-Douglas aggregate over industries, $\ln \mathcal{C}_t = \int_0^1 \ln C_{kt}\, dk$, the wage is the numeraire, and Π_t collects firm profits, rebated through a representative portfolio. Within an industry, consumption aggregates across the I firms with constant elasticity of substitution $\sigma > 1$, weighted by each firm's share $\omega_{ikt} \equiv e^{b_{ikt}} / \sum_j e^{b_{jkt}}$ of the industry's customer capital,

$$C_{kt} = \Big[\sum_{i=1}^{I} \omega_{ikt}^{1/\sigma}\, c_{ikt}^{\frac{\sigma-1}{\sigma}} \Big]^{\frac{\sigma}{\sigma-1}}. \tag{3}$$

Customer capital is the household's prior tendency to buy from a firm before observing prices, the product's image or its shelf presence.[9] Because customer capital enters only through the normalized weights ω, a uniform shift $b_{ikt} \to b_{ikt} + c$ leaves the bundle unchanged: only the *distribution* of customer capital across firms enters utility. In this setting, customer capital is purely allocative.

The household side delivers a CES demand curve for each firm, $c_{ikt} = \omega_{ikt}(p_{ikt}/P_{kt})^{-\sigma} C_{kt}$ with the customer-neutral price index $P_{kt} = (\sum_i \omega_{ikt}\, p_{ikt}^{1-\sigma})^{\frac{1}{1-\sigma}}$; the aggregate conditions that tie the industries together follow in Section 3.2. Appendix A.7 spells out the full household problem.

Firms. Firm i produces with labor, $c_{it} = e^{a_{it}/(\sigma-1)}\, l_{it}$, where a_{it} is its productivity scaled into market-share units.[10] Marketing η_{it} carries a convex labor cost $\frac{\kappa}{2}\eta_{it}^2$ and adds to the firm's stock of customer capital. The two stocks follow

$$\begin{aligned} a_{i,t+1} &= \delta_a\, a_{it} + \epsilon^a_{i,t+1}, \\ b_{i,t+1} &= \delta_b\, b_{it} + \eta_{it} + \epsilon^b_{i,t+1}, \end{aligned} \tag{4}$$

with retentions $\delta_a, \delta_b \in (0,1)$ and independent Gaussian innovations $\epsilon^a \sim N(0, \nu_a^2)$, $\epsilon^b \sim N(0, \nu_b^2)$. The primitives impose no dependence of customer capital on productivity;

[9] Appendix A.5 microfounds this demand system from rational inattention in consumption choices.

[10] The scaling makes a one percent productivity gap map into a one percent market-share gap, holding customer capital fixed; the quantitative analysis treats the scaling explicitly.

whatever alignment the model produces comes from firms' marketing choices.

Granularity. Firms are large relative to their market: their shocks move market aggregates, and they internalize their impact on the price index, competing in quantities as in Cournot. Amiti et al. (2019) document that large firms' pass-through behavior indicates the strong role of strategic complementarities.

3.2 Industry Equilibrium

An equilibrium of the aggregate economy is a path of industry prices, quantities, and marketing policies such that the household optimizes, the firms in each industry play the game defined in Sections 3.3 and 3.4, and the labor market clears,

$$\mathcal{L}_t = \int_0^1 \left(L_{kt}^{\text{prod}} + L_{kt}^{\text{mkt}}\right) dk, \tag{5}$$

with the two industry labor pools defined in Section 3.7. Household optimality reduces to three conditions, derived in Appendix A.7. Linear disutility equates the marginal value of wealth to the wage, so with the wage as numeraire the marginal value of wealth is one and labor supply is perfectly elastic; the labor market absorbs any aggregate labor demand (5) without moving relative prices. Log aggregation across industries makes expenditure shares constant across k, and with the marginal value of wealth at one the common level is pinned,

$$E_{kt} \equiv \sum_{i=1}^{I} p_{ikt} c_{ikt} = 1 \qquad \text{for every } k \text{ and } t, \tag{6}$$

so each industry's revenue is invariant to every aggregate and to every other industry's state. The portfolio condition prices firm profits at the household's discount factor β.

Together these conditions make the aggregate equilibrium block-recursive. Industry expenditure, the wage, and the discount rate are all constants, so the equilibrium of the aggregate economy is the equilibrium of a single industry, solved on its own state $(\mathbf{a}_k, \mathbf{b}_k)$ and replicated over the continuum. Sections 3.3 through 3.6 characterize the equilibrium of one industry, dropping the index k. Welfare is the one remaining aggregate object; it aggregates linearly in logs across the continuum, and Section 3.7 establishes that the

aggregate flow is the ergodic mean of the per-industry flow even though individual industries never settle at a steady state.

3.3 Static Equilibrium

Within a period, given the industry state $(\mathbf{a}, \mathbf{b})$, quantity competition delivers markup pricing, $p_{it} = \mu_{it}\, e^{-a_{it}/(\sigma-1)}$ with $\mu_{it} = \frac{\sigma}{\sigma-1}\frac{1}{1-s_{it}}$: markups rise with market share. Writing $z_{it} \equiv a_{it} + b_{it}$ for the *composite*, shares solve the fixed point

$$s_{it} = \frac{e^{z_{it}}\,(1-s_{it})^{\sigma-1}}{\sum_{j=1}^{I} e^{z_{jt}}\,(1-s_{jt})^{\sigma-1}}, \tag{7}$$

and profits are a convex function of the own share alone,

$$\pi_{it} = \frac{1}{\sigma} s_{it} + \frac{\sigma-1}{\sigma} s_{it}^2. \tag{8}$$

Profits depend on the state only through the composite z_{it}: for the firm, a unit of productivity and a unit of customer capital are perfect substitutes. Real consumption does not share this symmetry. Per unit of production labor,

$$\log \frac{C}{L} = \frac{1}{\sigma-1} \log \sum_i \omega_i\, e^{a_i}\, \mu_i^{1-\sigma}, \tag{9}$$

so productivity raises output directly while customer capital only reweights it across firms; by the level invariance of (3), among equally productive firms every configuration of demand yields the same surplus. Appendix A.1 derives the share fixed point, the marginal-profit matrix, and the output expressions behind (7)–(9).

3.4 Dynamic Equilibrium

Marketing is the model's only choice with dynamic consequences. We study Markov perfect equilibria: each firm's strategy $\eta_i(\mathbf{a}, \mathbf{b})$ depends on the payoff-relevant state alone. Taking rivals' strategies as given, firm i solves

$$V_i(\mathbf{a}, \mathbf{b}) = \max_{\eta_i} \left\{ \pi_i(\mathbf{z}) - \frac{\kappa}{2}\eta_i^2 + \beta\, \mathbb{E}\left[V_i(\mathbf{a}', \mathbf{b}') \right] \right\}, \tag{10}$$

subject to the laws of motion (4) for all firms, including the feedback of rivals' strategies into rivals' states.

Definition 1 *A* marketing equilibrium *is a profile of strategies* $\{\eta_i(\mathbf{a},\mathbf{b})\}_{i=1}^{I}$ *such that each* η_i *attains the maximum in* (10) *given the others.*

The equilibrium is characterized by the first-order and envelope conditions. Writing $\mathbf{x} \equiv (\mathbf{a},\mathbf{b})$ for the stacked state and $\nabla_k V_i \equiv \partial V_i / \partial b_k$,

$$\kappa\, \eta_i(\mathbf{x}) = \beta\, \mathbb{E}\big[\nabla_i V_i(\mathbf{x}')\big], \tag{11}$$

$$\nabla_i V_i(\mathbf{x}) = \phi_{ii}(\mathbf{z}) + \beta\delta_b\, \mathbb{E}\big[\nabla_i V_i(\mathbf{x}')\big] + \beta \sum_{k\neq i} \frac{\partial \eta_k(\mathbf{x})}{\partial b_i}\, \mathbb{E}\big[\nabla_k V_i(\mathbf{x}')\big], \tag{12}$$

where ϕ_{ii} is the marginal profit of own customer capital implied by (7)–(8), derived in Appendix A.1.[11] The last term in the envelope condition is the strategic feedback: firm i's state moves rivals' future marketing, and firm i internalizes that response. This term is absent under atomistic competition; the two-firm case shows it generates the legacy effects.

Computing this equilibrium is difficult on two counts. The strategic feedback makes each firm's policy depend on the derivatives of rivals' policies, a fixed point in a space of functions; and the state is the full cross-section $(\mathbf{a},\mathbf{b})$, so grid-based methods face a state space that grows exponentially in I. We take the two routes in turn: closed form at two firms, then a computational method that scales to the calibrated firm count.

3.5 Two-Firm Case

With two firms the game's economics is analytical. When the two retentions coincide ($\delta_a = \delta_b = \delta$), profits depend on the state only through the composite gap $\zeta \equiv z_1 - z_2$, and the equilibrium reduces to one dimension. The two-firm case establishes, in closed form, the three results the quantitative analysis scales up: the marketing incentive peaks at a strict lead, so competition rewards incumbency; the strategic feedback raises the persistence of market positions above its technological rate; and concentration is good

[11] In closed form, $\phi_{ik} = \pi'(s_i)\,\chi_i\,(\mathbb{I}_{i=k} - \hat{\chi}_k)$ with marginal profit $\pi'(s) = \frac{1}{\sigma} + 2\frac{\sigma-1}{\sigma}s$, share responsiveness $\chi_i = \big(\frac{1}{s_i} + \frac{\sigma-1}{1-s_i}\big)^{-1}$, and normalized responsiveness $\hat{\chi}_k = \chi_k / \sum_l \chi_l$. The own-policy term cancels by the first-order condition.

or bad for measured productivity depending on whether it is productivity-driven or demand-driven. Throughout, statements are exact unless labeled numerical; numerical statements use a fitted equilibrium (computed by the method of Section 3.6) at illustrative parameters. Appendix A.3 collects the proofs, the parameters, and every magnitude.

The share fixed point (7) solves in closed form at $I = 2$: the leader's share is logistic in the gap,

$$s_1(\zeta) = \frac{1}{1 + e^{-\zeta/\sigma}}, \tag{13}$$

with slope $s'(\zeta) = s(1-s)/\sigma$. The flow incentive to widen the gap combines how the share responds to the gap with how profit responds to the share, $h(\zeta) \equiv d\pi_1/d\zeta = s'(\zeta)\,\pi'(s)$.

Proposition 1 (Incentive peaks at a strict lead) *With profits (8), the flow incentive* $h(\zeta) = s(1-s)\left[1 + 2(\sigma-1)s\right]/\sigma^2$ *is maximized at the unique share*

$$s^* = \frac{(2\sigma - 3) + \sqrt{(2\sigma-3)^2 + 6(\sigma-1)}}{6(\sigma-1)} \;\in\; \left(\tfrac{1}{2}, 1\right),$$

with $s^* \to \frac{1}{2}$ *as* $\sigma \to 1$. *If profit were linear in the share (constant markups),* $h \propto s(1-s)$ *would peak exactly at the tie.*

Locally the result is a statement about profit curvature alone: since $s''(0) = 0$, the slope of the incentive at the tie is $h'(0) = s'(0)^2\,\pi''(\bar{s})$, so the sign of π'' classifies the peak. Convex profit, the Cournot case where markups rise with share, puts the peak on the leader's side; concave profit would put it on the laggard's side and generate catch-up instead. At conventional elasticities the peak sits well inside the leader's territory, and the fitted equilibrium policy, which aggregates discounted future flow incentives, peaks at a still larger lead (numerical). In the step-by-step innovation models of Aghion et al. (2005), rents are lowest when firms are neck and neck, so the incentive peaks at the tie and firms invest to escape competition; here interior duopoly rents survive the tie, and the same escape-competition logic pushes the peak to a strict lead, so investment defends leadership rather than contesting it.

Endogenous persistence. The gap follows the closed loop $\zeta' = \delta\zeta + \Delta\eta(\zeta) + \epsilon_\zeta$, where $\Delta\eta \equiv \eta_1 - \eta_2$ is the equilibrium marketing difference and $\epsilon_\zeta \sim N(0, \nu_\zeta^2)$ with $\nu_\zeta^2 = 2\nu^2$.

The slope $\gamma \equiv \Delta\eta'(0)$ measures the strategic feedback: whether a widening gap induces marketing that widens it further.

Proposition 2 (Feedback is positive) *To leading order in β, the feedback is the shock-weighted asymmetry of the flow incentive,*

$$\gamma = \frac{\beta\delta}{\kappa\nu_\zeta^2}\,\mathbb{E}\Big[\epsilon\big(h(\epsilon) - h(-\epsilon)\big)\Big] + O(\beta^2),$$

where

$$h(\epsilon) - h(-\epsilon) = \frac{2(\sigma-1)}{\sigma^2}\, s(1-s)(2s-1) \geq 0,$$

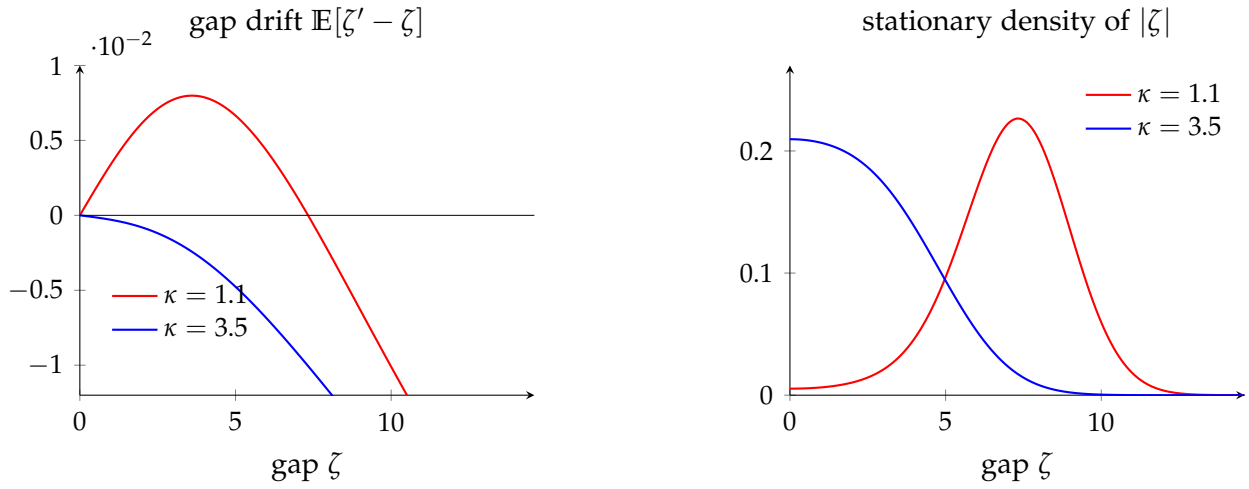
with strict inequality for $\sigma > 1$, and $\gamma = 0$ exactly when profit is linear in the share.

Positive feedback raises the persistence of market positions above the technological retention: the closed-loop gap behaves like an AR(1) with coefficient $\delta + \gamma$. At the illustrative parameters the fitted feedback is positive and modest, amplifying the stationary variance of the gap (numerical). This wedge between measured and technological persistence is the *legacy effect*; the calibration in Section 4.3 identifies it in the data. The feedback also admits a threshold: if γ exceeds the depreciation $1 - \delta$, the symmetric state turns unstable and the stationary distribution of the gap peaks at a strict lead; marketing alone then sustains concentration. Figure 1 displays the threshold as a comparative static in the marketing cost: at a high marketing cost the gap drift is negative everywhere and the stationary distribution is unimodal at the tie, while at a low marketing cost the drift turns positive near the tie and the distribution peaks when a firm is leading, so concentration can arise without a productivity gap.

Internalization is essential to the sign. An atomistic firm facing a variable-elasticity (Kimball) demand has, by the envelope theorem, $d^2 \ln \pi / dz^2$ proportional to the slope of its demand elasticity, so its proportional return to the state *declines* in size exactly when markups rise with size: the atomistic analogue of γ is nonpositive. Amplification requires firms large enough to perceive the wedge between the margin and the physical elasticity; quantity competition among granular firms provides that scale.

Section 3.7 defines three productivity indices: the equilibrium index Z^{eq}, output per unit of production labor; the frontier Z^*, the index with labor reallocated efficiently at the same state; and their ratio, the markup wedge $\mathcal{M}$. In the gap coordinates $\alpha \equiv a_1 - a_2$,

FIGURE 1: TWO-FIRMS



Notes: Closed loop of the composite gap under the leading-order policy of Proposition 2, at $\sigma = 3.9$, $\delta = 0.998$, $\beta = 0.5$, $\nu_a = 0.1$, $\nu_b = 0$; the low discount keeps the leading-order policy accurate and the near-unit retention makes the instability threshold visible. Left: expected one-period change of the gap. Right: implied stationary density of the absolute gap, unimodal at the tie for $\kappa = 3.5$ and peaked at a strict lead for $\kappa = 1.1$. Source: model computations.

$\beta_b \equiv b_1 - b_2$, $\zeta = \alpha + \beta_b$, they take closed forms:

$$Z^* = \left[\frac{\cosh(\zeta/2)}{\cosh(\beta_b/2)}\right]^{\frac{1}{\sigma-1}}, \tag{14}$$

$$\mathcal{M}(\zeta) = \frac{[\cosh(\zeta/2)]^{\frac{1}{\sigma-1}}}{\left[\cosh\left(\frac{\zeta}{2\sigma}\right)\right]^{\frac{\sigma}{\sigma-1}}},$$

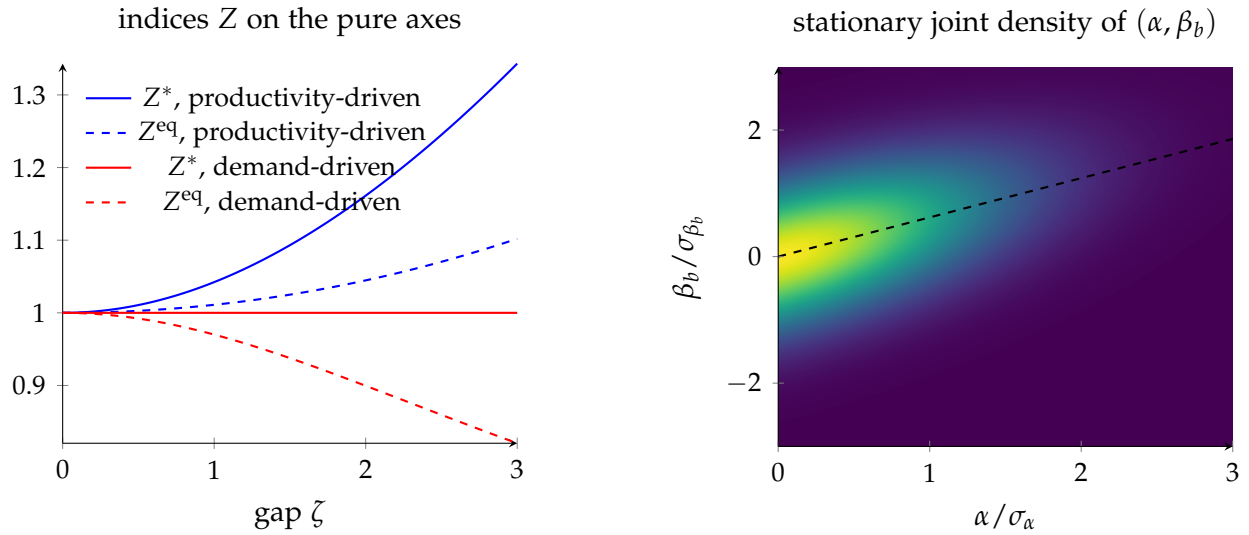
$$\ln Z^{\text{eq}} = \ln Z^* - \ln \mathcal{M}.$$

The markup wedge $\mathcal{M}$ depends on the composite alone, so it is invariant to whether a gap comes from productivity or from demand. All source dependence sits in the frontier Z^*, and the equilibrium index splits accordingly into an alignment term and a markup term. Underlying the closed forms is a property of Cournot pricing.

