Who Competes for Whom?
Monopsony in Ability-Segregated Labor Markets

Luca Lorenzini*
UCLA, Anderson School of Management

First version: February 2024. This version: May 2026
Link to the latest version

Abstract

I develop a general-equilibrium oligopsony model in which firms differ in their demand for worker ability, generating worker-specific monopsony power. Using matched employer-employee data for Italy and Germany, I document that high-ability workers are sorted into high-productivity firms, while low-ability workers are segregated into low-productivity firms. This demand heterogeneity weakens the link between firm productivity and monopsony power. Monopsony and associated welfare losses are largest for low- and high-ability workers, especially the latter. Labor-market power amplifies wage inequality: its widening of between-market wage dispersion outweighs its compression of assortative matching and top wages.

Labor-market power and wage inequality have long concerned economists and policymakers because of their implications for allocative efficiency and the distribution of welfare. A salient feature of labor markets is the presence of both worker and firm heterogeneity and nonrandom matching between the two. High-type workers tend to work for high-type firms (i.e., sorting), while workers of similar type tend to work together and therefore cluster in different firms (i.e., segregation).[1]

These patterns are well documented, but we know much less about what they imply for *who competes for whom* in the labor market: if sorting and segregation reflect systematic differences

*lucalorenzini@ucla.edu. I am deeply grateful to Hugo Hopenhayn, Romain Wacziarg, Jonathan Vogel, and Nico Voigtländer for their guidance, support, and feedback throughout this project. I also thank Daniel Haanwinckel, Simon Mongey, Richard Rogerson, Michael Rubens, Rafael Rubião, Gianluca Violante, Brian Wheaton, and many others, as well as participants at EUI, Princeton, and UCLA seminars and at several conferences for valuable comments and suggestions. All remaining errors are my own. I gratefully acknowledge financial support from the Center for Global Management at UCLA Anderson. This study uses data from the Italian Social Security Institute (INPS), accessed through the VisitINPS Scholars Program Type B under the project *Endogenous Oligopsony*, its previously circulated title. This study also uses the Sample of Integrated Employer–Employee Data (SIEED 7518) from the German Institute for Employment Research (IAB); remote data access was provided under project number *fdz2701*. This paper received the Consultaccount Award for Best Paper presented by a PhD student at the 2024 annual meeting of the Portuguese Economic Journal.

[1] See, for example, Card et al., 2013; Song et al., 2019; Lamadon et al., 2022.

in firms' demand for worker ability, then firms need not compete symmetrically for all workers. In this paper, I study how such demand heterogeneity shapes competition across firms, the efficiency costs of labor-market power, wage inequality, and which workers—low-, middle-, or high-ability—bear those costs.

To study these issues, I develop a quantitative general-equilibrium model of labor-market power in endogenously segregated labor markets. Workers and firms differ in type, and labor supply follows the preference-heterogeneity framework of Berger et al. (2022) (hereafter, BHM), in which workers first choose a local labor market (hereafter, LLM) and then a firm, subject to idiosyncratic taste shocks. As a result, firms face upward-sloping labor supply curves for each worker type and compete strategically for workers of different types within the LLM. Wages therefore reflect firm–worker-specific markdowns applied to marginal products, with markdowns determined by firms' market shares within the LLM for each worker type.[2] At the LLM level, worker-type-specific average markdowns are closely linked to worker-type-specific Herfindahl–Hirschman Indices (HHIs), which summarize the concentration of employment for each worker type across firms.

I model firm-level heterogeneity in the demand for worker ability through a production function that combines an aggregate capital–labor input with an endogenous productivity shifter that depends on workforce composition.[3] This structure generates a flexible, endogenous worker–firm assignment under a parsimonious common parameterization through a size–quality tradeoff and coworker interdependencies in production. Specifically, hiring affects output through two channels: a *size effect*, which raises output by expanding employment, and a *composition effect*, which changes realized productivity by altering workforce quality. Both channels are affected by coworker interdependencies in production. When the composition effect is sufficiently negative, a firm optimally chooses not to hire a low-ability worker, even at an arbitrarily low wage.

I derive theoretical results that characterize how worker–firm complementarities and composition effects shape sorting and the distribution of markdowns in three benchmark cases nested in the framework. The first is the BHM benchmark, obtained in the limit where worker ability does not affect firm realized productivity. In this case, more productive firms have identical workforce composition, are larger for every worker type, and therefore face greater labor-market distortions. In a second benchmark, worker ability affects production but scales output proportionally at all firms. Since ability affects all firms symmetrically, it generates no comparative advantage across firms: workforce composition is identical across firms, and after renormalizing firm productivity, the allocation and markdown distortions coincide with those in the BHM benchmark. In a third benchmark, worker and firm types are sufficiently strong complements, but composition effects remain limited. Concentration and welfare losses then rise monotonically with ability, and more productive firms remain more distorted.

[2] Let MPL denote the worker-specific marginal product of labor. I define the markdown μ by $w = \mu MPL$, so $\mu \in (0, 1]$ measures the share of marginal product paid to the worker. A lower μ therefore corresponds to greater labor-market power.

[3] Equivalently, this object can be interpreted as product quality, since the two concepts are isomorphic in this setting.

The full model shows how these benchmark predictions can fail. When composition effects are sufficiently strong, high-productivity firms may optimally avoid hiring low-ability workers because doing so lowers realized productivity. These workers can then sort disproportionately into smaller, lower-productivity firms. As a result, markdowns may become nonmonotonic in ability, and the link between firm productivity and labor-market distortions weakens relative to the benchmark cases: lower-productivity firms can acquire market power over the workers they disproportionately employ.

To discipline the model, I first document reduced-form patterns using matched employer–employee data from Italy and Germany. To measure heterogeneity, I estimate a two-way fixed-effects wage decomposition in the tradition of Abowd et al. (1999) (hereafter, AKM).[4] Using model-generated data, I show that the within-LLM ranking of AKM fixed effects closely tracks the ranking of latent types, thereby validating their use as *empirical proxies* for indirect inference. I define LLMs as the intersection of industry or occupation and commuting zone. Three empirical patterns emerge. First, high-ranked workers are strongly concentrated in high-ranked, larger firms, whereas low-ranked workers are disproportionately employed in low-ranked, smaller firms, although they remain more dispersed across firm types. Second, hiring standards rise with firm rank: the minimum worker fixed effect among new hires is higher at high-ranked firms. Third, worker-rank-specific concentration indices are nonmonotonic across the worker-rank distribution.

The empirical patterns motivate a calibration of the model, which serves as a measurement device for assessing the effects of labor-market power on three tightly linked objects: firm-level labor-market distortions and their implications for aggregate production, heterogeneous markdowns and the implied welfare losses across worker types, and the contribution of imperfect competition to wage dispersion. These objects are not directly observed in the data. HHI indices by AKM worker-fixed-effect rank provide only a noisy proxy for worker-type-specific markdowns, since worker fixed effects only indirectly measure latent ability and must be grouped into coarse bins. The calibrated model is therefore needed to translate these empirical patterns into measured markdowns, wedges, and counterfactual outcomes.

I calibrate the model by indirect inference using the industry-based definition of LLMs in the Italian data, which allows me to incorporate balance-sheet information. The key parameters are the within-market and across-market labor supply elasticities, together with the parameters governing the dispersion of worker and firm types and their interaction in the worker-level production function. I show that, up to a first-order log-linear approximation, the within-market labor supply elasticity is identified from idiosyncratic firm-level labor-demand shocks once one accounts for induced workforce composition effects and firms' strategic employment responses. I implement this strategy using unexpected worker deaths, which generate quasi-experimental shifts in firms' labor demand for surviving workers due to replacement needs. All other parameters are jointly calibrated by indirect inference.

[4]Following Abowd et al. (1999), wages are decomposed into worker and firm components. Firms are grouped by applying a K-means clustering procedure, in the spirit of Bonhomme et al. (2022).

For the across-market elasticity, I show that a firm's average wage equals its average marginal product multiplied by a labor-market-power firm-level wedge. This wedge summarizes the effect of markdowns on the labor share, equals one absent labor-market power, and can be measured from balance-sheet data in the spirit of De Loecker et al. (2020).[5] Firms in more concentrated markets face weaker competition and thus sustain wedges further below one. Conditional on the within-market elasticity, the across-market elasticity governs the sensitivity of markdowns to market concentration. I therefore identify this parameter from the empirical relationship between firm-level wedges and market concentration.[6] The remaining parameters are chosen to match a small set of moments from a simulated employer–employee panel.

The model also reproduces a broader set of untargeted patterns, including the joint distribution of worker and firm types, the relationship between hiring thresholds and measures of firm quality, heterogeneity in HHI indices across worker AKM ranks, and a range of wage-dispersion moments. As a benchmark, I compare the baseline model to a homogeneous-workers version that collapses to the BHM environment. I consider three versions of this benchmark: the original BHM calibration, a version that updates only firm heterogeneity, and a fully recalibrated version.

Wedges and aggregate production. Labor-market power reduces the labor share and raises the profit share by roughly 12 percentage points. At the firm level, firms in the top decile of the type distribution have an average markdown of 0.88, compared with 0.91 for firms in the bottom decile. Although within-firm heterogeneity prevents the firm-level wedge from coinciding exactly with the average markdown, the resulting discrepancy is quantitatively small. Interpreting wedges as average markdowns is therefore a good approximation in this application.[7] Turning to aggregate production, output in the baseline economy is 2.32% below the no-monopsony allocation, compared with 8.44% in the original BHM calibration and 2.47% in the fully recalibrated BHM benchmark. In the model, cross-market wage heterogeneity is driven primarily by heterogeneity in labor-market power, and the original BHM calibration substantially overpredicts this moment relative to the data. This suggests excessive heterogeneity in markdowns across markets and, in turn, excessive misallocation. After fully recalibrating the homogeneous-worker economy, a small gap remains relative to the baseline, consistent with labor-market segregation mitigating the efficiency costs of monopsony.

Markdowns and welfare by worker type. The median worker receives about 82.7% of their marginal product of labor. This share varies nonmonotonically with ability, ranging from about 81% in the upper part of the ability distribution to 83.5% in the lower part, and peaking at 83.7%, roughly two standard deviations below the mean of log ability. These markdowns imply heterogeneous welfare losses relative to an equilibrium without labor-market power. In consumption-equivalent terms, workers would require consumption increases of about 24% at the top of the

[5] More precisely, I show that, in this framework, the firm-level wedge admits a decomposition into the average markdown and a covariance term between markdowns and worker-level output. This result delivers an empirical counterpart of the wedge that is identifiable from firm-level cost shares and revenues using standard production-function methods; see, for example, De Ridder et al. (2026).

[6] This strategy parallels Edmond et al. (2023), who discipline markups using related cross-sectional variation.

[7] This interpretation appears, for instance, in Yeh et al. (2022).

ability distribution, 22.7% at the bottom, and roughly 20% for workers below the median. By contrast, entrepreneurs gain about 53% from labor-market power. These patterns reflect heterogeneity in the effective degree of competition across ability segments. High-ability workers face concentrated choice sets dominated by a small number of large, high-paying firms, while low-ability workers are excluded from most employers and are employed primarily by a limited set of smaller, lower-productivity firms. For workers at the bottom of the distribution, these losses are especially consequential from a redistributive perspective, because labor-market power further depresses a consumption level that is already low in the absence of monopsony.

Wage inequality. Eliminating markdowns affects wage dispersion through three channels. First, it raises wages disproportionately at the top of the worker distribution, because high-ability workers face more-concentrated effective labor markets and therefore larger equilibrium markdowns. Second, it strengthens assortative matching: in the baseline equilibrium, high-productivity firms impose relatively large markdowns on high-ability workers, which dampens their supply to those firms; absent markdowns, high-ability workers reallocate toward higher-productivity firms, while low-ability workers reallocate toward lower-productivity firms. As a result, assortative matching tends to increase when workers are paid their marginal product. Both channels steepen the wage schedule across worker types within LLMs. Third, it compresses the between-market component of wage inequality associated with differential labor-market power. In the calibration, this between-firm compression dominates the within-market steepening, so aggregate wage inequality falls: the standard deviation of log wages declines from 0.366 in the baseline equilibrium to 0.322.

Related Literature. This paper relates to three literatures: monopsony, misallocation, and assortative matching.

Monopsony and oligopsonistic labor markets. This paper relates to the literature on employer market power (Robinson, 1933; Manning, 2003). It is most closely related to BHM, which extends the tools developed in Atkeson et al. (2008) to the preference-heterogeneity framework of Card et al. (2018), in a homogeneous-labor environment, thereby generating firm-level heterogeneity in equilibrium markdowns, which in turn leads to misallocation. More generally, the framework developed by BHM has proved useful for rationalizing variation in markdowns arising from differences in firms' competitive environments (e.g., Yeh et al., 2022), strategic wage-setting interactions (e.g., Staiger et al., 2010), and incomplete pass-through of shocks to wages (e.g., Kline et al., 2019). A closely related contemporaneous paper is Bils et al. (2025), which, like this paper, allows markdowns to vary within firms.[8] They introduce worker–firm heterogeneity through exogenous firm-specific productivity draws, and production is linear in efficiency units, so coworkers do not interact in production. Moreover, their analysis is developed in symmetric oligopsony or asymmetric duopsony settings.

[8]This paper is also related to Volpe (2024), where firms are atomistic and heterogeneous markdowns arise from heterogeneity in workers' wage–amenity trade-offs and the resulting sorting of workers across firms.

I retain the richer general-equilibrium structure developed by BHM while introducing worker heterogeneity and allowing firms' labor demand to vary endogenously across worker types. Worker ability, firm productivity, and coworker composition jointly determine firms' demand for different workers and, therefore, who competes for whom in equilibrium. As a result, workers endogenously face different sets of competing employers, so the intensity of competition and the resulting markdowns vary with worker ability rather than being common across workers within a firm. Relative to BHM and Bils et al. (2025), this richer structure makes it possible to study questions that lie outside homogeneous-labor environments, including markdowns and welfare losses by worker ability, assortative matching and segregation, and the effects of labor-market power on wage inequality once the worker-heterogeneity and sorting margins emphasized by the recent literature (e.g., Song et al., 2019) are taken into account. At the same time, the framework still allows for the study of misallocation and the firm-size distribution.

This paper also relates to Felix (2026), who studies the impact of trade on labor-market power. Methodologically, the author shows how to estimate labor supply elasticities in a homogeneous-worker environment. I contribute by extending this approach to a framework with unobserved worker heterogeneity and by using unexpected worker deaths, constructed as in Jäger et al. (2024), as demand shocks to replacement hiring.[9]

Misallocation and aggregate costs of microdistortions. This paper also relates to the literature quantifying the aggregate output and welfare costs of microlevel distortions and misallocation (Harberger, 1954; Restuccia et al., 2008; Hsieh et al., 2009; Hopenhayn, 2014). The closest paper to mine is Edmond et al. (2023), which uses production-function-based methods to recover firm markups from firm-level observables and map them into aggregate productivity losses. In a similar spirit, I use the model as a measurement device to infer worker-specific labor wedges and map them into aggregate efficiency losses and the distribution of welfare. This paper is also related to the literature estimating market power from balance-sheet data (De Loecker et al., 2020; Yeh et al., 2022; De Ridder et al., 2026), especially Yeh et al. (2022), which applies this approach to estimate markdowns in the United States. Relative to Yeh et al., 2022, I show that, in the presence of worker heterogeneity, the estimation of average markdowns is contaminated by a covariance term. Reassuringly, this term is quantitatively negligible in the calibration.

Assortative matching. Finally, this paper relates to work on firm and worker heterogeneity, coworker interactions, and firm-specific wage components in matched employer–employee data (Abowd et al., 1999; Moretti, 2004; Mas et al., 2009; Card et al., 2013; Bender et al., 2018; Song et al., 2019; Lamadon et al., 2022), as well as to theoretical work on assortative matching between workers and firms (Becker, 1973; Shimer et al., 2000; Saint-Paul, 2001; Costinot et al., 2010; Helpman et al., 2010; Eeckhout et al., 2018). I contribute a parsimonious and tractable framework that jointly accommodates flexible labor-market segregation, the firm size distribution, and existing benchmark cases. I also derive theoretical results linking assumptions on technology to assorta-

[9] With linear production in efficiency units, losing one worker does not affect coworkers' marginal products. The fact that worker-death shocks affect coworker wages therefore suggests within-firm production interdependencies, for instance through decreasing returns to labor.

tive matching and markdowns by worker type. I further provide a set of empirical regularities on assortative matching, worker-type-specific concentration, and firms' hiring practices by pay premium, in data from Germany and Italy.

The rest of this paper proceeds as follows. Section 1 presents the model. Section 2 describes the data and empirical evidence. Section 3 outlines the calibration and estimation strategy. Section 4 presents the main quantitative results. Section 5 concludes.

1 The Model

In this section, I develop and characterize a general-equilibrium model of sorting and ability-based segregation in the labor market. Time is discrete and infinite, and the analysis focuses on the steady state. To simplify notation, I suppress time subscripts except when stating decision problems explicitly. All derivations and proofs are in Supplemental Appendix A.

1.1 Environment

Agents — The economy consists of households (workers) with discrete abilities $a \in \mathcal{A}$, where $\mathcal{A}$ is countable. Abilities are distributed according to the cumulative distribution function $F_a(a)$, with associated probability mass function $f_a(a)$. The economy consists of a continuum of LLMs, indexed by $j \in [0, 1]$. Each market draws a random number of firms, $m_j \sim F_m(m)$, which constitutes the sole ex ante source of heterogeneity across markets. Within market j, firms are indexed by i and draw latent baseline productivity z from $F_z(z)$.

Income and rents — Firms are owned by the representative entrepreneur, who receives profits as lump-sum payments and accumulates capital rented to firms. Workers supply labor and earn wage income, while the entrepreneur receives profits and capital rents.[10] There is no trade in bonds or other financial assets, so workers consume current labor income each period.

Notation — Subscripts (i, j) index firm i in market j, and the argument a indexes worker ability. Throughout, $x_{ij}(a)$ denotes the equilibrium value of variable x for a worker of ability a employed at firm (i, j); for example, $w_{ij}(a)$ denotes the equilibrium wage of an ability-a worker at firm (i, j).

1.2 Workers

Preferences are concave, and labor disutility follows the nested-CES specification in BHM. Utility is defined at the level of each ability group and expressed in per-capita terms by normalizing consumption and labor by the mass $f_a(a)$ of type-a workers. Since product-market power is not

[10] For tractability, I impose a sharp distinction between workers—who derive nearly all of their income from labor—and the entrepreneur, who receives profits. This assumption is consistent with evidence from the *Survey of Consumer Finances*, which shows that most households depend primarily on labor income, whereas a small subset derives most of its income from capital (see, for example, Berger et al., 2025).

central to the analysis, final goods are perfect substitutes, implying no markups.[11] The consumption good is the numéraire.

Let $\mathcal{S}_j(a)$ denote the set of firms offering jobs to workers of ability a in LLM j, as determined endogenously in equilibrium. The household allocates labor supply $n_{ijt}(a)$ across firms in its choice set and chooses consumption bundles $c_{ijt}(a)$ to maximize discounted utility, taking wages $\{w_{ijt}(a)\}$ and choice sets $\{\mathcal{S}_j(a)\}$ as given:

$$\begin{aligned} U_0(a) = \max_{\{n_{ijt}(a),\, c_{ijt}(a)\}_{t\geq 0}} & \sum_{t=0}^{\infty} \beta^t \left[\frac{1}{1-\sigma} \left(\frac{C_t(a)}{f_a(a)} \right)^{1-\sigma} - \frac{1}{1+\frac{1}{\varphi}} \left(\frac{N_t(a)}{f_a(a)} \right)^{1+1/\varphi} \right] \\ \text{s.t.} \quad C_t(a) = & \int_0^1 \sum_{i \in \mathcal{S}_j(a)} w_{ijt}(a)\, n_{ijt}(a)\, dj \end{aligned} \tag{1}$$

with $\eta \geq \theta$, where

$$C_t(a) \coloneqq \int_0^1 \sum_{i=1}^{m_j} c_{ijt}(a)\, dj, \qquad N_t(a) \coloneqq \left[\int_0^1 n_{jt}(a)^{\frac{\theta+1}{\theta}}\, dj \right]^{\frac{\theta}{\theta+1}}, \qquad n_{jt}(a) \coloneqq \left[\sum_{i \in \mathcal{S}_j(a)} n_{ijt}(a)^{\frac{\eta+1}{\eta}} \right]^{\frac{\eta}{\eta+1}}$$

Discussion. The nested-CES labor-supply system, microfounded by BHM, belongs to the broader class of *preference heterogeneity* models surveyed in Manning (2021) and Card (2022).[12] Workers choose employers based on wages and idiosyncratic taste shocks, which capture commuting costs, relocation costs, preferences for firm culture, and work environment, thereby generating imperfect substitutability across firms. Workers therefore need not supply labor to the highest-wage firm, generating an upward-sloping firm-level labor supply curve.

The parameter η governs substitution across firms within a market, while θ governs substitution across markets. When $\eta > \theta$, intramarket substitution is easier than intermarket substitution. As $\eta \to \infty$, firms within a market become perfect substitutes and workers allocate all labor to the highest-wage firm in $\mathcal{S}_j(a)$; as $\theta \to \infty$, markets become perfect substitutes. The atomistic monopsony model arises as the special case $\eta = \theta$.

As in Berger et al. (2025), I assume that income is pooled within ability groups, but not across ability types, through the presence of a representative household for each type a.

[11] Allowing for monopolistic competition with a constant markup would not alter any result. Firms would set prices as a constant markup over marginal cost and maximize a decreasing-returns revenue function rather than a production function. All equilibrium conditions remain unchanged.

[12] Following BHM, this labor-supply structure arises from workers' discrete choices along three margins: whether to work, which market to select, and which firm to select within a market. Correlated Gumbel taste shocks then yield the nested-CES structure, with θ governing dispersion across markets and η governing dispersion across firms within markets. Supplemental Appendix A.1 shows that this microfoundation is unchanged by the endogenous choice set $\mathcal{S}_j(a)$.

Optimality conditions. The first-order conditions imply the inverse labor-supply system

$$\left(\frac{N(a)}{f_a(a)}\right)^{\frac{1}{\varphi}+\sigma} = W(a)^{1-\sigma}, \qquad w_{ij}(a) = \left(\frac{n_{ij}(a)}{n_j(a)}\right)^{1/\eta} \left(\frac{n_j(a)}{N(a)}\right)^{1/\theta} W(a) \tag{2}$$

where $W(a)$ is the aggregate wage index. Define $w_j(a)n_j(a) = \sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)n_{ij}(a)$ and $W(a)N(a) = \int_0^1 w_j(a)n_j(a)\,dj$. Using these definitions and Equation (2),

$$w_j(a) = \left(\sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)^{1+\eta}\right)^{1/(1+\eta)}, \qquad W(a) = \left(\int_0^1 w_j(a)^{1+\theta}\,dj\right)^{1/(1+\theta)} \tag{3}$$

1.3 Representative Entrepreneur

The representative entrepreneur e has concave preferences over consumption of the final good and chooses next-period capital K_{t+1} and consumption bundles $c_{ijt}(e)$ to maximize discounted utility. The entrepreneur rents capital to firms, which demand k_{ijt}, and receives their profits π_{ijt} as lump-sum payments. Given an initial capital stock K_0, the problem is

$$U_0(e) = \max_{\{K_{t+1},\, c_{ijt}(e)\}_{t\geq 0}} \sum_{t=0}^{\infty} \beta^t \frac{C_t(e)^{1-\sigma}}{1-\sigma} \quad \text{s.t.} \quad C_t(e) + K_{t+1} - (1-\delta)K_t = \Pi_t + R_t K_t \tag{4}$$

where

$$C_t(e) \coloneqq \int_0^1 \sum_{i=1}^{m_j} c_{ijt}(e)\,dj, \qquad K_t \coloneqq \int_0^1 \sum_{i=1}^{m_j} k_{ijt}\,dj, \qquad \Pi_t \coloneqq \int_0^1 \sum_{i=1}^{m_j} \pi_{ijt}\,dj \tag{5}$$

Optimality condition. The steady-state entrepreneur's capital accumulation Euler equation is

$$1 = \beta(R + 1 - \delta) \tag{6}$$

1.4 Firms

Firm (i, j) draws an exogenous productivity type z_{ij} from F_z. A worker of ability $a \in \mathcal{A}$ employed by a firm of type z produces $\phi(a, z)$ units of output per unit of capital, where $\phi(a, z)$ is the worker–firm output function.

Let k_{ijt} denote capital. Total employment and the induced ability distribution are $h_{ijt} = \sum_{a\in\mathcal{A}} n_{ijt}(a)$ and $g_{ijt}(a) = n_{ijt}(a)/h_{ijt}$. I assume that the firm's production function is given by

$$y_{ijt} = \mathbb{E}_{g_{ijt}}[\phi(a, z_{ij})] \left(k_{ijt}^{1-\gamma} h_{ijt}^{\gamma}\right)^{\alpha}, \qquad \mathbb{E}_{g_{ijt}}[\phi(a, z_{ij})] \equiv \sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, g_{ijt}(a) \tag{7}$$

Thus, realized productivity is endogenous because it depends on workforce composition.

Quantitative specification. In the quantitative analysis, I assume that $\phi(a, z)$ takes the CES form

$$\phi(a,z) = \left[(1-\omega_a)\, z^{\frac{\rho-1}{\rho}} + \omega_a\, a^{\frac{\rho-1}{\rho}}\right]^{\frac{\rho}{\rho-1}}, \qquad \rho \leq 1,\ \omega_a \in [0,1] \tag{8}$$

For $\rho \in (0,1)$ and $\omega_a \in (0,1)$, $\phi(a,z)$ is strictly log-supermodular.

Special cases. The CES specification in Equation (8) nests several benchmark technologies:

1. **Cobb–Douglas benchmark.** When $\omega_a = 0$, worker output depends only on firm productivity, so $\phi(a,z) = z$ and
$$y = z\,(k^{1-\gamma}h^{\gamma})^{\alpha}$$
This nests the production side of BHM.[13]

2. **Multiplicative complementarities.** As $\rho \to 1$, the CES aggregator becomes log-linear:
$$\phi(a,z) = z^{1-\omega_a} a^{\omega_a} \qquad \Rightarrow \qquad y = z^{1-\omega_a}\, \mathbb{E}_g\left[a^{\omega_a}\right]\,(k^{1-\gamma}h^{\gamma})^{\alpha}$$
This is similar to Helpman et al. (2010), where output depends on firm productivity and average worker ability.

3. **Additive aggregation.** When $\alpha = \gamma = 1$, output is linear in labor efficiency units:
$$y = \sum_{a \in \mathcal{A}} \phi(a,z)\, n(a)$$
This yields the additive structure in Costinot et al. (2010).[14]

Discussion. Supplemental Appendix A.2 provides a microfoundation for Equation (7). Suppose worker-level output is $f(a,z,\kappa(a)) = \phi(a,z)\kappa(a)^{1-\gamma}$, so total output is $y = \sum_a \phi(a,z)\kappa(a)^{1-\gamma}$, where $\kappa(a)$ is the capital allocated to type-a workers. If managers cannot condition capital allocation on a, then Equation (7) follows.[15] Under this interpretation, $\phi(a,z)$ is the output of a type-a worker employed by a type-z firm with a standardized bundle of common inputs, and $\mathbb{E}_g[\phi(a,z)]$ is realized firm productivity. This mechanism builds on Kremer (1993), Saint-Paul (2001), and Helpman et al. (2010), and is closely related to theories of sorting and productivity interactions such as in Shimer et al. (2000), Costinot et al. (2010), and Eeckhout et al. (2018).

More broadly, the specification parsimoniously captures how workers interact in production and how workforce composition shapes firm productivity (e.g., Moretti, 2004; Gennaioli et al.,

[13] The production side coincides with BHM, but the income distribution does not. In BHM, a representative household receives wages, profits, and capital income. Here, entrepreneurs own capital and collect profits, while workers receive only wages. With income effects affecting the labor supply, equilibrium allocations may differ.

[14] As in Costinot et al. (2010), one can introduce downward-sloping demand with parameter $\zeta \neq \alpha$. IN that case, the results of Section 1.5 apply to marginal revenue product rather than to marginal product.

[15] For example, workers may share office space, equipment, or support staff that cannot be tailored to individual ability, so each worker effectively receives the same bundle of common inputs.

2013; Bender et al., 2018). A worker's marginal product, derived below in Equation 11, depends on her coworkers through two distinct channels. First, more productive coworkers raise realized firm productivity and, through this channel, tend to increase the marginal product of all workers, consistent with evidence on within-firm complementarities and peer or social-pressure effects (e.g., Ichino et al., 2000; Falk et al., 2006; Mas et al., 2009; Bandiera et al., 2010). Second, a worker's marginal product depends on how her own productivity compares with the average productivity of her coworkers. In that sense, the framework can also be viewed as a reduced-form way of capturing mechanisms through which workers' relative standing within the firm affects effective performance, for example through morale and effort responses to within-firm pay comparisons (e.g., Cohn et al., 2014; Breza et al., 2018). The model does not explicitly represent the underlying microstructure, but it captures its central implications in a tractable way. It generates rich patterns of sorting and segregation with a small number of interpretable parameters, nests standard benchmark environments, and provides a unified lens through which to interpret the empirical evidence presented in the rest of the paper.