Proposition 3 (Labor invariance) *At $I = 2$ the two firms hire identical production labor at every state: $\ell_i = \frac{\sigma-1}{\sigma} s_i(1 - s_i)$ and $s_2 = 1 - s_1$ make $\ell_1 = \ell_2$ an identity.*

Both firms equate marginal revenue to the same wage, so the leader's entire composite advantage exits through its price. The equilibrium never reallocates production labor toward the productive firm, so the markup wedge measures the productivity value of

Figure 2: Two-Firm Productivity Indices



Notes: Left: exact closed forms (14). Right: ergodic joint density of the standardized productivity and demand gaps at the baseline parameters ($\sigma_\alpha = 0.29$, $\sigma_{\beta_b} = 0.04$, correlation 0.62), folded onto the half plane; dashed line: $\mathbb{E}[\beta_b \mid \alpha]$. Source: model computations.

the reallocation that markups block.

Corollary 1 (Sign test) *Exactly,*

$$\frac{\partial \ln Z^{eq}}{\partial \zeta}\Big|_{\beta_b} = \frac{\tanh(\zeta/2\sigma)}{2(\sigma-1)} > 0,$$

while

$$\frac{\partial \ln Z^{eq}}{\partial \beta_b}\Big|_{\zeta} = -\frac{\tanh(\beta_b/2)}{2(\sigma-1)} < 0.$$

Concentration raises measured productivity when it is productivity-driven and lowers it when it is demand-driven, at identical shares and markups in the two cases. With customer capital shut down ($\mathbf{b} \equiv 0$), the frontier gain dominates the markup wedge at every gap, so productivity-driven concentration is always efficient; misallocation that hurts requires the demand dimension.

The concentration–productivity relationship has no fixed sign, and its slope inherits the sign of the alignment between demand and productivity. The left panel of Figure 2 traces the two indices along the pure axes: a productivity-driven gap raises both the frontier and the equilibrium index, while a demand-driven gap of the same size in shares and markups lowers the equilibrium index one for one with the markup wedge.

Alignment and misalignment. Marketing is the endogenous variable that correlates productivity with customer capital. With $\Delta\eta$ linearized to $\gamma\zeta$, the stationary covariance of the productivity and demand gaps is proportional to γ: zero under constant markups, positive here. The linear approximation is accurate: the fitted policy is close to linear over the states the economy visits. The resulting correlation between the productivity and demand gaps is strongly positive (the right panel of Figure 2).

Alignment holds on average, yet the ergodic distribution keeps recurring mass on misaligned states. Productivity moves with fresh innovations while the demand gap is a distributed lag of past marketing, so a productivity reversal leaves the old leader holding the accumulated customer capital. Exactly, the economy is less productive than its no-marketing counterpart ($Z^* < 1$) if and only if $\alpha(\alpha + 2\beta_b) < 0$, the entrenchment region. Customer capital tilts toward the less productive firm in a sizable minority of states, and only rarely is the market leader itself the less productive firm (numerical).

On the ergodic distribution the equilibrium index gains almost nothing over the no-marketing benchmark, alignment net of entrenchment, while the marketing contest absorbs a first-order share of revenue (Appendix A.3 reports the magnitudes). The social cost of marketing at two firms is dissipation rather than misallocation, and the quantitative analysis shows the same ordering survives at the calibrated firm count.

3.6 Computation

Beyond two firms the equilibrium has no closed form. We solve it with a method that exploits the Gaussian innovations to handle the two obstacles, the strategic feedback and the growing state space. The method works on the stacked state $\mathbf{x} = (\mathbf{a}, \mathbf{b})$ with its two retentions; when the retentions coincide the state reduces to the composite $\mathbf{z}$, as in the two-firm case. Derivations, the admissible basis families, and the accuracy audit against exact small-I solutions are in Appendix A.2.

Stein representation. Stein's lemma (Gaussian integration by parts) replaces derivatives of the value function inside an expectation with shock-weighted levels: for smooth f of polynomial growth,

$$\mathbb{E}\Big[\frac{\partial f}{\partial b_k'}(\mathbf{x}') \,\Big|\, \mathbf{x}\Big] = \frac{1}{\nu_b^2}\,\mathbb{E}\big[\epsilon_k^b\, f(\mathbf{x}') \,\big|\, \mathbf{x}\big]. \tag{15}$$

Applied to (11), it writes the equilibrium conditions in levels of V alone:

$$\kappa\,\eta_i(\mathbf{x}) = \frac{\beta}{\nu_b^2}\,\mathbb{E}\big[\epsilon_i^b\,V_i(\mathbf{x}')\,\big|\,\mathbf{x}\big], \tag{16}$$
$$V_i(\mathbf{x}) = \pi_i(\mathbf{a}+\mathbf{b}) - \frac{\kappa}{2}\eta_i(\mathbf{x})^2 + \beta\,\mathbb{E}\big[V_i(\mathbf{x}')\,\big|\,\mathbf{x}\big].$$

The first equation is the policy condition and the second the value recursion; following the terminology of reinforcement learning, we call them the *actor* and the *critic*. The expectation averages rivals' randomness into the level of V, and the covariance with the firm's own demand shock recovers the derivative, so the algorithm never takes a numerical derivative of a stored function. The second-order condition can also be written in levels, and we verify it at every solution.

Critic family. We approximate the value function by a sum of one-dimensional index functions of the stacked state, a linear-quadratic block plus M probit terms,

$$V_i(\mathbf{x}) \approx \theta_0 + \theta_1\,(u_i\cdot\mathbf{x}) + \theta_2\,(u_i\cdot\mathbf{x})^2 + \sum_{m=1}^{M}\theta_{m+2}\,\Phi\Big(\frac{d_m\cdot\mathbf{x}-\bar{x}_m}{w_m}\Big), \tag{17}$$

where an index $d_m\cdot\mathbf{x} = d_m^a\cdot\mathbf{a} + d_m^b\cdot\mathbf{b}$ loads the two stocks separately, and we restrict the directions in three ways. Directions are mean zero within each stock, because the equilibrium is invariant to a common shift in either stock and only relative positions matter. Profiles are *conjugate* to the Gaussian, so every expectation in (16) has a closed form and no numerical integration is needed. The approximant inherits the game's symmetry (invariance to permutations of the rivals) because each free direction is tied to the sum over its rival permutations. These restrictions do not reduce flexibility, since sums of index functions can approximate any continuous function. Because the method stores a coefficient vector rather than a function on a grid, it is unaffected by the dimension of the state.

Fitting as regression. With the actions held at the current iterate, the Bellman residual of (16) rearranges to a linear regression of current flow payoffs on *annuitized* basis functions, where each regressor is the basis function minus β times its expected next-period value. The algorithm iterates the regression and the action update to a joint fixed point

on a simulated ergodic panel, adding probit terms only when the residual remains large. Because the one-step-ahead term is computed in closed form, the regressors contain no simulation noise and the associated bias is absent. We audit every fit four ways: the Bellman residual out of sample, the distance to an exact small-I solver, the second-order condition, and a stationarity certificate requiring the closed-loop persistence of every pairwise gap $z_i - z_j$ to lie below one. The fitted M measures how far the equilibrium departs from its linear-quadratic benchmark. The fitted policies match exact solutions at small firm counts to a few parts in ten thousand; Appendix A.2 reports the audit (Figure A1 and Table A1).

3.7 Aggregation and Welfare

Because the near-unit-root persistence of Section 4.3 keeps each industry moving through its state space, individual industries never reach a steady state. The aggregate economy nonetheless has one, because industries are ex-ante identical and hit by independent shocks: at any date the cross-section of industry states is the ergodic distribution π of the industry equilibrium, and the aggregates are deterministic constants, $\ln \mathcal{C}_t = \mathbb{E}_\pi[\ln C(\mathbf{z})]$ and $\mathcal{L}_t = \mathbb{E}_\pi[L^{\text{prod}} + L^{\text{mkt}}]$. Aggregate welfare is therefore the ergodic mean of a per-industry flow, $\frac{1}{1-\beta}\,\mathbb{E}_\pi[W(\mathbf{z})]$ with W defined below, and every counterfactual in Section 5 is a difference of such ergodic means (Appendix A.7 states the argument). Within an industry, unit expenditure and markup pricing give the labor pools and output in closed form (Appendix A.4). Each firm's production labor is its revenue deflated by its markup, $\ell_i = s_i/\mu_i$, so the pool is the reciprocal of the share-weighted harmonic markup $\bar{\mu}_H \equiv (\sum_i s_i/\mu_i)^{-1}$ and falls one for one with concentration, $\text{HHI} \equiv \sum_i s_i^2$,

$$L^{\text{prod}}(\mathbf{z}) = \frac{1}{\bar{\mu}_H(\mathbf{z})} = \frac{\sigma - 1}{\sigma}\big(1 - \text{HHI}(\mathbf{z})\big), \tag{18}$$

because high markups at the largest firms reduce their hiring. Production labor depends on shares alone: with expenditure fixed at one, a productivity improvement lowers prices and raises quantities one for one in revenue terms, so it changes output per worker but not employment.[12] Marketing labor is $L^{\text{mkt}} = \frac{\kappa}{2}\sum_i \eta_i^2$, and equilibrium output is

[12]The property follows from unit expenditure, the log upper tier, and linear labor disutility. With wealth effects or elastic industry expenditure, productivity would also move labor and the accounting below would include an additional term.

$C^{\text{eq}} = \frac{\sigma-1}{\sigma}\left[\sum_j \omega_j\, e^{a_j}(1-s_j)^{\sigma-1}\right]^{\frac{1}{\sigma-1}}$.

The welfare accounting uses three productivity indices. The equilibrium index is output per unit of production labor; the frontier index is the same object with labor reallocated efficiently at the given state; and their ratio defines the markup wedge:

$$\begin{aligned} Z^{\text{eq}} &\equiv \frac{C^{\text{eq}}}{L^{\text{prod}}}, \\ Z^{*} &\equiv \Big[\sum_i \omega_i\, e^{a_i}\Big]^{\frac{1}{\sigma-1}}, \\ \mathcal{M} &\equiv \frac{Z^{*}}{Z^{\text{eq}}}. \end{aligned} \tag{19}$$

The frontier Z^* is the customer-capital-weighted mean of productivity. It rises when demand is allocated to productive firms, and it equals one when firms are equally productive, regardless of how demand is distributed. The markup wedge $\mathcal{M}$ measures the output lost because markups misallocate production labor across firms.

We measure welfare by household utility, the model-consistent metric. With unit expenditure, industry consumption is the reciprocal of the industry price index, which decomposes as $\ln C^{\text{eq}} = -\ln P = \ln Z^* - \ln \bar{\mu}_H - \ln \mathcal{M}$ (Appendix A.4), where the harmonic markup term captures the labor that markups withhold and the wedge $\mathcal{M}$ captures the misallocation of the remaining labor. The per-period flow is then

$$W(\mathbf{z}) = \ln C^{\text{eq}} - L^{\text{prod}} - L^{\text{mkt}} = \ln Z^* - \underbrace{\left(\ln \bar{\mu}_H + \ln \mathcal{M} + L^{\text{prod}}\right)}_{\text{net cost of market power}} - L^{\text{mkt}}. \tag{20}$$

We separate marketing's welfare effect into three channels: the position of demand on the productivity distribution ($\ln Z^*$), the net static cost of market power, and the labor used in marketing. The middle term nets a log consumption gain against a linear labor cost because $\ln \bar{\mu}_H = -\ln L^{\text{prod}}$ exactly, by (18); reported separately, the markup burden and the withheld labor would appear as two large offsetting entries. Because production labor does not respond to productivity, the frontier term involves no labor adjustment, and the split of any welfare change across the three channels is unique. Each term is a function of the state, so counterfactual tables difference ergodic means channel by channel and the columns sum to the welfare change exactly; the level and dispersion components of the markup burden reappear separately in the Hsieh–Klenow account-

ing of Section 4.3. A productivity index alone omits the last two channels, which the counterfactuals of Section 5 show account for most of the welfare effects.

The equilibrium has two externalities: markup dispersion misallocates production labor, and firms, which are paid on the composite, do not internalize the social value of allocating customer capital to productive firms. Section 5 derives the planner's solution and quantifies both.

4 Quantification

Given the model ingredients, we now take a closer look at the data in order to leverage the model machinery. We calibrate on the Compustat and Capital IQ panel of Section 2, quantifying the misallocation between demand and productivity and isolating the part the accumulation dynamics sustain.[13] We measure the static dispersion of firm size with the Hsieh–Klenow accounting, customer capital in place of capital, then use the joint law of motion of demand and productivity to split it into the component the dynamics generate and the permanent heterogeneity a stationary economy cannot. We calibrate the dynamic game to those dynamics and recompute the misallocation and the concentration inside the calibrated model.

4.1 Static Misallocation

A long literature interprets the gap between large firms and productive firms as allocative inefficiency, the output lost when inputs do not flow to their most productive use (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). Following Hsieh and Klenow (2009), we measure a revenue-productivity residual, $\log \hat{a}_{it} = \log R_{it} - \alpha_B \log b_{it}$, where $\alpha_B = 0.568$ is the firm-fixed-effects elasticity of revenue to customer capital, and decompose the variance of log revenue TFP into productivity dispersion, customer-capital dispersion, and an offsetting covariance.[14] Positive comovement between productivity and customer capital raises efficiency; independent or opposing movement leaves re-

[13] The NielsenIQ scanner data of Appendix C corroborate the customer-capital construct and the dynamics with an independent, structural measure of demand.

[14] Revenue TFP is $\text{TFPR}_i \equiv R_i/(l_i^{\alpha_L} b_i^{\alpha_B})$ with α_L the labor revenue elasticity, and $\sigma^2_{\text{TFPR}} = \vartheta_a^2 \text{Var}(\log \hat{a}) + \vartheta_b^2 \text{Var}(\log b) - 2\vartheta_a \vartheta_b \text{Cov}(\log \hat{a}, \log b)$, with loadings $(\vartheta_a, \vartheta_b) = (0.438, 0.189)$ at $\sigma = 4$ under constant returns; the TFP loss is $\exp[\frac{\sigma}{2}\sigma^2_{\text{TFPR}}] - 1$.

sources mismatched.

On the baseline sample, dispersion in customer capital exceeds dispersion in productivity ($\text{Var}(\log b) = 4.19$ against $\text{Var}(\log \hat{a}) = 1.98$), and a positive covariance (1.77) cancels part of both. Aggregated at $\sigma = 4$ (following (Hottman et al., 2016)), the implied static TFP loss is 61% (Table 2).

Table 2: Cross-section moments and static TFP loss

Panel A: cross-section moments (NAICS3×year demeaned)	
$\text{Var}(\log \hat{a})$	1.977
$\text{Var}(\log b)$	4.191
$\text{Cov}(\log \hat{a}, \log b)$	1.768
Panel B: static TFP loss	
σ^2_{TFPR}	0.237
TFP loss at $\sigma = 4$ baseline	**61**%

Notes: Imputed 5-consecutive sample, $N = 41{,}473$ firm-years across 5,926 firms. TFP loss $= (\exp[\frac{\sigma}{2}\sigma^2_{\text{TFPR}}] - 1) \times 100$ with $(\vartheta_a, \vartheta_b) = (0.438, 0.189)$ at $\sigma = 4$, derived under constant returns ($\nu = 1 - \omega$) from the empirical $\alpha_B = 0.568$.

4.2 Dynamic Misallocation

How much of this dispersion is a static fact of cross-firm heterogeneity, and how much is sustained by the dynamics of accumulation? This connects our misallocation question closer to the dynamic model. We estimate the joint evolution of productivity and customer capital as a bivariate first-order vector autoregression,

$$\log \hat{a}_{it} = \rho_{aa} \log \hat{a}_{i,t-1} + \varepsilon^A_{it}, \tag{21}$$

$$\log b_{it} = \rho_{ba} \log \hat{a}_{i,t-1} + \rho_{bb} \log b_{i,t-1} + \varepsilon^B_{it}, \tag{22}$$

estimated by random-effects GLS with year effects, which separates the permanent cross-firm component of dispersion from the component that shocks generate as they propagate through the system. On the baseline sample, productivity persistence is $\rho_{aa} = 0.910$, customer-capital persistence is $\rho_{bb} = 0.964$, and the cross-effect from productivity to demand is positive at $\rho_{ba} = 0.033$ (Table 3).[15]

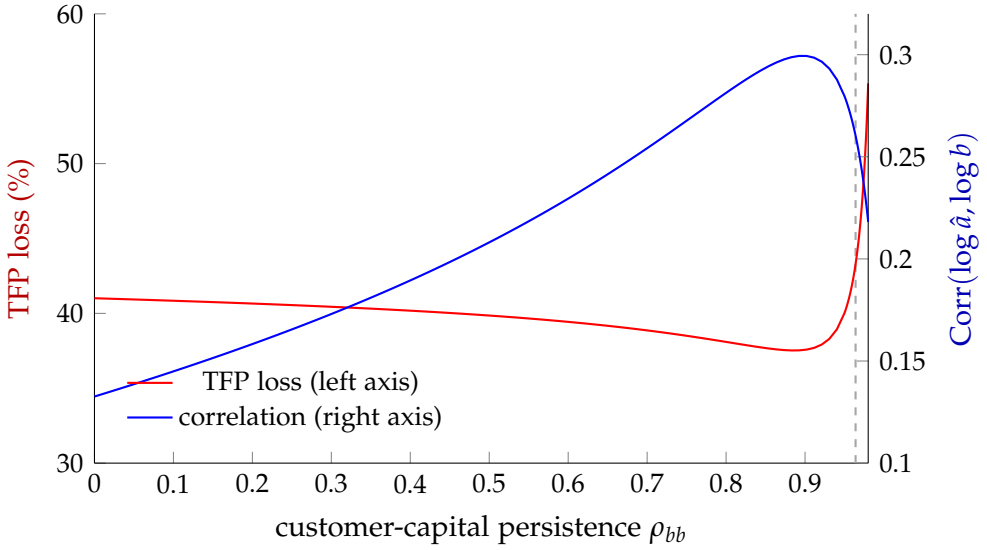
[15]The VAR persistence ($\rho_{bb} = 0.964$) is the RE-GLS estimate net of permanent firm heterogeneity, slightly below the simple industry-demeaned AR(1) of 0.98 in Table 1.

We then connect the estimated parameters to a stationary distribution to measure the misallocation driven by the dynamics and not the cross-firm fixed effects. Under stationarity, the VAR implies a closed-form steady-state dispersion, which we set against the data. The dynamic component alone generates a TFP loss of 43%, about seventy percent of the 61% static loss; the remainder is permanent firm heterogeneity that the VAR does not generate.

What are the implications of the coefficients for overall efficiency? Figure 3 traces how the loss and the productivity–demand correlation move with persistence. As ρ_{bb} rises, demand aligns more tightly with productivity, the correlation $\text{Corr}(\log \hat{a}, \log b)$ climbs, and the loss falls to a minimum near $\rho_{bb} \approx 0.9$. This indicates the fact that one wants demand to be relatively persistent, as firms with high demand tend to be more productive.

Past that point the persistence of b misaligns with a, so the loss turns up and diverges as persistence approaches one. The data estimate, $\rho_{bb} = 0.964$, lies on the rising branch. As $\rho_b b$ goes to 1, the misallocation explodes, as b becomes totally disconnected from the productivity process. Thus, what we termed legacy effects in the model add to measured persistence and raise the cost of demand misallocation at an increasing rate.

Figure 3: Demand Misallocation and the Productivity–Demand Correlation



Notes: Steady-state TFP loss (red, left axis) and the productivity–demand correlation $\text{Corr}(\log \hat{a}, \log b)$ (blue, right axis) as customer-capital persistence ρ_{bb} varies, with the other VAR primitives held at their baseline. The dashed line marks the data estimate, $\rho_{bb} = 0.964$. Source: author calculations.

The loss is stable within the sampling band and moves with the estimator choice:

Table 3: Bivariate VAR estimates and dynamic-component decomposition

	Estimate	SE
Panel A: RE-GLS VAR coefficients		
ρ_{aa} (productivity persistence)	0.910	(0.006)
ρ_{bb} (customer-capital persistence)	0.964	(0.002)
ρ_{ba} (cross-effect)	0.033	(0.003)
LR multiplier $\rho_{ba}/(1-\rho_{bb})$	0.917	
Panel B: dynamic-component stationary distribution vs. data		
	Dynamic	Data
$\mathrm{Var}(\hat{a})$	0.91	1.977
$\mathrm{Var}(b)$	1.60	4.191
$\mathrm{Cov}(\hat{a}, b)$	0.32	1.768
σ^2_{TFPR}	0.179	0.237
TFP loss	**43**%	61%

Notes: Panel A: RE-GLS with year effects and firm-clustered standard errors on the baseline marketing panel, 27,508 firm-year pairs across 3,862 firms. Panel B: dynamic-component stationary moments solve $\Sigma_{\mathrm{dyn}} = \Phi\,\Sigma_{\mathrm{dyn}}\,\Phi' + \Sigma_\varepsilon$ at the Panel A coefficients with the within-firm innovation covariance $(\hat{v}^2_A, \hat{v}^2_B, \hat{v}_{AB}) = (0.157, 0.092, 0.012)$; the loss applies the reduced-form index of Table 2, whose cross-section moments the Data column repeats.

within two standard errors of the estimate it runs 42 to 45 percent, read off the curve in Figure 3, and the Arellano–Bond point matches the baseline. The system-GMM point exceeds one, where no stationary loss exists.

Two forces. The persistence and cross-effect are equilibrium objects that net two forces the model separates. A positive cross-effect can arise because the demand system rewards productivity directly, which is efficient, or because a firm's own size makes further accumulation cheaper independently of productivity, which entrenches incumbents and is distortionary. The two are observationally equivalent in the cross-section yet carry opposite welfare implications, and the near-unit-root persistence is what makes the distinction first-order: close to the estimate the loss is steeply increasing in ρ_{bb}, so whether that persistence reflects productivity sorting or pure incumbency decides how much demand misallocation costs.

We now calibrate the model of Section 3 and quantify how much of the measured persistence and misallocation the size incentive sustains.

4.3 Model Calibration

The model has eight parameters. Three are set externally: the substitution elasticity $\sigma = 4$, the median firm-level estimate of Hottman et al. (2016) on a matching demand system; the discount factor $\beta = 0.96$; and the number of firms $I = 7$, consistent with the four to seven effective competitors per market documented in Section 2. The remaining five, the two retentions, the marketing cost curvature, and the two shock volatilities $(\delta_a, \delta_b, \kappa, \nu_a, \nu_b)$, are calibrated jointly to five moments: the two VAR persistences, the two VAR innovation variances, and the productivity-to-demand cross-effect ρ_{ba} (Table 4).

We compute the model's moments with the same estimator as the data. We simulate the model's ergodic panel, construct *measured* productivity inside it exactly as in the data, $\log \hat{a} = \log R - \alpha_B \log b$ with an in-panel firm-effect estimate of α_B, and run the same random-effects VAR on the simulated panel as on the Compustat panel of Section 4.2. The markup term therefore loads into measured productivity in the model just as oligopoly pricing loads it into the data.

Customer capital outlasts productivity as a technology, $\delta_b - \delta_a \approx 0.02$ in the fitted retentions; a composite-retention specification would force this gap to zero. On top of

Table 4: Calibration: Parameters and Moments

Parameter		Identification	Value
Substitution elasticity	σ	Hottman et al. (2016)	3.90
Discount factor	β	annual discounting	0.96
Number of firms	I	effective competitors	7
Productivity retention	δ_a	joint fit to $(\rho_{aa}, \rho_{bb}, \rho_{ba}, \hat{v}_A^2, \hat{v}_B^2)$	0.913
Customer-capital retention	δ_b		0.933
Marketing cost curvature	κ		3.30
Productivity shock	ν_a		0.481
Demand shock	ν_b		0.281
Moment		**Model**	**Data**
ρ_{aa} (targeted)		0.910	0.910
ρ_{bb} (targeted)		0.964	0.964
ρ_{ba} (targeted)		0.033	0.033
$\hat{v}_A^2$, $\hat{v}_B^2$ (targeted)		0.157, 0.092	0.157, 0.092

Notes: Moments computed on the measured $\hat{a} = \log R - \alpha_B \log b$ within the simulated ergodic panel, random-effects VAR as in the data. Data moments are the baseline VAR estimates of Table 3. Targeted moments are matched to within 2×10^{-4}; the out-of-sample Bellman residual at the solution is 1.1×10^{-2}. Second-order conditions verified at the solution.

that retention, the strategic feedback adds about 0.031 to the measured persistence of demand ($\rho_{bb} - \delta_b$), the general-I counterpart of the two-firm γ of Proposition 2. The calibration targets the cross-effect ρ_{ba} directly and fits it at 0.033. Measurement also attenuates: because markups load into $\hat{a}$, a structural productivity shock of 0.48 appears as a measured residual dispersion of 0.40.

4.4 Measured Misallocation through the Model

The measurement of Section 4.1 holds exactly inside the model. In the Cournot equilibrium, revenue satisfies $\log R_i = \alpha_B \log \mathfrak{B}_i + \log \hat{a}_i$ with $\mathfrak{B}_i$ the physical customer-capital stock, and the revenue-productivity index conflates true productivity with the oligopoly markup, $\log \hat{a}_i = a_i - (\sigma - 1) \log \mu_i$. The Hsieh–Klenow efficient benchmark reallocates the aggregate stock of customer capital each period, and the optimal allocation is the log-linear line $b_i^* \propto a_i$. The static loss then splits into the markup wedge and a *demand misallocation* term proportional to the variance of the distance from this line. That variance equals the variance of the log revenue product of customer capital, the quantity measured in the data.

Against that VAR-stationary column the calibrated game lines up (Table 5): it generates 1.16 of the VAR-stationary 1.60 in customer-capital dispersion and 0.68 of 0.91 in measured productivity, both far below the raw 4.19 and 1.98. Demand misallocation

Table 5: Static Loss: Model and Data

	Data	VAR-stationary	Model
$\mathrm{Var}(\log \hat{a})$	1.98	0.91	0.68
$\mathrm{Var}(\log \mathfrak{B})$	4.19	1.60	1.16
$\mathrm{corr}(a, b)$	0.61	0.26	0.21
TFP loss	61%	43%	26.8%
markup wedge			6.5%
demand misallocation			20.3%

Notes: Data column: the Compustat cross-section (Table 2). VAR-stationary column: the stationary distribution implied by the estimated VAR of Section 4.2; its loss uses the reduced-form index of the Data column.

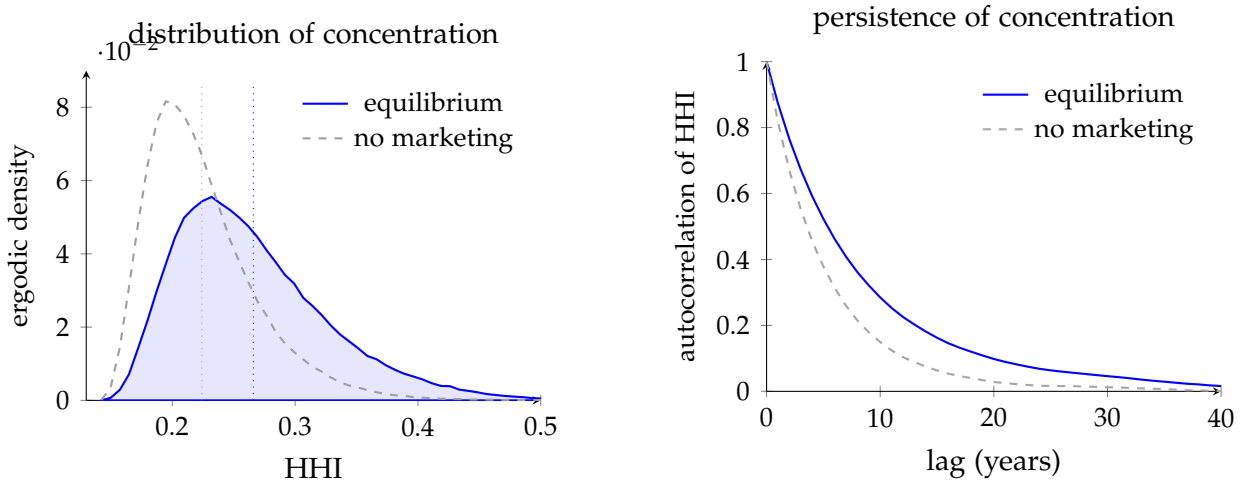
dominates the model's loss, 20.3 of its 26.8 percentage points, markups the rest. The correlation row puts the model's structural alignment, $+0.21$, next to the measured 0.26 of the VAR-stationary data. Markups enter $\hat{a}$ with a negative loading, so a measured correlation understates the structural alignment on both sides; the remaining gap comes from markup measurement rather than from the dynamics.

4.5 Concentration

Marketing raises both the level and the persistence of concentration. Against a no-marketing benchmark simulated on the same shock draws, the equilibrium raises expected concentration from an HHI of 0.224 to 0.266, moving the effective number of firms from 4.6 to 4.0, which the calibration never targeted. The larger change is dynamic: the closed marketing loop raises the annual persistence of the leadership gap from 0.92 to 0.95, a half-life of 14.1 rather than 8.2 years. Figure 4 displays both margins, the rightward shift of the ergodic concentration distribution and the slower decay of its autocorrelation.

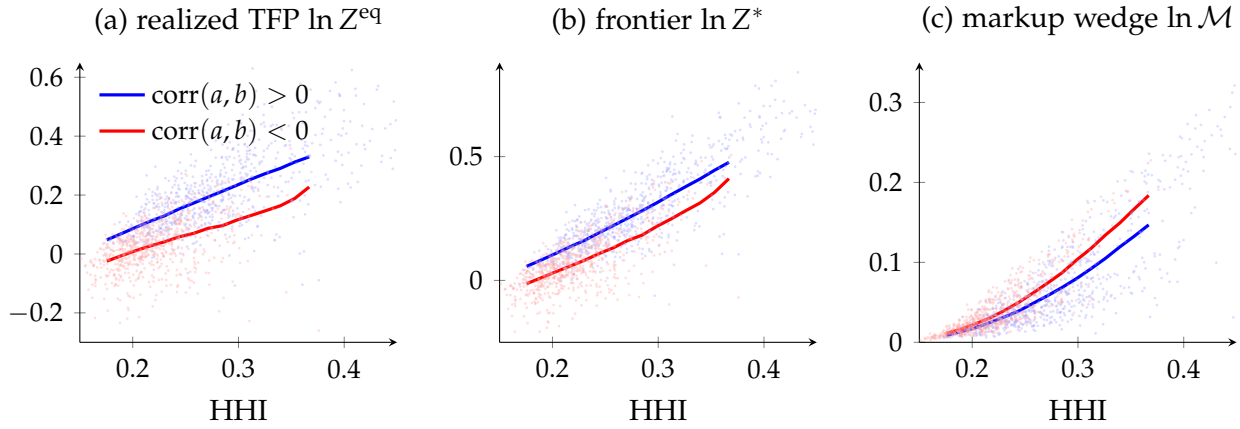
The cross-sectional relationship between concentration and productivity has no fixed sign, as the sign test of Corollary 1 already showed at two firms. At $I = 7$ the same pattern holds state by state (Figure 5): in states where the productive firms hold the demand, realized TFP rises with concentration, and in misaligned states it falls, while the markup component rises with concentration in both. The ergodic mean of the alignment is $+0.21$, so the economy's average slope is positive, again untargeted. Grouping states by their productivity dispersion instead isolates the same margin from the productivity side (Figure 6): the more productivity dispersion a state has, the steeper its concentration–productivity slope. This positive analysis suggests the costs of concentra-

Figure 4: Concentration in the Calibrated Game and Without Marketing



Notes: Calibrated ergodic panel and the no-marketing benchmark on common shock draws ($T = 200{,}000$). Left: ergodic density of the market HHI; dotted lines mark the means (0.224 without marketing, 0.266 in equilibrium). Right: autocorrelation of the HHI by lag. Source: model computations.

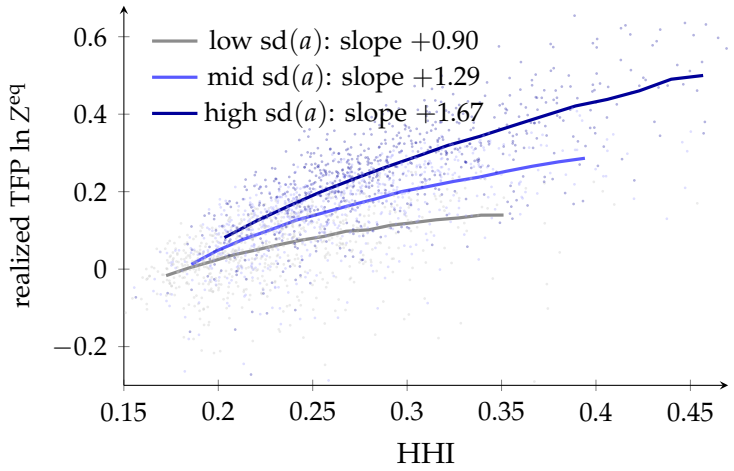
Figure 5: Concentration and Productivity in the Cross-Section



Notes: Calibrated model at $I = 7$, ergodic states ($T = 200{,}000$) split by the sign of the within-state cross-firm correlation of productivity and customer capital; thick lines are conditional means by HHI. Realized TFP (a) decomposes into the frontier (b) net of the markup wedge (c) by (19). Source: model computations.

tion are closely linked to the drivers; when markets have high and persistent productivity dispersion, concentration tends to be much more benign outcome than when dispersion is high.

Figure 6: Concentration and Productivity, States Grouped by Productive Dispersion



Notes: Ergodic states grouped into terciles of the within-state cross-firm standard deviation of productivity; thick lines are conditional means by HHI, legend slopes from a linear fit of $\ln Z^{eq}$ on HHI within each tercile. Source: model computations.

5 Counterfactuals and Policy

We use the calibrated model to measure what competition in demand costs the household and what a policymaker could recover. We characterize the planner's solution, compare the equilibrium to four counterfactuals on common shock draws (shutting marketing down, free marketing, a constrained planner under Cournot pricing, and the first best), and close with a size-based subsidy that backfires.

5.1 Planner's Problem

The planner chooses every firm's marketing at once, under one of two static regimes. The *constrained planner* inherits Cournot pricing and controls marketing only, so its gap to the equilibrium isolates the conduct of the marketing game; the *first-best planner* also allocates production labor, so its additional gap measures market power itself. Formally, for regime $r \in \{C, F\}$ the planner solves

$$V^P(\mathbf{x}) = \max_{\boldsymbol{\eta} \in \mathbb{R}^I} \left\{ W^r(\mathbf{x}) - \frac{\kappa}{2} \sum_{i=1}^{I} \eta_i^2 + \beta \, \mathbb{E}\left[V^P(\mathbf{x}') \,\middle|\, \mathbf{x}, \boldsymbol{\eta} \right] \right\}, \tag{23}$$

subject to the laws of motion (4) for every firm, with the regime flows

$$W^C(\mathbf{x}) = \ln C^{\text{eq}}(\mathbf{x}) - L^{\text{prod}}(\mathbf{x}), \tag{24}$$

$$W^F(\mathbf{x}) = \ln Z^*(\mathbf{x}) - 1. \tag{25}$$

The objective is household welfare, the flow (20) with the marketing bill written out, evaluated at the regime's statics. The constrained flow W^C is the household flow at the Cournot allocation; the first-best flow is its value with production labor allocated efficiently, $\max_{\ell,L}\{\ln C - L\} = \ln Z^* - 1$ (Appendix A.1). The two regimes bracket what policy can reach: conduct remedies alone, and conduct together with the output-side correction.

Level invariance makes the social value of customer capital zero-sum. A uniform shift in $\mathbf{b}$ leaves welfare unchanged, so the social flow value sums to zero across firms at every state, and the same holds for the discounted shadow values (Appendix A.4). The planner's marketing is therefore purely redistributive: negative marketing on one firm offsets positive marketing on another one-for-one, and at any symmetric state the planner does not market at all. In the equilibrium, by contrast, each firm's private marginal value of customer capital at the symmetric state is 0.039 while the social value is exactly zero, so marketing at the modal state redistributes customers without creating surplus.

The planner's first-order condition clarifies what a corrective policy must change. The planner is a single agent, so the envelope theorem removes the strategic-feedback block from the shadow-value recursion:

$$\kappa\, \eta_i^P(\mathbf{x}) = \beta\, \mathbb{E}\big[\partial_i V^P(\mathbf{x}')\big], \tag{26}$$

$$\partial_i V^P(\mathbf{x}) = g_i(\mathbf{x}) + \beta\delta_b\, \mathbb{E}\big[\partial_i V^P(\mathbf{x}')\big], \tag{27}$$

the same recursion as the equilibrium condition (11) with the private flow ϕ_{ii} replaced by the social flow g_i and no feedback term. In the first-best regime the social flow has a closed form,

$$g_i^F = \frac{\omega_i}{\sigma - 1}\left[\frac{e^{a_i}}{\sum_j \omega_j e^{a_j}} - 1\right], \tag{28}$$

positive exactly when firm i's productivity exceeds the customer-capital-weighted aver-

age, so marketing has social value only when it moves demand toward productive firms. Appendix A.4 derives the constrained-regime flow in closed form and verifies both flows against finite differences.

5.2 Implementing the Planner Solution

The planner's problem solves with the equilibrium's computational machinery: under the Stein representation it is the critic–actor pair

$$\kappa\,\eta_i^P(\mathbf{x}) = \frac{\beta}{\nu_b^2}\,\mathbb{E}\big[\epsilon_i^b\,V^P(\mathbf{x}')\,\big|\,\mathbf{x}\big], \tag{29}$$

$$V^P(\mathbf{x}) = W^r(\mathbf{x}) - \frac{\kappa}{2}\sum_{i=1}^{I}\eta_i^P(\mathbf{x})^2 + \beta\,\mathbb{E}\big[V^P(\mathbf{x}')\,\big|\,\mathbf{x}\big], \tag{30}$$

the same operator as the equilibrium system (16), solved by the same family and algorithm on the same laws of motion (4). The distance between the two problems has an exact expression in terms of the total private value.

Definition 2 (Total private value) $V^\Sigma(\mathbf{x}) \equiv \sum_{i=1}^{I} V_i(\mathbf{x})$, *the sum of the firms' equilibrium values.*

Summed across i, the firms' critics (16) yield a total that obeys the planner's recursion (30) with aggregate profit as its flow,

$$V^\Sigma(\mathbf{x}) = \underbrace{\tfrac{1}{\sigma} + \tfrac{\sigma-1}{\sigma}\,\mathrm{HHI}(\mathbf{z})}_{\text{aggregate profit }\Pi} - \frac{\kappa}{2}\sum_{i=1}^{I}\eta_i(\mathbf{x})^2 + \beta\,\mathbb{E}\big[V^\Sigma(\mathbf{x}')\,\big|\,\mathbf{x}\big], \tag{31}$$

with the full marketing bill in both. With labor in closed form (18), aggregate profit is linear in concentration with slope $\frac{\sigma-1}{\sigma}$. Two wedges separate the equilibrium from the planner. In the actor, each firm receives only its own component of V^Σ, so it ignores the value its marketing destroys at rivals. This is the *business-stealing wedge*, and by the zero-sum result above it accounts for all marketing at symmetric states. In the flow, the constrained-social objective is $W^C = \ln C^{\mathrm{eq}} - \frac{\sigma-1}{\sigma}(1-\mathrm{HHI})$; its concentration term cancels against the one in Π, leaving real consumption alone, $W^C - \Pi = \ln C^{\mathrm{eq}} - 1$. This *valuation wedge* is the consumer surplus firms cannot collect. The strategic feedback in

(12) is the shadow-value counterpart of the first wedge, and the flow gap $g_i - \phi_{ii}$ nets both; the whole correction therefore collapses into a single marketing-side schedule.