Firm's problem. Given the production function above, firms are infinitesimal in the aggregate economy but granular within LLMs. They take aggregate quantities $N_t(a)$ and $W_t(a)$ as given, but they internalize the effect of their own hiring on market-level quantities $n_{jt}(a)$ and $w_{jt}(a)$.

Taking R_t and competitors' employment as given,[16] firm (i,j) chooses capital k_{ijt}, total employment h_{ijt}, and the allocation of employment across worker types, $n_{ijt}(a) \geq 0$:

$$\begin{aligned} \pi_{ij,t} = \max_{\{n_{ij,t}(a)\}_{a\in\mathcal{A}},\, h_{ij,t},\, k_{ij,t}} & \left\{ \mathbb{E}_{g_{ij,t}}\left[\phi(a,z_{ij})\right] \left(k_{ij,t}^{1-\gamma} h_{ij,t}^{\gamma}\right)^{\alpha} - R_t k_{ij,t} - h_{ij,t}\, \mathbb{E}_{g_{ij,t}}\left[w_{ij,t}(a)\right] \right\} \\ \text{s.t.} \quad & n_{ij,t}(a) \geq 0 \;\forall a \in \mathcal{A}, \qquad h_{ij,t} = \sum_{a\in\mathcal{A}} n_{ij,t}(a), \qquad g_{ij,t}(a) = \frac{n_{ij,t}(a)}{h_{ij,t}} \end{aligned} \tag{9}$$

where wages satisfy the inverse labor-supply condition in Equation (2). The capital FOC implies

$$\left(\mathbb{E}_{g_{ij}}\left[\phi(a,z_{ij})\right]\right)^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\frac{\gamma\alpha}{1-\alpha(1-\gamma)}} = \frac{R\, k_{ij}}{(1-\gamma)\alpha} \tag{10}$$

Substituting out optimal capital, hiring an additional worker of type a affects output through two channels: a positive *size effect*, because it raises total employment h, and a *composition effect*, because it changes average workforce productivity.[17] The composition effect is negative whenever

[16] For notational simplicity, I suppress firm and competitor indices hereafter. Thus, when writing $w_{ij}(a)$, this should be understood as $w_{ij}\big(a, n_{ij}(a), n^{*}_{-ij}(a), N(a), W(a)\big)$, as implied by the inverse labor-supply system in (2).

[17] Formally, the two channels are

$$\frac{\partial y}{\partial h}\frac{dh}{dn(a)} \qquad \text{and} \qquad \frac{\partial y}{\partial \mathbb{E}_{g_{ij}}[\phi(a,z_{ij})]}\frac{d\,\mathbb{E}_{g_{ij}}[\phi(a,z_{ij})]}{dn(a)}$$

The size effect coincides with the marginal product under homogeneous labor with productivity $\mathbb{E}_{g_{ij}}[\phi(a,z_{ij})]$.

$\phi(a, z_{ij}) < \mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]$. Combining the two channels yields the marginal product of type-a labor:

$$MPL_{ij}\big(a \mid \mathbb{E}_{g_{ij}}[\phi(\cdot)], h_{ij}\big) = \overline{MPL}_{ij}\left[1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\phi(a, z_{ij})}{\mathbb{E}_{g_{ij}}\big[\phi(a, z_{ij})\big]}\right)\right],$$

$$\overline{MPL}_{ij} \equiv Z\alpha\gamma\Big(\mathbb{E}_{g_{ij}}\big[\phi(a, z_{ij})\big]\Big)^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\frac{\alpha-1}{1-\alpha(1-\gamma)}}, \qquad Z := \left(\frac{\alpha(1-\gamma)}{R}\right)^{\frac{(1-\gamma)\alpha}{1-(1-\gamma)\alpha}} \tag{11}$$

For expositional simplicity, I refer to this object as $MPL_{ij}(a)$. If $\alpha = \gamma = 1$, $MPL_{ij}(a) = \phi(a, z_{ij})$.

Under decreasing returns to labor ($\alpha\gamma < 1$), the negative composition effect may dominate the positive size effect, so hiring sufficiently low-ability workers can reduce, and even make negative, the marginal product of labor. The key implication is that the marginal product is an equilibrium object: it depends not only on the worker's own productivity $\phi(a, z_{ij})$ but also on the endogenous composition of the firm's workforce and on firm size. Intuitively, each worker contributes to output but also occupies a *spot*—a share of common firm resources such as capital. With decreasing returns, an additional spot carries a shadow cost because it dilutes resources per worker. Under constant returns to labor, this dilution effect vanishes, and $MPL_{ij}(a) = \phi(a, z_{ij})$.

Coworker interactions in production. I next characterize how coworker composition affects a worker's marginal product within the firm. This is the channel through which coworker interdependencies in production generate wage differences, sorting, and segregation across firms in equilibrium. Consider a change in the within-firm ability distribution g_{ij} that raises realized firm productivity $\mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]$, holding both firm size h_{ij} and worker ability a fixed. Under $\alpha\gamma < 1$, this affects marginal products through two channels.

First, an increase in $\mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]$ raises firm efficiency and, through the capital first-order condition, induces capital deepening. Since firm efficiency scales every worker's marginal product and capital and labor are complements, both forces raise the marginal product of all workers. This channel is consistent with the evidence in Mas et al. (2009): more productive coworkers raise individual productivity, measured as marginal product.

Second, an increase in $\mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]$ raises the value of productive capacity and hence the opportunity cost of allocating that capacity to a low-productivity worker. As average productivity rises, a given worker becomes a weaker component in production, lowering that worker's effective marginal product. This mechanism is reminiscent of the O-ring logic in Kremer (1993). Under constant returns to labor ($\alpha = \gamma = 1$), capital is absent and $MPL_{ij}(a) = \phi(a, z_{ij})$, so marginal products are independent of coworker composition. Under decreasing returns, by contrast, the negative O-ring force dominates for sufficiently low-ability workers, whereas the positive firm-efficiency and capital-deepening forces dominate for sufficiently high-ability workers.

More formally,

$$\frac{\partial MPL_{ij}(a)}{\partial \mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]} \;\propto\; \left[\phi(a, z_{ij}) - \mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]\,\frac{1-\alpha\gamma}{\alpha(1-\gamma)}\right]$$

where the omitted proportionality factor is strictly positive. Since $\phi(a, z_{ij})$ is strictly increasing in a, there exists at most one cutoff ability a^* such that the bracketed term is negative for all $a < a^*$ and positive for all $a > a^*$. Thus, productivity improvements driven by more productive coworkers lower the marginal product of low-ability workers but raise that of high-ability workers, if such a cutoff lies in the support of $\mathcal{A}$, with the effect becoming more positive as ability rises.

Firm wages. Given Equation (11) and the firm's static problem in Equation (9), I now characterize how firms map marginal products into wages. For worker types with positive marginal product, the first-order condition for $n_{ij}(a)$ implies a Lerner-type condition: the equilibrium wage equals an endogenous firm–worker-specific markdown $\mu_{ij}(a) \leq 1$ applied to the marginal product of labor. If $MPL_{ij}(a) \leq 0$, the firm does not hire that type, so $w_{ij}(a) = 0$. Proposition 1 summarizes the wage schedule.

Proposition 1. *Let $MPL_{ij}(a)$ be given by Equation* (11)*. A profit-maximizing firm's wage schedule satisfies*[18]*,*[19]

$$
\begin{aligned}
w_{ij}(a) &= \begin{cases} \mu_{ij}(a)\, MPL_{ij}(a) & \text{if } MPL_{ij}(a) > 0, \\ 0 & \text{if } MPL_{ij}(a) \leq 0, \end{cases} \\
\mu_{ij}(a) &= \frac{\epsilon_{ij}(a)}{\epsilon_{ij}(a) + 1}, \qquad \epsilon_{ij}(a) := \left[\frac{\partial \log w_{ij}(a)}{\partial \log n_{ij}(a)} \Big|_{n^*_{-ij}(a)} \right]^{-1}
\end{aligned}
\tag{12}
$$

where $\epsilon_{ij}(a)$ is the firm-specific labor-supply elasticity for worker type a. Under the specified preferences,

$$
\epsilon_{ij}(a) = \left[\frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta} \right) s_{ij}(a) \right]^{-1}, \qquad s_{ij}(a) = \frac{w_{ij}(a) n_{ij}(a)}{\sum_{i \in \mathcal{S}_j(a)} w_{ij}(a) n_{ij}(a)} \tag{13}
$$

Firm labor-market power. Proposition 1 implies that workers receive a fraction $\mu_{ij}(a)$ of their marginal product. Thus, $\mu_{ij}(a)$ is a markdown that directly measures the labor-market power exercised by firm (i, j) over worker type a. This wedge is monotonically related to the firm-specific labor-supply elasticity $\epsilon_{ij}(a)$, which depends on the firm's wage-bill market share $s_{ij}(a)$ and varies across firms, markets, and worker types. Because competition is defined with respect to a worker's relevant choice set, a firm may be small in the aggregate economy but large relative to the set of employers available to a given worker type. When the firm accounts for a substantial share of that set, it internalizes the worker's outside option, and the markdown is shaped by the lower elasticity of substitution θ. When its share is small ($s_{ij}(a) \approx 0$), the markdown is governed primarily by the within-market elasticity η, as in standard atomistic models.

[18] Equation (12) characterizes any profit-maximizing wage schedule. In the quantitative implementation, I impose (12) and solve the firm's problem as a fixed point in $(\mathbb{E}_{g_{ij}}[\phi(a, z_{ij})], h_{ij})$, iterating on employment choices $\{n_{ij}(a)\}$ until convergence; see Supplemental Appendix C.1.

[19] Because global concavity has not been established in general for the problem associated with (9), sufficiency of the first-order conditions is not guaranteed. I therefore conduct a grid-based numerical search for profitable unilateral deviations in the dual $(\mathbb{E}_{g_{ij}}[\phi(a, z_{ij})], h_{ij})$ formulation; see Supplemental Appendix C.2.

The next result shows that labor-market power shapes firm-level wages, profits, and the labor share through a single sufficient statistic. This statistic can be identified from observables using the same production-function logic used to measure markups and markdowns (e.g., De Loecker et al., 2020; Yeh et al., 2022), a connection I exploit in the empirical analysis.[20]

Proposition 2 (firm-level aggregation)**.** *Define the within-firm averages,*

$$\overline{\mu}_{ij} := \sum_{a \in \mathcal{A}} \mu_{ij}(a)\, g_{ij}(a), \qquad \overline{MPL}_{ij} := \sum_{a \in \mathcal{A}} MPL_{ij}(a)\, g_{ij}(a), \qquad \overline{w}_{ij} := \sum_{a \in \mathcal{A}} w_{ij}(a)\, g_{ij}(a)$$

the firm-level labor share, and labor-market-power wedge

$$ls_{ij} := \frac{h_{ij}\overline{w}_{ij}}{y_{ij}}, \qquad \widetilde{\psi}_{ij} := \overline{\mu}_{ij} + \frac{1}{\alpha\gamma}\, \mathrm{cov}_{g_{ij}}\left(\mu_{ij}(a), \frac{\phi(a, z_{ij})}{\mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]}\right)$$

where the covariance is taken with respect to $g_{ij}(a)$. Then $\widetilde{\psi}_{ij} \leq 1$ and

$$\overline{MPL}_{ij} = \alpha\gamma\, \frac{y_{ij}}{h_{ij}}, \qquad \frac{\pi_{ij}}{y_{ij}} = 1 - \alpha(1-\gamma) - \alpha\gamma\, \widetilde{\psi}_{ij},$$
$$\overline{w}_{ij} = \overline{MPL}_{ij}\, \widetilde{\psi}_{ij}, \qquad ls_{ij} = \alpha\gamma\, \widetilde{\psi}_{ij}$$

Without markdowns, $\widetilde{\psi}_{ij} = 1$, so the average wage equals the average marginal product. In this benchmark, the firm's labor and profit shares coincide with those implied by a Cobb–Douglas technology with homogeneous labor: labor receives a constant share $\alpha\gamma$ of revenue, capital absorbs $\alpha(1-\gamma)$, while profits receive the residual share $1-\alpha$.

Proposition 2 shows that, with oligopsony power, the effects of markdowns on firm-level wages, profits, and labor shares are summarized by the single wedge $\widetilde{\psi}_{ij}$. This wedge has two components. The first is the average markdown $\overline{\mu}_{ij}$: stronger monopsony power lowers $\overline{\mu}_{ij}$ and $\widetilde{\psi}_{ij}$, reducing the labor share and raising profits. The second is the covariance term, which captures how markdowns covary with workers' relative output $\phi(a, z_{ij})/\mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]$.[21] A more negative covariance means that the firm applies larger markdowns precisely to the workers who contribute most to output, amplifying the reduction in the labor share beyond what the average markdown alone would imply.

This characterization also clarifies what standard production-function methods identify. Empirically, a common approach measures "the" average markdown by comparing a firm's average marginal revenue product of labor to its average wage, using output, revenue, and input shares to infer the former (e.g., Yeh et al., 2022). In the present environment, where markdowns vary across

[20] This identification result exploits the Cobb–Douglas structure in (k, h) and the absence of product-market markups, which allow wedges to be read directly from the firm's labor share. Supplemental Appendix Lemma E.6 shows that the same logic extends to more general technologies and to settings with heterogeneous markups: as long as worker heterogeneity enters the marginal product of labor as $MPL_{ij,t}(a) = (\partial y_{ij,t}/\partial h_{ij,t})\psi_{ij,t}(a)$ for some function $\psi_{ij,t}(a)$ and firms use a flexible input, the wedge $\widetilde{\psi}_{ij,t}$ is recovered from the ratio of labor-input cost share to flexible-input cost share, together with the ratio of output elasticities, exactly as in Yeh et al. (2022).

[21] The factor $1/(\alpha\gamma)$ rescales this revenue-based covariance into a wedge in marginal-product units, since $\alpha\gamma$ is the effective output elasticity, and hence revenue share, of labor.

worker types, this comparison still recovers a meaningful object, but it is the model-consistent wedge $\widetilde{\psi}_{ij}$ rather than the simple average markdown $\overline{\mu}_{ij}$. The empirical wedge inferred from revenue, average wages, and employment therefore provides a sufficient statistic for how firm-level markdowns translate into firm-level profits and the labor share, incorporating both the level and composition effects of labor-market power.

Having characterized wages, marginal products, and firm-level wedges, I now define the steady-state general equilibrium of the economy.

Definition 1 (steady-state general equilibrium)**.** Given primitives, a steady-state general equilibrium is a collection $\{n_{ij}(a), k_{ij}, \pi_{ij}, w_{ij}(a),\, K, C(a), C(e), R\}_{i,j,a}$ such that, for each $a \in \mathcal{A}$, households solve their static problem given wages and choice sets, with steady-state allocations satisfying Equation (2); the return to capital satisfies Equation (6), with $K = \int_0^1 \sum_{i=1}^{m_j} k_{ij}\, dj$; each firm (i, j) maximizes profits taking R, aggregate objects, and labor supply as given, with steady-state choices satisfying (10)–(11) and (12); goods-market clearing holds, $\int_0^1 \sum_{i=1}^{m_j} y_{ij}\, dj = \sum_{a\in\mathcal{A}} C(a) + C(e) + \delta K$; budget constraints satisfy $C(a) = \int_0^1 \sum_{i\in\mathcal{S}_j(a)} w_{ij}(a) n_{ij}(a)\, dj$, $C(e) = \Pi + (R - \delta)K$; and all aggregator identities hold by definition.

Efficient benchmark. I compare the decentralized equilibrium to a steady-state general equilibrium (Definition 1) with no markdowns:

$$\mu_{ij}(a) = 1 \quad \text{and hence} \quad w_{ij}(a) = MPL_{ij}(a) \quad \forall (i, j, a)$$

which I refer to as a *competitive-production efficient equilibrium*. Proposition A.1 in Supplemental Appendix A.5 shows that, for some welfare weights, this equilibrium coincides with the planner's allocation in terms of production. Markdowns therefore create wedges between wages and marginal products: firms paying workers less than their marginal products attract too few of them relative to the competitive-production efficient allocation, misallocating labor toward other firms or leisure.

1.5 Market Equilibrium: Sorting, Segmentation, and Labor-Market Power

In this subsection, I characterize the within-market allocation of workers across firms and its implications for labor-market power, misallocation, and segregation. Taking the supply of each ability type as given, I characterize how assumptions on technology shape hiring patterns and how the resulting equilibrium maps into assortative matching, concentration, and distortions. The goal is twofold: to derive sharp analytical predictions under benchmark cases, and to use their empirical signatures to understand what departures from those benchmarks require and what they reveal about the underlying model parameters.

Within local market j, let $f_{ja}(a)$ denote the mass of type-a workers. I summarize the allocation

of workers across firms by employment shares

$$q_{ij}(a) := \frac{n_{ij}(a)}{f_{ja}(a)}, \qquad \sum_{i \in \mathcal{S}_j(a)} q_{ij}(a) = 1$$

Labor supply is given by (2), and wages satisfy (12).

I proceed in two steps. First, as in Felix (2026), I link markdowns to observables (Lemma 1). I then derive theoretical results for limiting benchmark cases (Lemmas 2 and 3, and Propositions 3 and 4). Second, I illustrate these mechanisms using a simple parametrization which, unless otherwise noted, is the one reported in Table 1. I draw 100 firms from a log-normal distribution with a mean of zero and standard deviation σ_z, and let $f_{ja}(a)$ follow a discretized log-normal distribution with 500 types, a mean of zero, and standard deviation σ_a. This parametrization is not a calibration; rather, it's a device to illustrate the forces identified by the analytical results and the additional patterns generated by the full model.

Table 1: Baseline Parametrization for the Illustrative Labor-Market Exercise

Parameter	ρ	ω_a	γ	α	η	θ	σ_a	σ_z
Value	0.35	0.85	0.70	0.94	10	0.50	0.33	0.40

Lemma 1 (ability-specific average inverse markdown)**.** *For ability type a in market j, let $\mu_{ij}(a)^{-1} = \mathrm{MPL}_{ij}(a)/w_{ij}(a)$ denote the inverse markdown. Then*

$$\sum_i s_{ij}(a)\, \mu_{ij}(a)^{-1} = 1 + \frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big) HHI_j(a) \tag{14}$$

where $HHI_j(a) = \sum_i s_{ij}(a)^2$ is the HHI for type-a workers in market j.

Intuitively, a higher $HHI_j(a)$ means more jobs in fewer firms and less competition.[22]

Technology, markdowns, and worker–firm allocation. A useful benchmark implication is that more productive firms employ more of every worker type and exert greater monopsony power:

Lemma 2 (more productive firms employ more workers and exert greater market power)**.** *Fix a local market j and ability type a, and suppose $\theta < \eta$. If $MPL_{ij}(a)$ is increasing in z_{ij}, then for any pair of firms with $z_{i'j} > z_{ij}$ such that both firms belong to $\mathcal{S}_j(a)$,*

$$q_{i'j}(a) > q_{ij}(a) \quad \textit{and} \quad \mu_{i'j}(a) < \mu_{ij}(a)$$

[22]The ability-specific average inverse markdown, $\sum_i s_{ij}(a)\mu_{ij}(a)^{-1}$, is a convenient summary statistic linking markdowns to concentration and can be recovered from observable wage-bill shares. It is not, however, a sufficient statistic for welfare: even if concentration were observed for every worker type, the structure of the model is still needed to map these wedges into welfare losses.

Lemma 2 is especially transparent in the BHM benchmark. Under $\omega_a = 0$, higher z raises the marginal product of every worker type uniformly, so more productive firms pay higher wages, attract a larger market share that is identical across worker types (i.e., there is no assortative matching), and exert greater labor-market power. Proposition 3 shows that the same allocation result extends to the multiplicative benchmark $\rho \to 1$: after a suitable rescaling of firm productivity, any equilibrium under $\omega_a = 0$ can be replicated under $\rho \to 1$. The two cases therefore imply identical firm shares and distortions when calibrated to the same data; the multiplicative case differs only in generating wage dispersion across worker types. Panel (a) of Figure 1 illustrates this result.[23]

Proposition 3 (invariance from $\omega_a = 0$ to $\rho \to 1$)**.** *Consider the benchmark cases $\omega_a = 0$ and $\rho \to 1$.*

(a) In the limit as $\rho \to 1$, there exists a no-sorting steady-state equilibrium such that $q_{ij}(a) \equiv q_{ij}$.

(b) Consider any steady-state equilibrium under $\omega_a = 0$ with productivities $\{z_{ij}^{(0)}\}$ and equilibrium market shares $\{q_{ij}^{(0)}\}$. For any $\omega_a \in [0, 1)$, under $\rho \to 1$, the same market shares can be replicated by rescaling firm productivities from z to $\widetilde{z}$, as defined in Supplemental Appendix A.6.7, provided that $MPL_{ij}^{(1)}(a) > 0$ for all employed types a.[24]

I now derive a sufficient condition for positive assortative matching, under which higher-ability workers are relatively more likely to work in more productive firms. Combined with Lemma 2, this also implies that higher-ability workers are employed in more-concentrated market segments and face greater distortions. I establish this result analytically in a two-firm benchmark; the numerical exercises show the same qualitative pattern with richer firm-size distributions, and Supplemental Appendix A.6.6 provides sufficient conditions for its extension to an arbitrary number of firms.[25]

Lemma 3 (positive assortative matching in a duopsony)**.** *Consider an LLM j with two firms, i and i', such that $z_{i'j} > z_{ij}$. Suppose $MPL_{ij}(a)$ is log-supermodular in (a, z_{ij}). Then, for any $a' > a$,*

$$\frac{q_{i'j}(a')}{q_{ij}(a')} > \frac{q_{i'j}(a)}{q_{ij}(a)}$$

Hence higher-ability workers are relatively more likely to work in the more productive firm.

Proposition 4 (market concentration and worker ability in a duopsony)**.** *Consider an LLM j with two firms, i and i', such that $z_{i'j} > z_{ij}$. Suppose $MPL_{ij}(a)$ is strictly increasing and strictly log-supermodular in (a, z_{ij}), so that, by Lemma 2, $q_{i'j}(a) > q_{ij}(a)$ for all a, and, by Lemma 3, the employment-*

[23] Any remaining heterogeneity in the figure is negligible and reflects numerical error.

[24] If $MPL_{ij}^{(1)}(a) \leq 0$ for some types, the same result holds on the subset of abilities for which $MPL_{ij}^{(1)}(a) > 0$.

[25] With more than two firms, the mapping from marginal products to wages is multidimensional because all wage ratios jointly determine wage-bill shares and markdowns, making closed-form comparative statics analytically intractable. The two-firm case isolates the interaction between complementarities in $MPL_{ij}(a)$ and endogenous markdowns. With constant markdowns ($\theta \to \eta$), the result extends immediately to any number of firms.

share ratio $q_{i'j}(a)/q_{ij}(a)$ is strictly increasing in a.[26] *Then, for any $a' > a$,*

$$HHI_j(a') > HHI_j(a)$$

To illustrate, consider the parametrization in Table 1, further imposing $\alpha = \gamma = 1$, so that $MPL_{ij}(a) = \phi(a, z_{ij})$. In this case, $MPL_{ij}(a)$ is increasing and log-supermodular in (a, z). By Lemmas 2 and 3, more productive firms employ more of every worker type and disproportionately attract higher-ability workers, even with heterogeneous markdowns, as shown in Figure 1b. Proposition 4 then links these sorting patterns to market structure: $HHI_j(a)$ is strictly increasing in ability, so higher-ability workers are employed in more-concentrated market segments and, by Lemma 1, face higher average inverse markdowns. Figure 1c illustrates this result. Because markdowns are shaped by type-specific market shares, and more productive firms employ more workers of every type, higher-z firms impose larger markdowns on all worker types, with stronger distortions for higher-ability workers. As a result, they are inefficiently compressed toward a smaller operating scale, as illustrated in Figure 1d.

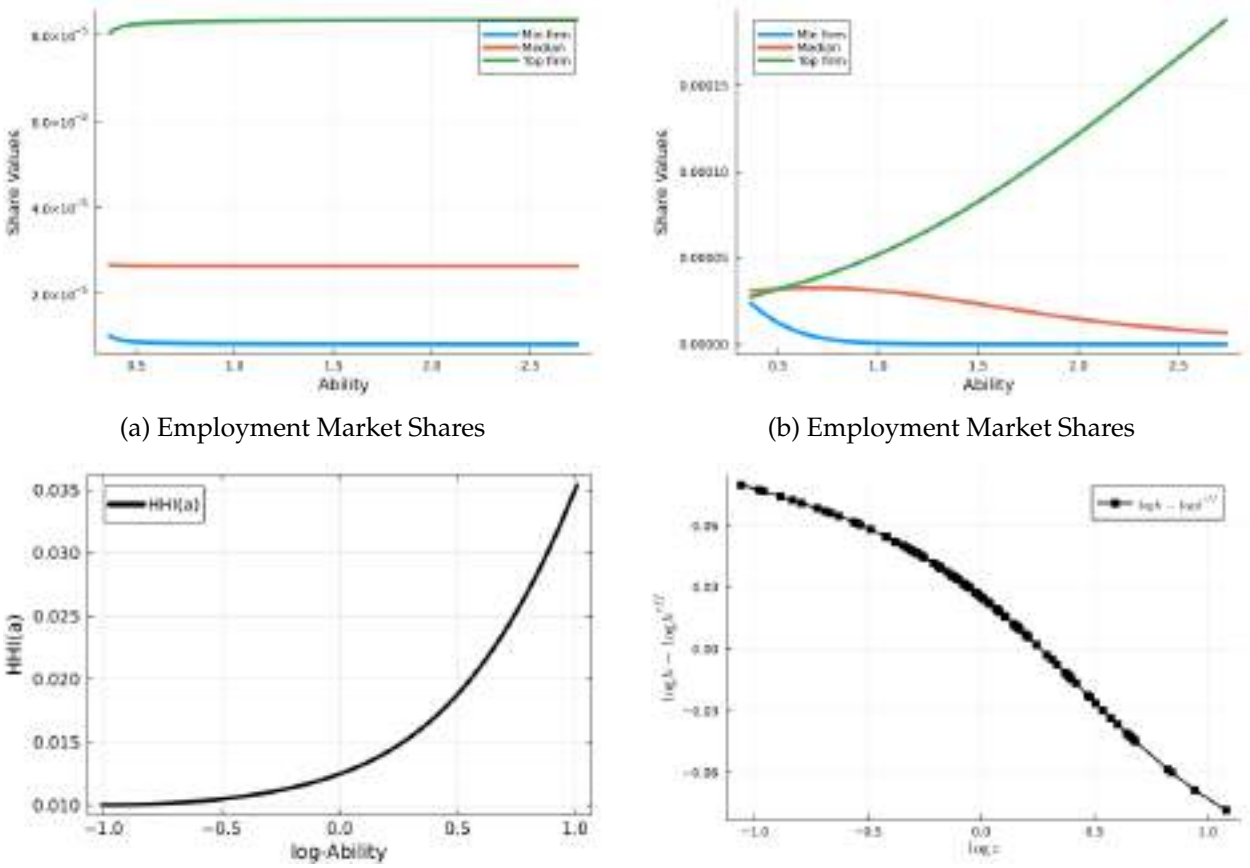
Market equilibrium under coworker interdependencies. Finally, consider the full calibration of Table 1. Now, the marginal product of low-ability workers can be higher in low-productivity firms than in high-productivity ones when the negative composition effect outweighs the positive size effect. Figure 2a reports employment market shares by firm type (the markdown profile is the mirror image) and shows that low-ability workers are segregated into low-type firms, even though these firms have less scope to pay high average wages. Figure 2b plots the minimum ability threshold for employment by firm type, showing that low-ability workers are rationed out of the most productive firms. The labor market thus exhibits both *sorting* and *segregation*. Competition becomes *localized*, as each firm competes primarily with firms of similar productivity targeting similar worker types. Figure 2c reports the corresponding $HHI_j(a)$ profile, which maps directly into the average markdown. At the bottom of the ability distribution, low-productivity firms internalize that low-ability workers have restricted choice sets and that they constitute a large share of the options actually available to these workers—even if they are small relative to the aggregate economy.

This concentration in the LLM allows these firms to set wages further below the efficient benchmark. At the top, high-productivity firms dominate the choice sets of high-ability workers: even though these workers face no rationing, offers from top firms crowd out outside options, generating high effective concentration analogous to Proposition 4.

Turning to production distortions, segmentation weakens the strong positive correlation between firm productivity and distortion that emerged in Benchmark 2, pushing toward an equalization of average markdowns across firms. Comparing Figure 1d to Figure 2d shows how firm-size distortions change once segregation emerges. As segregation intensifies (here, as ω_a increases),

[26] With $N > 2$ firms, the same conclusion holds if each adjacent employment-share ratio $q_{i+1,j}(a)/q_{ij}(a)$ is strictly increasing in a; see Supplemental Appendix Proposition A.2.

less productive firms may acquire greater labor-market power, which can flatten the relationship between productivity and size distortions. To the extent that this weakens the positive correlation between productivity and wedges, it may also reduce the aggregate efficiency cost of labor-market power. Output losses from misallocation are generally thought to be larger when distortions are more pronounced among highly productive firms and when wedges are more dispersed across firms; segmentation tends to attenuate both margins.[27]



(a) Employment Market Shares

(b) Employment Market Shares

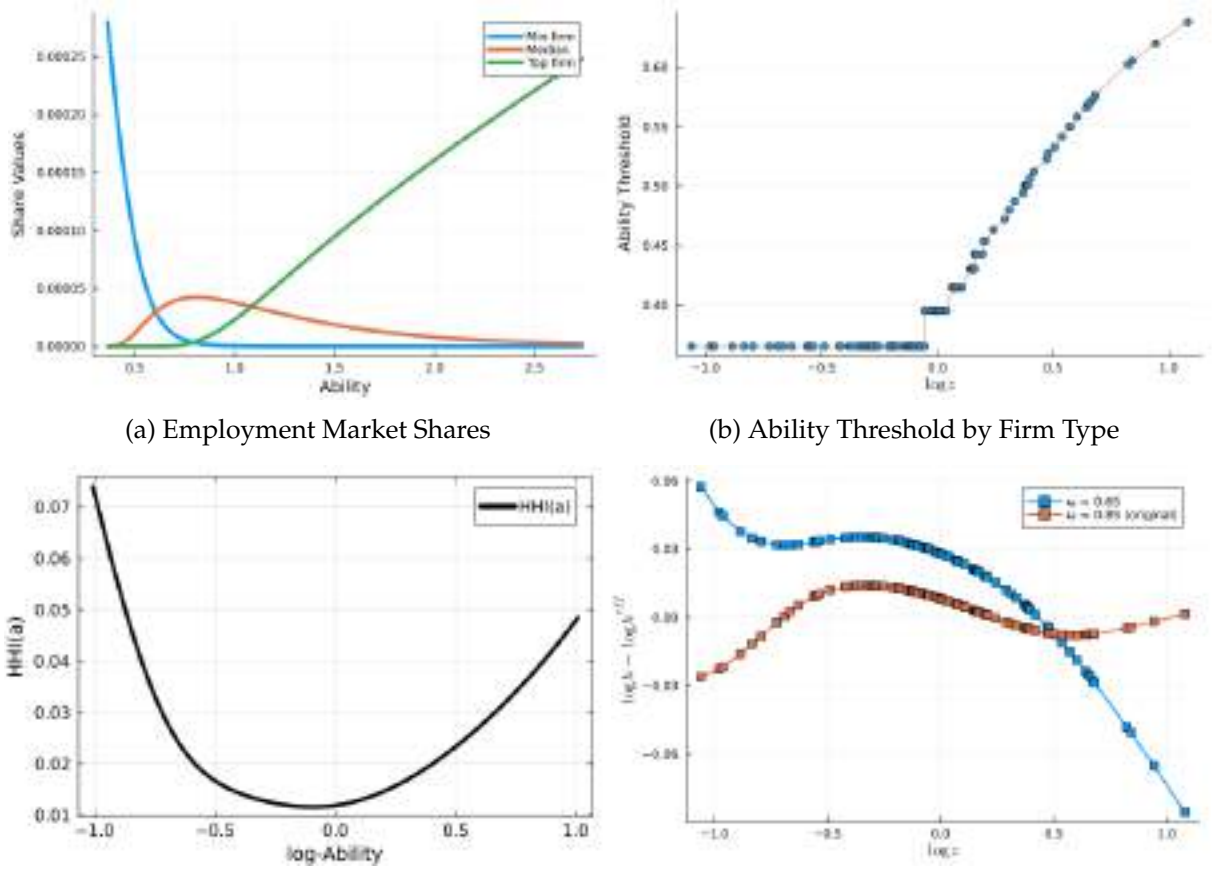
(c) HHI Index by Worker Log-Ability

(d) Size Distortion by Firm Type

Figure 1: Benchmark Equilibria: Market Shares, Size Distortions, and Concentration by Ability

Notes: This figure reports equilibrium outcomes for benchmark parameterizations of the model. Panels (a) and (b) display employment market shares by worker log-ability for three firms (bottom, median, and top of the productivity distribution) under, respectively, the multiplicative benchmark ($\rho \to 1$) and the sorting case with constant returns to labor ($\alpha = \gamma = 1$). Panel (c) reports the corresponding $HHI_j(a)$ across the ability distribution. Panel (d) shows firm-size distortions—the log deviation of firm employment from the efficient level—in the sorting case with constant returns to labor. Markdowns by worker ability mirror the market-share distribution across abilities, and average markdowns by ability mirror the pattern in $HHI_j(a)$; they are therefore omitted. When $\omega_a = 0$, market shares and distortions coincide, up to rescaling, with those in the $\rho \to 1$ case and thus are not reported; in both the $\omega_a = 0$ and $\rho \to 1$ environments, $HHI_j(a)$ is flat in ability. All simulations use the calibration in Table 1, varying one parameter at a time relative to the baseline.

[27] If markdowns were completely equalized, labor-market power would still distort the work–leisure margin but would not, holding aggregate labor supply fixed, generate misallocation *across* firms and markets.



(a) Employment Market Shares

(b) Ability Threshold by Firm Type

(c) HHI Index by Worker Log-Ability

(d) Size Distortion by Firm Type

Figure 2: Baseline Calibration: Market Shares, Distortions, Concentration, and Screening

Notes: This figure summarizes equilibrium outcomes at the baseline parameterization in Table 1. Panel (a) shows employment market shares by worker log-ability for three representative firms (bottom, median, and top of the productivity distribution). Panel (b) plots the ability threshold $\widetilde{a}_{ij}$—the minimum ability employed at firm (i, j)—against firm productivity z_{ij}, focusing on a single LLM j. Panel (c) displays the $HHI_j(a)$ across the worker log-ability distribution. Panel (d) reports firm-size distortions (log deviations of equilibrium employment from the efficient allocation) for the baseline ω_a and a lower value $\omega_a = 0.65$. Markdowns by worker ability mirror the market-share distribution across abilities, and average markdowns by ability mirror the pattern in $HHI_j(a)$, and are therefore omitted.

2 Empirical Evidence

In this section, I present the empirical evidence that guides the analysis, with a twofold interpretation. First, it documents model-free reduced-form facts on sorting and segregation in labor markets. Second, it interprets these facts through the lens of the model. In the calibration, I target a subset of these moments and reserve the remainder as validation moments.

2.1 Data Sources

Italian data. The main data come from the Italian social security administration (INPS), accessed through the VisitINPS Scholars program. They combine three sources. First, I use matched employer–employee panel data covering the universe of private-sector-dependent employees from 1974 through 2024. Weekly spells report labor income, broad occupation categories (blue-collar, white-collar, and managers), and full-time-equivalent (FTE) weeks worked. I use FTE weeks to construct employment, which closely matches the model's concept of hours worked, and I define the real weekly wage as annual spell earnings divided by FTE weeks worked and deflated to 2022 euros. I also merge demographic records. Spells are mapped to commuting zones using the official INPS crosswalk. Employers are identified by both an enterprise code and an activity–location code; I use the latter as the firm identifier, so *firm* henceforth refers to an industry–location unit. I construct an annual panel and restrict the sample to blue- and white-collar workers aged 20–65 in private, for-profit firms, excluding managers, apprentices, special categories, the self-employed, public-sector employees, agricultural workers, and contractors.[28] Second, I use an INPS extract that reports occupation and education coded according to the International Standard Classification of Occupations (ISCO), available from 2010 onward only for movers.[29] Third, I use firm-level financial information from the Italian credit-rating agency CERVED, merged through the enterprise identifier.[30] CERVED covers incorporated firms from 1996 through 2018, a subset of firms in the INPS data. I construct rolling sample windows; the baseline period is 2015–2019 in the main data and 2015–2018 in the merged balance-sheet sample. Additional details and summary statistics are provided in Supplemental Appendix B.

German data. I complement the Italian evidence with the German Sample of Integrated Employer–Employee Data (SIEED) from the Institute for Employment Research (IAB), a representative 1.5% sample of establishments with complete worker biographies and AKM worker and establishment fixed effects estimated on the full administrative data following Card et al. (2013). I restrict the sample to fully employed West German workers aged 20–60 with real daily wages above €10; I report cross-sectional results for 2010 through 2017. Additional details are provided in Supplemental Appendix B.6.

Local labor markets. I consider two standard definitions of an LLM: 3-digit industry by commuting zone, as in BHM and Yeh et al. (2022), and 3-digit occupation by commuting zone, as in Bils et al. (2025) and Felix (2026). Reporting both helps show robustness to market definition. Under the first, firms are establishments; under the second, they are establishment–occupation cells.[31]

[28] Managers are excluded because the model focuses on production labor, and it is natural to interpret managers as shifting z rather than production labor h. Moreover, managers may face different labor-supply and bargaining protocols that are reflected in distinct compensation-setting mechanisms.

[29] These variables are recorded only in mandatory employer notifications filed at contract initiation or modification and are therefore available only from 2010 onward.

[30] Because CERVED reports financials at the enterprise level, I assign each enterprise to the activity–location unit with the highest cumulative employment; this mapping is unique in 93% of enterprise–year observations.

[31] Under the occupation-based definition, firms active in multiple occupations are split into multiple units, increasing the number of firm identifiers.

In the Italian data, this yields 83,298 industry–commuting-zone markets and 65,696 occupation–commuting-zone markets. In the German data, I define LLMs as 3-digit occupations by 141 regions, yielding 14,550 markets.

2.1.1 Indirect-Inference Identification of Unobserved Types

A central challenge in studying labor-market sorting and wage determination is that key sources of heterogeneity are unobserved. I start by constructing model-generated employer–employee panels, as described in Supplemental Appendix E.4. On this synthetic panel, I estimate a two-way fixed-effects (Abowd et al., 1999) decomposition of log wages

$$\log w_{a,ij,t} = \alpha_a + \psi_{J(a,t)} + \epsilon_{a,ij,t} \tag{15}$$

where $\log w_{a,ij,t}$ denotes the log real wage of worker a in year t at firm ij, α_a is a worker-specific component capturing time-invariant wage premia, $\psi_{J(a,t)}$ is a firm-cluster-specific pay premium,[32] and $\epsilon_{a,ij,t}$ is an idiosyncratic residual. I estimate the AKM regression on the synthetic panel using simulated observation weights, which correspond to hours worked in the model. The framework developed in this paper does not impose the log-additive wage equation (15); wages may instead reflect heterogeneous markdowns, complementarities, and within-firm interdependencies. In Supplemental Appendix B.4, I show that in model-simulated data the within-LLM rankings of workers and firms implied by the AKM fixed effects closely align with the corresponding rankings of the structural types a and z.[33] The close alignment between AKM-based and structural rankings supports using AKM-based within-market rankings as *indirect-inference proxies* for latent type rankings, and the interpretation of the captured heterogeneity as within-industry or within-occupation heterogeneity. Intuitively, α_a captures persistent, portable worker wage premia, whereas $\psi_{J(a,t)}$ captures persistent pay premia associated with firm productivity and/or labor-market power.

In the model, higher ability raises individual wages, while higher firm types raise average wages. Because market power is determined primarily by the local competitive environment, within-market comparisons net out cross-market differences in wage-setting conditions and isolate the component of wage variation associated with the structural types a and z. Accordingly, within LLMs, the rankings of α_a and $\psi_{J(a,t)}$ closely track those of the corresponding structural types. Notably, I use the AKM decomposition purely as a measurement device to construct proxies for unobserved worker and firm types, rather than as a structural AKM-style variance de-

[32] $J(a,t)$ indexes the *firm cluster* in which worker a is employed at time t, following the approach of Bonhomme et al. (2022). Clusters are constructed using k-means in a pre-estimation stage. I estimate worker and firm effects using this clustered approach, which extends the standard Abowd et al. (1999) framework to mitigate limited-mobility bias. Only firms are discretized, while worker effects are estimated at the individual level, as in Lamadon et al. (2022). See Supplemental Appendix B.4 for details.

[33] The appendix reports results for the baseline calibration. I have also examined additional parameterizations (available upon request), and the ranking alignment is robust. The AKM decomposition is informative about both worker and firm heterogeneity only when the calibration allows for some worker productivity heterogeneity ($\omega_a > 0$); otherwise, wage dispersion is too limited to separately identify the two sets of fixed effects.

composition. I do not assume a one-to-one mapping between fixed effects and the model's latent types.

Next, I estimate the AKM regression (15) in the administrative data to construct the empirical moments used in the calibration, weighting each job spell by weeks worked to preserve comparability with the model. Supplemental Appendix B.4 reports additional details.

2.2 Empirical Facts

Because my calibration uses the industry-based definition of LLMs for Italy, I report those empirical facts in the main text. Corresponding results for Germany, occupation-based markets, and a full-time-only subsample are reported in Supplemental Appendices D.2, D.3, and D.4.[34]

2.2.1 Fact #1: Market Shares

I study how workers of a given AKM fixed effect allocate employment across firms with different wage policies, providing the empirical counterpart to the market shares implied by the model. Workers are ranked by their AKM fixed effects within LLMs, and the baseline firm ranking is the AKM firm fixed effect. For each year, I compute (from the worker perspective) the share of employment—measured in FTE weeks—that each decile of the within-year worker-rank distribution allocates to each decile of the firm-rank distribution. I assign deciles separately within each LLM and year, and I average the resulting matrices over 2015–2019 using FTE employment weights.[35] To ensure balanced decile partitions and to limit noise from very small markets, I restrict the sample to LLMs with at least 100 workers and 100 firms.

Figure 3a reports the employment matrix across worker and firm deciles when workers and firms are ranked, within each LLM, by their AKM worker and firm fixed effects, respectively. Figure 3b reports the analogous matrix based on coworker quality: for each worker, I compute the average AKM worker fixed effect of her coworkers, excluding the worker herself, assign that worker to the corresponding coworker-quality decile, and then compute employment shares analogously.

Relative to benchmarks that imply either homogeneous market shares or monotone sorting into higher-type firms, the data reveal substantial *segregation by worker rank*. Across both panels, top-decile workers are concentrated in top-ranked firms and among top-ranked coworkers, while

[34] As a robustness check, I re-estimate the AKM model and recompute the corresponding moments on a sample restricted to full-time workers, assigning each worker to the highest-paying firm in each year, as is standard in the AKM literature. In this specification, observations are not weighted by FTE weeks, so each worker receives equal weight. I do not impose this restriction in the main text, for two reasons. First, the model's employment concept depends on both the extensive and intensive margins, so the full sample weighted by FTE weeks provides a closer mapping between model and data. Second, the data report FTE weeks for each spell, allowing me to construct reliable measures of weekly wages and labor input. Restricting attention to full-time jobs and a single employer per year would discard this informative intensive-margin variation.

[35] Closely related matrices appear in papers such as Card et al. (2013). Relative to those papers, the most important difference for my purposes is that I classify worker and firm types *within* each LLM, the relevant competitive environment in the model.

bottom-decile workers are disproportionately concentrated in the bottom deciles, despite top firms being about twice the size of bottom firms. The top-top shares are 34% and 24%, while the bottom-bottom shares are 16% and 18% in panels (a) and (b), respectively. By comparison, the top-bottom shares are 5% and 6%, and the bottom-top shares are 11% and 9%.

Results are nearly identical when firms are ranked by average log wages and when LLMs are defined by occupation rather than by industry (Supplemental Appendices D.1 and D.3.1).[36] In the full-time-only sample, segregation is even stronger: bottom-ranked workers are more than four times as likely to work in bottom- rather than top-decile firms, while the pattern for top-ranked workers is similar (Supplemental Appendix D.4.1). Germany shows the same qualitative pattern (Supplemental Appendix D.2), suggesting that segregation is not peculiar to the Italian setting.

2.2.2 Fact #2: Hiring Thresholds

A key implication of the model in Section 1 is that, in the presence of within-firm interdependencies, higher-quality firms set higher hiring thresholds. To assess this implication empirically, I measure a firm's hiring threshold as the minimum worker fixed effect among its new hires and study how this measure varies with firm rank. I restrict the analysis to job-to-job transitions within LLMs with at least ten such firms and ten such workers, ensuring meaningful decile rankings. In the periods analyzed, roughly 30% of workers are newly hired each year in both the Italian and German data.

For each firm i in labor market j and year t, I define the hiring threshold $\widetilde{a}_{ij,t}$ as the minimum worker fixed effect among new hires, standardized by the standard deviation of the local worker fixed-effect distribution.[37]

I then estimate

$$\widetilde{a}_{ij,t} = \beta_0 + f(\text{Firm Decile}_{ij,t}) + \beta \log(\text{New Hires}_{ij,t}) + \gamma_m + \gamma_t + \varepsilon_{ij,t} \tag{16}$$

where $\widetilde{a}_{ij,t}$ is the standardized hiring threshold, $\text{Firm Decile}_{ij,t}$ is the firm's position in the local firm-rank distribution, $f(\cdot)$ denotes a full set of firm-decile dummies, $\text{New Hires}_{ij,t}$ is total hiring measured in FTE worked weeks, and γ_m and γ_t are market and year fixed effects.

I include $\log(\text{New Hires}_{ij,t})$ to absorb both temporary hiring surges, which may lower the minimum observed hire as firms become less selective (Carrillo-Tudela et al., 2023), and the mechanical fact that larger hiring cohorts generate lower minima because the dependent variable is a minimum-order statistic. Standard errors are two-way clustered by firm and market–year.

I rank firms using three measures: the AKM firm fixed effect, the average worker fixed effect among incumbent employees, and the average incumbent log wage. Figure 3c plots the relation-

[36] In the model, higher-type firms pay higher wages within an LLM, so a firm's average log wage is also a useful proxy for its latent type.

[37] As an alternative, I also use the average fixed effect among new hires, as in Carrillo-Tudela et al. (2023). Results are qualitatively similar, but I prefer the minimum-based measure because it more directly captures the firm's hiring threshold.

ship implied by Equation (16). Under the AKM-firm-effect and average-log-wage rankings, the relationship is strongly positive and approximately linear, with top-bottom hiring-threshold gaps of about 0.66 and 0.48 standard deviations, respectively. When firms are ranked by the average worker fixed effect among incumbents, the relationship is flatter and slightly declines in the upper tail, peaking in the ninth decile at about 0.05 standard devations.

Supplemental Appendices D.2.2, D.3.2, and D.4.2 show that the results are nearly identical under occupation-based markets and stronger in the full-time-only and German samples.

2.2.3 Fact #3: Concentration Indices by Worker AKM Rank

Lemma 1 shows that, within an LLM, the average markdown faced by workers of a given ability is a function of the wage-bill HHI for that ability group and the relevant labor supply elasticities. Proposition 4 further implies that the HHI profile across worker ability is informative about the underlying technological parameters, since it can be nonmonotonic in a only when marginal products are not increasing in z for all a.[38] HHI measures therefore provide an indirect summary of labor-market power across worker groups and of the associated welfare losses.

In the data, I approximate this object by dividing each LLM into ten deciles of the within-market AKM worker fixed-effect distribution and computing, for each decile, the LLM wage-bill HHI. I then aggregate these local HHIs to the national level using employment weights, where employment is measured in FTE weeks for the corresponding worker decile. Figure 3d reports this relationship when LLMs are defined by industry–commuting zones. For completeness, I also report heterogeneity by broad occupation by splitting the sample into white- and blue-collar workers and repeating the analysis separately for each group.[39]

Notably, this procedure yields only an approximate empirical counterpart to the theoretical relationship. First, AKM worker fixed effects are imperfect proxies for latent worker heterogeneity. Second, decile grouping both smooths within-group variation and introduces mechanical ranking error in small markets, where the local rank distribution is necessarily coarse. Nevertheless, this measure provides a natural empirical benchmark against which to assess the model and, through the model, to draw disciplined inferences about the underlying latent worker types.

Under the industry-based definition, the mean wage-bill HHI is about 0.27 overall, 0.28 for blue-collar workers, and 0.34 for white-collar workers. Across worker fixed-effect deciles, the overall HHI is about 0.28 in the bottom decile, declines to roughly 0.26 around the fourth and fifth deciles, and rises again to about 0.29 in the top decile. The relationship between worker AKM

[38] This pattern depends not only on the extent of segregation, but also on the dispersion of wage offers within each segment of the labor market, as characterized in Proposition A.1 in the Supplemental Appendix. What matters is not only how many firms hire workers in a given group, but also how dispersed their wage offers are, because the HHI depends on the realized coefficient of variation of job offers. In this sense, the empirical HHI is also informative about the distributional assumptions on worker and firm types.

[39] The five most frequent two-digit occupations among white-collar workers are *office clerks*, *commercial qualified professions*, *social science specialists*, *financial and administrative clerks*, and *research and education professionals*. Among blue-collar workers, the five largest occupations are *craft and skilled metal workers*, *drivers*, *elementary occupations*, *semi-skilled machine operators*, and *qualified professions in commerce and services*.

rank and concentration is therefore U-shaped, with higher concentration at both tails.[40]

Splitting the sample by broad occupation shows that concentration is higher among white-collar than blue-collar workers. Since white-collar jobs are also better paid on average, this helps explain the higher HHI for top-ranked workers in the pooled sample. Within both groups, the profile remains broadly U-shaped, with concentration lowest in the middle of the distribution and higher at both tails. Among white-collar workers, concentration is highest at the bottom of the fixed-effect distribution, indicating that low-ranked white-collar workers face especially concentrated labor markets. Supplemental Appendix D.3.3 shows a similar pattern under the occupation-based definition, but at lower levels. Supplemental Appendix D.4.3 reports the corresponding measures for the subsample of full-time workers.

2.3 Taking Stock

Viewed through the lens of the model, these findings are difficult to reconcile with the limiting parameterizations discussed in Section 1.5. Under homogeneous or purely multiplicative technologies, equilibrium allocations do not generate market shares or concentration that vary across worker ranks. In the constant-returns-to-scale-with-strong-complementarities benchmark, higher-ability workers sort into higher-type firms, but the segregation of lower-ranked workers into lower-ranked firms is weaker, and concentration rises monotonically with worker type rather than nonmonotonically. The joint presence of strong segregation, tighter hiring thresholds at higher-ranked firms, and nonmonotonic concentration across the worker-rank distribution instead points to the richer environment with the within-firm interdependencies and segmented competition analyzed in Section 1.

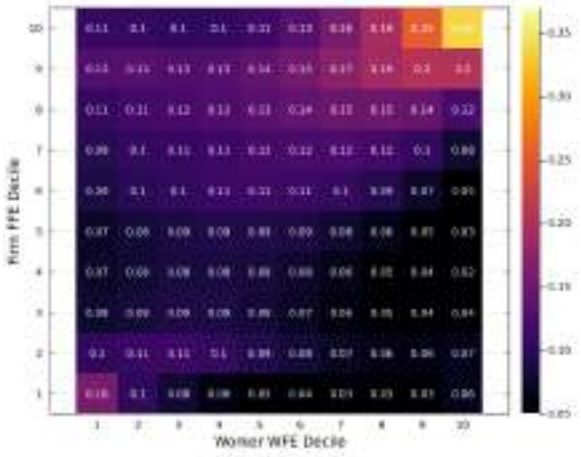
3 Calibration

I calibrate the model to a subset of informative empirical moments. The calibration combines externally calibrated parameters, parameters estimated directly from the data, and parameters disciplined by indirect inference. I then assess the model against a broad set of untargeted moments. Because several calibration targets use balance-sheet data, the baseline calibration defines LLMs as three-digit-industry–commuting-zone cells, as in BHM and Yeh et al. (2022).[41] Under this definition, each firm is assigned to a unique LLM. Although a full calibration is not feasible under the occupation–commuting-zone definition, the main empirical facts are similar under the two definitions, suggesting that this choice has little quantitative effect on the results.[42] Table E.8

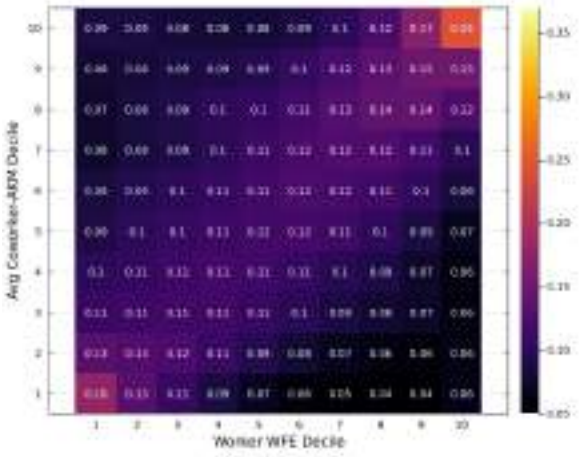
[40] Part of the higher HHI in the lowest worker decile may mechanically reflect very small LLMs, where worker ranks are coarsely measured and concentration is high. I nevertheless retain these markets because they account for an important share of overall concentration. Employment weighting partly mitigates the issue, and the same measurement problem is present in the model-generated data, so the model–data comparison remains informative.

[41] Defining LLMs by occupation would split firm identifiers across categories, preventing the use of balance-sheet information.

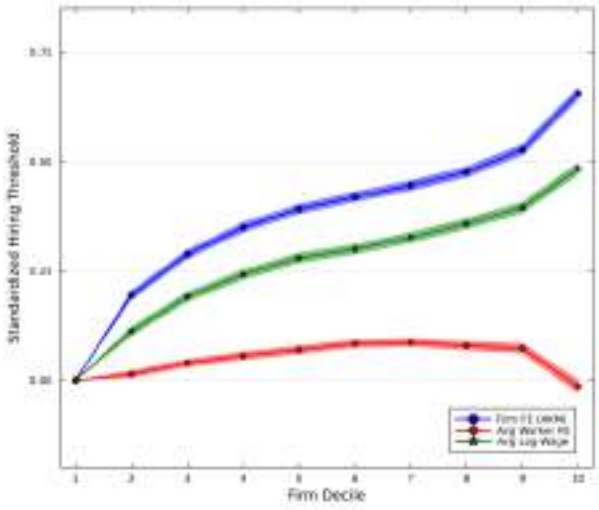
[42] Worker–firm allocation moments are stronger and smoother in the subsample restricted to full-time workers. Using the full FTE sample therefore provides a specification that is closer to the model and sets, if anything, more de-



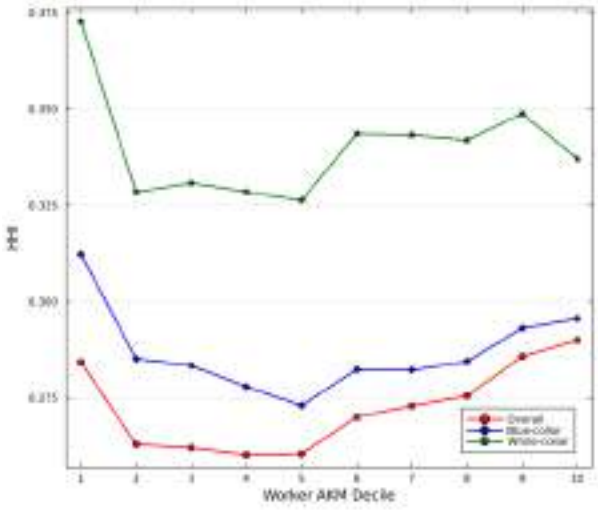
(a) Employment Shares Across Worker and Firm Fixed-Effect Deciles



(b) Employment Shares Across Worker and Coworker Fixed-Effect Deciles



(c) Hiring Thresholds by Firm Rank



(d) Wage-Bill HHI by Worker Fixed-Effect Decile

Figure 3: Segregation, Screening, and Concentration in Industry-Based Local Labor Markets

Notes: All panels use Italian data and define LLMs as industry–commuting-zone cells. Panels (a) and (b) report employment-share matrices by worker-fixed-effect decile, with columns defined by firm-fixed-effect deciles in panel (a) and by deciles of coworkers' average worker fixed effect in panel (b), excluding the worker herself. Deciles are assigned within each LLM and year, matrices are averaged over 2015–2019 using FTE employment weights, and the sample is restricted to LLMs with at least 100 workers and 100 firms. Panel (c) plots hiring thresholds by firm-rank decile from Equation (16); firms are ranked by the AKM firm fixed effect, average incumbent-worker fixed effect, or average incumbent log wage. Estimates control for $\log(\text{New Hires}_{ij,t})$ and include market and year fixed effects; ribbons show 95% confidence intervals with standard errors two-way clustered by firm and market–year. Panel (d) plots the wage-bill HHI across worker fixed-effect deciles for the full sample and separately for white-collar and blue-collar workers; the HHI is averaged across markets using FTE employment weights.

in the Supplemental Appendix reports the full set of parameter values used in the quantitative analysis.

3.1 Macro and Technology Parameters

A subset of parameters is fixed externally at standard values. I set the real interest rate to $R = 0.10$ and the Frisch-elasticity parameter to $\varphi = 0.5$, in line with conventional estimates (e.g., Chetty et al., 2011). I calibrate the curvature of utility from consumption to match a Marshallian labor supply elasticity of 0.06 (Keane, 2011), which implies that $\sigma = 0.83$. The depreciation rate is set to $\delta = 0.08$. The model is simulated over $M = 1000$ LLMs, each with up to 200 firms. The number of firms per market follows a mixed distribution with a point mass at one and a Pareto component for markets with more than one firm. The parameters of this distribution are estimated by maximum likelihood using the empirical distribution of firms across markets. Figure E.1 in the Supplemental Appendix E compares the empirical and model-implied distributions. Worker and firm types, a and z, are drawn from mean-zero log-normal distributions with standard deviations σ_a and σ_z, respectively; a is discretized on a 500-point percentile grid.