Lemma 1 (A marketing subsidy implements the planner) *For either regime $r \in \{C, F\}$ there is a state-dependent subsidy $\tau_i(\mathbf{x})$ per unit of marketing, financed lump sum, such that the planner's policies form a marketing equilibrium of the subsidized game, in which firm i's flow is $\pi_i + \tau_i(\mathbf{x})\,\eta_i - \frac{\kappa}{2}\eta_i^2$. The schedule equals the sum of the two wedges on the implemented path,*

$$\tau_i(\mathbf{x}) = \underbrace{\frac{\beta}{\nu_b^2}\,\mathbb{E}^P\left[\epsilon_i^b\,(V^P - \tilde{V}^\Sigma)(\mathbf{x}') \,\middle|\, \mathbf{x}\right]}_{\textit{valuation}} + \underbrace{\frac{\beta}{\nu_b^2}\sum_{k\neq i}\mathbb{E}^P\left[\epsilon_i^b\,\tilde{V}_k(\mathbf{x}') \,\middle|\, \mathbf{x}\right]}_{\textit{business stealing}}, \tag{32}$$

where $\tilde{V}_k$ is firm k's value in the subsidized game and $\tilde{V}^\Sigma$ their total; no re-equilibration term appears, because the continuation is the planner's own.

The derivation is in Appendix A.4. The instrument is marketing-side alone, leaving pricing, production, and the laws of motion untouched. Each component of (32) is a closed-form evaluation of fitted critics, so the two wedges can be drawn state by state; Section 5.4 does so after the counterfactual numbers.

5.3 Counterfactual Welfare

We perform four counterfactuals to study the process. Shutting marketing down ($\eta \equiv 0$) keeps Cournot pricing and lets both stocks evolve on their own. Free marketing is the frictionless benchmark: customer capital is reassigned each period to the efficient line $b_i^* = \frac{\alpha_B}{1-\alpha_B}a_i$ of (A15) at no resource cost, with labor allocated efficiently. The two planner rows are the regimes of the previous subsection, feasible marketing policies under Cournot pricing and under efficient labor allocation.

Shutting marketing down raises welfare by 5.2 percent (Table 6, $\eta \equiv 0$): the recovered bill (5.1) and the market-power relief (5.0) outweigh the lost alignment (4.9 points of frontier). The constrained planner does better by allowing some marketing. Its total gain is at least 8.8 points, at least 3.6 above the marketing ban. Redirecting rather than banning marketing turns the frontier loss into a 0.9-point gain and raises the alignment $\mathrm{corr}(a, b)$ by 0.06. Its policy is close to linear in the state: marketing rises with relative productivity and falls with inherited customer capital.

Table 6: Counterfactuals: Welfare Relative to the Equilibrium

	ΔW	$\Delta \ln Z^*$	$-\Delta\big(\ln \bar{\mu}_H + \ln \mathcal{M} + L^{\text{prod}}\big)$	$-\Delta L^{\text{mkt}}$	ΔHHI	$\Delta\text{corr}(a,b)$
equilibrium	0	0	0	0	0	0
$\eta \equiv 0$	+5.2	−4.9	+5.0	+5.1	−.041	−.21
free marketing	+50.2	+23.1	+22.0	+5.1	+.072	+.79
constrained planner	+8.8	+0.9	+3.3	+4.6	−.027	+.06
first-best planner	+30.4	+4.7	+22.0	+3.7	+.001	+.12

Notes: Entries are welfare changes in log points of flow utility, consumption-equivalent percent under log utility; the three split columns (the frontier, the net cost of market power, and the marketing bill) sum to ΔW state by state, by the identity (20). Ergodic means over 2×10^5 states at the calibrated parameters, common shock draws across scenarios. The marketing-bill column is expressed at the 5.1 percent data share of marketing. Planner rows use linear policy rules, so they are feasible lower bounds on the planner's value.

5.4 Corrective Wedges

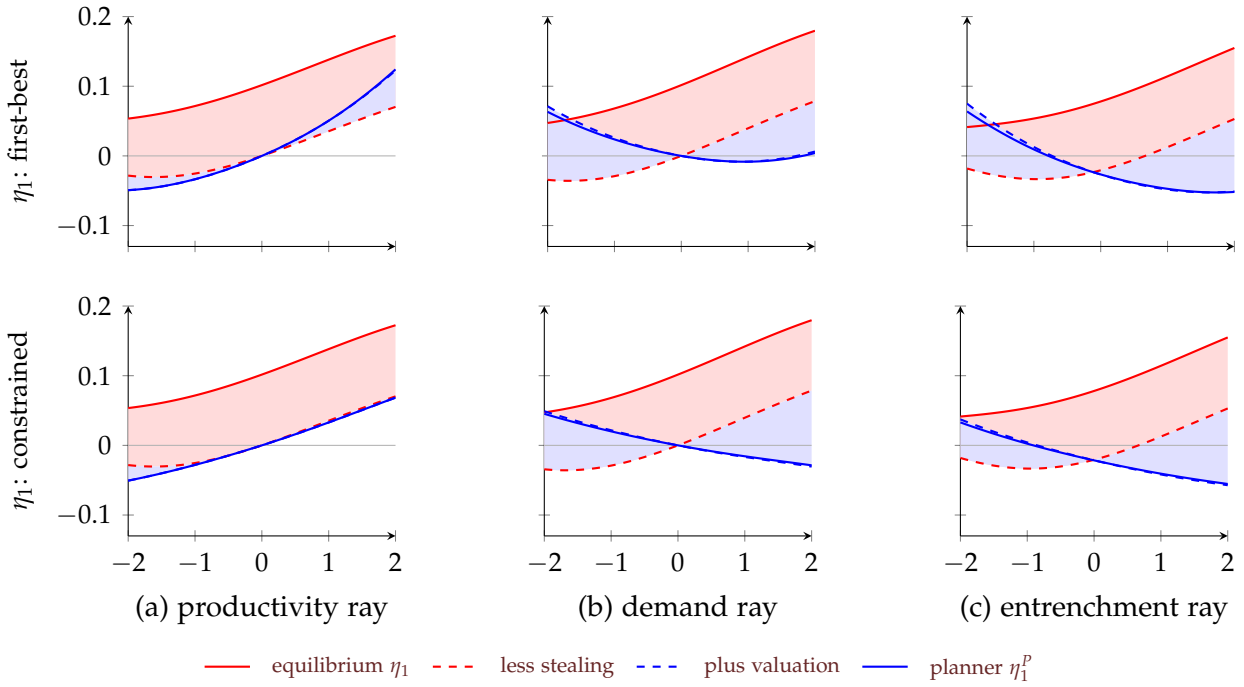
The correction is largest in states where demand is inherited. Figure 7 plots the equilibrium marketing rate against the planner's along three state rays, the first-best regime in the top row and the constrained regime beneath, with the gap split into the two wedges of Lemma 1. In every panel the red band removes business stealing and the blue band then adds the valuation wedge, landing on the planner's rate. At the symmetric state the fitted solutions reproduce the zero-sum result exactly. Both planners stop marketing, and all marketing is business stealing. Along the productivity ray, the planner keeps marketing the productive leader at a rate below the equilibrium's by roughly the stealing wedge, which is nearly flat across states. Along the demand and entrenchment rays, the equilibrium raises its marketing rate with the accumulated advantage while both planners set negative marketing for the entrenched incumbent.

The correction is therefore (a, b)-dependent, and an instrument must tell the two dimensions apart to implement it. The planner conditions on the split directly; a policymaker who observes only size cannot, and any payment that conditions on size rewards the composite, accumulated stock included. The next subsection shows what that constraint costs.

5.5 Size-Based Policy

Now consider a policymaker who observes size but not its source, and who pays a size-based output subsidy offsetting a fraction $\lambda \in [0, 1]$ of each firm's log markup (construction and closed forms in Appendix A.6). At $\lambda = 0$ the subsidy is off; at $\lambda = 1$

Figure 7: The Correction in Investment Rates: Equilibrium versus Planner



Notes: Firm 1's fitted marketing rate in the equilibrium (red) and under the planner (blue) along state rays from the symmetric state, top row first-best, bottom row constrained; horizontal axes: the ray deviation, (a) a_1 pushed, (b) b_1 pushed, (c) b_1 pushed with $a_1 < 0$. Red band: the business-stealing component of the gap; blue band: the valuation component; dashed: the equilibrium rate with each component applied in turn. Fitted solutions at the calibrated parameters. Source: model computations.

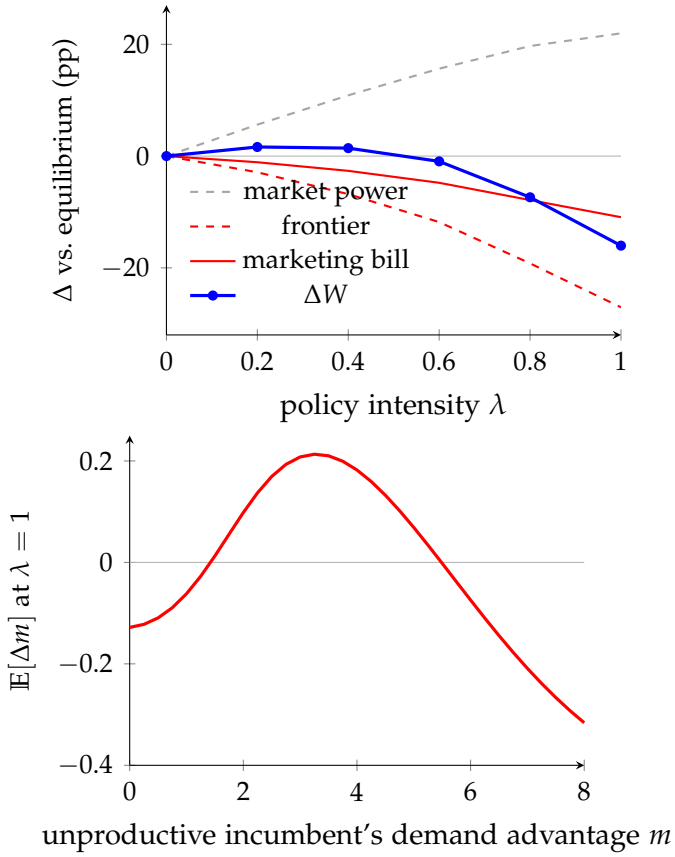
pricing is at marginal cost and the statics are the first-best allocation given $\mathbf{z}$, the same statics the first-best planner implements. The static allocation improves by construction at every λ, so the question is how the marketing game responds to the reshaped incentives. Every λ-economy is solved at the same discount $\beta = 0.96$ as the equilibrium, on common shock draws.

At $\lambda = 1$ welfare falls 16.0 points relative to the equilibrium.[16] The static gain is 22.0 points, but the subsidy rewards size, and firms acquire size through marketing: the private value of customer capital rises by half, the marketing bill triples, the frontier falls, and effective firms drop from four to one and a half.

Along the λ-path (Figure 8, left panel) the welfare change peaks near $\lambda = 0.2$ at $+1.6$ points, crosses zero near $\lambda = 0.55$, and falls steeply thereafter as intensified marketing spending consumes it. The optimal λ corrects a fifth of the markup. The right panel sets $\lambda = 1$: for an unproductive incumbent, the expected drift of its demand advantage turns

[16] The marketing-bill series is expressed at the 5.1 percent data share, with the residual marketing cost folded into the frontier series; levels are not directly comparable to Table 6.

Figure 8: The λ-Subsidy: Welfare Path and Entrenchment Drift



Notes: Left: welfare change and its three channels along the subsidy path, ergodic means of each λ-economy against the equilibrium, common shock draws; the marketing-bill series is at the 5.1 percent data share with the residual carried in the frontier series (footnote in the text). Right: expected one-period change of an unproductive incumbent's demand advantage m at $\lambda = 1$, its productivity gap fixed at -1, rivals symmetric. At $\lambda = 1$ the closed-loop AR(1) of the leadership gap is 0.98. Source: model computations.

positive over a wide range, so the subsidy sustains leads that the laissez-faire equilibrium never generates.

The planner's policy locates the social problem in the *size incentive* rather than in marketing itself; the planner still markets. Private returns rise with the accumulated stock while social returns fall with it, so instruments that reward size, output subsidies included, push marketing in the wrong direction, the mechanism behind the backfire above. And because the loss the game sustains is the entrenchment component of the measured misallocation, policies that only compress markups leave the dynamic distortion in place; the corrective margin is the matching between demand and productivity rather than the volume of marketing.

6 Conclusion

This paper studies market concentration through the lens of accumulated customer demand. In an oligopolistic setting, marketing can entrench incumbents. Positive productivity shocks in the past can lead to accumulated demand that can become self-sustaining. These legacy effects can misallocate the market and create a disconnection between firm size and productivity. The dynamics of accumulation sustain a significant share of the misallocation we measure between customer capital and productivity, and concentration reflects past marketing as much as current productivity. We term these forces *legacy effects*, which emerge through firms' spending on marketing having significant path dependency even if productivity falls. We develop this in an oligopolistic model to study standard policies.

Quantitatively, we develop some insights that may be useful for economists studying the market for competition. Standard size-based policies to alleviate markups may reward accumulating demand, which can entrench incumbents; correcting this classic static markup distortion can deepen the dynamic one. Policies aimed at concentration need to keep in mind the *sources* of concentration. Demand-driven concentration can be costly and persistent, whereas productivity-driven concentration tends to have initially positive consequences.

This paper does not speak to the endogenous feedback loop between demand and productivity, and the many roles of customer acquisition at different stages of the firm life cycle. These are exciting areas for future research in the world of market concentration, productivity, and development. We expect the demand side of firms will remain an important element for studies on a wide range of topics.

References

Afrouzi, H., A. Drenik, and R. Kim: 2023, 'Concentration, Market Power, and Misallocation: The Role of Endogenous Customer Acquisition'. Technical report, National Bureau of Economic Research. Forthcoming, Econometrica.

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt: 2005, 'Competition and Innovation: An Inverted-U Relationship'. *Quarterly Journal of Economics* **120**(2), 701–728.

Akcigit, U. and S. T. Ates: 2023, 'What happened to US business dynamism?'. *Journal of Political Economy* **131**(8), 2059–2124.

Amiti, M., O. Itskhoki, and J. Konings: 2019, 'International Shocks, Variable Markups, and Domestic Prices'. *The Review of Economic Studies* **86**(6), 2356–2402.

Arellano, M. and S. Bond: 1991, 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations'. *Review of Economic Studies* **58**(2), 277–297.

Argente, D., D. Fitzgerald, S. Moreira, and A. Priolo: 2025, 'How Do Entrants Build Market Share? The Role of Demand Frictions'. *American Economic Review: Insights*.

Argente, D., M. Lee, and S. Moreira: 2024, 'The Life Cycle of Products: Evidence and Implications'. *Journal of Political Economy* **132**(2).

Atkeson, A. and A. Burstein: 2008, 'Pricing-to-Market, Trade Costs, and International Relative Prices'. *American Economic Review* **98**(5), 1998–2031.

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

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

Bornstein, G.: 2019, 'Entry and Profits in an Aging Economy: The Role of Consumer Inertia'. *Review of Economic Studies*. Forthcoming.

Bornstein, G. and A. Peter: 2025, 'Nonlinear Pricing and Misallocation'. *American Economic Review* **115**(11), 3853–3908.

Bronnenberg, B. J., S. K. Dhar, and J.-P. Dubé: 2009, 'Brand History, Geography, and the Persistence of Brand Shares'. *Journal of Political Economy* **117**(1), 87–115.

Bronnenberg, B. J., J.-P. Dubé, and M. Gentzkow: 2012, 'The Evolution of Brand Preferences: Evidence from Consumer Migration'. *American Economic Review* **102**(6), 2472–2508.

Bronnenberg, B. J., J.-P. Dubé, and C. Syverson: 2022, 'Marketing Investment and Intangible Brand Capital'. *Journal of Economic Perspectives* **36**(3), 53–74.

Cavenaile, L., M. A. Celik, J. Perla, and P. Roldan-Blanco: 2023, 'A Theory of Dynamic Product Awareness and Targeted Advertising'. *Working Paper*.

Cavenaile, L., M. A. Celik, P. Roldan-Blanco, and X. Tian: 2025, 'Style over substance? advertising, innovation, and endogenous market structure'. *Journal of Monetary Economics* **149**, 103683.

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

Crouzet, N. and J. C. Eberly: 2019, 'Understanding weak capital investment: The role of market concentration and intangibles'. Technical report, National Bureau of Economic Research.

David, J. M. and V. Venkateswaran: 2019, 'The Sources of Capital Misallocation'. *American Economic Review* **109**(7), 2531–67.

De Loecker, J. and J. Eeckhout: 2018, 'Global Market Power'. National Bureau of Economic Research Working Paper 24768.

De Loecker, J., J. Eeckhout, and G. Unger: 2020, 'The Rise of Market Power and the Macroeconomic Implications*'. *Quarterly Journal of Economics* **135**(2), 561–644.

De Ridder, M.: 2024, 'Market power and innovation in the intangible economy'. *American Economic Review* **114**(1), 199–251.

Dinlersoz, E. M. and M. Yorukoglu: 2012, 'Information and Industry Dynamics'. *American Economic Review* **102**(2), 884–913.

Drozd, L. A. and J. B. Nosal: 2012, 'Understanding International Prices: Customers as Capital'. *American Economic Review* **102**(1), 364–395.

Edmond, C., V. Midrigan, and D. Y. Xu: 2023, 'How Costly Are Markups?'. *Journal of Political Economy* **131**(7), 1619–1675.

Einav, L., P. Klenow, J. D. Levin, and M.-G. Ravid: 2021, 'Customers and Retail Growth'. Working Paper.

Eslava, M., J. Haltiwanger, and N. Urdaneta: 2024, 'The size and life-cycle growth of plants: The role of productivity, demand, and wedges'. *Review of Economic Studies* **91**(1), 259–300.

Fitzgerald, D., S. Haller, and Y. Yedid-Levi: 2024, 'How Exporters Grow'. *The Review of Economic Studies* **91**(4), 2276–2306.

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

Foster, L., J. Haltiwanger, and C. Syverson: 2016, 'The Slow Growth of New Plants: Learning about Demand?'. *Economica* **83**(329), 91–129.

Gabaix, X.: 2011, 'The granular origins of aggregate fluctuations'. *Econometrica* **79**(3), 733–772.

Gourio, F. and L. Rudanko: 2014, 'Customer Capital'. *The Review of Economic Studies* **81**(3), 1102–1136.

Haskel, J. and S. Westlake: 2017, *Capitalism without capital: the rise of the intangible economy*. Princeton University Press.

He, B., L. I. Mostrom, and A. Sufi: 2024, 'Investing in Customer Capital'. Technical report, National Bureau of Economic Research.

Heise, S.: 2024, 'Firm-to-Firm Relationships and the Pass-Through of Shocks: Theory and Evidence'. *Review of Economics and Statistics*.

Hopenhayn, H. A.: 1992, 'Entry, Exit, and Firm Dynamics in Long Run Equilibrium'. *Econometrica* **60**(5), 1127–1150.

Hottman, C. J., S. J. Redding, and D. E. Weinstein: 2016, ' Quantifying the Sources of Firm Heterogeneity *'. *The Quarterly Journal of Economics* **131**(3), 1291–1364.

Hsieh, C.-T. and P. J. Klenow: 2009, 'Misallocation and Manufacturing TFP in China and India'. *Quarterly Journal of Economics* **124**(4), 1403–1448.

Hubmer, J. and L. Nord: 2025, 'Investment in Demand and Dynamic Competition for Customers'. Working paper. Available at SSRN: https://ssrn.com/abstract=5491506.

Ignaszak, M. and P. Sedlácek: 2022, 'Customer acquisition, business dynamism and aggregate growth'. Technical report, Working Paper.

Jovanovic, B.: 1982, 'Selection and the Evolution of Industry'. *Econometrica* **50**(3), 649–70.

Kaplan, G. and P. Zoch: 2020, 'Markups, Labor Market Inequality and the Nature of Work'. NBER Working Paper 26800, National Bureau of Economic Research.

Kehrig, M. and N. Vincent: 2021, 'The micro-level anatomy of the labor share decline'. *The Quarterly Journal of Economics* **136**(2), 1031–1087.

Klette, T. J. and S. Kortum: 2004, 'Innovating Firms and Aggregate Innovation'. *Journal of Political Economy* **112**(5), 986–1018.

Liu, E., A. Mian, and A. Sufi: 2022, 'Low Interest Rates, Market Power, and Productivity Growth'. *Econometrica* **90**(1), 193–221.

Olmstead-Rumsey, J.: 2022, 'Market concentration and the productivity slowdown'. *Working Paper*.

Pearce, J. and L. Wu: 2024, 'Brand Reallocation and Market Concentration'. Staff Report 1116, Federal Reserve Bank of New York.

Peters, M.: 2020, 'Heterogeneous Markups, Growth, and Endogenous Misallocation'. *Econometrica* **88**(5), 2037–2073.

Restuccia, D. and R. Rogerson: 2008, 'Policy Distortions and Aggregate Productivity with Heterogeneous Establishments'. *Review of Economic Dynamics* **11**(4), 707–720.

Wu, L.: 2024, 'Partially directed search in the labor market'. *Journal of Economic Theory* **220**, 105858.

Appendix

The Appendix contains two sections: the Model Appendix (A) collects the model's proofs and derivations, and Appendix B covers the estimation, moment conditions, and robustness.

A Model Appendix

This appendix collects the proofs and derivations for the model of Section 3: the static-equilibrium objects (Appendix A.1), the Stein representation and the critic family behind the computational method (Appendix A.2), the two-firm results (Appendix A.3), the welfare accounting, the planner's solution, and the Hsieh–Klenow bridge (Appendix A.4), the rational-inattention microfoundation (Appendix A.5), the policy extensions (Appendix A.6), and aggregation with the household problem (Appendix A.7).

A.1 Static Equilibrium: Shares, Marginal Profits, and Output

Share Jacobian. Totally differentiating the share fixed point (7) with respect to a perturbation db_i, holding the other states constant, gives for the focal firm

$$\frac{ds_i}{s_i} = db_i - (\sigma-1)\frac{ds_i}{1-s_i} - \sum_{\ell=1}^{I} s_\ell \Big[db_\ell - (\sigma-1)\frac{ds_\ell}{1-s_\ell}\Big],$$

and the same expression without the db_i term for every rival. Guessing a solution proportional to the share responsiveness $\chi_i \equiv \left(\frac{1}{s_i} + \frac{\sigma-1}{1-s_i}\right)^{-1}$, with normalized responsiveness $\hat{\chi}_i \equiv \chi_i / \sum_\ell \chi_\ell$, and verifying,

$$\begin{aligned} \frac{\partial s_i}{\partial b_i} &= \chi_i\,(1-\hat{\chi}_i), \\ \frac{\partial s_j}{\partial b_i} &= -\chi_j\,\hat{\chi}_i \quad (j \neq i), \end{aligned} \tag{A1}$$

which satisfies the adding-up check $\sum_j \partial s_j/\partial b_i = \chi_i - \chi_i = 0$. The responsiveness χ_i is hump-shaped in the share: firms at intermediate shares respond most to customer capital, dominant and marginal firms least. Because profits (8) depend on the own share alone, with marginal profit $\pi'(s) = \frac{1}{\sigma} + 2\frac{\sigma-1}{\sigma}s$, the marginal-profit matrix of customer

capital is

$$\phi_{ik}(\mathbf{z}) \equiv \frac{\partial \pi_i}{\partial b_k} = \pi'(s_i)\,\chi_i\,(\mathbb{I}_{i=k} - \hat{\chi}_k), \tag{A2}$$

the flow term of the envelope condition (11).

Labor pools. With unit expenditure, firm i's revenue is s_i and its production labor is revenue over the markup, $\ell_i = s_i/\mu_i = \frac{\sigma-1}{\sigma} s_i(1-s_i)$; summing across firms delivers (18).

Equilibrium output. Under markup pricing $p_i = \mu_i e^{-a_i/(\sigma-1)}$ with $w = 1$ and $E = 1$, revenue $p_i c_i = s_i$ gives $c_i = \frac{\sigma-1}{\sigma}(1-s_i)\,s_i\,e^{a_i/(\sigma-1)}$. Substituting into the aggregator (3) and using the share fixed point in the form $e^{z_i/\sigma} = s_i^{1/\sigma}(1-s_i)^{-(\sigma-1)/\sigma} N^{1/\sigma}$, where $N(\mathbf{z}) \equiv \sum_j e^{z_j}(1-s_j)^{\sigma-1}$,

$$\omega_i^{1/\sigma} c_i^{\frac{\sigma-1}{\sigma}} = \left(\tfrac{\sigma-1}{\sigma}\right)^{\frac{\sigma-1}{\sigma}} s_i \left[\frac{N}{\sum_j e^{b_j}}\right]^{1/\sigma},$$

so the shares sum out and

$$C^{\mathrm{eq}}(\mathbf{z}) = \frac{\sigma-1}{\sigma}\left[\frac{N(\mathbf{z})}{\sum_j e^{b_j}}\right]^{\frac{1}{\sigma-1}} = \frac{\sigma-1}{\sigma}\left[\sum_j \omega_j\, e^{a_j}(1-s_j)^{\sigma-1}\right]^{\frac{1}{\sigma-1}}. \tag{A3}$$

The frictionless allocation $\ell_i^* \propto e^{z_i}$ (maximizing C at fixed total labor) delivers $C^* = (\sum_j \omega_j e^{a_j})^{1/(\sigma-1)} L = Z^* L$, the frontier index of (19) times labor.

A.2 Computation: Stein Representation and the Critic Family

Level representation. Stein's lemma (15) applied to the first-order and envelope conditions (11) gives the critic–actor system (16) (the *critic* is the value recursion, the *actor* the policy condition) on the stacked state $\mathbf{x} = (\mathbf{a}, \mathbf{b})$. Higher derivatives reduce to levels in the same way, with Hermite weights, $\mathbb{E}[\partial^2 f/\partial b_k' \partial b_i' \,|\, \mathbf{x}] = \mathbb{E}[(\epsilon_k^b \epsilon_i^b - \nu_b^2 \mathbb{I}_{k=i}) f(\mathbf{x}') \,|\, \mathbf{x}]/\nu_b^4$; differentiating the actor equation and applying this identity shows that the strategic slopes are shock-weighted levels to leading order in the policy slopes,

$$\frac{\partial \eta_k(\mathbf{x})}{\partial b_i} \approx \frac{\beta \delta_b}{\kappa \nu_b^4}\,\mathbb{E}\big[(\epsilon_k^b \epsilon_i^b - \nu_b^2 \mathbb{I}_{k=i})\, V_k(\mathbf{x}') \,\big|\, \mathbf{x}\big].$$

Interiority (the second-order condition) also takes a level form, $\frac{\beta}{\nu_b^2}\,\mathbb{E}[((\epsilon_i^b)^2/\nu_b^2 - 1) V_i(\mathbf{x}') \,|\, \mathbf{x}] < \kappa$, which we verify at every solution. Because $\mathbf{x}'$ contains $\boldsymbol{\eta}(\mathbf{x})$, the action vector at each

state solves a small fixed point per evaluation; the squared policy term keeps the pair from being a monotone contraction, so we iterate a damped critic–actor alternation under a discount-continuation schedule.

Index-function reduction. Let T_w denote the Gaussian smoothing operator, $(T_w f)(y) \equiv \mathbb{E}[f(y + w\xi)]$ with $\xi \sim N(0,1)$, which satisfies the semigroup property $T_w \circ T_{w'} = T_{\sqrt{w^2+w'^2}}$. For a direction $d = (d^a, d^b)$, mean zero within each stock, and an absolutely continuous profile f of polynomial growth, the index $d \cdot \mathbf{x}'$ is univariate Gaussian conditional on $\mathbf{x}$, $d \cdot \mathbf{x}' \sim N(d \cdot \mathbf{x}^e,\, v_d^2)$ with the expected next state $\mathbf{x}^e(\mathbf{x}) \equiv (\delta_a \mathbf{a},\ \delta_b \mathbf{b} + \boldsymbol{\eta}(\mathbf{x}))$ and shock loading $v_d^2 \equiv \nu_a^2 |d^a|^2 + \nu_b^2 |d^b|^2$, and

$$\begin{aligned} \mathbb{E}\left[f(d \cdot \mathbf{x}') \,\middle|\, \mathbf{x} \right] &= (T_{v_d} f)(d \cdot \mathbf{x}^e), \\ \mathbb{E}\left[\epsilon_k^b \, f(d \cdot \mathbf{x}') \,\middle|\, \mathbf{x} \right] &= \nu_b^2 \, d_k^b \, (T_{v_d} f)'(d \cdot \mathbf{x}^e). \end{aligned} \tag{A4}$$

Hence if the profile family is closed under T_w with an explicit parameter map (*conjugacy*), every expectation entering the critic and the actor is an evaluation of a member of the same family, and the multivariate problem requires only this one-dimensional property. Polynomials are conjugate through Gaussian moments (degree two is the linear-quadratic block, which serves as the warm start), and probits are conjugate through cdf smoothing,

$$\begin{aligned} \mathbb{E}\left[\Phi\left(\tfrac{d \cdot \mathbf{x}' - \bar{x}}{w} \right) \middle| \mathbf{x} \right] &= \Phi\left(\frac{d \cdot \mathbf{x}^e - \bar{x}}{\sqrt{w^2 + v_d^2}} \right), \\ \mathbb{E}\left[\epsilon_k^b \, \Phi\left(\tfrac{d \cdot \mathbf{x}' - \bar{x}}{w} \right) \middle| \mathbf{x} \right] &= \frac{\nu_b^2 \, d_k^b}{\sqrt{w^2 + v_d^2}} \, \Phi'\left(\frac{d \cdot \mathbf{x}^e - \bar{x}}{\sqrt{w^2 + v_d^2}} \right), \end{aligned} \tag{A5}$$

so with the family (17) the actor is an explicit function of the coefficients. The derivative of a probit is a Gaussian density, which is how the sigmoid terms of the value function generate the hump-shaped policies of the two-firm analysis.

Symmetry and orbit-tying. The primitives are exchangeable, so symmetric-equilibrium values are equivariant, $V_{\varrho(i)}(\varrho \cdot \mathbf{x}) = V_i(\mathbf{x})$ for every permutation ϱ of the firms (acting on both stocks at once); equivariance across firms is imposed by construction (firm j's members are firm i's with indices transposed), leaving the rival-permutation invariance $V_i(\varrho \cdot \mathbf{x}) = V_i(\mathbf{x})$ for ϱ fixing i. A single index member $f(d \cdot \mathbf{x})$ with non-constant f

satisfies this only if the rival weights of d are all equal (the own-index direction) or the orbit is a reflection pair with f even (the rival-gap quadratic), because invariance forces $f(d \cdot \mathbf{x}) = f(d' \cdot \mathbf{x})$ for every d' in the orbit, two non-parallel directions would force f to be constant, and permutations preserve the norm. Free directions therefore enter only through the group average over rival permutations,

$$G_d(\mathbf{x}) = \sum_{\varrho} f(d_\varrho \cdot \mathbf{x}), \tag{A6}$$

the orbit sum with one tied coefficient. The orbit sum is invariant by construction; conjugacy is preserved because $|d_\varrho| = |d|$, so the smoothing acts term by term at a common width and commutes with the symmetrization; and no flexibility is lost, because the group average is a norm-nonexpansive linear projection onto the invariant functions, so density of the untied family implies density of the tied family in the invariant class. The main pattern we use is the distinguished-rival direction (own weight, one rival's weight, and a common weight on the rest), whose orbit has size $I - 1$, so its cost is linear in I; it covers the firm's own position, each bilateral rivalry, and the rest-of-market aggregate.

Fitting as regression. With the actions frozen at the current iterate ($\bar{\boldsymbol{\eta}}$, $\mathbf{x}^e = (\delta_a \mathbf{a},\ \delta_b \mathbf{b} + \bar{\boldsymbol{\eta}}(\mathbf{x}))$), the Bellman residual of (16) rearranges to the linear model

$$\pi_i(\mathbf{a}+\mathbf{b}) - \tfrac{\kappa}{2}\bar{\eta}_i(\mathbf{x})^2 = \theta_0(1-\beta) + \sum_m \theta_m \left[f_m(d_m \cdot \mathbf{x}) - \beta \left(T_{v_{d_m}} f_m\right)(d_m \cdot \mathbf{x}^e) \right] - R(\mathbf{x};\theta), \tag{A7}$$

a weighted least-squares regression of flow payoffs on annuitized basis functions, with each regressor equal to the basis function minus β times its smoothed one-step-ahead value. The algorithm alternates the regression with the action update until the two reach a joint fixed point. This is the least-squares temporal-difference regression of reinforcement learning, except that the next-period term is computed in closed form rather than sampled; the double-sampling (errors-in-variables) bias of that estimator is therefore absent, and weighted least squares is exact. The steps are: (1) simulate an ergodic panel under an initial policy (the linear-quadratic solution), keeping an out-of-sample subsample; (2) minimize the weighted Bellman residual by Levenberg–Marquardt with analytic Jacobians, multi-started from the linear-quadratic null with a ridge on θ; (3) resimulate under the fitted policy and refit until the sample and coefficients are jointly stable; (4) audit: the out-of-sample residual, the distance to the exact small-I solver (which also calibrates what residual level counts as accurate where no exact solve exists, since a Bellman residual bounds the value error only through $\|R\|/(1-\beta)$), the interiority check, and

the stationarity certificate on every pairwise gap; (5) on failure, restart along a shorter discount ladder, or add the single member most correlated with the residual. The smallest M that passes the audit measures the equilibrium's nonlinearity: $M = 0$ means the linear-quadratic benchmark suffices, and each additional member adds one sigmoidal component to the value function.

Conjugate families and the basis used. Other families are also conjugate under T_w but less useful in practice: exponentials are ill conditioned on wide ergodic supports, and rank-based symmetrizations are not conjugate, because sorting is only piecewise linear in the shock. The calibrated solves use the linear-quadratic block on the own- and rival-mean indices of each stock together with orbit-tied probit members (A6) on the pairwise composite gaps $(z_i - z_j)/2$ (the distinguished-rival pattern), adding members while the out-of-sample residual remains large.

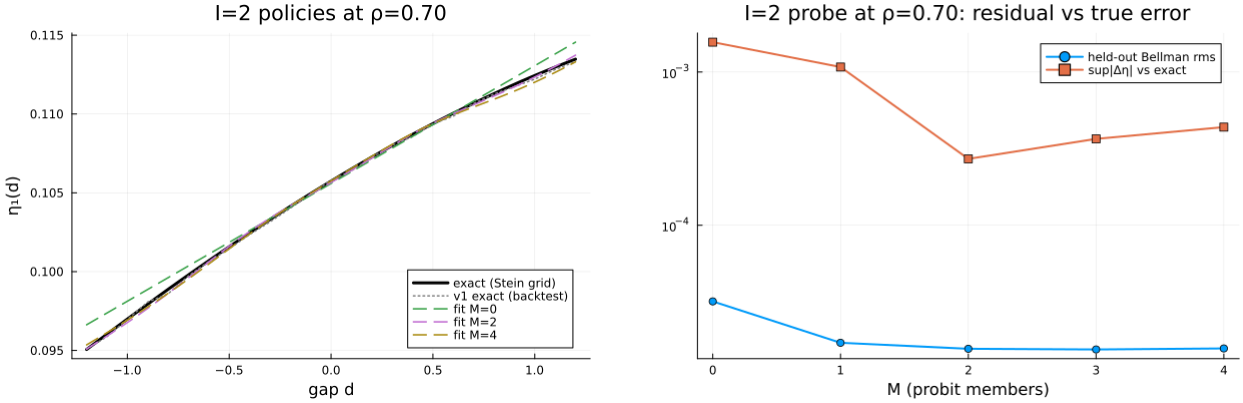
Approximation accuracy. The audit against exact solutions runs on a test game with a common retention ($\sigma = 3.9$, $\delta = 0.87$, $\kappa = 1.1$, $\nu = 0.1$), where exact solvers exist at small firm counts: a one-dimensional gap solver at $I = 2$ and a two-dimensional collocation solver at $I = 3$ (its own permutation-symmetry error 2.7×10^{-4}). Table A1 reports the fitted family against these benchmarks and its out-of-sample Bellman residuals through $I = 5$; Figure A1 overlays the fitted and exact policies at $I = 2$ and traces the out-of-sample residual as probit members are added. The fitted policies are within a few parts in ten thousand of the exact ones, and the residual flattens after two members, so the measured nonlinearity of this game is $M = 2$. At $\beta = 0.95$ the audits pass at $I = 2$ and 3 while the closed loop nears a unit root at $I \geq 4$, where the stationarity certificate is the binding audit; the calibrated split-retention solves at $\beta = 0.96$ report their own out-of-sample residuals with each results table.

Table A1: Accuracy of the Critic Family against Exact Solutions

	exact benchmark	out-of-sample rms residual	sup $\lvert\Delta\eta\rvert$ vs. exact
$I = 2$	one-dimensional solve	1.6×10^{-5}	2.8×10^{-4}
$I = 3$	two-dimensional solve	1.7×10^{-5}	5.9×10^{-4}
$I = 4$	—	3.0×10^{-5}	—
$I = 5$	—	4.7×10^{-5}	—

Notes: Test game $\sigma = 3.9$, $\delta = 0.87$, $\kappa = 1.1$, $\nu = 0.1$, discount anchored at $\rho = 0.70$ where the exact solvers converge; orbit-tied pairwise-gap probit basis; second-order conditions verified at every row. No exact solver exists beyond $I = 3$; there the out-of-sample residual and the stationarity certificate provide the audit.

Figure A1: The $I = 2$ Audit: Policies and the Residual Ladder



Notes: Left: fitted marketing policies at $M = 0, 2, 4$ probit members against the exact solutions at $I = 2$. Right: the out-of-sample Bellman residual and the true policy error sup $|\Delta\eta|$ against the exact solution as members are added; both flatten after $M = 2$, and the residual tracks the true error. Test-game parameters as in Table A1. Source: model computations.

A.3 Two-Firm Case: Proofs

Numerical statements in Section 3.5 and in this appendix use the fitted equilibrium at the illustrative parameters $\sigma = 3.9$, $\delta = 0.87$, $\kappa = 1.1$, $\beta = 0.95$, $\nu_a = 0.1$, $\nu_b = 0$, simulated over $T = 200{,}000$ periods; the comparative statics in Figures 1 and 2 state their own parameters in the notes.

Logistic shares (13). The fixed point implies $\ln \frac{s_1}{s_2} - (\sigma - 1) \ln \frac{1-s_1}{1-s_2} = \zeta$; with $s_2 = 1 - s_1$ the left side is $\sigma \ln \frac{s_1}{1-s_1}$. Invert and differentiate; $s'' = s'(1 - 2s)/\sigma$ vanishes at the tie.

Proof of Proposition 1. By (13), $\zeta \mapsto s$ is a strictly increasing bijection, so maximize $s(1-s)[1 + 2(\sigma - 1)s]$ over s. Its derivative, $1 - 2s + 4(\sigma - 1)s - 6(\sigma - 1)s^2$, is a concave parabola, positive at $s = \frac{1}{2}$ (value $\frac{\sigma-1}{2}$) and negative at $s = 1$ (value $-1 - 2(\sigma - 1)$): exactly one root in $(\frac{1}{2}, 1)$, the stated s^*. Globally, the reflection identity $h(\zeta) - h(-\zeta) = s(1-s)[\pi'(s) - \pi'(1-s)]/\sigma$ keeps the leader branch above the laggard branch pointwise when $\pi'' > 0$. At $\sigma = 3.9$ the peak share is $s^* = 0.64$, and the fitted equilibrium policy peaks at the larger lead $\zeta \approx 4.7$. The incentive peak is a property of the policy profile rather than the ergodic distribution: the gap drifts by $(\delta - 1)\zeta + \Delta\eta(\zeta)$, and at the baseline parameters the fitted policy slope never exceeds 0.03 while depreciation pulls at $1 - \delta = 0.13$, so the drift is strictly negative for $\zeta > 0$ and the stationary law is unimodal at the tie. An interior mode requires $\Delta\eta(\bar{\zeta}) = (1 - \delta)\bar{\zeta}$ with the policy crossing from above, the threshold displayed in Figure 1.