From Equation 10, the capital share of value added equals $(1-\gamma)\alpha$. I therefore calibrate $(1-\gamma)\alpha$ to match the aggregate capital share computed from firm-level balance-sheet data from CERVED, computed with user cost of capital taken from the Penn World Table (Feenstra et al., 2015). This yields an aggregate capital share of 0.21.[43] I then recover the effective labor output elasticity, $\alpha\gamma$, from three-digit sector-level production-function estimates and aggregate them using employment weights (see Supplemental Appendix E.2 for details). These two moments jointly pin down (α, γ), yielding $\gamma = 0.78$ and $\alpha = 0.94$.[44]

3.1.1 Within-Market Inverse Labor-Supply Elasticity

Next, I estimate η. Under homogeneous-worker assumptions, Equation (2) collapses to a firm-level inverse labor-supply relationship in total employment, and the oligopsonistic term can be absorbed by market–time fixed effects, as in Felix (2026). Here, however, Equation (2) holds at the worker-type level, so the oligopsonistic term varies with (a, j, t) and cannot be absorbed by a tractable set of market–time fixed effects.[45] I therefore derive a firm-level approximation by linearizing the aggregate inverse labor-supply schedule induced by the type-level system around the equilibrium path: absent the shock, the firm would continue to evolve along its equilibrium

manding and thus conservative targets.

[43] For comparison, the corresponding aggregate capital share in the United States is about 18% (Barkai, 2020).

[44] BHM calibrate $\alpha = 0.940$ and $\gamma = 0.808$ for the U.S. economy.

[45] Even abstracting from measurement error from the use of AKM fixed effects as proxies for ability, absorbing the dependence of the oligopsonistic term on (a, j, t) would require an extremely high-dimensional set of fixed effects—for example, market $\times$ time $\times$ worker—which is computationally infeasible and would leave little residual variation for identification.

path, and I study an exogenous departure from that path. The discrete-time analogue[46] is

$$\Delta\overline{\log w}_{ij,t+1} = \frac{1}{\eta}\,\Delta \log h_{ij,t+1} + \underbrace{\sum_{a} \Delta g_{ij,t}(a)\ \log w_{ij,t+1}(a)}_{\text{composition effect}} - \beta_{\eta\theta}\ \underbrace{\mathbb{E}_{g_{ij,t}}\big[\lambda_{ij,t}(a)\,\Delta \log w_{ij,t+1}(a)\big]}_{\text{oligopsonistic effect}} + \varepsilon_{ij,t+1} \tag{17}$$

where $\Delta x_{t+1} \equiv x_{t+1} - x_t$. $\lambda_{ij,t}(a)$ denotes the pass-through from firm ij's wage to the local-market wage index $w_j(a)$, and $\beta_{\eta\theta} = \frac{\theta-\eta}{\eta}$.

Estimating Equation (17) raises three issues. First, the object of interest is the *local* slope of inverse labor supply around the equilibrium at time t, so identification requires a small, plausibly exogenous firm-level labor-demand shock and a focus on the impact response, rather than large shocks that move employment far from the initial equilibrium.[47] Second, firm-level demand shock can change workforce composition, so that $\Delta\overline{\log w}_{ij,t+1}$ may partly reflect reweighting across worker types rather than movements along a fixed inverse labor-supply curve. Third, the oligopsonistic term emphasized by BHM depends on the market CES wage index and is itself affected by the same firm-level demand shock used for identification. Once omitted, it enters the error term and becomes correlated with the instrument, biasing the estimate of $1/\eta$.

I address these concerns in three ways. First, I use unexpected deaths of nonmanagerial workers as a quasi-random shift in firm-specific demand for coworkers and replacement hires. Following Jäger et al. (2024), I match each firm–deceased-worker pair to a placebo nondeceased firm–worker pair based on worker and firm characteristics, then I remove both workers from the sample in all periods.[48] I then estimate the impact response of coworkers' and replacement hires' FTE weeks and wages, which provides a local movement along the survivor-based inverse labor-supply curve.

Second, I control explicitly for composition changes induced by the shock, so that identification of $1/\eta$ comes from within-composition variation in wages. In practice, I control for the deceased worker's preshock position in the firm's wage distribution, and for a high-dimensional set of percentile-bin fixed effects in the change in the firm's average remaining worker AKM fixed effect, where the fixed effects are estimated using only preshock data to avoid contamination.

Third, I allow the slope of inverse labor supply to vary with the firm's wage-bill share in the local market or restrict attention to firms with preshock shares below approximately 1%, where oligopsonistic feedback through the local wage index is plausibly second-order. This procedure allows me to recover the approximately atomistic inverse elasticity. In particular, under this strategy, I obtain a reliable estimate of η under four conditions: (i) the death shock is relevant and

[46] See Supplemental Appendix E.3.1 for the derivation and additional details.

[47] For example, a demand shock induced by a large, persistent increase in import competition (a "China shock") unfolds over a long time and may generate adjustments that go beyond the local log-linear approximation in Equation (17).

[48] Following Jäger et al. (2024), I further restrict the sample to death shocks occurring in firms whose maximum total employment over the sample period remains below a given threshold. This is also consistent with the model: identification of η should be driven by firms that are approximately atomistic in their LLMs, whereas oligopsonistic wage-setting concerns are most pronounced for the very largest firms.

satisfies the exclusion restriction; (ii) the firms driving the estimate are approximately atomistic in their local markets; (iii) the main composition effects induced by the shock are captured by the included controls; and (iv) the death shocks are small enough that the log-linear approximation in Equation (17) is locally accurate. Supplemental Appendix E.3.1 describes the construction of the death shock, the sample restrictions that make conditions (i)–(iv) plausible—including tests of parallel pretrends—and a range of robustness checks.

This procedure yields an inverse labor-supply elasticity of $\hat{\eta} = 13.2$, slightly above the benchmark estimate reported by BHM for the United States ($\hat{\eta} = 10.85$).

3.1.2 Indirect-Inference Identification of Remaining Parameters

The remaining parameters are the across-market elasticity of substitution θ, the dispersion of worker and firm heterogeneity (σ_a, σ_z), and the parameters governing worker–firm complementarities in production (ρ, ω_a). I discipline these objects jointly via indirect inference. This subsection discusses the empirical moments used in the calibration.

Across-market elasticity (θ)

Conditional on η, cross-firm variation in markdowns is governed by the gap $\eta - \theta$ and reflected in the firm-level wage wedge $\widetilde{\psi}_{ij,t}$ from Proposition 2. In large, competitive markets, $\widetilde{\psi}_{ij,t}$ is governed mainly by η and converges to $\eta/(\eta+1)$ in the atomistic limit. In small, concentrated markets, it is governed increasingly by θ and converges to $\theta/(\theta+1)$ when a single firm employs all workers in the market. Therefore, I calibrate θ by indirect inference, targeting the cross-sectional relationship between $\widetilde{\psi}_{ij,t}$ and LLM concentration, measured by market HHI.[49] I extend the identification of $\widetilde{\psi}_{ij,t}$ to a richer production environment with variable markups and output elasticities (see Lemma E.6 in Supplemental Appendix E). I construct $\widetilde{\psi}_{ij,t}$ from observable cost shares and estimate its relationship with concentration in regressions with firm and year fixed effects, where the firm fixed effects absorb time-invariant production-function elasticities and wedges in the flexible input.[50] I choose θ so that the model-implied mapping from local market shares to $\widetilde{\psi}_{ij,t}$ matches this empirical relationship. To address the mechanical correlation induced by the use of the wage bill in both $\widetilde{\psi}_{ij,t}$ and the wage-bill HHI, I also instrument the wage-bill HHI with a revenue-based HHI. For robustness, I estimate an analogous specification using the firm's wage-bill share, instrumented by its revenue share. Because the two regressions imply approximately the same gradient in the model, this second specification provides an additional overidentifying restriction on θ and serves as a validation check. See Supplemental Appendix E.3.2 for details.

[49] This strategy is analogous in spirit to Edmond et al. (2023), who infer demand elasticities by matching the relationship between markups and market shares in oligopolistic settings.

[50] Following De Ridder et al. (2026), one can study how wedges or markups co-move with observables without fully estimating a production function, provided that output elasticities can be controlled for as functions of observables. Here, firm fixed effects absorb time-invariant production elasticities and material wedges, while year fixed effects capture aggregate shocks.

Remaining parameters: (σ_a, σ_z) **and** (ρ, ω_a)

As discussed in Section 1.5, different values of (ρ, ω_a) imply different patterns of sorting and segregation, while (σ_a, σ_z) shape the dispersion of wages and firm size. As $\sigma_a \to 0$, within-firm wage heterogeneity vanishes; as $\sigma_z \to 0$, firms become homogeneous within LLMs, compressing firm-size differences. I calibrate these parameters by indirect inference, jointly with θ. For any candidate parameter vector, I solve for the steady-state equilibrium, simulate a synthetic worker–firm panel, and choose parameters to minimize a weighted quadratic distance between simulated and empirical moments, targeting a relatively small set of moments and treating the remaining empirical facts as untargeted validation moments; see Supplemental Appendix E.4.1.

Table F.1 in the Supplemental Appendix reports the targeted moments and their model-implied counterparts: (i) the standard deviation of firm log employment (1.50 in the data vs. 1.51 in the model); (ii) the standard deviation of AKM worker fixed effects (0.261 vs. 0.260); (iii) the share of top-decile workers employed by top-decile firms (0.337 vs. 0.336); (iv) the gradient of hiring thresholds with respect to firm rank (0.045 vs. 0.045)[51]; and (v) the labor-wedge gradient with respect to wage-bill HHI (0.414 vs. 0.416), which disciplines θ. Heuristically, the standard deviation of worker fixed effects primarily disciplines σ_a and ω_a, while the standard deviation of log employment is informative about σ_z. The share of top-decile workers employed by top-decile firms is informative about wage, and thus effective productivity, differences between top- and lower-ranked firms, helping identify ω_a and ρ. The hiring-threshold gradient with respect to average incumbent log wages is informative about segregation, further disciplining ρ in combination with ω_a. Finally, the labor-wedge–HHI gradient disciplines θ. All parameters are estimated jointly to match this set of moments. Additional details on the calibration algorithm, simulation design, and convergence criteria are provided in Supplemental Appendix E.4.

3.1.3 Calibration Results and Model Validation

Discussion of calibrated parameters. The calibration yields $\theta \approx 1.51$, $\rho \approx 0.41$, $\omega_a \approx 0.81$, $\sigma_a \approx 0.23$, and $\sigma_z \approx 0.75$. These values imply strong complementarities in production and a high weight on worker ability, which together are key to matching the observed patterns of sorting and segregation. When ρ moves closer to one or ω_a moves closer to zero, the model no longer generates the hiring-threshold gradient observed in the data. At the same time, a high ω_a prevents the employment share of top-ability workers in top-type firms from becoming counterfactually large. If firm type z received too much weight, wage differences between high- and low-type firms would become too large, causing top-ability workers to concentrate excessively in a small set of top-type firms and pushing the "top–top" share well above its empirical counterpart. Given

[51] This corresponds to estimating a linear analogue of Equation (16) rather than a fully nonparametric specification. Specifically, I target the coefficient β_1 in

$$\widetilde{a}_{ij,t} = \beta_0 + \beta_1 \text{Firm Decile}_{ij,t} + \beta \log(\text{New Hires}_{ij,t}) + \gamma_m + \gamma_t + \varepsilon_{ij,t}$$

the high weight placed on ability, only modest worker heterogeneity, σ_a, is needed to match the observed dispersion in AKM worker fixed effects. Cross-firm productivity heterogeneity, governed by σ_z, instead provides the main source of dispersion in firm size and employment, and implies that a relatively small set of high-productivity firms attracts a disproportionate share of top-ability workers. The estimate $\theta \approx 1.51$, combined with my estimate of η, implies an atomistic markdown of $\mu_{\text{atom}} \approx 0.93$, close to the BHM value of $\mu_{\text{atom}} \approx 0.90$, and a full-monopsonist markdown of $\mu_{\text{mon}} \approx 0.60$. By comparison, BHM report an across-market elasticity of $\theta \approx 0.42$ for the United States, implying a full-monopsonist markdown of $\mu_{\text{mon}} \approx 0.30$. Related evidence from Yeh et al. (2022) implies that firms in one-firm markets have inverse markdowns about 0.10–0.25 higher than firms with LLM shares between 0 and 0.10. Overall, the elasticities implied by my calibration lie between the relatively low across-market elasticity in BHM and the smaller monopsony premia implied by Yeh et al. (2022), and they are quantitatively closer to the latter.

Homogeneous-workers benchmark. As a benchmark, I consider the BHM version of the model (i.e., $\omega_a = 0$), where only σ_z, and θ remain to be disciplined by the data. I calibrate these parameters to match the slope of $\widetilde{\psi}_{ij,t}$ with respect to the market HHI and the standard deviation of log employment.[52] The resulting homogeneous-workers calibration is $(\sigma_z, \theta) = (0.143, 1.54)$.

I implement the recalibration sequentially. The first calibration (BHM1) uses the parameter values reported in BHM, $(\eta, \theta, \sigma_z) = (10.85, 0.42, 0.31)$. The second (BHM2) replaces only σ_z with my calibrated value, while holding η and θ fixed at their BHM values. Finally, BHM3 also replaces η and θ with the values from my preferred homogeneous-workers calibration. This sequence helps isolate the role of firm heterogeneity and labor-supply elasticities in driving the results and allows me to evaluate each specification against the data.[53]

Model validation. To validate the model, I compare its implications for a set of untargeted moments; details are reported in Supplemental Appendix F. I first replicate the worker-side employment matrix of Figure 3a. The mean absolute error is 2.9 percentage points, and the correlation between model and data cell entries is 0.82. For the coworker employment matrix in panel (b), the mean absolute error is 1.5 percentage points and the correlation is 0.69. For the analogous matrix when firms are ranked by average log wages (Supplemental Appendix D.1), the mean absolute error is 3.5 percentage points and the correlation is 0.88.[54]

[52] BHM use the labor share of GDP to discipline the dispersion of firm types. I do not target this moment here because the model abstracts from variable markups, which would further reduce the labor share and therefore constitute an important omitted factor in calibrating this parameter.

[53] Relative to BHM, these benchmark calibrations still differ in two respects. First, I keep fixed the empirical distribution of the number of firms per LLM used in the baseline calibration, which differs nontrivially from that in BHM. Second, I also hold fixed the remaining baseline parameters, such as α and γ. In principle, both differences can affect quantitative results. For the comparisons emphasized here, the distribution of firms per market matters for the labor share of GDP but has only limited effects on misallocation, while α and γ are quantitatively close to their BHM counterparts.

[54] As a benchmark, I repeat the same exercise for the homogeneous-workers BHM3 specification. Since worker heterogeneity is shut down, there is no meaningful AKM decomposition on simulated data, so I compare latent worker and firm types in the model to the empirical matrices constructed from AKM ranks. The mean absolute errors are 9.4 and 9.8 percentage points, with correlations of 0.0003 and 0.0014, for Figure 3a and the wage-ranked matrix in Supplemental Appendix D.1, respectively.

For the replication of Figure 3c, the mean absolute deviations are 0.20, 0.12, and 0.06, and the corresponding upper-tail gaps are 0.40, 0.15, and 0.01, all in standard deviations of worker AKM fixed effects, for rankings based on AKM firm effects, average incumbent worker fixed effects, and average incumbent log wages, respectively; the replication also preserves the nonmonotonic pattern for average incumbent-worker fixed effects. For the replication of Figure 3d, the average gap is 0.04 HHI points and the correlation with the empirical series is 0.98.

Table F.2 in the Supplemental Appendix reports additional untargeted moments. Among these, the between-local-labor-market component of wage variance is particularly important. In the model, this component reflects cross-market differences in average wages driven by markdowns. Therefore, overpredicting it would indicate excessive dispersion in markdowns across LLMs. The variance of log wages is 0.134 in the model, compared with 0.154 in the data. Decomposing wage variance into within-firm, between-firm-within-local-labor-market, and between-local-labor-market components, the within-firm term is 0.065 in the model vs. 0.063 in the data, the between-firm-within-market term is 0.016 vs. 0.037, and the between-local-labor-market term is 0.054 vs. 0.054. The employment-weighted standard deviation of firm fixed effects is 0.23 in the model vs. 0.185 in the data, while the covariance between worker and firm fixed effects is 0.006 vs. 0.018.[55] Finally, the model implies an aggregate labor share of 0.61, compared with 0.52 in the Italian national accounts (Feenstra et al., 2015).[56]

Table F.2 in the Supplemental Appendix reports these moments for the BHM calibrations as well, whenever they are defined. Relative to the data, which imply a standard deviation of log employment of 1.50, a labor share of 0.52, and a variance of log wages of 0.154, BHM1 implies corresponding values of 2.56, 0.44, and 0.35. BHM2 implies values of 1.53, 0.47, and 0.24. BHM3 implies values of 1.50, 0.605, and 0.059. For the between-local-labor-market component of wage variance, BHM1–BHM3 imply 0.35, 0.24, and 0.056, respectively, vs. 0.054 in the data; for the between-firm-within-market component, they imply 0.00044, 0.0017, and 0.0026, vs. 0.037.

4 Results

In this section, I address three quantitative questions. First, which firms are on average most distorted, and how do these distortions affect aggregate production? Second, how does labor-market power vary across the worker-ability distribution, how large are the associated welfare losses and entrepreneurial gains, and how are these effects distributed across workers? Third, to what extent do markdowns distort assortative matching and contribute to wage inequality?

I begin by using the model to characterize the distribution of the firm-level labor-share wedge $\tilde{\psi}_{ij}$. In principle, these wedges can be inferred from the data using cost shares (as in Lemma E.6

[55] The baseline model generates worker–firm covariance through sorting within LLMs, but it abstracts from systematic sorting of worker and firm types across markets. In the data, the AKM covariance also reflects across-market sorting, e.g., across occupations, sectors, or geographies.

[56] I view this discrepancy as broadly reassuring, given that the model abstracts from variable markups, which would further reduce the labor share.

in the Supplemental Appendix), but the resulting estimates are difficult to interpret. Their level is not cleanly identified from firm-level balance-sheet data (De Ridder et al., 2026), while their dispersion may reflect not only true markdown heterogeneity but also other distortions, such as measurement error, implicit taxes, and adjustment costs. The structural model instead delivers a clean distribution of wedges that isolates labor-market power. I also decompose $\tilde{\psi}_{ij}$ into the average markdown $\overline{\mu}_{ij}$, and quantify the resulting effect on aggregate production. I then use the calibrated model to measure the average markdown by worker ability, $\mu(a)$. By Lemma 1, $\mu(a)$ could in principle be recovered from the data using worker-ability HHI indices. In practice, however, empirical proxies based on HHI indices and AKM worker fixed effects are necessarily coarse and noisy: ability is observed only indirectly through AKM fixed effects, and HHI indices are computed from binned groups within local markets. The model instead delivers $\mu(a)$ directly at the structural ability level.[57] Comparing the realized equilibrium to the counterfactual without markdowns then shows how labor-market power shapes assortative matching, wage inequality, and the welfare distribution. In particular, I compute welfare gains and losses as the percentage change in per-capita consumption—$\lambda(a)$ for workers and $\lambda(e)$ for the entrepreneur—that makes agents indifferent between the observed steady state and the production-efficient equilibrium. Finally, I assess how measures of wage inequality would change once the inefficiencies induced by markdowns are corrected.

Firm-level wedges and aggregate production. In the baseline calibration, the mean of $\tilde{\psi}_{ij}$ is 0.898, the revenue-weighted mean is 0.833, the median is 0.921, and the standard deviation is 0.059.[58] As a result, monopsony generates lower labor share and higher aggregate profit share relative to the production-efficient benchmark (60.7% vs. 73% and 18.5% vs. 6.3%, respectively). Supplemental Appendix Table G.1 reports average wedges and average markdowns by firm-type decile in the baseline calibration. In the baseline economy, $\tilde{\psi}_{ij}$ declines with firm type, albeit somewhat nonmonotonically, from 0.914 in the bottom decile to 0.880 in the top decile. Moreover, $\tilde{\psi}_{ij} - \overline{\mu}_{ij}$ is positive in the bottom six deciles and negative in the top four, ranging from 0.001 in the bottom deciles to -0.003 in the top decile. This implies that the covariance term in $\tilde{\psi}_{ij}$ adds an extra labor-share wedge for high-type firms, reflecting relatively greater market power over workers in the upper tail of their within-firm ability distribution. Reassuringly, these deviations are quantitatively small, so interpreting $\tilde{\psi}_{ij}$ as an average markdown, as in Yeh et al. (2022), is a good approximation.

Supplemental Appendix Table G.2 reports total output losses and their decomposition in the baseline calibration and in the three homogeneous-worker benchmarks, BHM1–BHM3.[59] Output is 2.32% below the efficient level in the baseline economy, compared with 8.44% in BHM1, 5.82% in BHM2, and 2.47% in BHM3. The model validation showed that BHM1 and BHM2 sub-

[57] Any two environments that generate the same equilibrium market shares, for given labor-supply elasticities, imply the same distribution of $\mu(a)$. The measurement therefore depends on the model-implied equilibrium allocation rather than on the particular mechanism that generates it.

[58] From Proposition 2, $ls_{ij} = \tilde{\psi}_{ij}\alpha\gamma$. Because the aggregate labor share is a revenue-weighted average of firm-level labor shares, the aggregate wedge equals the revenue-weighted mean of $\tilde{\psi}_{ij}$.

[59] I measure output losses as $1 - \frac{Y}{Y_{eff}}$, where Y is realized aggregate output and Y_{eff} is efficient aggregate output.

stantially overpredict both firm-size dispersion and cross-market wage dispersion. Because cross-market wage dispersion in the model is driven primarily by heterogeneity in markdowns, this finding suggests that BHM1 and BHM2 generate excessive dispersion in firm-level markdowns and, therefore, excessive misallocation.[60] Despite substantial differences between the two models, the effect of monopsony on aggregate output is remarkably similar in the baseline model and in the fully recalibrated BHM3, though somewhat smaller in the former, consistent with segregation reducing the aggregate costs of markdowns.

Average markdowns and welfare distribution by worker ability. I now use the model to characterize how labor-market power varies across the worker-ability distribution. Figure 4a reports the employment-weighted average markdown by worker log ability. The median worker receives about 82.7% of their marginal product. The wage share of marginal product varies systematically across ability, and is mildly nonmonotone: it peaks at about 83.7% for workers roughly two standard deviations below the mean of the log-ability distribution, remains around 83.5% at the very bottom, and falls to about 81% at the top.[61]

Figure 4b reports welfare gains and losses by worker log ability relative to the production-efficient allocation. Labor-market power lowers welfare both by reducing aggregate output through misallocation and by redistributing income from workers to entrepreneurs through the gap between wages and marginal products. Entrepreneurs benefit substantially: in the baseline calibration, they would need to give up roughly 53% of consumption to be indifferent between the decentralized allocation and the efficient benchmark.

Workers instead experience large and heterogeneous welfare losses. These range from about 20.5% to 24.5% and are U-shaped in worker ability: workers at the bottom would require about 22.7% more consumption to attain the utility they would obtain under the production-efficient allocation, workers at the top about 24%, and middle-ability workers about 20.5%.[62]

This pattern reflects weaker effective competition at both tails of the ability distribution. Low-ability workers are hired by only a limited set of low-productivity firms, while high-ability workers are concentrated in the most productive firms, which compete with only a handful of similarly productive rivals. As a result, workers at the extremes of the ability distribution face weaker outside options than middle-ability workers, generating particularly low take-home shares at the top and large welfare losses at both tails. These losses are especially consequential for low-ability workers, who already consume less even when paid their marginal product.

Assortative matching, segregation, worker reallocation, and wage inequality. I next use the model to characterize worker allocation across firms in the decentralized equilibrium, how it changes without labor-market power, and the implications for wage inequality.

Figure 4c plots, for each worker type, the expected employer type, $\ln z$, using employment

[60] See the model-validation section and Table F.2 in the Supplemental Appendix. Consistent with this interpretation, the standard deviation of $\tilde{\psi}_{ij}$ falls from 0.137 in BHM1 to 0.059 in BHM3.

[61] This result complements Bils et al. (2025), who also emphasize that labor-market power is stronger for workers at the top of the distribution.

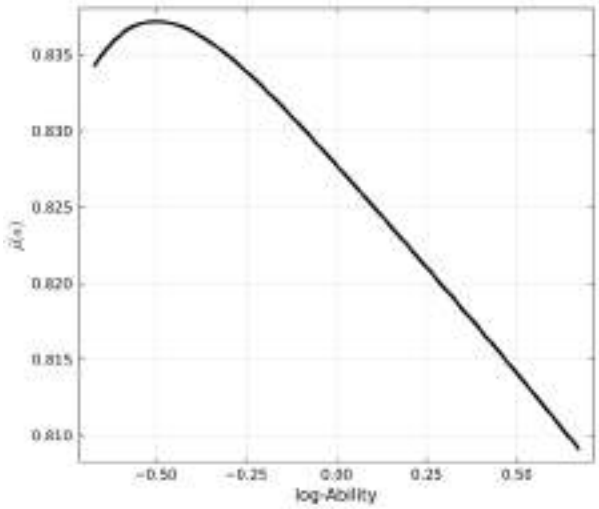
[62] The welfare profile is more curved than the profile of average markdowns because welfare depends on the full wage index faced by each worker type, not only on its average markdown.

weights across all jobs. It compares the decentralized equilibrium with markdowns to the efficient allocation without labor-market power. In both allocations, expected firm type is increasing and approximately linear in worker ability, consistent with positive assortative matching and ability segregation: workers in the lower part of the ability distribution are employed, on average, by firms below the median firm type, whereas top workers are employed by firms roughly one standard deviation above the median.

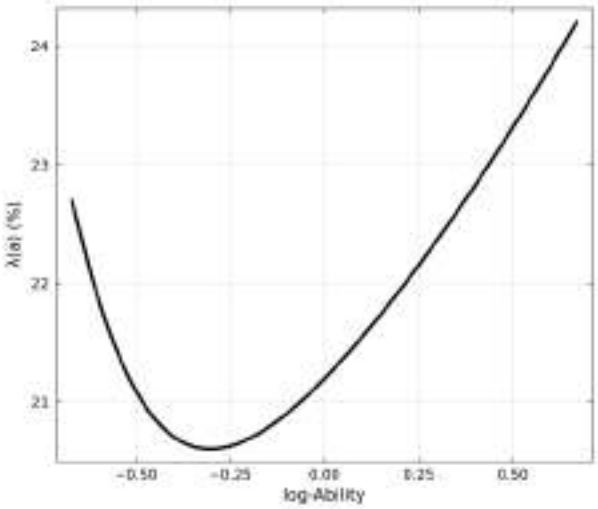
Figure 4d reports the change in global employment shares by worker and firm decile between the efficient and decentralized allocations. Red indicates reallocation toward a given cell, while blue indicates reallocation away from it. Removing markdowns tends to shift employment toward higher-type firms, although the pattern differs across worker types. In Figure 4d, high-type workers are inefficiently allocated in lower-type employers and reallocate back toward more productive firms once markdowns are removed. Conversely, in the decentralized allocation, low-type workers are somewhat overrepresented in both the highest- and lowest-type firms and, absent markdowns, reallocate toward middle-type firms, which exert market power over them. From Figure 4c, expected firm type rises for high-ability workers and falls for low-ability workers; assortative matching therefore strengthens.[63] The largest reallocation effect arises for workers in the top ability decile, whose employment share in top firms is about 2 percentage points higher under the efficient allocation. Labor-market power therefore dampens segregation and weakens positive assortative matching because it is strongest when high-type firms hire high-type workers, making these matches especially underpaid.

These patterns have direct implications for wage dispersion. Supplemental Appendix Table G.3 shows how overall wage inequality changes in the absence of markdowns and decomposes it into several components. Markdowns are largest for the highest-ability workers, who also tend to earn the highest wages, and removing markdowns strengthens sorting by reallocating them toward more productive firms. Consistent with this, the variance of worker fixed effects rises from 0.068 to 0.070, the covariance between worker and firm fixed effects rises from 0.006 to 0.008, and the between-firm-within-local-labor-market component rises from 0.016 to 0.021. In isolation, both forces would raise wage inequality. Yet overall wage inequality—measured by the standard deviation of log wages—falls by 0.044 log points. The reason is that removing markdowns sharply compresses the between-local-labor-market component of wage variance, which falls from 0.054 to 0.019. This indicates that cross-market differences in firms' wage-setting policies, driven by labor-market power, are an important source of wage inequality. Eliminating markdowns therefore compresses average wages across LLMs by enough to more than offset the increase in within-market wage dispersion.

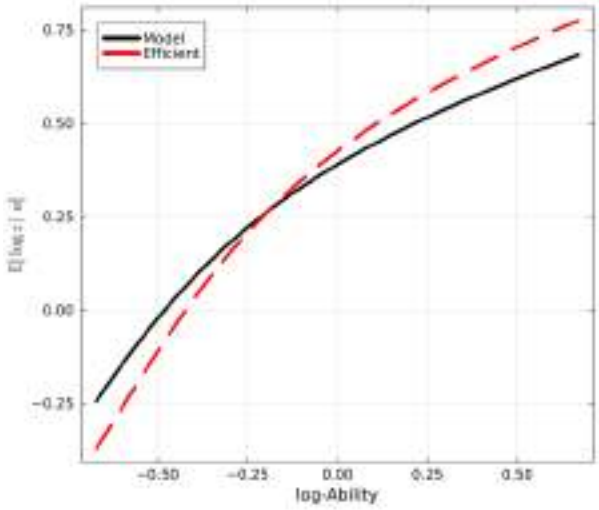
[63] Some reallocation toward low-paying firms echoes Volpe (2024), who finds that labor-market power reallocates labor away from low-paying firms and toward high-paying firms. In that framework, the mechanism operates through worker preference heterogeneity: low-paying firms attract workers with less wage-elastic labor supply and therefore retain labor-market power through a composition effect. Here, by contrast, the mechanism operates through heterogeneity in the competitive environment, and the reallocation result is reversed for high-ability workers.



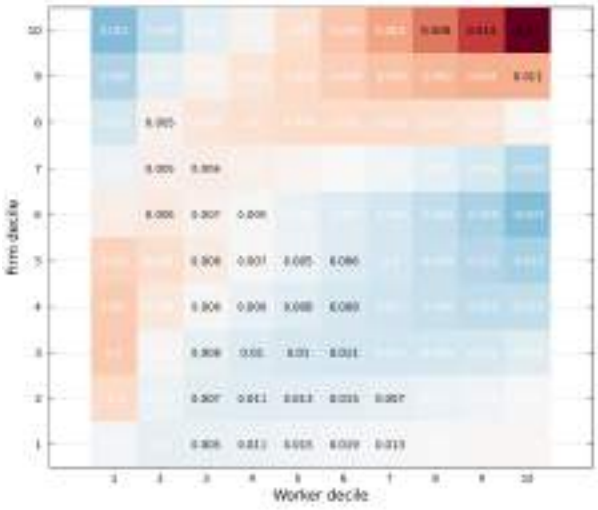
(a) Employment-Weighted Markdown by Worker Log Ability



(b) Welfare Loss by Worker Log Ability



(c) Expected Firm Type by Worker Log Ability



(d) Worker Reallocation Across Firms

Figure 4: Worker Ability, Markdowns, Welfare, and Reallocation

Notes: All panels report model-implied objects from the baseline quantification. Panel (a) plots the employment-weighted average markdown $\mu(a)$ by worker log ability, where markdowns are defined as wages relative to marginal products and weights are given by the steady-state employment allocation. Panel (b) plots welfare losses by worker log ability, measured as the consumption-equivalent percentage change that equalizes steady-state utility under labor-market power and under the production-efficient allocation. Panel (c) plots the expected employer type, measured by $\log z$, by worker log ability in the decentralized and efficient allocations. Panel (d) reports worker reallocation across firms, measured as the change in employment shares across worker and firm deciles between the decentralized and efficient allocations; firms are ranked globally by latent productivity. Red indicates reallocation toward a given cell, while blue indicates reallocation away from it.

5 Conclusion

In this paper, I develop a quantitative general-equilibrium model of oligopsony in which heterogeneity in firms' demand for worker ability generates sorting and segregation across firms. As a result, competition becomes localized, with firms competing primarily against other firms that target similar segments of the ability distribution. The key element is a production technology in which firm productivity is endogenous and depends on workforce composition, while still allowing the model to nest standard labor-market benchmarks as special cases. Using matched employer–employee data for Italy and Germany, I document three reduced-form patterns consistent with the mechanisms implied by this structure: (i) low- (high-) paid workers are disproportionately employed in low- (high-) paying, smaller (larger) firms; (ii) high-paying firms set higher hiring thresholds; and (iii) concentration indices are U-shaped across the worker-pay distribution.

I then use the model, calibrated to the Italian data, as a measurement device. Under the baseline calibration, labor-market power generates aggregate output losses of 2.32%, far below benchmarks in the literature. Reassuringly, although conventional methods for estimating the average markdown are contaminated by worker heterogeneity, the resulting discrepancy is quantitatively small in the model. Markdowns vary with worker ability. Workers receive between 81% and 83.7% of their marginal product, with the lowest share at the top of the ability distribution. Workers experience large and heterogeneous welfare losses, ranging from about 20.5% to 24% in consumption-equivalent terms. These losses are largest at the top of the ability distribution, where outside options are concentrated among a small number of top-paying firms, but they are also substantial at the bottom (22.7%), where workers are excluded from most high-productivity firms and are therefore concentrated among a narrow set of small, low-productivity employers. Finally, labor-market power increases wage inequality. On one hand, it dampens the wages of the highest-ability workers, who face the largest markdowns, and weakens assortative matching. On the other hand, it generates substantial between-market wage dispersion. In the quantitative analysis, this latter force dominates: absent markdowns, the standard deviation of log wages falls by 0.044 log points.

Taken together, these results show that heterogeneity in firms' demand for workers of different ability fundamentally alters who competes with whom in the labor market. It weakens the tight link between firm size, productivity, and market power; amplifies wage inequality; generates systematic heterogeneity in monopsony across worker types; and creates sizable heterogeneous welfare losses through workers' unequal exposure to imperfect competition.

References

Abowd, John M, Francis Kramarz, and David N Margolis (1999). "High Wage Workers and High Wage Firms". In: *Econometrica* 67.2, pp. 251–334.

Atkeson, Andrew and Ariel Burstein (2008). "Pricing-to-market, trade costs, and international relative prices". In: *American Economic Review* 98.5, pp. 1998–2031.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2010). "Social Incentives in the Workplace". In: *The Review of Economic Studies* 77.2, pp. 417–458.

Barkai, Simcha (2020). "Declining Labor and Capital Shares". In: *The Journal of Finance* 75.5, pp. 2421–2463.

Becker, Gary S (1973). "A Theory of Marriage: Part I". In: *Journal of Political Economy* 81.4, pp. 813–846.

Bender, Stefan, Nicholas Bloom, David Card, John Van Reenen, and Stefanie Wolter (2018). "Management Practices, Workforce Selection, and Productivity". In: *Journal of Labor Economics* 36.S1, S371–S409.

Berger, David, Kyle Herkenhoff, and Simon Mongey (2022). "Labor Market Power". In: *American Economic Review* 112.4, pp. 1147–93.

— (2025). "Minimum Wages, Efficiency, and Welfare". In: *Econometrica* 93.1, pp. 265–301.

Bils, Mark, Barış Kaymak, and Kai-Jie Wu (Aug. 2025). *Robinson Meets Roy: Monopsony Power and Comparative Advantage*. NBER Working Paper 34165. National Bureau of Economic Research.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2022). "Discretizing Unobserved Heterogeneity". In: *Econometrica* 90.2, pp. 625–643.

Breza, Emily, Supreet Kaur, and Yogita Shamdasani (2018). "The Morale Effects of Pay Inequality". In: *The Quarterly Journal of Economics* 133.2, pp. 611–663.

Card, David (2022). "Who Set Your Wage?" In: *American Economic Review* 112.4, pp. 1075–90.

Card, David, Ana Rute Cardoso, Jörg Heining, and Patrick Kline (2018). "Firms and Labor Market Inequality: Evidence and Some Theory". In: *Journal of Labor Economics* 36.S1, S13–S70.

Card, David, Jörg Heining, and Patrick Kline (2013). "Workplace Heterogeneity and the Rise of West German Wage Inequality". In: *The Quarterly Journal of Economics* 128.3, pp. 967–1015.

Carrillo-Tudela, Carlos, Hermann Gartner, and Leo Kaas (2023). "Recruitment Policies, Job-Filling Rates, and Matching Efficiency". In: *Journal of the European Economic Association* 21.6, pp. 2413–2459.

Chetty, Raj, Adam Guren, Day Manoli, and Andrea Weber (2011). "Are Micro and Macro Labor Supply Elasticities Consistent? A Review of Evidence on the Intensive and Extensive Margins". In: *American Economic Review* 101.3, pp. 471–475.

Cohn, Alain, Ernst Fehr, Benedikt Herrmann, and Frédéric Schneider (2014). "Social Comparison and Effort Provision: Evidence from a Field Experiment". In: *Journal of the European Economic Association* 12.4, pp. 877–898.

Costinot, Arnaud and Jonathan Vogel (2010). "Matching and Inequality in the World Economy". In: *Journal of Political Economy* 118.4, pp. 747–786.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger (2020). "The Rise of Market Power and the Macroeconomic Implications". In: *The Quarterly Journal of Economics* 135.2, pp. 561–644.

De Ridder, Maarten, Basile Grassi, and Giovanni Morzenti (2026). "The Hitchhiker's Guide to Markup Estimation: Assessing Estimates from Financial Data". In: *Econometrica* 94.1, pp. 137–168.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu (2023). "How Costly Are Markups?" In: *Journal of Political Economy* 131.7, pp. 1619–1675.

Eeckhout, Jan and Philipp Kircher (2018). "Assortative Matching With Large Firms". In: *Econometrica* 86.1, pp. 85–132.

Falk, Armin and Andrea Ichino (2006). "Clean Evidence on Peer Effects". In: *Journal of Labor Economics* 24.1, pp. 39–58.

Feenstra, Robert C, Robert Inklaar, and Marcel P Timmer (2015). "The Next Generation of the Penn World Table". In: *American Economic Review* 105.10, pp. 3150–82.

Felix, Mayara (Mar. 2026). *Trade, Labor Market Concentration, and Wages*. NBER Working Paper 35018. National Bureau of Economic Research.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer (2013). "Human Capital and Regional Development". In: *The Quarterly Journal of Economics* 128.1, pp. 105–164.

Harberger, Arnold C (1954). "The Welfare Loss from Monopoly". In: *American Economic Review* 44.2, pp. 77–87.

Helpman, Elhanan, Oleg Itskhoki, and Stephen Redding (2010). "Inequality and Unemployment in a Global Economy". In: *Econometrica* 78.4, pp. 1239–1283.

Hopenhayn, Hugo A. (Aug. 2014). *On the Measure of Distortions*. NBER Working Paper 20404. National Bureau of Economic Research.

Hsieh, Chang-Tai and Peter J Klenow (2009). "Misallocation and Manufacturing TFP in China and India". In: *The Quarterly Journal of Economics* 124.4, pp. 1403–1448.

Ichino, Andrea and Giovanni Maggi (2000). "Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm". In: *The Quarterly Journal of Economics* 115.3, pp. 1057–1090.

Jäger, Simon, Jörg Heining, and Nathan Lazarus (2024). "How Substitutable Are Workers? Evidence from Worker Deaths". American Economic Review, conditionally accepted.

Keane, Michael P (2011). "Labor Supply and Taxes: A Survey". In: *Journal of Economic Literature* 49.4, pp. 961–1075.

Kline, Patrick, Neviana Petkova, Heidi Williams, and Owen Zidar (2019). "Who Profits from Patents? Rent-Sharing at Innovative Firms". In: *The Quarterly Journal of Economics* 134.3, pp. 1343–1404.

Kremer, Michael (1993). "The O-ring Theory of Economic Development". In: *The Quarterly Journal of Economics* 108.3, pp. 551–575.

Lamadon, Thibaut, Magne Mogstad, and Bradley Setzler (2022). "Imperfect Competition, Compensating Differentials, and Rent Sharing in the US Labor Market". In: *American Economic Review* 112.1, pp. 169–212.

Manning, Alan (2003). "The Real Thin Theory: Monopsony in Modern Labour Markets". In: *Labour Economics* 10.2, pp. 105–131.

— (2021). "Monopsony in Labor Markets: A Review". In: *ILR Review* 74.1, pp. 3–26.

Mas, Alexandre and Enrico Moretti (2009). "Peers at Work". In: *American Economic Review* 99.1, pp. 112–145.

Moretti, Enrico (2004). "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions". In: *American Economic Review* 94.3, pp. 656–690.

Restuccia, Diego and Richard Rogerson (2008). "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments". In: *Review of Economic Dynamics* 11.4, pp. 707–720.

Robinson, Joan (1933). *The Economics of Imperfect Competition*. London: Macmillan and Co., Ltd.

Saint-Paul, Gilles (2001). "On the Distribution of Income and Worker Assignment under Intrafirm Spillovers, with an Application to Ideas and Networks". In: *Journal of Political Economy* 109.1, pp. 1–37.

Shimer, Robert and Lones Smith (2000). "Assortative Matching and Search". In: *Econometrica* 68.2, pp. 343–369.

Song, Jae, David J Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter (2019). "Firming Up Inequality". In: *The Quarterly Journal of Economics* 134.1, pp. 1–50.

Staiger, Douglas O, Joanne Spetz, and Ciaran S Phibbs (2010). "Is There Monopsony in the Labor Market? Evidence from a Natural Experiment". In: *Journal of Labor Economics* 28.2, pp. 211–236.

Volpe, Oscar (2024). *Job Preferences, Labor Market Power, and Inequality*. Working paper.

Yeh, Chen, Claudia Macaluso, and Brad Hershbein (2022). "Monopsony in the US labor market". In: *American Economic Review* 112.7, pp. 2099–2138.

Supplemental Appendix

Contents

A Theory

A.1 Derivation of Nested-CES Labor Supply

This appendix derives the nested-CES labor-supply system used in the main text. The microfoundation follows Berger et al. (2022) (BHM), extended to allow the choice set of firms to depend on worker ability a, denoted $\mathcal{S}_j(a)$ in local labor market j.

Setup. For each ability type a, there is a unit measure of ex-ante identical individuals indexed by $l \in [0,1]$. Individual l has heterogeneous preferences over firms (i,j), summarized by an idiosyncratic taste shock $\zeta_{lij}(a)$. Conditional on supplying hours $h_{lij}(a)$ to firm (i,j), the worker's disutility index is

$$\log v_{lij}(a) = \log h_{lij}(a) - \zeta_{lij}(a),$$

so a higher $\zeta_{lij}(a)$ makes the match more attractive for given hours.

The shock vector $\{\zeta_{lij}(a)\}_{i,j}$ follows a nested Gumbel distribution. For expositional simplicity, consider the discrete-market case:

$$F(\zeta) = \exp\left(-\sum_{j=1}^{J} \left(\sum_{i=1}^{m_j} e^{-(1+\eta)\zeta_{ij}} \right)^{\frac{1+\theta}{1+\eta}} \right),$$

where $\eta > \theta > 0$ govern substitutability *across firms* within a market and *across markets*, respectively. In the continuum-of-markets case used in the main text, sums over j are replaced by integrals.

Each individual receives income $Y_l(a)$, drawn from a distribution F_Y independently of $\zeta_{lij}(a)$. Conditional on choosing firm (i,j), hours satisfy

$$w_{ij}(a)\, h_{lij}(a) = Y_l(a).$$

Individual problem. Worker l of ability a chooses the employer that minimizes disutility:

$$\min_{(i,j)\in\mathcal{S}(a)} \big\{ \log h_{lij}(a) - \zeta_{lij}(a) \big\} \equiv \max_{(i,j)\in\mathcal{S}(a)} \big\{ -\log h_{lij}(a) + \zeta_{lij}(a) \big\},$$

where $\mathcal{S}(a) = \bigcup_j \{(i,j) : i \in \mathcal{S}_j(a)\}$ is the type-a choice set.

Using $h_{lij}(a) = Y_l(a)/w_{ij}(a)$, this becomes

$$\max_{(i,j)\in\mathcal{S}(a)} \left\{ \log w_{ij}(a) - \log Y_l(a) + \zeta_{lij}(a) \right\}.$$

Hence the systematic component of utility is $\log w_{ij}(a)$, while $\log Y_l(a)$ is common across firms and $\zeta_{lij}(a)$ generates the nested-logit structure.

Choice probabilities. By standard nested-logit results (McFadden, 1974), the probability that a worker of type a chooses firm $(i,j) \in \mathcal{S}_j(a)$ is

$$p_{ij}(a) = \frac{w_{ij}(a)^{1+\eta}}{\sum_{i' \in \mathcal{S}_j(a)} w_{i'j}(a)^{1+\eta}} \cdot \frac{\left(\sum_{i' \in \mathcal{S}_j(a)} w_{i'j}(a)^{1+\eta}\right)^{\frac{1+\theta}{1+\eta}}}{\displaystyle\int_0^1 \left(\sum_{k \in \mathcal{S}_{j'}(a)} w_{kj'}(a)^{1+\eta}\right)^{\frac{1+\theta}{1+\eta}} dj'}. \tag{A.1}$$

Because $Y_l(a)$ is independent of the taste shocks, this probability is the same for all individuals of ability a.

Wage indices. Define the market- and aggregate-level wage indices

$$w_j(a) = \left[\sum_{i \in \mathcal{S}_j(a)} w_{ij}(a)^{1+\eta}\right]^{\frac{1}{1+\eta}}, \qquad W(a) = \left[\int_0^1 w_j(a)^{1+\theta}\, dj\right]^{\frac{1}{1+\theta}}.$$

These are the within-market and across-market CES wage aggregators for type-a workers.

Employment aggregation. Let $\overline{Y}(a) \coloneqq \int Y_l(a)\, dF_Y(Y_l(a))$ denote aggregate labor income for type a. Total employment of ability a at firm (i,j) is

$$n_{ij}(a) = \int p_{ij}(a)\, h_{lij}(a)\, dF_Y(Y_l(a)) = p_{ij}(a)\, \frac{\overline{Y}(a)}{w_{ij}(a)},$$

where the second equality uses $h_{lij}(a) = Y_l(a)/w_{ij}(a)$ and independence of $Y_l(a)$ and $\zeta_{lij}(a)$.

Let

$$S_j(a) \coloneqq \sum_{i' \in \mathcal{S}_j(a)} w_{i'j}(a)^{1+\eta}.$$

Substituting (A.1) yields

$$n_{ij}(a) = \overline{Y}(a)\, w_{ij}(a)^{\eta}\, S_j(a)^{\frac{1+\theta}{1+\eta}-1} \Big/ \int_0^1 S_{j'}(a)^{\frac{1+\theta}{1+\eta}} dj'.$$

Using $w_j(a) = S_j(a)^{1/(1+\eta)}$ and

$$W(a)^{1+\theta} = \int_0^1 S_j(a)^{\frac{1+\theta}{1+\eta}} dj,$$

this simplifies to

$$n_{ij}(a) = \overline{Y}(a)\, w_{ij}(a)^{\eta}\, w_j(a)^{\theta-\eta} / W(a)^{1+\theta}.$$

Define the nested-CES employment aggregators

$$n_j(a) = \left[\sum_{i \in \mathcal{S}_j(a)} n_{ij}(a)^{\frac{\eta+1}{\eta}} \right]^{\frac{\eta}{\eta+1}}, \qquad N(a) = \left[\int_0^1 n_j(a)^{\frac{\theta+1}{\theta}} dj \right]^{\frac{\theta}{\theta+1}}.$$

A straightforward calculation gives

$$n_j(a) = \frac{\overline{Y}(a)}{W(a)^\theta} w_j(a)^\theta \quad \Rightarrow \quad N(a) = \frac{\overline{Y}(a)}{W(a)}.$$

Hence aggregate labor income satisfies

$$\overline{Y}(a) = W(a)N(a).$$

Substituting $\overline{Y}(a) = W(a)N(a)$ back into $n_{ij}(a)$ yields

$$n_{ij}(a) = \left(\frac{w_{ij}(a)}{w_j(a)} \right)^\eta \left(\frac{w_j(a)}{W(a)} \right)^\theta N(a), \tag{A.2}$$

which coincides with the firm-level labor-supply system in the main text.

Representative-household formulation. The labor-supply system implied by individual discrete choice is equivalent to the solution of a representative-household problem for each ability type a. Conditional on aggregate labor $N(a)$, the representative type-a household chooses $\{n_{ij}(a)\}$ to maximize labor income,

$$\max_{\{n_{ij}(a)\}} \int_0^1 \sum_{i \in \mathcal{S}_j(a)} w_{ij}(a)\, n_{ij}(a)\, dj,$$

subject to

$$N(a) = \left[\int_0^1 n_j(a)^{\frac{\theta+1}{\theta}} dj \right]^{\frac{\theta}{\theta+1}}, \qquad n_j(a) = \left[\sum_{i \in \mathcal{S}_j(a)} n_{ij}(a)^{\frac{\eta+1}{\eta}} \right]^{\frac{\eta}{\eta+1}}, \qquad \eta > \theta > 0.$$

The first-order conditions recover the nested-CES allocation in (A.2). Combined with the top-level disutility from total labor $N(a)$ in the main text, they deliver the inverse labor-supply system in (2).

A.2 Production Function: Microfoundation

This subsection provides a microfoundation for the production technology used in the main text, building on Saint-Paul (2001), Helpman et al. (2010), and Eeckhout et al. (2018). For a given firm, suppress indices and write y, k, h, and z for output, capital, total employment, and firm type.

Consider a firm that produces a single output y with a team of workers of heterogeneous

abilities $a \in \mathcal{A}$, supervised or coordinated by a manager of type z. As in Eeckhout et al. (2018), each worker produces according to

$$f(a, z, \chi) = \phi(a, z)\, \chi^{\omega_\chi},$$

where χ is the worker's share of a common firm-level resource and ω_χ is the elasticity of output with respect to that resource. Unlike Eeckhout et al. (2018), the firm cannot allocate χ differentially across workers based on ability a: each worker receives the same bundle of common resources.

For concreteness, suppose the firm rents a stock of common space or equipment k at price R and allocates it equally across workers. Let total employment be $h = \sum_{a \in \mathcal{A}} n(a)$, so each worker receives $\chi = k/h$. To match the baseline production function in the main text, impose $\omega_\chi = 1 - \gamma$. Worker-level production is then

$$f(a, z, \chi) = \phi(a, z) \left(\frac{k}{h}\right)^{1-\gamma}.$$

Aggregating across workers gives

$$\begin{aligned} y &= \sum_{a \in \mathcal{A}} \phi(a, z) \left(\frac{k}{h}\right)^{1-\gamma} n(a) \\ &= \left(\frac{k}{h}\right)^{1-\gamma} \sum_{a \in \mathcal{A}} \phi(a, z)\, n(a) \\ &= k^{1-\gamma} h^{\gamma} \sum_{a \in \mathcal{A}} \phi(a, z)\, \frac{n(a)}{h} = \mathbb{E}_{g(a)}[\phi(a, z)] \left(k^{1-\gamma} h^{\gamma}\right), \end{aligned} \tag{A.3}$$

where

$$g(a) = \frac{n(a)}{h}, \qquad \mathbb{E}_{g(a)}[\phi(a, z)] = \sum_{a \in \mathcal{A}} \phi(a, z)\, g(a)$$

denotes realized firm productivity.

In the main text, I introduce an additional curvature parameter $\alpha \in (0, 1]$ governing decreasing returns in the composite $k^{1-\gamma} h^{\gamma}$, so final output is

$$y = \mathbb{E}_{g(a)}[\phi(a, z)] \big(k^{1-\gamma} h^{\gamma}\big)^{\alpha}.$$

The microfoundation above corresponds to the special case $\alpha = 1$. Allowing $\alpha < 1$ parsimoniously captures additional decreasing returns at the firm level, such as managerial span-of-control constraints or a downward-sloping demand curve.

Relation to existing production functions. Combining (A.3) with the CES specification for $\phi(a, z)$ in (8),

$$\phi(a, z) = \left[(1 - \omega_a)\, z^{\frac{\rho-1}{\rho}} + \omega_a\, a^{\frac{\rho-1}{\rho}}\right]^{\frac{\rho}{\rho-1}}, \qquad \rho \leq 1,\ \omega_a \in [0, 1],$$

nests several benchmark cases as special or limiting forms:

- **Cobb–Douglas benchmark.** If $\omega_a = 0$, worker output depends only on firm type, and

$$\phi(a, z) = z \quad \Longrightarrow \quad y = z\, k^{1-\gamma} h^{\gamma},$$

  which corresponds, for $\alpha = 1$, to the production function in Berger et al. (2022).

- **Multiplicative complementarities.** As $\rho \to 1$, the CES aggregator becomes log-linear:

$$\phi(a, z) = z^{1-\omega_a} a^{\omega_a}, \qquad y = z^{1-\omega_a} k^{1-\gamma} h^{\gamma} \left(\sum_{a \in \mathcal{A}} a^{\omega_a} g(a) \right),$$

  similar to Helpman et al. (2010), where output depends on the interaction between firm productivity and the appropriately weighted average worker ability.

- **Additive aggregation.** If worker output does not depend on the shared resource, $\omega_\chi = 0$ (equivalently, $\gamma = 1$ in the parameterization above), then each worker produces $\phi(a, z)$ independently of k, and

$$y = \sum_{a \in \mathcal{A}} \phi(a, z)\, n(a),$$

  corresponding to the additive formulation in Costinot et al. (2010).

A.3 Firm Wage Structure

This subsection proves Proposition 1. I first derive the firm-level labor-supply elasticity and show that the derivative of the CES employment index equals the wage-bill share. I then substitute out optimal capital, rewrite revenue as a function of total worker-specific productivity units and total employment, and use the Lagrangian to characterize the implied wage schedule.

Step 1: Firm-level labor-supply elasticity and CES index. Recall the inverse labor-supply system for ability-a workers in market j:

$$w_{ij}(a) = W(a) \left(\frac{n_{ij}(a)}{n_j(a)} \right)^{1/\eta} \left(\frac{n_j(a)}{N(a)} \right)^{1/\theta},$$

where the CES employment index in market j is

$$n_j(a) = \left[\sum_{k \in \mathcal{S}_j(a)} n_{kj}(a)^{\frac{\eta+1}{\eta}} \right]^{\frac{\eta}{\eta+1}}.$$

Define the firm's wage-bill share for ability-a workers:

$$s_{ij}(a) := \frac{w_{ij}(a)\, n_{ij}(a)}{\sum_{k\in\mathcal{S}_j(a)} w_{kj}(a)\, n_{kj}(a)}.$$

(i) Wage-bill share. Using the inverse labor supply,

$$w_{ij}(a)n_{ij}(a) = W(a)\, N(a)^{-1/\theta} n_j(a)^{1/\theta - 1/\eta} n_{ij}(a)^{1+1/\eta},$$

and likewise for all firms in $\mathcal{S}_j(a)$. Hence

$$s_{ij}(a) = \frac{n_{ij}(a)^{\frac{\eta+1}{\eta}}}{\sum_{k\in\mathcal{S}_j(a)} n_{kj}(a)^{\frac{\eta+1}{\eta}}}. \tag{A.4}$$

(ii) Derivative of the CES employment index. Let

$$S(a) := \sum_{k\in\mathcal{S}_j(a)} n_{kj}(a)^{\frac{\eta+1}{\eta}}, \qquad n_j(a) = S(a)^{\frac{\eta}{\eta+1}}.$$

Then

$$\frac{\partial S(a)}{\partial n_{ij}(a)} = \frac{\eta+1}{\eta} n_{ij}(a)^{1/\eta},$$

and

$$\frac{\partial n_j(a)}{\partial n_{ij}(a)} = \frac{\eta}{\eta+1} S(a)^{\frac{\eta}{\eta+1}-1} \frac{\partial S(a)}{\partial n_{ij}(a)} = S(a)^{\frac{\eta}{\eta+1}-1} n_{ij}(a)^{1/\eta}.$$

Using $n_j(a) = S(a)^{\eta/(\eta+1)}$, we have $S(a)^{\frac{\eta}{\eta+1}-1} = n_j(a)/S(a)$, and therefore

$$\frac{\partial n_j(a)}{\partial n_{ij}(a)} = \frac{n_j(a)}{S(a)}\, n_{ij}(a)^{1/\eta}.$$

Thus

$$\begin{aligned}\frac{\partial \log n_j(a)}{\partial \log n_{ij}(a)} &= \frac{n_{ij}(a)}{n_j(a)} \frac{\partial n_j(a)}{\partial n_{ij}(a)} = \frac{n_{ij}(a)^{1+1/\eta}}{S(a)} \\ &= \frac{n_{ij}(a)^{\frac{\eta+1}{\eta}}}{\sum_{k\in\mathcal{S}_j(a)} n_{kj}(a)^{\frac{\eta+1}{\eta}}} = s_{ij}(a),\end{aligned}$$

where the last equality uses (A.4). Hence

$$\frac{\partial \log n_j(a)}{\partial \log n_{ij}(a)} = s_{ij}(a).$$

(iii) Firm-level elasticity. Taking logs of the inverse labor supply,

$$\log w_{ij}(a) = \log W(a) + \frac{1}{\eta}\big(\log n_{ij}(a) - \log n_j(a)\big) + \frac{1}{\theta}\big(\log n_j(a) - \log N(a)\big),$$

so

$$\begin{aligned}\frac{\partial \log w_{ij}(a)}{\partial \log n_{ij}(a)} &= \frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big)\frac{\partial \log n_j(a)}{\partial \log n_{ij}(a)} \\ &= \frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big) s_{ij}(a).\end{aligned}$$

By definition,

$$\epsilon_{ij}(a) := \left(\frac{\partial w_{ij}(a)}{\partial n_{ij}(a)}\frac{n_{ij}(a)}{w_{ij}(a)}\right)^{-1} = \left(\frac{\partial \log w_{ij}(a)}{\partial \log n_{ij}(a)}\right)^{-1},$$

and therefore

$$\epsilon_{ij}(a) = \left[\frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta}\right) s_{ij}(a)\right]^{-1}, \tag{A.5}$$

which coincides with equation (13) in the main text. Since $\frac{1}{\theta} - \frac{1}{\eta} > 0$, the elasticity is decreasing in $s_{ij}(a)$: a larger share implies a less elastic firm-level labor supply and hence a larger markdown distortion.

Step 2: Output after capital choice and job slots. Throughout, $\mathcal{A}$ denotes the finite grid of abilities, $g_{ij}(a)$ the within-firm employment shares, and

$$\Phi_{ij} = \sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, g_{ij}(a)$$

the corresponding endogenous productivity term. Recall the firm production function

$$y_{ij} = \Phi_{ij}\left(k_{ij}^{1-\gamma} h_{ij}^{\gamma}\right)^{\alpha}, \qquad \Phi_{ij} := \sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, g_{ij}(a), \quad g_{ij}(a) = \frac{n_{ij}(a)}{h_{ij}},$$

with $0 < \alpha \leq 1$ and $0 < \gamma \leq 1$. Let R denote the rental rate of capital.