Proof of Proposition 2. To first order in β, the first-order condition gives $\kappa\eta_1(\zeta) = \beta\,\mathbb{E}[h(\delta\zeta + \epsilon_\zeta)] + O(\beta^2)$. Differencing across the two firms, differentiating at the tie, and applying Stein's lemma together with the reflection $s(-\zeta) = 1 - s(\zeta)$ yields the displayed formula, with $h(\epsilon) - h(-\epsilon) = \frac{2(\sigma-1)}{\sigma^2} s(1-s)(2s-1) \geq 0$ and equality exactly when π is linear in s. Locally, $h'(0) = s'(0)^2\pi''(\bar{s})$ by the logistic lemma: the feedback is the shock-weighted asymmetry of the flow incentive. Fitted equilibria put γ at 0.017 at $\beta = 0.70$ and 0.024 at $\beta = 0.95$ (closed-loop persistence $\delta + \gamma = 0.894$, variance amplification $(1-\delta^2)/(1-(\delta+\gamma)^2) = 1.22$).

Atomistic benchmark. Let a firm be atomistic under a Kimball demand with local elasticity $\sigma(x)$ at relative size x, constant marginal cost, the state entering as an effective price pe^{-z}, and the price set optimally. The envelope theorem gives $d\ln\pi/dz = \sigma(x)$ exactly, hence

$$\frac{d^2\ln\pi}{dz^2} = \sigma'(x)\cdot\frac{d\ln x}{dz}, \qquad \text{where} \qquad \frac{d\ln x}{dz} = \frac{\sigma(\sigma-1)}{\sigma - 1 + \varepsilon_\mu} > 0,$$

with ε_μ the superelasticity. An atomistic markup varies only through $\sigma'(x)$, so the proportional return to the state declines in size if and only if markups rise with size: the atomistic analogue of γ is nonpositive whenever markups are (weakly) increasing in size. Amplification requires the wedge between the perceived and the physical elasticity that only internalization creates; under Cournot–CES the physical elasticity is constant while the margin rises with the share.

Proof of Proposition 3 and the closed forms (14). The wage bill is $w\ell_i = \frac{\sigma-1}{\sigma} s_i(1-s_i)$ (Appendix A.1); $s_2 = 1 - s_1$ makes $s_1(1-s_1) = s_2(1-s_2)$ an identity, so $\ell_1 = \ell_2$ at every state. (Bertrand instead gives $\ell_1/\ell_2 = (\sigma s_1 + s_2)/(\sigma s_2 + s_1) > 1$; at $I \geq 3$ the map $\ell_i \propto s_i(1-s_i)$ is humped and differs from every efficient allocation generically.) For the indices, the planner's problem is strictly concave with $\ell_i^* \propto e^{z_i}$, the equilibrium sets $\ell_i = L/2$, and the common constants cancel in the ratios, leaving the cosh expressions in (14). The exact facts follow by differentiation: $(\ln\mathcal{M})'(\zeta) = [\tanh(\zeta/2) - \tanh(\zeta/2\sigma)]/[2(\sigma-1)] > 0$ for $\zeta > 0$; only Z^* depends on the source of the gap, and at general I it equals the customer-capital-weighted mean of productivity; and $\ln Z^{\text{eq}} = \ln Z^* - \ln\mathcal{M}$. Corollary 1 is the partial-derivative computation, and the

no-customer-capital statement follows from

$$\frac{d\ln Z^*}{d\alpha} - \frac{d\ln \mathcal{M}}{d\alpha} = \frac{\tanh(\alpha/2\sigma)}{2(\sigma-1)} > 0 \quad \forall\, \alpha > 0,$$

so the frontier gain dominates the markup wedge pointwise and $Z^{\text{eq}} \geq 1$ whenever $\mathbf{b} \equiv 0$.

Ergodic joint distribution. In gap coordinates the dynamics are exact: $\alpha' = \delta\alpha + \epsilon_\alpha$, $\zeta' = \delta\zeta + \Delta\eta(\zeta) + \epsilon_\alpha + \epsilon_\beta$, $\beta_b = \zeta - \alpha$; with $\nu_b = 0$ the demand gap has the representation $\beta_{b,t} = \sum_{k\geq 0} \delta^k \Delta\eta(\zeta_{t-1-k})$, accumulated past marketing. Replacing $\Delta\eta$ by its linearization $\gamma\zeta$ yields a Gaussian system with closed-loop persistence $\rho_\zeta = \delta + \gamma$ and stationary moments

$$\begin{aligned} \sigma_\alpha^2 &= \frac{2\nu_a^2}{1-\delta^2}, & \sigma_{\alpha\zeta} &= \frac{2\nu_a^2}{1-\delta\rho_\zeta}, \\ \sigma_\zeta^2 &= \frac{2(\nu_a^2+\nu_b^2)}{1-\rho_\zeta^2}, & \sigma_{\alpha\beta_b} &= \frac{2\nu_a^2\,\delta\gamma}{(1-\delta\rho_\zeta)(1-\delta^2)}, \end{aligned} \tag{A8}$$

so the sign of the alignment covariance is the sign of γ: zero under constant markups, positive here. Simulation of the fitted nonlinear policy ($T = 200{,}000$) matches the closed forms, with $\text{sd}(\beta_b)$ 0.047 against 0.044, $\text{corr}(\alpha, \beta_b)$ 0.618 against 0.620, and $\text{corr}(\alpha, \zeta)$ 0.993 against 0.994; the linearization is tight because the fitted policy is near-linear on the ergodic support. The entrenchment region follows from the closed forms: $Z^* < 1$ if and only if $\alpha(\alpha + 2\beta_b) < 0$. Simulated frequencies, with the Gaussian arcsine values $\Pr = \frac{1}{2} - \arcsin(\text{corr})/\pi$ in parentheses:

$$\Pr(\alpha\beta_b < 0) = 0.288\ (0.287), \qquad \Pr(\text{sign}\,\zeta \neq \text{sign}\,\alpha) = 0.037\ (0.035).$$

Evaluating the exact indices on the simulated distribution gives $\mathbb{E}[\ln Z^*] = 0.0042$, $\mathbb{E}[\ln \mathcal{M}] = 0.0031$, $\mathbb{E}[\ln Z^{\text{eq}}] = 0.0010$, against 0.0009 in the no-marketing benchmark on the same shocks: a net allocative gain of 1.1×10^{-4}, while the ergodic marketing bill absorbs roughly 13% of revenue. To second order, each term reproducing its simulated value within 2%,

$$\mathbb{E}\left[\ln Z^{\text{eq}}\right] \approx \frac{\sigma_\alpha^2 + 2\sigma_{\alpha\beta_b}}{8\sigma(\sigma-1)} - \frac{\sigma_{\beta_b}^2}{8\sigma}\,: \tag{A9}$$

productivity dispersion helps, damped by $\sigma(\sigma-1)$ because markups withhold most of the gain; alignment helps with the same coefficient; demand-gap variance is pure loss.

A.4 Welfare, the Planner, and the Hsieh–Klenow Bridge

Household closure. The representative household's first-order conditions with log consumption utility and linear labor disutility equalize expenditure across industries and pin it at one in every state, with the wage as numeraire and the portfolio discounting at β (Appendix A.7); the welfare-relevant flow per industry is then (20). Unit expenditure makes industry consumption the reciprocal of the industry price index, and combining (A3) with (18) gives the decomposition used in the text, $\ln C^{\text{eq}} = \ln Z^* - \ln \bar{\mu}_H - \ln \mathcal{M}$: divide and multiply (A3) by $Z^* L^{\text{prod}}$ and identify the three factors. The closure requires exactly these preferences; with curvature in labor disutility the labor level would respond and the accounting would include an additional margin.

Zero-sum social value of customer capital. At every state, $\sum_i \partial W / \partial b_i = 0$: a uniform shift $\mathbf{b} \to \mathbf{b} + c\mathbf{1}$ leaves the weights ω unchanged, hence shares, labor pools, C^{eq}, and Z^* unchanged (the normalization in (3) is chosen to deliver this property), so W is constant along the shift; differentiate at $c = 0$. Because the shadow-value recursion below is linear with common decay, the zero-sum property extends to the discounted shadow values. This has two consequences: with unrestricted marketing the planner's policies sum to zero across firms, and at any symmetric state the planner does not market at all, while each firm's private flow value there is $\phi_{ii} = \pi'(\bar{s})\,\chi(\bar{s})\,(I-1)/I = 0.0387$ at the calibration and the social value is zero to machine precision, so marketing at the modal symmetric state is entirely business stealing.

Planner's first-order condition. The planner controls all policies simultaneously, so the envelope theorem removes the cross-policy terms that generate the equilibrium's strategic feedback, leaving the recursion (26), which is the equilibrium's Stein problem with the private flow ϕ_{ii} replaced by the social flow g_i and without the feedback block, and is therefore easier to compute. In the first-best regime the flow is (28), obtained by differentiating $\ln Z^*$, with the productivity gradient $\partial \ln Z^* / \partial a_i = \omega_i e^{a_i} / [(\sigma - 1) \sum_j \omega_j e^{a_j}] > 0$. In the constrained (Cournot) regime, using the share Jacobian (A1),

$$g_i^C = \frac{1}{\sigma - 1}\Big[\frac{\partial \ln N}{\partial z_i} - \omega_i\Big] + \frac{2(\sigma-1)}{\sigma}\Big[s_i\chi_i - \hat{\chi}_i \sum_j s_j \chi_j\Big], \tag{A10}$$

$$\frac{\partial \ln N}{\partial z_i} = s_i - (\sigma - 1)\Big[\frac{s_i \chi_i}{1 - s_i} - \hat{\chi}_i \sum_j \frac{s_j \chi_j}{1 - s_j}\Big],$$

where the second bracket is the markup-withholding term. Tilting the state toward large firms lowers production-labor demand, which raises flow utility through lower labor disutility; this gives the planner a mild preference for concentration that does not appear in the log index $\ln Z^{\text{eq}}$. All closed forms are verified against central finite differences (maximum gradient error 2.5×10^{-10}; zero-sum residual 10^{-14}). Since $\eta \equiv 0$ is feasible for the planner, the planner's value weakly exceeds the no-marketing benchmark in both regimes; the strict gain comes from aligning customer capital with productivity.

Derivation of Lemma 1. In the subsidized game firm i's critic and actor are

$$\tilde{V}_i(\mathbf{x}) = \pi_i(\mathbf{z}) + \tau_i(\mathbf{x})\,\eta_i(\mathbf{x}) - \tfrac{\kappa}{2}\eta_i(\mathbf{x})^2 + \beta\,\mathbb{E}\big[\tilde{V}_i(\mathbf{x}') \,\big|\, \mathbf{x}\big], \tag{A11}$$

$$\kappa\,\eta_i(\mathbf{x}) = \tau_i(\mathbf{x}) + \frac{\beta}{\nu_b^2}\,\mathbb{E}\big[\epsilon_i^b\,\tilde{V}_i(\mathbf{x}') \,\big|\, \mathbf{x}\big]. \tag{A12}$$

Fix the candidate policies at $\boldsymbol{\eta}^P$ and define the pair $(\boldsymbol{\tau}, \tilde{\mathbf{V}})$ jointly: $\tilde{V}_i$ as the value of (A11) under the planner's policies, and

$$\tau_i(\mathbf{x}) \equiv \kappa\,\eta_i^P(\mathbf{x}) - \frac{\beta}{\nu_b^2}\,\mathbb{E}^P\big[\epsilon_i^b\,\tilde{V}_i(\mathbf{x}') \,\big|\, \mathbf{x}\big]. \tag{A13}$$

Given $\boldsymbol{\tau}$ the values solve a discounted linear recursion; given the values $\boldsymbol{\tau}$ is explicit; the composite map is a contraction at discount β, so the pair exists and is unique. By construction the subsidized first-order condition (A12) holds at $\boldsymbol{\eta}^P$ in every state, and the second-order condition is unchanged because the subsidy is linear in η_i and the curvature κ is untouched, so the planner's policies form a marketing equilibrium of the subsidized game. Substituting the planner's actor (29) into (A13) gives $\tau_i = \frac{\beta}{\nu_b^2}\,\mathbb{E}^P[\epsilon_i^b(V^P - \tilde{V}_i)(\mathbf{x}') \,|\, \mathbf{x}]$, and adding and subtracting the subsidized total $\tilde{V}^\Sigma \equiv \sum_k \tilde{V}_k$ delivers the two-wedge form (32). Lump-sum financing through the household budget leaves (6) and the statics unchanged.

Hsieh–Klenow mapping. Define the physical customer-capital stock $\mathfrak{B}_i \equiv e^{b_i/\alpha_B}$, so the demand shifter is $\mathfrak{B}_i^{\alpha_B}$ with $\alpha_B \in (0,1)$ the revenue elasticity of customer capital. In the Cournot equilibrium $R_i = s_i$ and, from the share fixed point,

$$\ln R_i = \alpha_B\,\ln \mathfrak{B}_i + \ln \hat{a}_i + \text{const}, \qquad \text{with} \quad \ln \hat{a}_i \equiv a_i - (\sigma - 1)\ln \mu_i(s_i), \tag{A14}$$

verified in the simulated equilibrium to 10^{-15}. The estimating equation of Section 4.1 therefore holds exactly in the model, with three implications: the revenue-productivity index $\hat{a}$ conflates true productivity and the oligopoly markup (under monopolistic competition $\hat{a}_i \propto e^{a_i}$ exactly); the labor revenue product equals the markup, $R_i/\ell_i = \mu_i$; and the customer-capital revenue product is $\ln(R_i/\mathfrak{B}_i) = \ln \hat{a}_i - \frac{1-\alpha_B}{\alpha_B} b_i + \text{const}$, which is the object whose dispersion the data measure.

Free-b planner and the two-wedge decomposition. Let the planner reallocate the aggregate stock freely each period, $\sum_i \mathfrak{B}_i = \bar{\mathfrak{B}}$, with labor allocated efficiently. The optimum is log-linear in productivity,

$$b_i^* = \frac{\alpha_B}{1-\alpha_B} a_i + \text{const} \qquad \Longleftrightarrow \qquad \mathfrak{B}_i^* \propto e^{a_i/(1-\alpha_B)}, \tag{A15}$$

perfect alignment with slope $\alpha_B/(1-\alpha_B)$, and the optimized TFP is the Hsieh–Klenow power mean $A^* = [\sum_i e^{a_i/(1-\alpha_B)}]^{(1-\alpha_B)/(\sigma-1)}$. Letting A^L denote measured TFP with labor reallocated efficiently but **b** held fixed, the static loss splits exactly into two nonnegative wedges,

$$\ln \frac{A^*}{A^{\text{eq}}} = \underbrace{\ln \mathcal{M}}_{\text{markup wedge}} + \underbrace{\ln \frac{A^*}{A^L}}_{\text{demand misallocation}}, \tag{A16}$$

where $\ln \mathcal{M} = \ln A^L - \ln A^{\text{eq}}$ coincides with the markup wedge of (19) (the weight normalizations cancel). For (a_i, b_i) jointly normal and demeaned within the market, to second order (exact as $I \to \infty$ under log-normality),

$$\ln \frac{A^*}{A^L} = \frac{\alpha_B\,(1-1/I)}{2(1-\alpha_B)(\sigma-1)} \operatorname{Var}\Big(a_i - \frac{1-\alpha_B}{\alpha_B}\, b_i\Big), \tag{A17}$$

the variance of the distance from the free-b line (A15), which by (A14) is the variance of the log revenue product of customer capital up to the markup term, the empirical object of Section 4.1. A Monte Carlo check at the ergodic moments puts the formula within 3% of the exact loss at $I = 7$. This calculation is in output units. Since $\alpha_B < 1$ makes $\sum_i \mathfrak{B}_i^{\alpha_B}$ Schur-concave, the free-b reallocation concentrates stocks and lowers the weight normalizer $\sum_i e^{b_i}$; in household units, where consumption is weighted by the normalized ω, the reallocation gain exceeds the output gain, so the Hsieh–Klenow loss is a lower bound on the welfare-unit static loss.

VAR-stationary bridge. The comparison of the model with the VAR-stationary column of Table 5 requires two reconciliations, and both hold exactly in the simulated panel. The first concerns granularity: with $I = 7$ firms per market, the within-market cell mean contains the firm itself, so demeaned variances understate leave-one-out dispersion by the factor $(I/(I-1))^2 = 1.36$, an identity that holds pointwise in the panel. The second concerns the estimator: running the data's random-effects VAR inside the model's ergodic panel returns $\rho_{aa} = 0.907$, $\rho_{bb} = 0.970$, $\rho_{ba} = +0.031$ with innovation variances 0.161 and 0.092, so the data's own procedure recovers the calibration targets. Feeding these estimates into the stationary Lyapunov equations reproduces the panel's measured-productivity variance exactly (ratio 1.000) and its customer-capital variance within seven percent (0.934). The covariance is the one object the extrapolation misses: the implied value is $+0.110$ against $+0.001$ in the panel (leave-one-out units). The gap arises because markups load into measured productivity, so the productivity innovation is negatively correlated with the lagged stock, $\mathbb{E}[\varepsilon^A_{it}\, b_{i,t-1}] = -0.014$, a term the VAR's orthogonality restriction sets to zero. The same mechanism explains why the model's measured covariance, reported in the notes to Table 5, turns negative while the structural alignment $\text{corr}(a, b) = +0.21$ shown in the table stays positive.

A.5 Rational Inattention and Customer Base

With multiple firms, a rational inattention framework, adapted from Wu (2024), maps the consumer's choice across firms' product bundles to customer capital. Households have limited attention to allocate across competing options: each household h chooses the probability q_{hi} of purchasing from firm i by solving

$$\max_{q_{hi}, c_{hi}} \sum_i q_{hi} \log c_{hi} - \psi \sum_i q_{hi} \log \frac{q_{hi}}{\bar{q}_i},$$

where c_{hi} is consumption and

$$\bar{q}_i = \frac{\exp(b_i)}{\sum_j \exp(b_j)}$$

is the default attention allocation based on customer capital b_i. The choice respects the adding-up constraint $\sum_i q_{hi} = 1$ and the unit-budget constraint $c_{hi} = \frac{1}{p_i}$.

The solution is the expenditure share

$$q_i = \frac{\exp(b_i)c_i^{1/\psi}}{\sum_j \exp(b_j)c_j^{1/\psi}} = \frac{\exp(b_i)p_i^{-1/\psi}}{\sum_j \exp(b_j)p_j^{-1/\psi}},$$

where ψ is the cost of directing choice towards the more productive firms. The formulation generates expected utility $\log\left(\frac{\sum_i e^{b_i}c_i^{1/\psi}}{\sum_i e^{b_i}}\right)^{\psi}$, which is equivalent to the baseline CES utility when $\psi = \frac{\sigma}{\sigma-1}$, yielding

$$\log\left(\frac{\sum_i e^{b_i}c_i^{\frac{\sigma-1}{\sigma}}}{\sum_i e^{b_i}}\right)^{\frac{\sigma}{\sigma-1}}.$$

The model further relates consumption to firm productivity through $c_{it} = \exp\left(\frac{a_{it}}{\sigma-1}\right) l_{it}$, where a_{it} is firm-specific productivity and l_{it} is labor input.

Customer capital b_i is the stock of marketing and brand-building investment that shifts the default attention allocation $\bar{q}_i$ in the firm's favor, lowering consumers' cognitive cost of choosing its products. Marketing therefore pays even when it leaves product quality and characteristics unchanged: it operates on the attention allocation of boundedly rational consumers. Firms with larger customer capital maintain higher market shares and can charge premium prices, a value of brand recognition beyond productivity, in line with a growing literature on the centrality of customers in market share.