Define total efficiency units of labor as

$$\sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, n_{ij}(a) = h_{ij}\, \Phi_{ij}.$$

Substituting $\Phi_{ij} = \big(\sum_a \phi(a, z_{ij}) n_{ij}(a)\big)/h_{ij}$ into the production function gives

$$y_{ij} = \frac{\sum_a \phi(a, z_{ij})\, n_{ij}(a)}{h_{ij}}\left(k_{ij}^{1-\gamma} h_{ij}^{\gamma}\right)^{\alpha} = \Big(\sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, n_{ij}(a)\Big) k_{ij}^{\alpha(1-\gamma)} h_{ij}^{\alpha\gamma-1}.$$

Given $\{n_{ij}(a)\}_a$ and h_{ij}, the firm chooses k_{ij} to solve

$$\max_{k_{ij}} \Big(\sum_a \phi(a, z_{ij})\, n_{ij}(a)\Big) k_{ij}^{\alpha(1-\gamma)} h_{ij}^{\alpha\gamma-1} - R k_{ij}.$$

The first-order condition implies

$$R k_{ij} = \alpha(1-\gamma)\, y_{ij},$$

so at the optimum capital costs absorb a constant fraction $\alpha(1-\gamma)$ of revenue. Solving for k_{ij} and substituting back into the production function yields

$$y_{ij} = Z \Big(\sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, n_{ij}(a)\Big)^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\frac{\alpha\gamma-1}{1-\alpha(1-\gamma)}}, \qquad Z > 0, \tag{A.6}$$

where Z depends only on (α, γ, R). Note that $1-\alpha(1-\gamma) > 0$ for $0 < \alpha \leq 1$ and $0 < \gamma \leq 1$.

It is convenient to subtract capital costs and work with revenue net of capital,

$$\widetilde{y}_{ij} := \big(1-\alpha(1-\gamma)\big) y_{ij} = \widetilde{Z} \Big(\sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, n_{ij}(a)\Big)^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\nu},$$

where

$$\widetilde{Z} := \big(1-\alpha(1-\gamma)\big) Z > 0, \qquad \nu := \frac{\alpha\gamma - 1}{1-\alpha(1-\gamma)}.$$

Under decreasing returns to labor, $\alpha\gamma < 1$, so $\nu < 0$. Holding $\{n_{ij}(a)\}_a$ fixed,

$$\frac{\partial \widetilde{y}_{ij}}{\partial h_{ij}} = \nu\, \widetilde{Z} \Big(\sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, n_{ij}(a)\Big)^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\nu-1} < 0 \quad \text{if } \alpha\gamma < 1.$$

Thus, conditional on $\sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, n_{ij}(a)$, expanding the number of job slots reduces net revenue: although more slots raise employment capacity, each worker receives a smaller share of the common resource. Under decreasing returns, this dilution effect dominates.

Step 3: Lagrangian, shadow cost of job slots, and marginal products. Using the reduced-form representation, the firm's static problem can be written as

$$\pi_{ij} = \max_{\{n_{ij}(a)\}_a,\, h_{ij}} \left\{ \widetilde{y}_{ij}(\{n_{ij}(a)\}_a, h_{ij}) - \sum_{a\in\mathcal{A}} w_{ij}(a)\, n_{ij}(a) \right\}, \qquad h_{ij} = \sum_{a\in\mathcal{A}} n_{ij}(a), \quad n_{ij}(a) \geq 0. \tag{A.7}$$

Formally,

$$\widetilde{y}_{ij}(\{n_{ij}(a)\}_a, h_{ij}) = \widetilde{Z} \Big(\sum_{a\in\mathcal{A}} \phi(a, z_{ij})\, n_{ij}(a)\Big)^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\nu}, \qquad \nu = \frac{\alpha\gamma-1}{1-\alpha(1-\gamma)} < 0 \text{ if } \alpha\gamma < 1.$$

The Lagrangian is

$$\mathcal{L} = \widetilde{y}_{ij}(\{n_{ij}(a)\}_a, h_{ij}) - \sum_a w_{ij}(a)\, n_{ij}(a) + \lambda\left(\sum_a n_{ij}(a) - h_{ij}\right) + \sum_a \varphi(a)\, n_{ij}(a),$$

with multiplier λ on the identity $h_{ij} = \sum_a n_{ij}(a)$ and multipliers $\varphi(a) \geq 0$ enforcing $n_{ij}(a) \geq 0$.

FOC with respect to h_{ij}. Differentiating $\mathcal{L}$ with respect to h_{ij} gives

$$\frac{\partial \mathcal{L}}{\partial h_{ij}} = \frac{\partial \widetilde{y}_{ij}}{\partial h_{ij}} - \lambda = 0 \quad \Rightarrow \quad \lambda = \frac{\partial \widetilde{y}_{ij}}{\partial h_{ij}}.$$

Under $\alpha\gamma < 1$, we have $\frac{\partial \widetilde{y}_{ij}}{\partial h_{ij}} < 0$, so $\lambda < 0$. The multiplier λ is the shadow value of a job slot: expanding h_{ij} raises the number of slots but, holding $\{n_{ij}(a)\}_a$ fixed, reduces the net value of a given bundle of efficiency units because common resources are spread more thinly across workers.

FOC with respect to $n_{ij}(a)$. Differentiating with respect to $n_{ij}(a)$,

$$\frac{\partial \mathcal{L}}{\partial n_{ij}(a)} = \frac{\partial \widetilde{y}_{ij}}{\partial n_{ij}(a)} + \lambda - w_{ij}(a) - \frac{\partial w_{ij}(a)}{\partial n_{ij}(a)}\, n_{ij}(a) + \varphi(a) = 0.$$

The first term,

$$\frac{\partial \widetilde{y}_{ij}}{\partial n_{ij}(a)} = \widetilde{Z}\, \frac{1}{1-\alpha(1-\gamma)} \Big(\sum_b \phi(b, z_{ij})\, n_{ij}(b)\Big)^{\frac{1}{1-\alpha(1-\gamma)}-1} h_{ij}^{\nu}\, \phi(a, z_{ij}),$$

captures the direct effect of hiring an additional worker of type a on net revenue through total efficiency units; the shadow value λ captures the indirect effect through the job-slot constraint. It is natural to define the marginal product of type-a labor as

$$MPL_{ij}(a) := \frac{\partial \widetilde{y}_{ij}}{\partial n_{ij}(a)} + \lambda. \tag{A.8}$$

A direct calculation using (A.6) reproduces the marginal-product expression in equation (11) in the main text: $MPL_{ij}(a)$ can be decomposed into a positive size effect and a composition effect that is negative whenever $\phi(a, z_{ij})$ lies below firm-average productivity.

Substituting (A.8) into the FOC yields

$$MPL_{ij}(a) - w_{ij}(a) - \frac{\partial w_{ij}(a)}{\partial n_{ij}(a)}\, n_{ij}(a) + \varphi(a) = 0, \qquad \varphi(a) \geq 0, \quad \varphi(a) n_{ij}(a) = 0. \tag{A.9}$$

Step 4: Wage schedule and Lerner condition. Using the firm-specific labor-supply elasticity,

$$\epsilon_{ij}(a) = \left(\frac{\partial w_{ij}(a)}{\partial n_{ij}(a)} \frac{n_{ij}(a)}{w_{ij}(a)}\right)^{-1},$$

we can rewrite

$$\frac{\partial w_{ij}(a)}{\partial n_{ij}(a)}\, n_{ij}(a) = \frac{w_{ij}(a)}{\epsilon_{ij}(a)},$$

so (A.9) becomes

$$MPL_{ij}(a) - w_{ij}(a)\Big(1 + \frac{1}{\epsilon_{ij}(a)}\Big) + \varphi(a) = 0. \tag{A.10}$$

We now classify optimal choices by the sign of $MPL_{ij}(a)$.

Case 1: $n_{ij}(a) = 0$**.**

At $n_{ij}(a) = 0$, the inverse labor supply implies $w_{ij}(a) = 0$, and the term $\frac{\partial w_{ij}(a)}{\partial n_{ij}(a)}\, n_{ij}(a)$ is also zero. Equation (A.9) then simplifies to

$$MPL_{ij}(a) + \varphi(a) = 0 \quad \Rightarrow \quad MPL_{ij}(a) \leq 0,$$

since $\varphi(a) \geq 0$. Thus, if a worker type is not employed by the firm, its marginal product at that firm must be weakly negative, and the equilibrium wage is $w_{ij}(a) = 0$.

Case 2: $MPL_{ij}(a) > 0$**.**

If $MPL_{ij}(a) > 0$, Case 1 implies that $n_{ij}(a) = 0$ is impossible, so necessarily $n_{ij}(a) > 0$. Complementary slackness then implies $\varphi(a) = 0$. Equation (A.10) reduces to

$$MPL_{ij}(a) = w_{ij}(a)\Big(1 + \frac{1}{\epsilon_{ij}(a)}\Big),$$

so

$$w_{ij}(a) = \frac{\epsilon_{ij}(a)}{1 + \epsilon_{ij}(a)}\, MPL_{ij}(a) =: \mu_{ij}(a)\, MPL_{ij}(a), \qquad \mu_{ij}(a) := \frac{\epsilon_{ij}(a)}{1 + \epsilon_{ij}(a)} \in (0, 1). \tag{A.11}$$

Step 5: Wage structure. Combining the two cases, the equilibrium wage schedule is

$$w_{ij}(a) = \begin{cases} \dfrac{\epsilon_{ij}(a)}{1 + \epsilon_{ij}(a)}\, MPL_{ij}(a), & \text{if } MPL_{ij}(a) > 0, \\ 0, & \text{if } MPL_{ij}(a) \leq 0, \end{cases}$$

with $\epsilon_{ij}(a)$ given by (A.5). This is exactly the wage structure stated in Proposition 1 in the main text. □

A.4 Proof of Proposition 2

This appendix derives the expressions for the average marginal product, average wage, profits, and labor share stated in Proposition 2. Throughout, $\mathcal{A}$ denotes the finite grid of abilities, $g_{ij}(a)$ the within-firm employment shares, and

$$\Phi_{ij} = \sum_{a \in \mathcal{A}} \phi(a, z_{ij})\, g_{ij}(a)$$

the corresponding endogenous productivity term.

Step 1: Average marginal product. Start from the firm's production function before substituting out capital:

$$y_{ij} = \Phi_{ij}\big(k_{ij}^{1-\gamma} h_{ij}^{\gamma}\big)^{\alpha}.$$

Holding k_{ij} and the composition $g_{ij}(a)$, and hence Φ_{ij}, fixed, the partial derivative of output with respect to total employment h_{ij} is

$$\frac{\partial y_{ij}}{\partial h_{ij}} = \Phi_{ij}\,\alpha\big(k_{ij}^{1-\gamma} h_{ij}^{\gamma}\big)^{\alpha-1}\big(k_{ij}^{1-\gamma}\gamma h_{ij}^{\gamma-1}\big) = \alpha\gamma\,\frac{y_{ij}}{h_{ij}}.$$

Define the firm-level average marginal product of labor as

$$\overline{MPL}_{ij} := \frac{\partial y_{ij}}{\partial h_{ij}} = \alpha\gamma\,\frac{y_{ij}}{h_{ij}}.$$

Step 2 shows that this coincides with the employment-weighted average of individual marginal products, $\sum_a g_{ij}(a) MPL_{ij}(a)$.

Step 2: Decomposition of $MPL_{ij}(a)$. From equation (11) in the main text,

$$MPL_{ij}(a) = \overline{MPL}_{ij}\,\psi_{ij}(a), \qquad \psi_{ij}(a) := 1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\phi(a, z_{ij})}{\Phi_{ij}}\right).$$

Thus each worker's marginal product is a scalar multiple of the firm's average marginal product. Averaging over g_{ij} gives

$$\sum_{a\in\mathcal{A}} g_{ij}(a) MPL_{ij}(a) = \overline{MPL}_{ij} \sum_{a\in\mathcal{A}} g_{ij}(a)\psi_{ij}(a).$$

Using the definition of $\psi_{ij}(a)$ and $\sum_a g_{ij}(a)\phi(a, z_{ij}) = \Phi_{ij}$,

$$\sum_{a} g_{ij}(a)\psi_{ij}(a) = 1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\Phi_{ij}}{\Phi_{ij}}\right) = 1,$$

and therefore

$$\sum_{a\in\mathcal{A}} g_{ij}(a) MPL_{ij}(a) = \overline{MPL}_{ij} = \alpha\gamma\,\frac{y_{ij}}{h_{ij}}.$$

This proves the first identity in Proposition 2.

Step 3: Average wage. For any employed skill type, that is, any a with $g_{ij}(a) > 0$, the firm's optimality conditions imply $MPL_{ij}(a) \geq 0$ and the wage satisfies

$$w_{ij}(a) = \mu_{ij}(a) MPL_{ij}(a), \qquad \mu_{ij}(a) \in (0, 1].$$

Types with $MPL_{ij}(a) \leq 0$ are not employed, so $g_{ij}(a) = 0$ and drop out of all averages.

The firm's average wage is

$$\overline{w}_{ij} = \sum_{a \in \mathcal{A}} g_{ij}(a)\, w_{ij}(a) = \sum_{a \in \mathcal{A}} g_{ij}(a)\, \mu_{ij}(a) MPL_{ij}(a).$$

Using the decomposition from Step 2,

$$\overline{w}_{ij} = \overline{MPL}_{ij} \sum_{a \in \mathcal{A}} g_{ij}(a)\, \mu_{ij}(a) \psi_{ij}(a).$$

Define

$$\widetilde{\psi}_{ij} := \sum_{a \in \mathcal{A}} g_{ij}(a)\, \mu_{ij}(a) \psi_{ij}(a),$$

so that

$$\overline{w}_{ij} = \overline{MPL}_{ij}\, \widetilde{\psi}_{ij}.$$

Step 4: Decomposition of $\widetilde{\psi}_{ij}$. Start from

$$\widetilde{\psi}_{ij} = \sum_{a \in \mathcal{A}} g_{ij}(a)\, \mu_{ij}(a)\, \psi_{ij}(a), \qquad \psi_{ij}(a) = 1 - \frac{1}{\alpha\gamma}\Big(1 - \frac{\phi(a, z_{ij})}{\Phi_{ij}}\Big),$$

with $\Phi_{ij} = \sum_a \phi(a, z_{ij})\, g_{ij}(a)$ and $\overline{\mu}_{ij} = \sum_a g_{ij}(a) \mu_{ij}(a)$.

Compute $\mu_{ij}(a)\psi_{ij}(a)$ term by term:

$$\begin{aligned} \mu_{ij}(a)\psi_{ij}(a) &= \mu_{ij}(a) \left[1 - \frac{1}{\alpha\gamma}\Big(1 - \frac{\phi(a, z_{ij})}{\Phi_{ij}}\Big)\right] \\ &= \mu_{ij}(a) - \frac{1}{\alpha\gamma}\, \mu_{ij}(a) + \frac{1}{\alpha\gamma}\, \mu_{ij}(a) \frac{\phi(a, z_{ij})}{\Phi_{ij}}. \end{aligned}$$

Averaging over g_{ij} yields

$$\begin{aligned} \widetilde{\psi}_{ij} &= \sum_a g_{ij}(a)\mu_{ij}(a) - \frac{1}{\alpha\gamma} \sum_a g_{ij}(a)\mu_{ij}(a) + \frac{1}{\alpha\gamma\Phi_{ij}} \sum_a g_{ij}(a)\mu_{ij}(a)\phi(a, z_{ij}) \\ &= \Big(1 - \frac{1}{\alpha\gamma}\Big)\overline{\mu}_{ij} + \frac{1}{\alpha\gamma\Phi_{ij}}\, \mathbb{E}_{g_{ij}}[\mu\phi], \end{aligned}$$

where

$$\mathbb{E}_{g_{ij}}[\mu\phi] := \sum_a g_{ij}(a)\mu_{ij}(a)\phi(a, z_{ij}).$$

Now use

$$\mathbb{E}_{g_{ij}}[\mu\phi] = \mathbb{E}_{g_{ij}}[\mu]\mathbb{E}_{g_{ij}}[\phi] + \mathrm{cov}_{g_{ij}}(\mu, \phi),$$

together with $\mathbb{E}_{g_{ij}}[\phi] = \Phi_{ij}$, to obtain

$$\mathbb{E}_{g_{ij}}[\mu\phi] = \overline{\mu}_{ij}\,\Phi_{ij} + \mathrm{cov}_{g_{ij}}(\mu, \phi).$$

Substituting back,

$$\begin{aligned}\widetilde{\psi}_{ij} &= \Big(1 - \frac{1}{\alpha\gamma}\Big)\overline{\mu}_{ij} + \frac{1}{\alpha\gamma\Phi_{ij}}\left(\overline{\mu}_{ij}\,\Phi_{ij} + \mathrm{cov}_{g_{ij}}(\mu, \phi)\right) \\ &= \overline{\mu}_{ij} + \frac{1}{\alpha\gamma\Phi_{ij}}\,\mathrm{cov}_{g_{ij}}(\mu, \phi) = \overline{\mu}_{ij} + \frac{1}{\alpha\gamma}\,\mathrm{cov}_{g_{ij}}\left(\mu, \frac{\phi}{\Phi_{ij}}\right).\end{aligned}$$

Hence

$$\widetilde{\psi}_{ij} = \overline{\mu}_{ij} + \frac{1}{\alpha\gamma}\,\mathrm{cov}_{g_{ij}}\left(\mu_{ij}(a), \frac{\phi(a, z_{ij})}{\mathbb{E}_{g_{ij}}[\phi(a, z_{ij})]}\right),$$

which is the expression stated in Proposition 2.

Upper bound. For any employed ability type, the marginal product is nonnegative by the firm's optimality conditions, so $\psi_{ij}(a) \geq 0$ whenever $g_{ij}(a) > 0$. Since $0 \leq \mu_{ij}(a) \leq 1$, we have

$$g_{ij}(a)\,\mu_{ij}(a)\,\psi_{ij}(a) \leq g_{ij}(a)\,\psi_{ij}(a) \quad \text{for all } a.$$

Summing over a gives

$$\widetilde{\psi}_{ij} = \sum_a g_{ij}(a)\mu_{ij}(a)\psi_{ij}(a) \leq \sum_a g_{ij}(a)\psi_{ij}(a).$$

But

$$\sum_a g_{ij}(a)\psi_{ij}(a) = 1 - \frac{1}{\alpha\gamma}\Big(1 - \frac{\Phi_{ij}}{\Phi_{ij}}\Big) = 1,$$

so $\widetilde{\psi}_{ij} \leq 1$.

Special case. If $\mu_{ij}(a) \equiv 1$ for all a, then

$$\widetilde{\psi}_{ij} = \sum_a g_{ij}(a)\psi_{ij}(a) = 1 - \frac{1}{\alpha\gamma}\Big(1 - \frac{\Phi_{ij}}{\Phi_{ij}}\Big) = 1,$$

as required.

Step 5: Profits. Using the capital FOC, $Rk_{ij} = \alpha(1-\gamma)y_{ij}$, profits can be written as

$$\pi_{ij} = y_{ij} - Rk_{ij} - \sum_a w_{ij}(a)n_{ij}(a) = [1 - \alpha(1-\gamma)]y_{ij} - h_{ij}\overline{w}_{ij}.$$

Substituting $y_{ij} = h_{ij}\overline{MPL}_{ij}/(\alpha\gamma)$ and the expression for $\overline{w}_{ij}$,

$$\pi_{ij} = h_{ij}\overline{MPL}_{ij}\left[\frac{1-\alpha(1-\gamma)}{\alpha\gamma} - \widetilde{\psi}_{ij}\right] = \left[1-\alpha(1-\gamma)-\alpha\gamma\,\widetilde{\psi}_{ij}\right] y_{ij}.$$

Absent markdowns, that is, when $\widetilde{\psi}_{ij} = 1$, the profit share reduces to the standard Cobb–Douglas expression $1-\alpha$.

Step 6: Labor share. The firm's labor share is

$$ls_{ij} = \frac{h_{ij}\overline{w}_{ij}}{y_{ij}} = \frac{h_{ij}\overline{MPL}_{ij}\widetilde{\psi}_{ij}}{h_{ij}\overline{MPL}_{ij}/(\alpha\gamma)} = \alpha\gamma\,\widetilde{\psi}_{ij}.$$

In the absence of markdowns, the labor share collapses to the constant $\alpha\gamma$, as in the homogeneous Cobb–Douglas benchmark.

Conclusion. The expressions for $\overline{MPL}_{ij}$, $\overline{w}_{ij}$, π_{ij}/y_{ij}, and ls_{ij} in Proposition 2 follow directly from the steps above. □

A.5 No production misallocation without markdowns

Throughout, $C_t(a)$ and $N_t(a)$ denote total consumption and total labor supplied by workers of ability a at date t. The planner evaluates per-capita utility within each ability group, so welfare is aggregated as

$$\sum_{t=0}^{\infty}\beta^t\left[\sum_{a\in\mathcal{A}}\psi(a)\,f_a(a)\,U\left(\frac{C_t(a)}{f_a(a)},\frac{N_t(a)}{f_a(a)}\right)+\psi(e)\,U\big(C_t(e)\big)\right].$$

Proposition A.1 (No production misallocation without markdowns)**.** *In any competitive-production efficient equilibrium, there is no production misallocation relative to the planner's problem. Wages equal marginal products of labor at every firm–worker pair, the return to capital equals its marginal product, and the allocation of labor and capital across leisure, firms, and markets satisfies the planner's first-order conditions for production.*[1]

Proof. To study production efficiency, consider a dynamic planner that chooses, for each date t, type-specific consumption $\{C_t(a)\}_{a\in\mathcal{A}}$, entrepreneur consumption $C_t(e)$, labor allocations $\{n_{ijt}(a)\}_{i,j,a}$, firm-level capital allocations $\{k_{ijt}\}_{i,j}$, and next period aggregate capital K_{t+1}. The planner solves

$$\max_{\{C_t(a),\,C_t(e),\,n_{ijt}(a),\,k_{ijt},\,K_{t+1}\}_{t\geq 0}}\sum_{t=0}^{\infty}\beta^t\left[\sum_{a\in\mathcal{A}}\psi(a)\,f_a(a)\,U\left(\frac{C_t(a)}{f_a(a)},\frac{N_t(a)}{f_a(a)}\right)+\psi(e)\,U\big(C_t(e)\big)\right]$$

[1]This is a production-efficiency statement: for some choice of Pareto weights $\{\psi(a),\psi(e)\}$, the competitive-production equilibrium implements the planner's allocation of production factors. Alternative Pareto weights may change the distribution of consumption across types, but they do not alter the absence of wedges between factor prices and marginal products.

subject to, for every t,

$$\sum_{a\in\mathcal{A}} C_t(a) + C_t(e) + K_{t+1} - (1-\delta)K_t = \int_0^1 \sum_{i=1}^{m_j} y_{ijt}\, dj,$$

$$\int_0^1 \sum_{i=1}^{m_j} k_{ijt}\, dj = K_t,$$

together with the production technology and the nested-CES labor aggregators defining $N_t(a)$.

Let λ_t denote the multiplier on the aggregate resource constraint and ξ_t the multiplier on the capital-allocation constraint. Define

$$c_t(a) \equiv \frac{C_t(a)}{f_a(a)}, \qquad n_t(a) \equiv \frac{N_t(a)}{f_a(a)}.$$

Planner's FOCs for consumption and labor. The first-order condition with respect to $C_t(a)$ is

$$\psi(a)\, f_a(a)\, U_1\big(c_t(a), n_t(a)\big)\frac{1}{f_a(a)} - \lambda_t = 0,$$

that is,

$$\psi(a)\, U_1\big(c_t(a), n_t(a)\big) = \lambda_t.$$

Next, the first-order condition with respect to $n_{ijt}(a)$ is

$$\psi(a)\, f_a(a)\, U_2\big(c_t(a), n_t(a)\big)\frac{1}{f_a(a)}\frac{\partial N_t(a)}{\partial n_{ijt}(a)} + \lambda_t \frac{\partial y_{ijt}}{\partial n_{ijt}(a)} = 0,$$

or equivalently,

$$\psi(a)\, U_2\big(c_t(a), n_t(a)\big)\frac{\partial N_t(a)}{\partial n_{ijt}(a)} + \lambda_t\, MPL_{ijt}(a) = 0.$$

Dividing by the consumption FOC gives

$$-\frac{U_2\big(c_t(a), n_t(a)\big)}{U_1\big(c_t(a), n_t(a)\big)}\frac{\partial N_t(a)}{\partial n_{ijt}(a)} = MPL_{ijt}(a).$$

Hence the planner's intratemporal labor condition is

$$\mathrm{MRS}_{ijt}(a) = MPL_{ijt}(a),$$

where

$$\mathrm{MRS}_{ijt}(a) \equiv -\frac{U_2\big(c_t(a), n_t(a)\big)}{U_1\big(c_t(a), n_t(a)\big)}\frac{\partial N_t(a)}{\partial n_{ijt}(a)}.$$

Decentralized households. In the decentralized economy, the representative household of type a solves the labor-supply problem in the main text, taking wages and choice sets as given. Its

intratemporal first-order conditions imply

$$\mathrm{MRS}_{ijt}(a) = w_{ijt}(a),$$

which is equivalent to the inverse labor-supply system in (2).

By definition of a competitive-production efficient equilibrium, firms set no markdowns, so

$$w_{ijt}(a) = MPL_{ijt}(a) \qquad \forall (i, j, t, a).$$

Combining the planner's condition with the household condition therefore shows that the decentralized allocation satisfies the planner's first-order condition for labor at every firm–worker pair.

Capital across firms. The planner's first-order condition with respect to firm-level capital k_{ijt} is

$$\lambda_t \frac{\partial y_{ijt}}{\partial k_{ijt}} - \xi_t = 0,$$

that is,

$$\lambda_t \, MPK_{ijt} = \xi_t.$$

Hence the planner equalizes the marginal product of capital across firms:

$$MPK_{ijt} = \frac{\xi_t}{\lambda_t} \qquad \forall (i, j, t).$$

In the decentralized equilibrium, firms choose capital taking the rental rate R_t as given, so their first-order condition implies

$$MPK_{ijt} = R_t \qquad \forall (i, j, t).$$

Therefore the decentralized equilibrium also equalizes the marginal product of capital across firms and satisfies the planner's static efficiency condition for the cross-firm allocation of capital.

Aggregate capital accumulation. Finally, the planner's first-order condition with respect to K_{t+1} is

$$-\lambda_t + \beta \Big[\xi_{t+1} + (1 - \delta) \lambda_{t+1} \Big] = 0.$$

Using the planner's capital-allocation condition, this becomes

$$\lambda_t = \beta \, \lambda_{t+1} \Big(MPK_{t+1} + 1 - \delta \Big),$$

where MPK_{t+1} denotes the common marginal product of capital across firms at date $t + 1$.

In the decentralized economy, the representative entrepreneur chooses aggregate capital in-

tertemporally and satisfies the main text Euler equation (6). Since firms satisfy $MPK_{ijt} = R_t$ at every firm, the decentralized equilibrium also satisfies the planner's intertemporal condition for aggregate capital accumulation.

Taken together, the decentralized allocation satisfies the planner's first-order conditions for labor, for the cross-firm allocation of capital, and for aggregate capital accumulation. Equivalently, there are no wedges between factor prices and marginal products. In this sense, there is no production misallocation relative to the planner's problem. □

A.6 Proofs for Market Equilibrium Results

A.6.1 Proof of Lemma 1: Markdowns and Concentration

The proof adapts a proposition in Felix, 2026 from homogeneous to heterogeneous labor.

Proof. Under the nested-CES structure, Proposition 1 from the main text implies

$$\mu_{ij}(a) = \frac{\epsilon_{ij}(a)}{1+\epsilon_{ij}(a)}, \qquad \epsilon_{ij}(a) = \left[\frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big) s_{ij}(a)\right]^{-1}.$$

Therefore

$$\mu_{ij}(a)^{-1} = 1 + \epsilon_{ij}(a)^{-1} = 1 + \frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big) s_{ij}(a) = 1 + \frac{1}{\eta}\big[1 - s_{ij}(a)\big] + \frac{1}{\theta} s_{ij}(a).$$

Taking the wage-bill-weighted average and using $\sum_i s_{ij}(a) = 1$ and $HHI_j(a) = \sum_i s_{ij}(a)^2$ gives

$$\sum_i s_{ij}(a)\, \mu_{ij}(a)^{-1} = 1 + \frac{1}{\eta} \sum_i s_{ij}(a)\big[1 - s_{ij}(a)\big] + \frac{1}{\theta} \sum_i s_{ij}(a)^2.$$

Since

$$\sum_i s_{ij}(a)[1 - s_{ij}(a)] = 1 - \sum_i s_{ij}(a)^2 = 1 - HHI_j(a),$$

this becomes

$$1 + \frac{1}{\eta}\big[1 - HHI_j(a)\big] + \frac{1}{\theta} HHI_j(a) = 1 + \frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big) HHI_j(a),$$

which is the rightmost expression in main text (14). □

The next lemma refines the link between markdowns and market structure by decomposing the Herfindahl–Hirschman index into two transparent components. For a given worker type, concentration depends both on the number of firms effectively in the choice set and on the dispersion of their effective wage offers; the lemma provides an exact finite-sample decomposition that makes this explicit.