A.6 Policy Extensions

We first consider a subsidy for a policymaker who observes firms' demand and productivity differences separately. The optimal policy that corrects the markup distortions removes firms' revenue incentives and reimburses each firm with the full social surplus it creates, through the gross transfer

$$T(q) = \frac{\sigma}{\sigma-1}\frac{e^{b_i/\sigma}q^{\frac{\sigma-1}{\sigma}}}{\sum_j \exp(a_j+b_j)} - \frac{e^{b_i/\sigma}q_j^{\frac{\sigma-1}{\sigma}}}{\sum_{j\neq i}\exp(b_j+a_j) + e^{b_i/\sigma}q^{\frac{\sigma-1}{\sigma}}}.$$

With the other firms all pricing at marginal cost, the post-transfer problem of the focal firm i is

$$\max_q \frac{\sigma}{\sigma-1}\frac{e^{b_i/\sigma}q^{\frac{\sigma-1}{\sigma}}}{\sum_j \exp(b_j+a_j)} - qe^{a_i/(1-\sigma)}.$$

The optimal choice of quantity is $q_i^* = \frac{e^{a_i+b_i\frac{\sigma}{\sigma-1}}}{\sum_j \exp(b_j+a_j)}$ and the price is the marginal cost $p_i^* = e^{a_i/(1-\sigma)}$. This implements the static optimal allocation. In the subsidized equilibrium, every firm receives the share of surplus it creates:

$$T_i^* = \frac{1}{\sigma-1}\frac{e^{a_i+b_i}}{\sum_j \exp(a_j+b_j)}.$$

We again linearize the post-transfer payoff around the equal-demand state $(\alpha, \mathbf{0})$:

$$T_j^* \approx \frac{1}{\sigma-1}\bar{s}_i^*(\alpha)\left(b_i - \sum_j \bar{s}_j^*(\alpha) b_j\right) + \frac{1}{\sigma-1}\frac{\exp(a_i)}{\sum_j \exp(a_j)}.$$

λ-subsidy: partial markup offset from size alone. When the policymaker observes only the composite z_i, the implementable instrument of Section 5 is a size-based output subsidy that offsets the fraction λ of each firm's log markup, $\mu_i \to \mu_i^{1-\lambda}$. Prices $p_i = \mu_i^{1-\lambda} e^{-a_i/(\sigma-1)}$ give the share fixed point

$$s_i^\lambda \propto e^{z_i}\left(1-s_i^\lambda\right)^{(1-\lambda)(\sigma-1)}, \tag{A18}$$

the Cournot system at the pseudo-elasticity $\sigma_\lambda = 1 + (1-\lambda)(\sigma-1)$: $\lambda = 0$ reproduces Cournot and $\lambda = 1$ is marginal-cost pricing, the first-best statics given $\mathbf{z}$ (softmax shares). The static aggregates are closed form,

$$\ln C_\lambda = (1-\lambda)\ln\tfrac{\sigma-1}{\sigma} + \frac{\ln N_\lambda}{\sigma-1}, \qquad N_\lambda = \sum_j \omega_j e^{a_j}\left(1-s_j^\lambda\right)^{(1-\lambda)(\sigma-1)},$$
$$L_\lambda^{\text{prod}} = \sum_i s_i^\lambda \left[\tfrac{(\sigma-1)(1-s_i^\lambda)}{\sigma}\right]^{1-\lambda},$$

and both endpoints are certified against the Cournot statics and the softmax limit to machine precision. The subsidy schedule leaves firms the Cournot rent shape at the subsidized shares (rent-preserving incidence), $\pi_i^\lambda = s_i^\lambda/\sigma + \frac{\sigma-1}{\sigma}(s_i^\lambda)^2$, so the dynamic game is the marketing equilibrium of Section 3.4 under the reshaped prize, solved by the method of Section 3.6 at the same β as the equilibrium for every λ. The moat-drift diagnostic in the right panel of Figure 8 is $\mathbb{E}[\Delta m] = (\delta_b - 1)m + (\eta_1 - \eta_{\text{rival}})$ at the fitted policies, with the rivals held symmetric.

A.7 Aggregation and Household Problem

The household makes static consumption and labor decisions and a dynamic savings decision. The household has flow utility $\log \mathcal{C}_t - \mathcal{L}_t$ and holds assets ω_t, with the budget constraint

$$\dot{\omega}_t = r_t \omega_t + \mathcal{L}_t + \mathbf{\Pi}_t - \int_0^1 \sum_{i=1}^{I} p_{ikt} c_{ikt} dk,$$

where aggregate consumption is Cobb–Douglas across product groups, $\log \mathcal{C}_t = \int_0^1 \log C_{kt}\, dk$, with C_{kt} the industry bundle of (3). The wage is normalized to 1.

First-order conditions. Let Λ_t denote the multiplier on the flow budget. The consumption condition for variety i in industry k is $\partial \ln C_{kt} / \partial c_{ikt} = \Lambda_t\, p_{ikt}$; multiplying by c_{ikt} and summing within the industry, homogeneity of degree one of the aggregator (3) gives $\sum_i (\partial \ln C_{kt} / \partial c_{ikt})\, c_{ikt} = 1$, so industry expenditure is $E_{kt} = 1/\Lambda_t$: equal across industries, the Cobb–Douglas constant-share result. The labor condition is $1 = \Lambda_t w_t$; with the wage as numeraire, $\Lambda_t = 1$ and hence $E_{kt} = 1$, which is (6). The savings condition is the Euler equation $\Lambda_t = \beta(1 + r_{t+1})\Lambda_{t+1}$, so $1 + r_{t+1} = 1/\beta$ and the representative portfolio prices profit streams at the household's discount factor. Linear disutility makes labor supply perfectly elastic at the numeraire wage, the aggregate labor condition of Section 3.2: the clearing condition (5) determines $\mathcal{L}_t$ residually, with no feedback into prices. The share Jacobian this consumption problem induces is (A1) in Appendix A.1.

Linear aggregation over non-settling industries. Industries are ex-ante identical, their states are independent, and each industry's state follows the same ergodic Markov equilibrium with invariant distribution π. By the law of large numbers across the continuum, cross-sectional averages equal ergodic means at every date: $\ln \mathcal{C}_t = \int_0^1 \ln C_{kt}\, dk = \mathbb{E}_\pi[\ln C(\mathbf{z})]$ and $\mathcal{L}_t = \mathbb{E}_\pi[L^{\text{prod}} + L^{\text{mkt}}]$ almost surely, both constant over time although no industry's state converges. Aggregate welfare is therefore $\sum_t \beta^t (\ln \mathcal{C}_t - \mathcal{L}_t) = \frac{1}{1-\beta}\, \mathbb{E}_\pi[W(\mathbf{z})]$ with W the per-industry flow (20), the object every counterfactual in Section 5 differences.

Multi-product extension. For the empirical analysis, we model multi-product firms with constant elasticity of substitution across products and decreasing returns in pro-

duction (as in Hottman et al., 2016). Firm-level consumption is

$$c_{it} = \left(\sum_{u=1}^{U_i} \varsigma_{iut} \, d_{iut}^{\frac{\sigma_u}{\sigma_u - 1}} \right)^{\frac{\sigma_u - 1}{\sigma_u}}, \quad \sigma_u > \sigma$$

where ς_{iut} are product appeals, $\sigma_u > \sigma$ implies constant markups across product lines, and production at the firm-group level exhibits decreasing returns to scale: $c_{ikt} = e^{a_{ikt}/(\sigma-1)} l_{ikt}^{\alpha_L}$ with α_L the labor revenue elasticity.

B Estimation Appendix

This appendix details our two-step estimation procedure for recovering the structural parameters of our model. We first estimate the autoregressive parameters across a bracket of panel estimators, then recover variance components through GMM moment matching.

B.1 Model Structure

Our empirical model decomposes both productivity and customer capital into persistent and transitory components. For firm j at time t:

$$a_{jt} = a_{jt}^P + \tau_{jt}^a + \bar{a}_j$$
$$b_{jt} = b_{jt}^P + \tau_{jt}^b + \bar{b}_j$$

where τ_{jt}^a and τ_{jt}^b are the transitory shocks, independent of the current and past persistent shocks; $\bar{a}_j$ and $\bar{b}_j$ are the firm fixed effects; ε_{jt}^a and ε_{jt}^b are the innovations of the persistent components. Here we take ρ_{aa}, ρ_{ba}, ρ_{bb} as already recovered from Arellano-Bond and recover the variances $\left(\sigma_{aT}^2, \sigma_{bT}^2, \sigma_{aP}^2, \sigma_{bP}^2\right)$. We will recover these parameters from the auto-covariance function of the first difference. More precisely, we denote

$$\Sigma_k^x = Cov\left(x_{jt} - x_{jt-1}, x_{jt-k} - x_{jt-k-1}\right), \quad x = a, b$$

It is useful to write out the values in terms of past realizations. More precisely, for any $k \geq 1$

$$a_{jt}^P = \rho_{aa}^k a_{jt-k}^P + error$$

where the error is independent to a^P_{jt-k}. We can do the same with b:

$$b^P_{jt} = \rho^k_{bb} b^P_{jt-k} + \left(\sum_{l=0}^{k-1} \rho^l_{bb} \rho_{ba} \rho^{k-l-1}_{aa} \right) a^P_{jt-k} + error$$

For notation simplicity, let us define

$$Q_k = \sum_{l=0}^{k-1} \rho^l_{bb} \rho_{ba} \rho^{k-l-1}_{aa}$$

Thus

$$b^P_{jt} = \rho^k_{bb} b^P_{jt-k} + Q_k a^P_{jt-k} + error$$

Productivity To start, we impose stationarity and denote the stationary variance of a^P as $Var(a^P)$:

$$Var\left(a^P\right) = \rho^2_{aa} Var\left(a^P\right) + \sigma^2_{aP} \implies Var\left(a^P\right) = \frac{\sigma^2_{aP}}{1-\rho^2_{aa}}$$

We start with $k = 0$:

$$\begin{aligned}
\Sigma^a_0 = Var\left(a_{jt} - a_{jt-1}\right) &= Var\left(a^P_{jt} - a^P_{jt-1} + \tau^a_{jt} - \tau^a_{jt-1}\right) \\
&= Var\left(a^P_{jt} - a^P_{jt-1}\right) + Var\left(\tau^a_{jt} - \tau^a_{jt-1}\right) \\
&= Var\left(\rho_{aa} a^P_{jt-1} + \varepsilon^a_{jt} - a^P_{jt-1}\right) + 2\sigma^2_{aT} \\
&= (1-\rho_{aa})^2 Var\left(a^P\right) + \sigma^2_{aP} + 2\sigma^2_{aT}
\end{aligned}$$

where the first equality uses the definition, the second equality uses the fact that τ^a is independent of the persistent components, the third equality uses the definition of the transitory variance, and the last equality uses the fact that ε^a_{jt} is independent of past values. Plugging the stationary variance, we have

$$\Sigma^a_0 = \frac{2}{1+\rho_{aa}} \sigma^2_{aP} + 2\sigma^2_{aT}$$

Now we move on to $k = 1$:

$$\begin{aligned}\Sigma_1^a =& Cov\left(a_{jt}^P - a_{jt-1}^P + \tau_{jt}^a - \tau_{jt-1}^a, a_{jt-1}^P - a_{jt-2}^P + \tau_{jt-1}^a - \tau_{jt-2}^a\right)\\ =& Cov\left(a_{jt}^P - a_{jt-1}^P, a_{jt-1}^P - a_{jt-2}^P\right) + Cov\left(\tau_{jt}^a - \tau_{jt-1}^a, \tau_{jt-1}^a - \tau_{jt-2}^a\right)\\ =& Cov\left(a_{jt}^P - a_{jt-1}^P, a_{jt-1}^P - a_{jt-2}^P\right) - \sigma_{aT}^2\\ =& Cov\left(a_{jt}^P, a_{jt-1}^P\right) - Var\left(a^P\right) - Cov\left(a_{jt}^P, a_{jt-2}^P\right) + Cov\left(a_{jt-1}^P, a_{jt-2}^P\right) - \sigma_{aT}^2\end{aligned}$$

where the first equality uses the independence of τ^a from the persistent values. Using the sequential form,

$$Cov\left(a_{jt}^P, a_{jt-1}^P\right) = Cov\left(a_{jt-1}^P, a_{jt-2}^P\right) = \rho_{aa} Var(a^P), \qquad Cov\left(a_{jt}^P, a_{jt-2}^P\right) = \rho_{aa}^2 Var(a^P),$$

and plugging these values we have

$$\Sigma_1^a = -\frac{1-\rho_{aa}}{1+\rho_{aa}}\sigma_{aP}^2 - \sigma_{aT}^2$$

Now we move on to $k \geq 2$:

$$\begin{aligned}\Sigma_k^a =& Cov\left(a_{jt}^P - a_{jt-1}^P + \tau_{jt}^a - \tau_{jt-1}^a, a_{jt-k}^P - a_{jt-k-1}^P + \tau_{jt-k}^a - \tau_{jt-k-1}^a\right)\\ =& Cov\left(a_{jt}^P - a_{jt-1}^P, a_{jt-k}^P - a_{jt-k-1}^P\right)\\ =& \rho_{aa}^k Var\left(a^P\right) - \rho_{aa}^{k-1} Var\left(a^P\right) - \rho_{aa}^{k+1} Var\left(a^P\right) + \rho_{aa}^k Var\left(a^P\right)\\ =& -\rho_{aa}^{k-1}\frac{1-\rho_{aa}}{1+\rho_{aa}}\sigma_{aP}^2\end{aligned}$$

This gives us the general formula

$$\Sigma_k^a = \begin{cases} \frac{2}{1+\rho_{aa}}\sigma_{aP}^2 + 2\sigma_{aT}^2 & k = 0 \\ -\frac{1-\rho_{aa}}{1+\rho_{aa}}\sigma_{aP}^2 - \sigma_{aT}^2 & k = 1 \\ -\rho_{aa}^{k-1}\frac{1-\rho_{aa}}{1+\rho_{aa}}\sigma_{aP}^2 & k > 1 \end{cases} \tag{A19}$$

Demand Similarly, we start by imposing stationarity:

$$\left(1-\rho_{bb}^2\right) Var\left(b^P\right) = \rho_{ba}^2 Var\left(a^P\right) + 2\rho_{ba}\rho_{bb}\, Cov\left(a^P, b^P\right) + \sigma_{bP}^2$$

$$Cov(a^P, b^P) = \rho_{aa}\rho_{ba} Var(a^P) + \rho_{aa}\rho_{bb} Cov(a^P, b^P)$$

With $Var(a^P)$ known from earlier steps, these two equations pin down $Var\left(b^P\right)$ and $Cov(a^P, b^P)$. From here on, we treat the two values as known.

$k = 0$

$$\begin{aligned}
\Sigma_0^b &= Var\left(b_{jt}^P - b_{jt-1}^P + \tau_{jt}^b - \tau_{jt-1}^b\right) \\
\text{(independence of } \tau^b\text{)} &= Var\left(b_{jt}^P - b_{jt-1}^P\right) + Var\left(\tau_{jt}^b - \tau_{jt-1}^b\right) \\
\text{(expand terms)} &= Var\left(\rho_{ba} a_{jt-1}^P + (\rho_{bb}-1) b_{jt-1}^P + \varepsilon_{jt}^b\right) + 2\sigma_{bT}^2 \\
\text{(expand terms)} &= \rho_{ba}^2 Var(a^P) + (1-\rho_{bb})^2 Var\left(b^P\right) + \sigma_{bP}^2 - 2\rho_{ba}(1-\rho_{bb}) Cov(a^P, b^P) + 2\sigma_{bT}^2
\end{aligned}$$

For higher-order terms we use the sequential form of b_{jt}^P derived above, whose error terms are independent of the values from $t-k$.

$k = 1$

$$\begin{aligned}
\Sigma_1^b =& Cov\left(b_{jt}^P - b_{jt-1}^P + \tau_{jt}^b - \tau_{jt-1}^b, b_{jt-1}^P - b_{jt-2}^P + \tau_{jt-1}^b - \tau_{jt-2}^b\right) \\
\text{(independence of } \tau^b\text{)} =& Cov\left(b_{jt}^P - b_{jt-1}^P, b_{jt-1}^P - b_{jt-2}^P\right) - \sigma_{bT}^2 \\
\text{(expand terms)} =& Cov\left(b_{jt}^P, b_{jt-1}^P\right) - Var\left(b_{jt-1}^P\right) - Cov\left(b_{jt}^P, b_{jt-2}^P\right) + Cov\left(b_{jt-1}^P, b_{jt-2}^P\right) - \sigma_{bT}^2
\end{aligned}$$

Now we can use the sequential form.

$$\begin{aligned}
Cov\left(b_{jt}^P, b_{jt-1}^P\right) = Cov\left(b_{jt-1}^P, b_{jt-2}^P\right) &= Cov\left(\rho_{bb} b_{jt-1}^P + Q_1 a_{jt-1}^P + error, b_{jt-1}^P\right) \\
&= Q_1 Cov\left(a^P, b^P\right) + \rho_{bb} Var\left(b^P\right)
\end{aligned}$$

$$\begin{aligned} Cov\left(b_{jt}^P, b_{jt-2}^P\right) &= Cov\left(\rho_{bb}^2 b_{jt-2}^P + Q_2 a_{jt-2}^P + error, b_{jt-2}^P\right) \\ &= Q_2 Cov\left(a^P, b^P\right) + \rho_{bb}^2 Var\left(b^P\right) \end{aligned}$$

Plugging in

$$\Sigma_1^b = -\left(1-\rho_{bb}\right)^2 Var\left(b^P\right) + \left(2Q_1 - Q_2\right) Cov\left(a^P, b^P\right) - \sigma_{bT}^2$$

$k > 1$

$$\begin{aligned} \Sigma_k^b =& Cov\left(b_{jt}^P - b_{jt-1}^P + \tau_{jt}^b - \tau_{jt-1}^b, b_{jt-k}^P - b_{jt-k-1}^P + \tau_{jt-k}^b - \tau_{jt-k-1}^b\right) \\ \text{(independence of } \tau^b\text{)} =& Cov\left(b_{jt}^P - b_{jt-1}^P, b_{jt-k}^P - b_{jt-k-1}^P\right) \\ =& Cov\left(b_{jt}^P, b_{jt-k}^P\right) - Cov\left(b_{jt-1}^P, b_{jt-k}^P\right) - Cov\left(b_{jt}^P, b_{jt-k-1}^P\right) + Cov\left(b_{jt-1}^P, b_{jt-k-1}^P\right) \\ =& \left(2\rho_{bb}^k - \rho_{bb}^{k-1} - \rho_{bb}^{k+1}\right) Var\left(b^P\right) + \left(2Q_k - Q_{k-1} - Q_{k+1}\right) Cov(a^P, b^P) \\ =& -\rho_{bb}^{k-1}\left(1-\rho_{bb}\right)^2 Var(b^P) + \left(2Q_k - Q_{k-1} - Q_{k+1}\right) Cov(a^P, b^P) \end{aligned}$$

This gives us the general formula

$$\Sigma_k^b = \begin{cases} \rho_{ba}^2 Var(a^P) + (1-\rho_{bb})^2 Var(b^P) + \sigma_{bP}^2 - 2\rho_{ba}(1-\rho_{bb}) Cov(a^P, b^P) + 2\sigma_{bT}^2 & k = 0 \\ -\left(1-\rho_{bb}\right)^2 Var\left(b^P\right) + \left(2Q_1 - Q_2\right) Cov\left(a^P, b^P\right) - \sigma_{bT}^2 & k = 1 \\ -\rho_{bb}^{k-1}\left(1-\rho_{bb}\right)^2 Var(b^P) + \left(2Q_k - Q_{k-1} - Q_{k+1}\right) Cov(a^P, b^P) & k > 1 \end{cases} \tag{A20}$$

B.2 Moment Conditions for Elasticity Estimation

Let

$$\begin{aligned} a_{it}^*(\alpha_L) &= \Delta p_{it} - \frac{\alpha_L}{1+\alpha_L} \Delta s_{it}, \\ b_{it}^*(\sigma) &= \Delta s_{it} + (\sigma - 1) \Delta p_{it}, \end{aligned}$$

and define innovation residuals

$$u_{a,it} = a^*_{it} - \rho_{aa}\, a^*_{i,t-1},$$
$$u_{b,it} = b^*_{it} - \rho_{ba}\, a^*_{i,t-1} - \rho_{bb}\, b^*_{i,t-1}.$$

Let $Z_{i,t-1}$ denote instruments measurable at $t-1$ (e.g. $L^2\Delta p_{it}, L^2\Delta s_{it}, L^3\Delta p_{it}, L^3\Delta s_{it}$). The dynamic restrictions require

$$\mathbb{E}[Z_{i,t-1}\, u_{a,it}] = 0, \tag{A21}$$
$$\mathbb{E}[Z_{i,t-1}\, u_{b,it}] = 0, \tag{A22}$$
$$\mathbb{E}[u_{a,it}\, u_{b,it}] = 0. \tag{A23}$$

Equations (A21)–(A22) enforce that the quasi-differenced demand and cost shocks are orthogonal to lagged information, while (A23) imposes no contemporaneous correlation between the structural innovations. Together with the variance restrictions

$$\mathbb{E}\left[(a^*_{it})^2\right] = V_a, \tag{A24}$$
$$\mathbb{E}\left[(b^*_{it})^2\right] = V_b, \tag{A25}$$

(with V_a and V_b as nuisance parameters), this yields five moments to identify $(\sigma, \alpha_L, \rho_{aa}, \rho_{ba}, \rho_{bb})$.