Lemma A.1 (Finite-sample decomposition of concentration)**.** *Fix a market j and an ability type a. Let $m_j(a)$ denote the number of firms that hire type a in market j, and let $\mathrm{CV}_j(a)$ denote the coefficient of variation of $w_{ij}(a)^{1+\eta}$ across those firms. Then*

$$HHI_j(a) = \frac{1}{m_j(a)}\big[1 + \mathrm{CV}_j(a)^2\big].$$

Hence, for a given worker type, market concentration reflects both the size of the worker choice set and the dispersion in effective wage offers across firms.

Proof. Under the nested-CES labor-supply system (see Appendix A.1),

$$s_{ij}(a) = \frac{w_{ij}(a)^{1+\eta}}{\sum_{i' \in \mathcal{S}_j(a)} w_{i'j}(a)^{1+\eta}}.$$

Define

$$X_{ij}(a) := w_{ij}(a)^{1+\eta}, \qquad \bar{X}_j(a) := \frac{1}{m_j(a)} \sum_{i \in \mathcal{S}_j(a)} X_{ij}(a),$$

and let

$$V_j(a) := \frac{1}{m_j(a)} \sum_{i \in \mathcal{S}_j(a)} \big(X_{ij}(a) - \bar{X}_j(a)\big)^2$$

denote the corresponding sample variance. Then

$$s_{ij}(a) = \frac{X_{ij}(a)}{\sum_{i'} X_{i'j}(a)},$$

and hence

$$HHI_j(a) = \sum_{i \in \mathcal{S}_j(a)} s_{ij}(a)^2 = \frac{\sum_{i \in \mathcal{S}_j(a)} X_{ij}(a)^2}{\big(\sum_{i \in \mathcal{S}_j(a)} X_{ij}(a)\big)^2}.$$

Write $S := \sum_{i \in \mathcal{S}_j(a)} X_{ij}(a) = m_j(a)\bar{X}_j(a)$. Then

$$HHI_j(a) = \frac{\sum_i X_{ij}(a)^2}{m_j(a)^2\, \bar{X}_j(a)^2}.$$

Using the decomposition of the sample second moment,

$$\frac{1}{m_j(a)} \sum_i X_{ij}(a)^2 = \bar{X}_j(a)^2 + V_j(a),$$

we obtain

$$HHI_j(a) = \frac{\bar{X}_j(a)^2 + V_j(a)}{m_j(a)\, \bar{X}_j(a)^2} = \frac{1}{m_j(a)} \left(1 + \frac{V_j(a)}{\bar{X}_j(a)^2}\right) = \frac{1}{m_j(a)}\big[1 + \mathrm{CV}_j(a)^2\big],$$

where $\mathrm{CV}_j(a) := \sqrt{V_j(a)}/\bar{X}_j(a)$ is the coefficient of variation of $X_{ij}(a) = w_{ij}(a)^{1+\eta}$. This identity

holds exactly for any finite choice set $\mathcal{S}_j(a)$. □

A.6.2 Proof of Lemma 2: Monotonicity in firm productivity

A.6.3 Proof of Lemma 2: Monotonicity in firm productivity

Proof. Fix an ability type a and consider two firms i and i' with $z_{i'j} > z_{ij}$, and suppose both firms belong to $\mathcal{S}_j(a)$.

Step 1: Employment shares. We first show that $q_{i'j}(a) \geq q_{ij}(a)$ by contradiction.

Within market j, the CES structure in the firm nest implies that, for type a,

$$q_{ij}(a) = \frac{n_{ij}(a)}{f_{ja}(a)} = \frac{w_{ij}(a)^{\eta}}{\sum_{k \in \mathcal{S}_j(a)} w_{kj}(a)^{\eta}},$$

so, holding other wages fixed, $q_{ij}(a)$ is strictly increasing in $w_{ij}(a)$.

Suppose instead that

$$q_{i'j}(a) < q_{ij}(a).$$

By strict monotonicity of $q_{ij}(a)$ in $w_{ij}(a)$, this implies

$$w_{i'j}(a) < w_{ij}(a).$$

The wage-bill share for type a at firm i is

$$s_{ij}(a) = \frac{w_{ij}(a) n_{ij}(a)}{\sum_{k \in \mathcal{S}_j(a)} w_{kj}(a) n_{kj}(a)} = \frac{w_{ij}(a)^{1+\eta}}{\sum_{k \in \mathcal{S}_j(a)} w_{kj}(a)^{1+\eta}},$$

which is also strictly increasing in $w_{ij}(a)$. Hence

$$w_{i'j}(a) < w_{ij}(a) \quad \Longrightarrow \quad s_{i'j}(a) < s_{ij}(a).$$

Under the nested-CES labor-supply structure, Proposition 1 implies

$$\mu_{ij}(a) = \frac{\epsilon_{ij}(a)}{1 + \epsilon_{ij}(a)}, \qquad \epsilon_{ij}(a) = \left[\frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta} \right) s_{ij}(a) \right]^{-1},$$

so when $\theta < \eta$, both $\epsilon_{ij}(a)$ and $\mu_{ij}(a)$ are strictly decreasing in $s_{ij}(a)$. Therefore

$$s_{i'j}(a) < s_{ij}(a) \quad \Longrightarrow \quad \mu_{i'j}(a) > \mu_{ij}(a).$$

By assumption, $MPL_{ij}(a)$ is strictly increasing in firm productivity, so $z_{i'j} > z_{ij}$ implies

$$MPL_{i'j}(a) > MPL_{ij}(a).$$

Using the wage equation $w_{ij}(a) = \mu_{ij}(a)MPL_{ij}(a)$, we obtain

$$\frac{w_{i'j}(a)}{w_{ij}(a)} = \frac{\mu_{i'j}(a)}{\mu_{ij}(a)} \cdot \frac{MPL_{i'j}(a)}{MPL_{ij}(a)} \; > \; 1,$$

because both factors on the right-hand side are strictly greater than one. This contradicts $w_{i'j}(a) < w_{ij}(a)$.

Hence

$$q_{i'j}(a) \geq q_{ij}(a).$$

If instead $q_{i'j}(a) = q_{ij}(a)$, then the CES demand system implies

$$w_{i'j}(a) = w_{ij}(a).$$

This in turn implies

$$s_{i'j}(a) = s_{ij}(a),$$

and therefore

$$\mu_{i'j}(a) = \mu_{ij}(a).$$

But since $MPL_{ij}(a)$ is strictly increasing in z_{ij} and $z_{i'j} > z_{ij}$,

$$MPL_{i'j}(a) > MPL_{ij}(a).$$

Using again $w_{ij}(a) = \mu_{ij}(a)MPL_{ij}(a)$, we obtain

$$w_{i'j}(a) = \mu_{i'j}(a)MPL_{i'j}(a) > \mu_{ij}(a)MPL_{ij}(a) = w_{ij}(a),$$

a contradiction. Hence

$$q_{i'j}(a) > q_{ij}(a).$$

Step 2: Markdown ordering. Since $q_{ij}(a)$ is strictly increasing in $w_{ij}(a)$, the inequality $q_{i'j}(a) > q_{ij}(a)$ implies

$$w_{i'j}(a) > w_{ij}(a).$$

Because $s_{ij}(a)$ is strictly increasing in $w_{ij}(a)$, it follows that

$$s_{i'j}(a) > s_{ij}(a).$$

Under $\theta < \eta$, $\mu_{ij}(a)$ is strictly decreasing in $s_{ij}(a)$, so

$$s_{i'j}(a) > s_{ij}(a) \quad \Longrightarrow \quad \mu_{i'j}(a) < \mu_{ij}(a).$$

Combining Steps 1 and 2 yields

$$q_{i'j}(a) > q_{ij}(a) \quad \text{and} \quad \mu_{i'j}(a) < \mu_{ij}(a).$$

□

A.6.4 Proof of Lemma 3: Positive assortative matching

Proof. Fix a market j and two firms i and i' with $z_{i'j} > z_{ij}$. For a given ability type a, define the wage and employment-share ratios

$$R(a) := \frac{w_{i'j}(a)}{w_{ij}(a)}, \qquad \mathcal{R}(a) := \frac{q_{i'j}(a)}{q_{ij}(a)}.$$

Under the within-market CES structure and the definition $q_{ij}(a) = n_{ij}(a)/f_{ja}(a)$, we have

$$q_{ij}(a) = \kappa(a)\, \frac{w_{ij}(a)^{\eta}}{\sum_{k\in\{i,i'\}} w_{kj}(a)^{\eta}},$$

where $\kappa(a)$ is common across firms and captures the total mass of type-a workers available to market j. Hence

$$\mathcal{R}(a) = \frac{q_{i'j}(a)}{q_{ij}(a)} = \left(\frac{w_{i'j}(a)}{w_{ij}(a)}\right)^{\eta} = R(a)^{\eta}. \tag{A.12}$$

Step 1: Implicit equation for the wage ratio. The wage condition implies, for each a,

$$w_{ij}(a) = \mu_{ij}(a)\, MPL_{ij}(a), \qquad w_{i'j}(a) = \mu_{i'j}(a)\, MPL_{i'j}(a),$$

so

$$R(a) = \frac{w_{i'j}(a)}{w_{ij}(a)} = \frac{\mu_{i'j}(a)}{\mu_{ij}(a)} \cdot \frac{MPL_{i'j}(a)}{MPL_{ij}(a)}.$$

Define

$$\Delta(a) := \log\left(\frac{MPL_{i'j}(a)}{MPL_{ij}(a)}\right), \qquad \Psi(a) := \frac{\mu_{i'j}(a)}{\mu_{ij}(a)}.$$

Taking logs yields

$$\log R(a) - \log \Psi(a) = \Delta(a). \tag{A.13}$$

Step 2: Wage-bill shares and markdowns as functions of R. At this step, the duopsony restriction in Lemma 3 is crucial. Because market j contains only the two firms i and i', the only employers of type a are i and i'. Hence, for a given a, all wage-bill shares can be written as functions of the single relative wage $R(a)$.

Write

$$w_{ij}(a) = w(a), \qquad w_{i'j}(a) = R(a)\, w(a).$$

The wage-bill shares for type a are then

$$s_{ij}(a) = \frac{w(a)^{1+\eta}}{w(a)^{1+\eta} + (R(a)\,w(a))^{1+\eta}} = \frac{1}{1 + R(a)^{1+\eta}},$$

$$s_{i'j}(a) = \frac{(R(a)\,w(a))^{1+\eta}}{w(a)^{1+\eta} + (R(a)\,w(a))^{1+\eta}} = \frac{R(a)^{1+\eta}}{1 + R(a)^{1+\eta}}.$$

Thus $s_{ij}(a)$ and $s_{i'j}(a)$ depend on a only through $R(a)$.

Under the nested-CES labor-supply structure (Proposition 1), markdowns depend on firm-specific wage-bill shares via

$$\mu_{ij}(a) = \mu\big(s_{ij}(a)\big), \qquad \mu_{i'j}(a) = \mu\big(s_{i'j}(a)\big),$$

for a common function $\mu(\cdot)$ with $\mu'(s) \leq 0$. Hence the markdown ratio can be expressed as

$$\Psi(a) = \frac{\mu_{i'j}(a)}{\mu_{ij}(a)} = \frac{\mu\big(s_{i'j}(a)\big)}{\mu\big(s_{ij}(a)\big)} =: \Psi\big(R(a)\big),$$

so (A.13) becomes

$$\log R(a) - \log \Psi\big(R(a)\big) = \Delta(a). \tag{A.14}$$

Step 3: Monotonicity of the markdown ratio in R. Define

$$s_i(R) := \frac{1}{1 + R^{1+\eta}}, \qquad s_{i'}(R) := \frac{R^{1+\eta}}{1 + R^{1+\eta}},$$

so that $\Psi(R) = \mu(s_{i'}(R))/\mu(s_i(R))$. Differentiating with respect to R,

$$\frac{ds_i}{dR} = -\frac{(1+\eta)R^{\eta}}{(1 + R^{1+\eta})^2} < 0, \qquad \frac{ds_{i'}}{dR} = \frac{(1+\eta)R^{\eta}}{(1 + R^{1+\eta})^2} > 0.$$

By the chain rule,

$$\frac{d}{dR}\log \Psi(R) = \frac{\mu'(s_{i'}(R))}{\mu(s_{i'}(R))}\frac{ds_{i'}}{dR} - \frac{\mu'(s_i(R))}{\mu(s_i(R))}\frac{ds_i}{dR}.$$

Because $\mu'(s) \leq 0$, $\mu(s) > 0$, $\frac{ds_{i'}}{dR} > 0$, and $\frac{ds_i}{dR} < 0$, each term on the right-hand side is weakly nonpositive, and the second is strictly negative whenever μ' is strictly negative on the relevant support. Hence

$$\frac{d}{dR}\log \Psi(R) \leq 0,$$

so $\log \Psi(R)$, and therefore $\Psi(R)$, is weakly decreasing in R.

Define

$$H(R) := \log R - \log \Psi(R).$$

Then

$$H'(R) = \frac{1}{R} - \frac{d}{dR}\log\Psi(R) \geq \frac{1}{R} > 0 \qquad \text{for all } R > 0,$$

so $H(R)$ is strictly increasing on $(0, \infty)$.

Step 4: Comparative statics across abilities. Log-supermodularity of $MPL_{ij}(a)$ in (a, z_{ij}) and $z_{i'j} > z_{ij}$ imply that

$$\frac{MPL_{i'j}(a)}{MPL_{ij}(a)}$$

is strictly increasing in a. Equivalently, $\Delta(a)$ is strictly increasing in a. For any $a' > a$, subtracting (A.14) at a' and a gives

$$H\big(R(a')\big) - H\big(R(a)\big) = \Delta(a') - \Delta(a) > 0.$$

Since H is strictly increasing, it follows that $R(a') > R(a)$.

Step 5: From wage ratios to employment shares. Using (A.12) and $\eta > 0$,

$$\frac{q_{i'j}(a')}{q_{ij}(a')} = \mathcal{R}(a') = R(a')^{\eta} > R(a)^{\eta} = \mathcal{R}(a) = \frac{q_{i'j}(a)}{q_{ij}(a)}.$$

Thus higher-ability workers are relatively more likely to work at the more productive firm. □

A.6.5 Proof of Proposition 4: Market concentration and worker ability in a duopsony

Proof. Fix a local labor market j and suppress the index j for brevity. There are two firms, labeled 1 and 2, with $z_2 > z_1$. Assume $MPL_i(a)$ is strictly increasing and strictly log-supermodular in (a, z_i). Then, by Lemma 2, wages are ordered by productivity:

$$w_2(a) > w_1(a) \qquad \text{for all employed ability types } a.$$

For a given ability type a, define the wage ratio

$$R(a) := \frac{w_2(a)}{w_1(a)} > 1.$$

Step 1: From positive assortative matching to an increasing wage ratio. Under the within-market CES structure, the employment share of type a at firm i can be written as

$$q_i(a) := \frac{n_i(a)}{f_{ja}(a)} = \kappa(a)\,\frac{w_i(a)^{\eta}}{w_1(a)^{\eta} + w_2(a)^{\eta}},$$

where $\kappa(a)$ is common across firms for a given a. Hence

$$\frac{q_2(a)}{q_1(a)} = \left(\frac{w_2(a)}{w_1(a)}\right)^{\eta} = R(a)^{\eta}. \tag{A.15}$$

By Lemma 3, strict log-supermodularity of $MPL_i(a)$ in (a, z_i) and $z_2 > z_1$ imply that the employment-share ratio $q_2(a)/q_1(a)$ is strictly increasing in a. By (A.15), $R(a)$ must therefore be strictly increasing in a:

$$\frac{q_2(a')}{q_1(a')} > \frac{q_2(a)}{q_1(a)} \quad \Longrightarrow \quad R(a')^\eta > R(a)^\eta \quad \Longrightarrow \quad R(a') > R(a), \qquad \forall a' > a.$$

Step 2: Wage-bill shares, the HHI, and monotonicity in $R(a)$. Because the market is a duopsony, wage-bill shares can be written as functions of the single wage ratio $R(a)$. For type a,

$$s_1(a) = \frac{w_1(a)n_1(a)}{w_1(a)n_1(a) + w_2(a)n_2(a)}, \qquad s_2(a) = \frac{w_2(a)n_2(a)}{w_1(a)n_1(a) + w_2(a)n_2(a)}.$$

Using $n_i(a) \propto w_i(a)^\eta$ and defining $v_i(a) := w_i(a)^{1+\eta}$, we obtain

$$s_i(a) = \frac{v_i(a)}{v_1(a) + v_2(a)}.$$

Since

$$v_1(a) = w_1(a)^{1+\eta}, \qquad v_2(a) = w_2(a)^{1+\eta} = R(a)^{1+\eta} w_1(a)^{1+\eta},$$

the wage-bill shares can be written purely as functions of $R(a)$:

$$s_1(a) = \frac{1}{1 + R(a)^{1+\eta}}, \qquad s_2(a) = \frac{R(a)^{1+\eta}}{1 + R(a)^{1+\eta}}.$$

The Herfindahl–Hirschman index for type a is therefore

$$HHI(a) = s_1(a)^2 + s_2(a)^2 = \frac{1}{\big(1 + R(a)^{1+\eta}\big)^2} + \frac{R(a)^{2(1+\eta)}}{\big(1 + R(a)^{1+\eta}\big)^2} = \frac{1 + R(a)^{2(1+\eta)}}{\big(1 + R(a)^{1+\eta}\big)^2}.$$

Define $t(a) := R(a)^{1+\eta} > 1$ and write

$$HHI(a) = \frac{1 + t(a)^2}{\big(1 + t(a)\big)^2} =: \widetilde{H}\big(t(a)\big).$$

A direct derivative gives

$$\widetilde{H}'(t) = \frac{2(t-1)}{(1+t)^3}.$$

Thus $\widetilde{H}'(t) > 0$ for all $t > 1$.

Since $R(a)$ is strictly increasing in a and $R(a) > 1$ for all employed types, $t(a) = R(a)^{1+\eta}$ is strictly increasing and satisfies $t(a) > 1$. Hence, for any $a' > a$,

$$R(a') > R(a) \quad \Longrightarrow \quad t(a') > t(a) \quad \Longrightarrow \quad HHI(a') = \widetilde{H}\big(t(a')\big) > \widetilde{H}\big(t(a)\big) = HHI(a).$$

Restoring the market index, this shows that in each local labor market the wage-bill concentration index is strictly increasing in ability:

$$HHI_j(a') > HHI_j(a) \qquad \text{for all } a' > a.$$

□

A.6.6 Sufficient condition for $HHI_j(a)$ increasing in a for $N > 2$

Proposition 4 shows that, in a duopsony, strict positive assortative matching (PAM) in employment shares across ability implies that the Herfindahl–Hirschman index is strictly increasing in ability. This subsection extends that logic to markets with $N > 2$ firms.

Rather than deriving PAM from technology, as in Lemma 3, I take as a sufficient condition that relative employment shares are log-supermodular in firm and worker types: for each adjacent pair of firms, the ratio of employment shares is strictly increasing in ability. Under nested-CES labor supply, this adjacent-PAM condition pins down how effective wage ratios vary with ability. I then show that the HHI is strictly increasing in ability whenever all adjacent ratios vary in this way.

I begin with auxiliary algebraic lemmas describing how the HHI depends on adjacent effective-wage ratios. Throughout, I suppress the market index j.

Lemma A.2 (Head–tail inequality)**.** *Let $\{r_i\}_{i=1}^{N-1}$ be positive numbers with $r_i \geq 1$, and define*

$$x_1 := 1, \qquad x_i := \prod_{m=1}^{i-1} r_m \quad (i = 2, \ldots, N).$$

Then $x_1 \leq x_2 \leq \cdots \leq x_N$.

Fix $k \in \{1, \ldots, N-1\}$ and set $t := r_k$. For each $j > k$, write

$$x_j = t\, y_j,$$

where $y_j := x_j/t$.

Define the head and tail index sets

$$\mathcal{H} := \{1, \ldots, k\}, \qquad \mathcal{T} := \{k+1, \ldots, N\},$$

and the corresponding aggregates

$$H_1 := \sum_{i \in \mathcal{H}} x_i, \quad T_1 := \sum_{j \in \mathcal{T}} y_j, \qquad A_0 := \sum_{i \in \mathcal{H}} x_i^2, \quad B_0 := \sum_{j \in \mathcal{T}} y_j^2.$$

Then

$$\frac{A_0}{H_1} \leq \frac{B_0}{T_1}.$$

Proof. Because each $r_i \geq 1$, the sequence $\{x_i\}$ is nondecreasing:

$$x_1 \leq x_2 \leq \cdots \leq x_N.$$

Head. For all $i \leq k$, $x_i \leq x_k$. Hence

$$x_i^2 \leq x_k x_i \quad \Rightarrow \quad A_0 = \sum_{i \in \mathcal{H}} x_i^2 \leq x_k \sum_{i \in \mathcal{H}} x_i = x_k H_1,$$

so

$$\frac{A_0}{H_1} \leq x_k.$$

Tail. For $j > k$,

$$x_j \geq x_{k+1} = x_k r_k = x_k t,$$

so

$$y_j = \frac{x_j}{t} \geq \frac{x_k t}{t} = x_k.$$

Thus

$$y_j^2 \geq x_k y_j \quad \Rightarrow \quad B_0 = \sum_{j \in \mathcal{T}} y_j^2 \geq x_k \sum_{j \in \mathcal{T}} y_j = x_k T_1,$$

which implies

$$\frac{B_0}{T_1} \geq x_k.$$

Combining the two inequalities yields

$$\frac{A_0}{H_1} \leq x_k \leq \frac{B_0}{T_1},$$

and therefore

$$\frac{A_0}{H_1} \leq \frac{B_0}{T_1}.$$

□

Lemma A.2 is purely algebraic: it compares the average of squares in the head of a nondecreasing sequence to that in the tail. The next lemma uses this inequality to show that the HHI is monotone in each adjacent effective-wage ratio when one ratio varies at a time.

Lemma A.3 (HHI monotonicity in each adjacent ratio)**.** *Fix positive numbers $v_1, \ldots, v_N$ and define wage-bill shares*

$$s_i := \frac{v_i}{\sum_{k=1}^N v_k}, \qquad HHI := \sum_{i=1}^N s_i^2.$$

Let the adjacent ratios be

$$r_i := \frac{v_{i+1}}{v_i} \geq 1, \qquad i = 1, \ldots, N-1,$$

and construct the normalized sequence

$$x_1 := 1, \qquad x_i := \prod_{m=1}^{i-1} r_m \quad (i = 2, \dots, N).$$

Then $v_i = c\, x_i$ for some $c > 0$, so that

$$s_i = \frac{x_i}{\sum_{k=1}^{N} x_k}, \qquad HHI = \frac{\sum_{i=1}^{N} x_i^2}{\left(\sum_{i=1}^{N} x_i\right)^2} =: H(r_1, \dots, r_{N-1}).$$

Fix $k \in \{1, \dots, N-1\}$ and treat all other ratios $\{r_\ell\}_{\ell \neq k}$ as constants. Then

$$\frac{\partial H}{\partial r_k} \;\geq\; 0,$$

with strict inequality whenever $r_k > 1$.

Proof. Fix k and write $t := r_k$. As in Lemma A.2, the sequence $\{x_i\}_{i=1}^{N}$ has the form:

- For $i \leq k$ (the head), x_i does not depend on t.
- For $i > k$ (the tail), $x_i = t\, y_i$, where y_i is independent of t.

Let

$$H_1 := \sum_{i \leq k} x_i, \qquad T_1 := \sum_{i > k} y_i, \qquad A_0 := \sum_{i \leq k} x_i^2, \qquad B_0 := \sum_{i > k} y_i^2.$$

Then

$$HHI = \frac{A_0 + t^2 B_0}{(H_1 + t T_1)^2} =: H(t).$$

Differentiating with respect to t yields

$$H'(t) = \frac{2\big(t H_1 B_0 - T_1 A_0\big)}{(H_1 + t T_1)^3}.$$

By Lemma A.2, we have

$$\frac{A_0}{H_1} \;\leq\; \frac{B_0}{T_1} \quad \Longleftrightarrow \quad H_1 B_0 \;\geq\; A_0 T_1.$$

Hence

$$t H_1 B_0 - T_1 A_0 = (t-1) H_1 B_0 + \big(H_1 B_0 - A_0 T_1\big) \;\geq\; 0,$$

so $H'(t) \geq 0$ for all $t \geq 1$.

Moreover, if $t > 1$, then $(t-1) H_1 B_0 > 0$ because $H_1 > 0$ and $B_0 > 0$. Therefore

$$t H_1 B_0 - T_1 A_0 > 0,$$

and hence $H'(t) > 0$ whenever $t > 1$.

Therefore, holding all other adjacent ratios fixed,

$$\frac{\partial H}{\partial r_k} = H'(t) \;\geq\; 0,$$

with strict inequality whenever $r_k > 1$. □

Lemma A.3 establishes coordinatewise monotonicity of the HHI in the adjacent ratios. The next lemma shows that if all adjacent ratios are weakly increasing in ability, then $HHI(a)$ is weakly increasing in ability.

Lemma A.4 (HHI monotonicity under adjacent-ratio monotonicity)**.** *Let $\mathcal{A}$ be a finite or countable subset of $\mathbb{R}$ endowed with the usual order. For each $a \in \mathcal{A}$, consider positive numbers $v_1(a), \ldots, v_N(a)$, define*

$$s_i(a) := \frac{v_i(a)}{\sum_{k=1}^{N} v_k(a)}, \qquad HHI(a) := \sum_{i=1}^{N} s_i(a)^2,$$

and adjacent ratios

$$r_i(a) := \frac{v_{i+1}(a)}{v_i(a)} \geq 1, \qquad i = 1, \ldots, N-1.$$

Construct the normalized sequence

$$x_1(a) := 1, \qquad x_i(a) := \prod_{m=1}^{i-1} r_m(a) \quad (i \geq 2),$$

so that, for some positive scalar $c(a)$, $v_i(a) = c(a)\, x_i(a)$ and

$$HHI(a) = \frac{\sum_{i=1}^{N} x_i(a)^2}{\left(\sum_{i=1}^{N} x_i(a)\right)^2} =: H\big(r_1(a), \ldots, r_{N-1}(a)\big).$$

Assume:

(i) For each i, the map $a \mapsto r_i(a)$ is weakly increasing: if $a' > a$, then $r_i(a') \geq r_i(a)$.

(ii) On the domain $\{r_i \geq 1\}$, the function H is weakly increasing in each coordinate, that is, $\partial H / \partial r_i \geq 0$ for all i.

Then $HHI(a)$ is weakly increasing in a.

Moreover, suppose that for any $a' > a$ there exists at least one index i such that

$$r_i(a') > r_i(a) \quad \textit{and} \quad \frac{\partial H}{\partial r_i}\Big(r_1(a'), \ldots, r_{N-1}(a')\Big) > 0.$$

Then $HHI(a)$ is strictly increasing in a, that is, $HHI(a') > HHI(a)$ for all $a' > a$.

Proof. Fix $a, a' \in \mathcal{A}$ with $a' > a$ and define

$$\mathbf{r}(a) := (r_1(a), \ldots, r_{N-1}(a)), \qquad \mathbf{r}(a') := (r_1(a'), \ldots, r_{N-1}(a')).$$

By assumption (i),

$$r_i(a') \geq r_i(a) \qquad \text{for all } i = 1, \ldots, N-1.$$

To compare $H(\mathbf{r}(a'))$ and $H(\mathbf{r}(a))$, introduce the intermediate vectors

$$\mathbf{r}^{(j)} := (r_1(a'), \ldots, r_j(a'), r_{j+1}(a), \ldots, r_{N-1}(a)), \qquad j = 0, \ldots, N-1,$$

so that $\mathbf{r}^{(0)} = \mathbf{r}(a)$ and $\mathbf{r}^{(N-1)} = \mathbf{r}(a')$.

At each step $j \to j+1$, only the $(j+1)$-th coordinate increases weakly, while all others remain unchanged. By assumption (ii) and Lemma A.3, H is weakly increasing in each coordinate, so

$$H(\mathbf{r}^{(j+1)}) \geq H(\mathbf{r}^{(j)}) \qquad \text{for } j = 0, \ldots, N-2.$$

Chaining these inequalities yields

$$HHI(a') = H(\mathbf{r}^{(N-1)}) \geq H(\mathbf{r}^{(0)}) = HHI(a),$$

establishing weak monotonicity.

For strict monotonicity, fix $a' > a$ and assume there exists an index i such that $r_i(a') > r_i(a)$ and $\partial H / \partial r_i(\mathbf{r}(a')) > 0$. Holding all other ratios at their weakly larger values in $\mathbf{r}(a')$, increasing the i-th coordinate from $r_i(a)$ to $r_i(a')$ strictly raises H. Thus

$$HHI(a') = H(\mathbf{r}(a')) > H(\mathbf{r}(a)) = HHI(a).$$

□

We can now translate the adjacent-PAM condition into monotonicity of the HHI under nested-CES labor supply.

Proposition A.2 (HHI monotonicity from adjacent PAM, $N > 2$)**.** *Fix a local labor market j with $N \geq 2$ firms. For each ability type a, let $q_{ij}(a) = n_{ij}(a)/f_{ja}(a)$ denote the employment share of type a at firm i, and let $w_{ij}(a)$ denote the corresponding wage.*

Assume:

(a) *$MPL_{ij}(a)$ is strictly increasing in (a, z_{ij}), and hence, by Lemma 2, wages can be ordered by productivity:*

$$w_{1j}(a) \leq \cdots \leq w_{Nj}(a), \qquad \textit{for all employed ability types } a.$$

(b) Adjacent PAM in employment shares: *for each $i = 1, \ldots, N-1$, the relative employment-share*

ratio

$$\frac{q_{i+1,j}(a)}{q_{ij}(a)}$$

is strictly increasing in a.