B.3 First Stage: Arellano–Bond Estimation

We first estimate the autoregressive parameters $(\rho_{aa}, \rho_{ba}, \rho_{bb})$ using the Arellano–Bond GMM estimator (Arellano and Bond, 1991). The estimator uses lagged levels as instruments for first differences to address the correlation between the fixed effects and the regressors.

B.3.1 Productivity Persistence

For productivity, we estimate:

$$\Delta a_{jt} = \rho_{aa}\Delta a_{jt-1} + \Delta u^a_{jt}$$

where $\Delta u^a_{jt} = \Delta\varepsilon^a_{jt} + \Delta\tau^a_{jt}$ combines the persistent and transitory innovations. The moment conditions are:

$$E[a_{jt-s}\Delta u^a_{jt}] = 0 \quad \text{for } s \geq 2$$

B.3.2 Customer Capital Dynamics

For customer capital, we estimate:

$$\Delta b_{jt} = \rho_{ba}\Delta a_{jt-1} + \rho_{bb}\Delta b_{jt-1} + \Delta u^b_{jt}$$

where $\Delta u^b_{jt} = \Delta\varepsilon^b_{jt} + \Delta\tau^b_{jt}$. The moment conditions are:

$$E[x_{jt-s}\Delta u^b_{jt}] = 0 \quad \text{for } x \in \{a, b\}, s \geq 2$$

B.3.3 Implementation and Results

Because the system is near a unit root in b, no single panel estimator is reliable on its own, and we report the persistence parameters across a bracket of five (Table B1): pooled OLS, the within estimator, the random-effects GLS used in Section 4.2, and the two dynamic-panel GMM estimators, implemented as two-step efficient GMM with lagged levels instrumenting the differenced equations and Windmeijer-corrected standard errors.

Table B1: Persistence Parameters across Panel Estimators

	ρ_{aa}	ρ_{bb}	ρ_{ba}
pooled OLS	0.981 (0.003)	0.981 (0.002)	0.022 (0.002)
firm fixed effects (within)	0.639 (0.015)	0.725 (0.010)	0.019 (0.007)
random-effects GLS (marketing panel)	0.898 (0.007)	0.958 (0.002)	0.032 (0.003)
Arellano–Bond (difference GMM)	0.632 (0.076)	0.963 (0.051)	−0.022 (0.041)
Blundell–Bond (system GMM)	0.930 (0.028)	1.164 (0.029)	−0.021 (0.016)

Notes: Marketing panel (He–Mostrom–Sufi flows, our stock), firms with five consecutive years, firm-clustered standard errors in parentheses.

The ranking $\rho_{bb} > \rho_{aa}$ holds in all five rows. Pooled OLS, which loads the firm effects onto the lag, and the within estimator, which has the short-T downward bias, bracket the random-effects estimate from above and below. Near the unit root the differenced-equation instruments are weak: the Arellano–Bond estimate of ρ_{bb}, 0.963 with a standard error of 0.051, is consistent both with the baseline estimate and with unity, and the Blundell–Bond level moments push it above one. The positive cross-effect is a level-information result: the level estimators put ρ_{ba} between 0.019 and 0.032, while both GMM estimators attenuate it to an insignificant −0.02. Validity of the GMM moment conditions rests on sequential exogeneity of the innovations and no serial correlation in the errors; the Hansen test of overidentifying restrictions ($p = 0.42$) and the test for second-order serial correlation in the differenced residuals ($p = 0.38$) support both. The

random-effects row applies the estimator of Section 4.2 to the marketing panel, where it reads 0.898/0.958/0.032; the imputed sample gives 0.904/0.963/0.033, and the baseline panel of the main text, which observes reported zeros, gives 0.910/0.964/0.033 (Table 3). What matters here is the bracket across estimators rather than any single point. The second stage treats these first-stage estimates as known.

Identification Concerns. We consider four identification concerns.
A. Reverse Causality. The primary identification concern is potential feedback from customer capital to productivity. While our baseline specification assumes productivity evolves independently, firms might adjust their production processes in response to brand-related shocks. An expanded system allows for bidirectional effects:

$$\text{Baseline:} \quad a_{jt} = \rho_{aa} a_{jt-1} + \bar{a}_j + \varepsilon^a_{jt} \tag{A26}$$

$$\text{Alternative:} \quad a_{jt} = \rho_{aa} a_{jt-1} + \rho_{ab} b_{jt-1} + \bar{a}_j + \varepsilon^a_{jt} \tag{A27}$$

The cross-coefficient ρ_{ab} bounds the feedback from customer capital to productivity.
B. Unobserved Heterogeneity. A second concern is that time-varying shocks might simultaneously affect both productivity and customer capital, violating our moment conditions. For example, quality improvements could drive both measures. An instrumental-variables approach leverages industry-level variation:

$$\text{First Stage:} \quad \Delta a_{jt} = \varphi_1 Z_{jt} + e_{jt} \tag{A28}$$

$$\text{Second Stage:} \quad \Delta b_{jt} = \rho_{ba} \Delta \hat{a}_{jt-1} + \rho_{bb} \Delta b_{jt-1} + e'_{jt} \tag{A29}$$

where Z_{jt} includes industry-level shifters of productivity that are plausibly exogenous to firm-specific customer capital.
C. Validity of Lag Instruments. The Arellano-Bond approach relies crucially on the validity of lagged levels as instruments. The Hansen test of overidentifying restrictions and the test for second-order serial correlation, reported above, support the lag instruments.
D. Non-linear Dynamics. Our linear AR(1) specification may miss important non-linearities in the evolution of productivity and customer capital. An expanded specification allows for state-dependent parameters:

$$\Delta b_{jt} = \rho_{ba}(s_{jt-1})\, \Delta a_{jt-1} + \rho_{bb}(s_{jt-1})\, \Delta b_{jt-1} + \Delta \varepsilon^b_{jt} \tag{A30}$$

where the coefficients are flexible functions of the lagged market share.

B.4 Second Stage: GMM Moment Matching

Given these first-stage estimates, we recover the variance parameters $(\sigma^2_{aT}, \sigma^2_{bT}, \sigma^2_{aP}, \sigma^2_{bP})$ by matching the empirical autocovariances of first differences to the theoretical functions (A19) and (A20) derived in the model-structure section above.

B.4.1 Implementation Details

Key numerical considerations:

- Near unit root in customer capital ($\rho_{bb} \approx 1$) requires careful handling of variance terms
- Regularization in matrix inversions:

$$A = \begin{bmatrix} 1 - \rho^2_{bb} + 1e-6 & -\rho_{ba} \\ -\rho_{aa}\rho_{ba} & 1 - \rho_{aa}\rho_{bb} \end{bmatrix}$$

- Treatment of initial conditions

B.4.2 Results

Our second-stage estimates are:

$$\begin{aligned} \sigma_{aP} &= 0.5534 \text{ (Productivity innovation std)} \\ \sigma_{aT} &= 0.3348 \text{ (Transitory productivity shock std)} \\ \sigma_{bP} &= 2.1619 \text{ (Customer capital innovation std)} \\ \sigma_{bT} &= 0.0034 \text{ (Transitory customer-capital shock std)} \end{aligned}$$

These components price the permanent layer of the proxies on the superseded first stage; σ_{bP} is therefore not comparable to the calibrated demand shock $\nu_b = 0.281$, which prices only the within-firm innovation the baseline VAR targets ($\hat{v}^2_B = 0.092$). The estimates reveal three main messages. Customer capital innovations have larger variance than productivity innovations. Productivity has significant transitory components while customer capital is primarily persistent. The model fits productivity autocovariance well but shows some deviation in customer capital autocovariance at longer lags.

B.5 Identification Discussion

The separate identification of variance components comes from a few forces. First, transitory shocks $(\sigma^2_{aT}, \sigma^2_{bT})$ are primarily identified by $k = 1$ autocovariances. Second, persistent shock variances $(\sigma^2_{aP}, \sigma^2_{bP})$ are identified by decay rates at higher lags. Third, cross-persistence ρ_{ba} helps identify the relative contribution of productivity versus customer capital shocks.

B.6 Different Selection Criteria

For our main results, we use all firms and do not condition on a balanced panel. Qualitatively, we find very similar results when we condition on a balanced panel.

B.7 Local Projections for Correlations

In this analysis, we examine the relationship between customer capital (b) and cost productivity (a) using local projections. Local projections provide a flexible approach to estimating impulse response functions without imposing the dynamic restrictions inherent in vector autoregressions.

We estimate the following regression for each horizon h:

$$b_{i,t+h} = \beta^h_0 + \beta^h_1 a_{i,t} + \beta^h_2 b_{i,t-1} + \beta^h_3 a_{i,t-1} + \beta^h_4 b_{i,t-2} + \beta^h_5 a_{i,t-2} + \varepsilon^h_{i,t}$$

where $b_{i,t}$ represents customer capital and $a_{i,t}$ represents cost productivity for firm i at time t. The regression is weighted by the firm's share of the product group.

After estimating these local projections, we construct predicted values of customer capital based on the estimated coefficients:

$$\hat{b}_{i,t+h} = \begin{cases} \hat{\beta}^h_1 a_{i,t} & \text{if } t = 0 \\ \hat{\beta}^h_1 a_{i,t} + \hat{\beta}^h_2 \hat{b}_{i,t+h-1} & \text{if } t > 0 \end{cases}$$

The correlation between the predicted customer capital $(\hat{b}_{i,t+1})$ and the realized future cost productivity $(a_{i,t+1})$ is 0.71, indicating a strong positive relationship between the predicted values of customer capital and future cost productivity. This framework allows us to purge the contemporaneous negatively correlated shocks between the two to focus on the dynamics of their relationship consistent with the model.

B.8 Customer-Capital Investment Momentum

The near-unit-root persistence of the customer-capital stock in Section 4.2 is the cumulation of momentum in the underlying investment, which we document with a growth-form distributed lag. We regress the growth of the customer-capital stock on its own lag and on $K+1$ lags of productivity growth,

$$\Delta \log b_{it} = \rho \, \Delta \log b_{i,t-1} + \sum_{k=0}^{K} \beta_k \, \Delta \log \hat{a}_{i,t-k} + \varepsilon_{it}, \tag{A31}$$

on firms with five consecutive years of data in the merged Compustat/Capital IQ panel, with productivity measured by the revenue-based index $\log \hat{a}_{it} = \log R_{it} - \alpha_B \log b_{it}$.

The own-growth coefficient is positive, $\rho = +0.038$, and stable across the productivity-lag horizon (Table B2). A convex adjustment cost cannot produce this: David and Venkateswaran (2019) show that under quadratic adjustment the conditional autocorrelation of investment growth satisfies $\rho \leq 0$, and physical-capital investment on the same firms mean-reverts. Positive own-autocorrelation in customer-capital investment therefore reflects propagation beyond convex adjustment: a firm that grew its customer base last period keeps growing it, independently of its productivity. Compounded period over period, this momentum sustains the near-unit-root stock persistence ρ_{bb} that drives the dynamic misallocation in Section 4.2. The productivity-growth terms shift from a negative contemporaneous coefficient to a positive one at the three-year lag, consistent with a cross-effect that builds over several years rather than within the year.

Table B2: Customer-Capital Growth: Distributed-Lag Robustness

	$K=0$	$K=1$	$K=2$	$K=3$
$\Delta \log b_{t-1}$	+0.038	+0.038	+0.037	+0.037
	(0.018)	(0.018)	(0.018)	(0.018)
$\Delta \log \hat{a}_t$	-0.042	-0.042	-0.042	-0.043
	(0.027)	(0.028)	(0.028)	(0.028)
$\Delta \log \hat{a}_{t-1}$		-0.005	-0.004	-0.003
		(0.017)	(0.017)	(0.017)
$\Delta \log \hat{a}_{t-2}$			+0.004	+0.001
			(0.017)	(0.017)
$\Delta \log \hat{a}_{t-3}$				+0.035
				(0.014)
R^2	0.004	0.004	0.004	0.006
Observations		5,380		
Firms		1,383		

Notes: Distributed-lag regressions of customer-capital stock-growth on its own lag and K+1 lags of productivity-growth: $\Delta \log b_t = \rho \Delta \log b_{t-1} + \sum_{k=0}^{K} \beta_k \Delta \log \hat{a}_{t-k} + \varepsilon$. The productivity index $\log \hat{a}_{it} = \log R_{it} - \alpha_B \log b_{it}$ is the residual from the labor-substituted revenue equation, with α_B estimated by OLS of NAICS3×year-demeaned $\log R$ on demeaned $\log b$ in the broader Capital IQ panel. NAICS3×year fixed effects implemented by demeaning all log levels by industry-year prior to differencing. Sample: firms with five consecutive years of customer capital, physical capital, and $\hat{a}$ on the merged Compustat/Capital IQ panel; dollar variables deflated to 2017 real dollars. Standard errors clustered at the firm level in parentheses. Under convex adjustment cost alone (David and Venkateswaran 2019), the conditional autocorrelation of investment growth satisfies $\rho \leq 0$; here $\rho > 0$ across all K, robust to productivity-lag horizon.

C Scanner-Data Corroboration

Section 2 measured customer capital from its observable input, marketing expenditure. That measure is directly customer-capital relevant but partial: it captures the demand a firm builds by spending rather than the demand it commands. NielsenIQ Retail Measurement Services (RMS) scanner data report prices and quantities at the barcode level within narrowly defined product groups, so that, under the structural demand system of Hottman et al. (2016), we decompose market shares into cost productivity, markups,

and a residual demand component, and identify that residual with customer capital, the demand state b of the model in Section 3. The two strategies are complementary: Section 2 observes what firms spend to build demand, while here we recover, under a structural assumption, the demand they command from data on prices and quantities. Agreement between the two measures supports the construct: the residual-demand measure shows the same persistence and size-correlation as the marketing-based stock. In product-level data, residual demand explains roughly four times as much as productivity in the cross-sectional variation of firm size, and demand is more persistent than productivity; the gross and net flows of customer capital, through product entry, continuation, and exit, give an independent check on the decay (δ) and marketing-cost (κ) parameters the calibration recovers from the Compustat dynamics (Section 4).

C.1 Data

Residual demand is a useful output-side metric of customer capital: given prices, the demand a firm commands within a product group identifies the stock the model calls b. We measure it with barcode-level data from the Kilts-NielsenIQ Retail Measurement Services, which cover prices and quantities for a broad set of products, brands, and firms across U.S. consumer-product categories, and we merge GS1 company-level information to assign each barcode to a firm (Argente et al., 2024). The product-level detail also lets us decompose the gross and net flows of customer capital.

For each product group k, we observe sales and prices at the firm-year level. Following Hottman et al. (2016), we focus on price and market share variation within product groups to control for differences across categories. Consistent with the model, we focus primarily on the largest firms in each product group; approximately 10 firms per group have a market share above 1% of the national share.

Our final sample includes 423 firms per product group on average, with substantial variation in firm size: the stark difference between mean sales (almost \$6M) and median sales (\$12,000) illustrates the skewness of firm size. The largest firm in a typical product group captures 27% market share, while the second largest firm captures 13%. Market share drops off quickly beyond the top firms; the 10th largest firm captures only a little over 1% share.

C.2 Residual Demand and Firm Size

The structural demand system of Hottman et al. (2016) decomposes the cross-sectional variance of log firm size within a product group into three orthogonal sources: residual demand (the output-side counterpart of customer capital b), cost productivity (a), and markups. On the scanner panel the decomposition puts the bulk of the variation on demand: residual demand accounts for 0.80 of the variance of firm size against 0.20 for productivity (Table C1), roughly four to one. This is the output-side counterpart of Fact 1, where the marketing-based stock tracks size far more closely than productivity does.

Table C1: Variance of Firm Size: Residual Demand vs. Productivity (NielsenIQ)

	residual demand b	productivity a
share of the variance of log firm size	0.80	0.20

Notes: Share of the within-product-group cross-sectional variance of log firm size attributed to residual demand versus cost productivity in the Hottman et al. (2016) structural decomposition, NielsenIQ scanner panel; markups absorb the remainder. Source: RMS NielsenIQ, authors' calculations.

C.3 Relative Persistence of Demand and Productivity

We adapt the static Hottman et al. (2016) demand system to a dynamic setting. In the reduced form, share growth is more persistent than price growth ($\hat{\phi}_s > \hat{\phi}_p$), so the demand component of a firm's position decays more slowly than the cost component.

Identification. We combine the static slope conditions of the CES demand system with dynamic orthogonality restrictions in the spirit of Arellano and Bond (1991): observable proxies for productivity and demand changes are built from price and share growth, and their quasi-differenced innovations are required to be orthogonal to lagged information (the full moment conditions are in Appendix B.2). Fixing the substitution elasticity at the value the quantitative analysis uses throughout, $\sigma = 3.9$ from Hottman et al. (2016), the demand-elasticity parameter ω and the reduced-form persistences $(\hat{\phi}_s, \hat{\phi}_p)$ map into the structural persistence of demand and productivity,

$$\rho_b = \frac{(1+\omega)\phi_s + \omega(\sigma-1)\phi_p}{1+\sigma\omega}, \qquad \rho_a = \frac{\omega(\sigma-1)\phi_s + (1+\omega)\phi_p}{1+\sigma\omega}.$$

Demand is more persistent than productivity whenever $\phi_s > \phi_p$ and $\omega(\sigma - 2) < 1$; both hold on the scanner panel. The residual-demand stock is close to a unit root and outlives productivity, $\rho_b = 0.95$ against $\rho_a = 0.90$ (Table C2), the output-side counterpart of Fact 2 and the same ordering the marketing-based stock delivers on the input side.

Table C2: Structural Persistence of Demand and Productivity (NielsenIQ)

	demand ρ_b	productivity ρ_a
scanner residual-demand system	0.95	0.90

Notes: Structural persistences from the dynamic Hottman et al. (2016) demand system on the NielsenIQ panel, with $\sigma = 3.9$ imposed. Source: RMS NielsenIQ, authors' calculations.

C.4 Dynamics of Demand

We now decompose the forces driving demand fluctuations at the firm level. Firm-level demand changes through three channels: product entry, continuation, and exit. We decompose both sales growth and residual demand into these margins.

New product entry explains approximately 23% of the variation in aggregate sales, consistent with Argente et al. (2024). In the model, the two core drivers of demand changes are endogenous investment (new product entry) and natural decay (continuing and exiting products). We group the latter two together and decompose residual demand as

$$\Delta b_{ikt} = \underbrace{\frac{N_{entry}}{N_{ikt}} \cdot b_{entry}}_{\text{entry}} + \underbrace{\frac{N_{cont}}{N_{ikt}} \cdot (b_{cont,t} - b_{cont,t-1}) - \frac{N_{exit}}{N_{ik,t-1}} \cdot b_{exit}}_{\text{decay}}. \tag{A32}$$

New products require active investment by firms. Given limited consumer attention, entry crowds out existing products. Projecting the creation and decay flows of customer capital on the lagged stock for the top 10 firms in each product group, the slopes identify how the gross flows scale with firm customer capital, a check on the decay rate and marketing cost of the calibrated model.

Larger firms have higher marketing intensity: a firm with one log point higher customer capital introduces 0.07 log points more new product demand. These firms also face higher decay: 0.13 log points per log point of customer capital. The decay slope directly identifies δ; the creation slope serves as an out-of-sample test for the model's endogenous investment mechanism.

These findings corroborate the marketing-based measure: demand heterogeneity is central to firm size and persistent.