(c) The within-market labor-supply structure is nested CES as in Appendix A.1, so that, for each a,

$$q_{ij}(a) = \frac{w_{ij}(a)^{\eta}}{\sum_{k=1}^{N} w_{kj}(a)^{\eta}}.$$

Then the wage-bill Herfindahl–Hirschman index $HHI_j(a)$ is strictly increasing in a.

Proof. Fix market j and suppress the index j. For each ability a, write $w_i(a)$ and $q_i(a)$ for wages and employment shares, ordered so that $w_1(a) \leq \cdots \leq w_N(a)$.

Step 1: Adjacent PAM implies adjacent effective-wage ratios increasing in a. Under nested-CES labor supply,

$$q_i(a) = \frac{w_i(a)^{\eta}}{\sum_{k=1}^{N} w_k(a)^{\eta}}.$$

Hence, for each i,

$$\frac{q_{i+1}(a)}{q_i(a)} = \left(\frac{w_{i+1}(a)}{w_i(a)}\right)^{\eta}.$$

By assumption (a) and Lemma 2,

$$R_i(a) := \frac{w_{i+1}(a)}{w_i(a)} \geq 1.$$

By assumption (b), $q_{i+1}(a')/q_i(a') > q_{i+1}(a)/q_i(a)$ whenever $a' > a$, so the wage ratio

$$R_i(a) := \frac{w_{i+1}(a)}{w_i(a)} \geq 1$$

is strictly increasing in a for each i.

Define effective wages $v_i(a) := w_i(a)^{1+\eta}$ and adjacent effective-wage ratios

$$r_i(a) := \frac{v_{i+1}(a)}{v_i(a)} = R_i(a)^{1+\eta} \geq 1.$$

Because $R_i(a)$ is strictly increasing in a, each $r_i(a)$ is strictly increasing in a.

Step 2: Apply Lemma A.4. For each fixed a, define wage-bill shares

$$s_i(a) = \frac{v_i(a)}{\sum_k v_k(a)}, \qquad HHI(a) := \sum_{i=1}^{N} s_i(a)^2.$$

By construction,

$$HHI(a) = H\big(r_1(a), \ldots, r_{N-1}(a)\big),$$

with H as in Lemma A.3. Lemma A.3 implies that H is weakly increasing in each coordinate on

the domain $\{r_i \geq 1\}$, and we have shown that, for each i, the map $a \mapsto r_i(a)$ is strictly increasing and satisfies $r_i(a) \geq 1$.

Now fix $a' > a$. Since each $r_i(a)$ is strictly increasing in a, we have

$$r_i(a') > r_i(a) \geq 1 \qquad \text{for all } i = 1, \ldots, N-1.$$

In particular, there exists at least one index i such that $r_i(a') > r_i(a)$, and because $r_i(a') > 1$, Lemma A.3 also gives

$$\frac{\partial H}{\partial r_i}\Big(r_1(a'), \ldots, r_{N-1}(a')\Big) > 0.$$

Thus the strictness condition in Lemma A.4 is satisfied, so

$$HHI(a') > HHI(a) \qquad \text{for all } a' > a.$$

Restoring the market index j, we conclude that the wage-bill concentration index is strictly increasing in a in any market satisfying adjacent PAM in employment shares. □

A.6.7 Proof of Proposition 3: Equilibrium Invariance

Proof. Part (a).

Homogeneous labor ($\omega_a = 0$). When $\phi(a, z) = z$, worker ability does not enter production or firms' first-order conditions. Take any steady-state equilibrium of the corresponding one-type model with productivities $\{z_{ij}\}$ and firm shares $\{q_{ij}\}$. Construct a many-type allocation by setting $q_{ij}(a) \equiv q_{ij}$ for all a, so that $g_{ij}(a) = f_{ja}(a)$ at every firm. Because neither marginal products nor the wage-setting condition depend on a under this technology, and because the nested-CES labor-supply system depends only on relative wages within each type, the same firm-level objects satisfy the equilibrium conditions type by type. Thus any one-type equilibrium can be interpreted as a many-type equilibrium with no sorting.

Multiplicative technology ($\rho \to 1$). Now consider the multiplicative case

$$\phi(a, z) = z^{1-\omega_a} a^{\omega_a}, \qquad y_{ij} = z_{ij}^{1-\omega_a}\, \mathbb{E}_{g_{ij}}[a^{\omega_a}]\, \big(k_{ij}^{1-\gamma} h_{ij}^{\gamma}\big)^{\alpha}.$$

Fix $\{z_{ij}\}$ and consider a no-sorting allocation with $q_{ij}(a) \equiv q_{ij}$ and $g_{ij}(a) = f_{ja}(a)$ for all (i, j). At such an allocation,

$$\mathbb{E}_{g_{ij}}[a^{\omega_a}] = \overline{a^{\omega_a}} := \mathbb{E}_{f_{ja}}[a^{\omega_a}],$$

so the production function becomes

$$y_{ij} = \big(z_{ij}^{1-\omega_a}\, \overline{a^{\omega_a}}\big)\, \big(k_{ij}^{1-\gamma} h_{ij}^{\gamma}\big)^{\alpha}.$$

This is exactly the production structure of a one-type model with firm productivities

$$\widetilde{z}_{ij} := z_{ij}^{1-\omega_a}\,\overline{a^{\omega_a}}.$$

Take any steady-state equilibrium of this one-type model with productivities $\{\widetilde{z}_{ij}\}$, firm shares $\{q_{ij}\}$, wages $\{\widetilde{w}_{ij}\}$, and markdowns $\{\widetilde{\mu}_{ij}\}$. Evaluate the multiplicative many-type environment at the no-sorting allocation $q_{ij}(a) \equiv q_{ij}$ and $g_{ij}(a) = f_{ja}(a)$. Using the general marginal-product formula (11), the marginal product of type-a labor at firm (i,j) factorizes as

$$MPL_{ij}(a) = \overline{MPL}_{ij}\,\Psi(a),$$

where $\overline{MPL}_{ij}$ is the average marginal product in the one-type equilibrium, and

$$\Psi(a) = 1 - \frac{1}{\alpha\gamma}\left(1 - \frac{a^{\omega_a}}{\overline{a^{\omega_a}}}\right)$$

depends only on a, not on (i,j).

Define wages for each ability type by

$$w_{ij}(a) := \widetilde{\mu}_{ij}\,MPL_{ij}(a) = \Psi(a)\,\widetilde{\mu}_{ij}\,\overline{MPL}_{ij} = \Psi(a)\,\widetilde{w}_{ij}.$$

For any fixed a, wages are therefore equal to the one-type wage profile $\{\widetilde{w}_{ij}\}$ scaled by the common factor $\Psi(a)$. Under nested-CES labor supply, this common factor cancels from all within-type relative wages, so the type-a allocation across firms is unchanged:

$$q_{ij}(a) = q_{ij}.$$

Since the wage-setting condition $w_{ij}(a) = \widetilde{\mu}_{ij}MPL_{ij}(a)$ holds by construction, the allocation with $q_{ij}(a) \equiv q_{ij}$ and wages $\{w_{ij}(a)\}$ satisfies the equilibrium conditions in the multiplicative many-type environment. This yields a no-sorting equilibrium for the multiplicative technology and completes part (a).

Part (b).

Take any steady-state equilibrium under homogeneous labor ($\omega_a = 0$) with productivities $\{z_{ij}^{(0)}\}$, firm shares $\{q_{ij}\}$, wages $\{w_{ij}^{(0)}\}$, and markdowns $\{\mu_{ij}^{(0)}\}$. Interpret this as a no-sorting allocation in a many-type environment by setting

$$q_{ij}(a) \equiv q_{ij} \quad \Rightarrow \quad g_{ij}(a) = f_{ja}(a) \qquad \text{for all } a.$$

Now consider the multiplicative $\rho \to 1$ technology with firm productivities $\{z_{ij}^{(1)}\}$ chosen so that

$$z_{ij}^{(1)\,1-\omega_a}\,\overline{a^{\omega_a}} = z_{ij}^{(0)}.$$

At the no-sorting allocation just described, realized productivity satisfies

$$\mathbb{E}_{g_{ij}}[\phi(a, z_{ij}^{(1)})] = z_{ij}^{(1)\,1-\omega_a}\,\overline{a^{\omega_a}} = z_{ij}^{(0)},$$

so

$$y_{ij}^{(1)} = z_{ij}^{(0)}\big(k_{ij}^{1-\gamma}h_{ij}^{\gamma}\big)^{\alpha} = y_{ij}^{(0)}, \qquad \overline{MPL}_{ij}^{(1)} = \alpha\gamma\,\frac{y_{ij}^{(1)}}{h_{ij}} = \overline{MPL}_{ij}^{(0)}.$$

Using again (11), the marginal product of type a in the $\rho \to 1$ economy factorizes as

$$MPL_{ij}^{(1)}(a) = \overline{MPL}_{ij}^{(0)}\,\Psi(a),$$

with the same $\Psi(a)$ as above, common across firms. Define candidate wages by

$$w_{ij}^{(1)}(a) := \mu_{ij}^{(0)}\,MPL_{ij}^{(1)}(a) = \Psi(a)\,\mu_{ij}^{(0)}\,\overline{MPL}_{ij}^{(0)} = \Psi(a)\,w_{ij}^{(0)}.$$

Thus, for each ability a, wages are an ability-specific factor $\Psi(a)$ times the homogeneous-labor wage profile.

Because this scaling factor is common across firms, the nested-CES labor-supply system implies

$$q_{ij}^{(1)}(a) = q_{ij}.$$

Wage-bill shares for type a therefore coincide with their homogeneous-labor counterparts:

$$s_{ij}^{(1)}(a) = s_{ij}^{(0)}.$$

Hence the markdowns implied by Proposition 1 satisfy

$$\mu_{ij}^{(1)}(a) = \mu_{ij}^{(0)} \qquad \text{for all } (i, j, a).$$

Thus, with firm productivities $\{z_{ij}^{(1)}\}$ chosen as above, the homogeneous-labor equilibrium allocation with shares $\{q_{ij}\}$ and wedges $\{\mu_{ij}^{(0)}\}$ can be interpreted as a steady-state equilibrium of the multiplicative $\rho \to 1$ economy, with no sorting across firms and ability-specific wage levels determined by $\Psi(a)$. □

B Data

B.1 Main Data Source: INPS administrative dataset

B.1.1 Data sources and coverage

The main microdata source for the empirical analysis is the administrative worker–firm panel extracted from the archives of the Istituto Nazionale della Previdenza Sociale (INPS) and accessed through the VisitINPS Scholars program. The VisitINPS panel covers private-sector dependent

employment from 1974 to 2024 and records, for each administrative spell: (i) a harmonized employer identifier, (ii) contract start and end dates, (iii) full-time-equivalent (FTE) weeks of contribution, (iv) gross taxable earnings, (v) contract type and qualification codes, and (vi) municipality-level location for both employer and employee. Worker demographics (year of birth, sex, citizenship) and mortality outcomes are drawn from the INPS anagraphic registry and merged to the spell data. Commuting-zone identifiers are imputed using the official crosswalk.

B.1.2 Overview of the cleaning pipeline

Data processing follows four main steps. First, I ingest and harmonize the annual spell files year by year. Second, I standardize identifiers and string variables and construct person–spell and firm–year aggregates. Third, I define a consistent firm unit for analysis and merge enterprise-level balance-sheet information. Fourth, I construct and deflate pay measures, using FTE worked weeks as the main employment and exposure measure.

B.1.3 Key steps and choices

Spell-to-annual panel and job-year definition. The raw data are observed at the spell level. I convert them into an annual panel by constructing worker–firm–year observations based on FTE worked weeks. When a worker holds multiple spells with the *same* firm within a calendar year, I aggregate those spells by summing FTE worked weeks and taxable earnings and then compute the average annual wage for that worker–firm–year cell. When a worker is employed by *different* firms within the same year, I retain all worker–firm–year observations together with their corresponding FTE worked weeks. This preserves multi-firm careers and allows employment to be measured using FTE worked weeks.

Wage construction and deflation. Annual nominal compensation for each worker–firm–year observation is constructed as the sum of taxable wages reported to INPS over the year. All nominal values are deflated to 2022 euros using the total consumer price index from FRED, following Feenstra et al. (2015).

Firm definition. The INPS data provide two employer identifiers. The first is the enterprise identifier, which aggregates all establishments belonging to the same legal entity, even if they operate in different industries or commuting zones. The second is the *location–commodity sector* identifier, which uniquely identifies an industry–commuting-zone combination. In the empirical analysis, I define the firm as the location–commodity sector unit and treat the location–commodity sector identifier as the firm ID. When merging enterprise-level financials from CERVED, where only the enterprise identifier is available, I assign each enterprise to the location–commodity sector unit with the largest cumulative employment, as described in Appendix B.3. In specifications where local labor markets are defined by occupation, I further refine the firm identifier to the intersection of the location–commodity sector unit and the three-digit occupation code.

Occupation and qualification harmonization. Qualification codes are harmonized into a compact set of labor categories: manual workers, white-collar staff, managers, apprentices, and special categories. Observations with highly specific or non-comparable qualifications, such as aviation-specific categories or managerial positions outside the analysis sample, are excluded according to the rules detailed below. This harmonization improves comparability over time and aligns the occupational classification with the model's interpretation of worker types.

B.1.4 Sample exclusions and final panel

The core empirical sample retains private-sector dependent employees in for-profit entities, including corporations, partnerships, profit-oriented cooperatives, and sole proprietorships. I exclude public-sector employers, non-profit organizations, religious and educational institutions, managers (see note below), apprentices, and specialized categories such as aviation. I also drop observations with implausibly low weekly wages (below €50). Firm-level employment is measured as the headcount of selected job spells in each year.

B.2 ISCO occupation and education extract

B.2.1 Data description

A separate INPS extract provides occupation and education information coded to the International Standard Classification of Occupations (ISCO). Employers are required to report ISCO codes at contract initiation and whenever contracts are modified, starting in 2010. As a result, analyses using occupation or education variables rely on the movers subsample. The ISCO information is available at the spell level and is merged to the INPS microdata accordingly.

B.3 Italian CERVED balance-sheet data

B.3.1 Data description

Firm-level financial information comes from the CERVED database, which reports annual balance-sheet variables for the universe of incorporated Italian businesses from 1996 to 2018. Key variables include total assets, revenues, value added, wage bill, legal form, year of foundation, and firm status indicators (active, suspended, closed). Because CERVED is organized at the enterprise level whereas the analysis is conducted at the location–commodity sector level, I assign each enterprise to a single location–commodity sector by selecting the location–commodity identifier with the highest cumulative employment across years for that enterprise. This creates a unique mapping from enterprises to location–commodity units and assigns enterprise financials to the local labor market unit used in the analysis. The approximation is empirically tight: in 93% of enterprise–year observations, the enterprise is associated with a single location–commodity identifier. Firm financials are then merged to the INPS matched employer–employee microdata by year and the resulting firm identifier.

B.3.2 Constructing firm-level variables

After assigning each enterprise to its predominant location–commodity identifier, I construct firm-level variables by combining information from the INPS panel and CERVED balance sheets. When CERVED data are used, I measure the wage bill and employment from the INPS annual worker–firm panel by summing individual wages and FTE weeks worked at the enterprise level. This ensures comparability with value added and capital, which are also measured at the enterprise level. Value added is taken directly from CERVED, while the capital stock is proxied by the sum of tangible and intangible fixed assets reported on the balance sheet. Intermediate input expenditures are then recovered residually from the accounting identity defining value added as revenues minus intermediate inputs.

B.4 AKM estimation and construction of worker and firm types

This section describes how I recover empirical proxies for the latent worker and firm types, a and z, that appear in the model. Because these objects are not directly observed in administrative data, I use a two-way fixed-effects wage decomposition in the spirit of Abowd et al. (1999), Card et al. (2013), Song et al. (2019), and Bonhomme et al. (2022). The resulting worker and firm pay components serve as *indirect-inference* measures of underlying heterogeneity. I then show, using simulated data from the baseline calibration, that their within–local-labor-market rankings closely track the corresponding rankings of the structural types a and z. Let $\log w_{a,ij,t}$ denote the log real wage of worker a in firm ij and year t.

Clustering firms into latent pay types. Direct estimation of a firm fixed effect ψ_{ij} for every employer is difficult in the Italian INPS data because worker mobility is limited and many firms contribute few mobility links. To address this, I follow Bonhomme et al. (2022) and discretize employers into a finite number of latent firm types before estimating the AKM model. For each firm ij, I compute the 10th, 50th, and 90th percentiles of $\log w_{a,ij,t}$ across all workers employed at ij during the analysis window. I standardize these statistics and apply a K-means clustering algorithm with $K = 50$. Each employer ij is then assigned to a latent firm-type index $g(ij) \in \{1, \ldots, K\}$. The resulting firm-cluster pay premium serves as the empirical indirect proxy for the model's productivity or pay type z_{ij}.

Two-way fixed-effects estimation on firm clusters. Given the residualized log wages and the discretized employer labels, I estimate the additive wage model

$$\log w_{a,ij,t} = \alpha_a + \psi_{g(J(a,t))} + \epsilon_{a,ij,t}, \tag{B.1}$$

where α_a is a worker-specific component and $\psi_{g(J(a,t))}$ is the firm-type premium associated with cluster $g(J(a,t))$. Here $J(a,t)$ denotes the employer of worker a in year t.

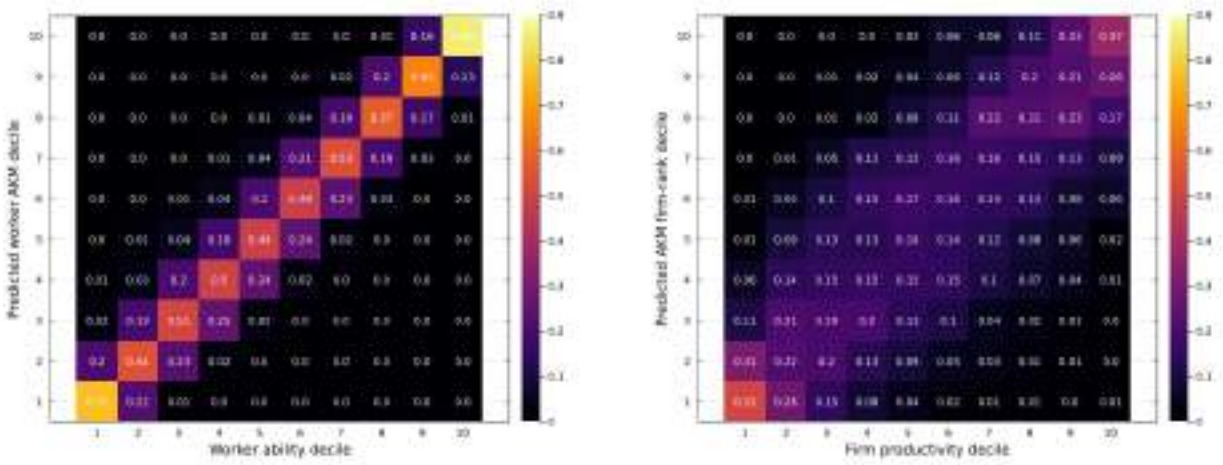
Most AKM applications include time-varying controls such as age or experience in order to isolate life-cycle wage profiles and interpret the remaining fixed effects as residual heterogeneity. Here, however, the goal is not to decompose wage inequality per se, but to recover worker and firm types that summarize persistent determinants of productivity and pay. For that purpose, it is natural to allow the worker effect α_a to absorb systematic age-related components of wages rather than partialing them out. This choice aligns the empirical types with the model's notion of ability: a worker is high type if she systematically commands higher productivity and wages, whether this reflects skill, experience, or other persistent attributes.

I therefore omit age and experience controls and estimate (B.1) by high-dimensional fixed-effects regression, absorbing worker and firm-type identifiers. A previous version of the paper (available upon request) reports specifications that additionally include time-varying observables; the resulting AKM-based moments are very similar.

Interpretation. The estimated worker effects $\hat{\alpha}_a$ provide a nonparametric measure of persistent worker heterogeneity and serve as empirical proxies for the structural worker types a. Similarly, the firm-type premia $\hat{\psi}_k$ summarize systematic pay differences across latent firm clusters. These pay differences are driven primarily by firm type and the competitive environment. It is therefore natural to interpret the within–local-labor-market ranking of $\hat{\psi}_k$ as an empirical proxy for the ranking of the model's firm heterogeneity z_{ij}. Although (B.1) is a linearized representation of the richer wage-setting environment in the model, it provides a tractable way to recover stable worker and firm components that closely track the structural types.

To illustrate this correspondence, Figures B.1a and B.1b report the distribution of AKM-based decile assignments conditional on the true structural deciles in a simulated panel generated from the model at the baseline calibration; details of the simulation appear in later sections. For each structural-type decile, I compute the share of workers or firms assigned to each AKM decile. The mass of these distributions lies heavily along the 45-degree diagonal: workers and firms in a given structural decile are most likely to be assigned to the same AKM decile. When misclassification occurs, it is almost entirely to adjacent deciles. This indicates that the AKM procedure provides a valid indirect proxy for the *ranking* of types within a local labor market. The alignment is sharper for workers than for firms.

This strong diagonal structure supports the interpretation of $\hat{\alpha}_a$ and $\hat{\psi}_k$ as model-consistent *indirect-inference proxies* for the latent worker and firm ranks (a, z_{ij}) used throughout the empirical analysis.



(a) Workers (b) Firms

Figure B.1: Share of latent-type deciles assigned to each AKM decile.

Notes: In each panel, the horizontal axis reports the decile of the underlying latent structural type in the simulated data, while the vertical axis reports the decile assigned by the AKM statistic. Each cell therefore shows, for a given true structural decile, the share of observations assigned to each AKM decile. Darker shading indicates a lower share of observations in that cell.

B.5 Additional descriptive statistics (period 2015–2019)

This subsection reports detailed descriptive statistics for the baseline period 2015–2019. Unless otherwise noted, the tables use the full cleaned worker–firm panel; whenever statistics are computed on the restricted AKM estimation sample, this is stated in the table notes. The tables cover contract types, working-time status, qualification categories, reasons for termination, the weekly wage distribution, legal-form composition and the retained for-profit sample, firm age, coverage of AKM fixed-effect measures, wage dispersion and variance decompositions, employment-size distributions, and selected AKM fixed-effect percentiles.

Employment contract, working-time status, and qualification (2015–2019). Table B.1 reports the distribution of contract types, working-time status, and broad qualification categories over 2015–2019. Permanent contracts account for about 69.2% of main-episode worker–year observations, fixed-term contracts for 27.4%, and seasonal contracts for 3.4%. Manual workers are the largest qualification group (63.9%), followed by white-collar employees (32.6%) and managers (3.5%). Managers are reported for completeness but excluded from the main analysis sample. Roughly two-thirds of main-episode jobs are full-time and one-third part-time.

Table B.1: Contract type, working-time status, and qualification (2015–2019)

Variable / category	Freq.	Percent
Contract type		
Permanent	58,728,827	69.17%
Fixed-term	23,255,215	27.39%
Seasonal	2,923,735	3.44%
Working-time status		
Full time	57,064,633	67.21%
Part time	27,842,762	32.79%
Qualification / occupational category		
Manual workers	54,251,901	63.90%
White-collar staff	27,695,352	32.62%
Managers & *quadro*	2,960,524	3.49%

Notes: Tabulations are based on the cleaned annual worker–firm panel for 2015–2019. Each observation is a worker–firm–year job spell, so workers may contribute multiple observations within a year if they hold multiple jobs. Employment and exposure are measured in full-time-equivalent (FTE) worked weeks in the underlying data, but the table reports unweighted counts of job spells. Contract type, working-time status, and qualification refer to spell characteristics. Managers are included here for completeness but excluded thereafter. Small differences in totals across panels reflect variable-specific missing values.

Weekly wage distribution (2015–2019). Table B.2 summarizes the distribution of real weekly wages over 2015–2019, in 2022 euros. The median weekly wage is about €486 and the mean about €575, indicating a moderately right-skewed distribution.

Table B.2: Weekly wage: mean and median (2015–2019)

Statistic	Value
Observations	76,248,512
Mean (weekly wage, €)	574.66
Median (weekly wage, €)	486.05

Notes: Weekly wages are computed as total annual earnings divided by total FTE weeks for each worker–firm–year spell and deflated to 2022 euros using the consumer price index described in the main text. The table reports unweighted summary statistics across all worker–firm–year observations with non-missing earnings and duration over 2015–2019.

Legal form (enterprise) and retained for-profit sample. Table B.3 reports enterprise legal-form frequencies from the CERVED–INPS linkage and identifies the subset retained for the baseline sample, namely private for-profit entities. Corporations are the dominant legal form (65.6% of enterprise–year observations), followed by partnerships (8.6%), for-profit cooperatives (9.1%), and

sole proprietorships or family businesses (9.8%). Applying the for-profit filter removes about 6.6 million enterprise–year observations from the raw panel.

Table B.3: Legal forms (grouped) and sample selection

Grouped legal form	Freq.	Percent (of full sample)
Corporations	63,105,413	65.65%
Partnerships	8,254,092	8.59%
Cooperatives (for-profit)	8,772,231	9.13%
Sole proprietorships / family business	9,415,906	9.80%
Other / public / non-profit	6,582,065	6.85%
Kept (private, for-profit forms)	89,547,642	93.15%

Notes: The table reports enterprise legal forms in the matched CERVED–INPS panel over 2015–2019 (96,129,707 enterprise–year observations). "Corporations," "Partnerships," "Cooperatives (for-profit)," and "Sole proprietorships / family business" correspond directly to the grouped categories in the raw legal-form classification. "Other / public / non-profit" aggregates non-profit cooperatives, consortia, public entities, religious entities, associations and foundations, educational and cultural entities, foreign entities, and residual legal forms and missings. The baseline sample retains only private for-profit entities—corporations, partnerships, for-profit cooperatives, and sole proprietorships / family businesses—yielding 89,547,642 enterprise–year observations; the remaining 6,582,065 are excluded.

Firm age (2015–2019). Table B.4 reports the distribution of firm age in the matched INPS–CERVED panel for the private for-profit sample. The median firm is 14 years old and the mean is 16.6 years, indicating substantial heterogeneity in firm longevity.

Table B.4: Firm age: mean and median (years, 2015–2019)

Statistic	Value
Observations (firm-years)	75,421,469
Mean firm age	16.65
Median firm age	14

Notes: Statistics are computed on firm–year observations in the matched INPS–CERVED panel over 2015–2019 after restricting to private for-profit enterprises as defined in Table B.3. Firm age is measured as years since the recorded year of foundation. Each firm–year receives equal weight.

Wage dispersion and variance decomposition (2015–2019). Table B.5 summarizes wage dispersion and decomposes the variance of log weekly wages into within- and between-firm components over 2015–2019. The standard deviation of log real weekly wages is 0.393; residualizing wages on a flexible Mincer specification reduces it to 0.319, and additionally absorbing occupation fixed effects reduces it further to 0.315. Using raw log wages, the within- and between-firm components are 0.063 and 0.091 when firms are defined at the employer (location–commodity) level. Defining

firms instead as occupation–employer cells lowers the within component to 0.044 and raises the between component to 0.112, while leaving total variance essentially unchanged at 0.154. Thus between-unit dispersion accounts for about 60% of wage variance under the employer definition and more than 70% under the occupation–firm definition.

Table B.5: Wage dispersion and variance decomposition (2015–2019)

Measure	**Within variance**	**Between variance**	**Total variance**
Raw log wage, firm as employer	0.0629	0.0911	0.1541
Raw log wage, firm as occupation–firm	0.0441	0.1120	0.1541

Notes: The table decomposes the variance of log real weekly wages into within-unit and between-unit components for 2015–2019. Wages are deflated to 2022 euros. "Firm as employer" uses the location–commodity sector identifier as the firm ID; "firm as occupation–firm" uses the intersection of that firm ID with the 3-digit occupation code. For each year, within- and between-unit components are computed from squared deviations of log wages around firm-year (or occupation–firm-year) means and then averaged over 2015–2019 using FTE worked weeks as frequency weights. The implied total variance of log weekly wages is 0.1541, corresponding to a standard deviation of 0.393. Residualized dispersion measures cited in the text are obtained by first regressing log wages on a flexible Mincer specification (age, tenure, education, and year dummies), with and without occupation fixed effects, and then computing the standard deviation of the residuals. These residual-based measures are not further decomposed in the table.

Employment size (firm and occupation–firm) and residualization. Table B.6 reports employment dispersion at the firm and occupation–firm levels over 2015–2019. Employment is measured in FTE workers constructed from weekly contribution records. At the employer level, the median firm employs about 1.3 FTE workers, the mean is 6.1, and the standard deviation of log employment is 1.50, indicating a highly skewed size distribution. At the occupation–firm level, the median cell employs 1 FTE worker and the mean is 2.2, with a standard deviation of log employment of 1.24. Residualizing log employment on flexible age profiles and sector–age interactions modestly reduces dispersion, especially at the firm level.

Table B.6: Employment distribution and dispersion (2015–2019)

	Firm-level employment	**Occupation–firm employment**
Median employment (FTE)	1.31	1.00
Mean employment (FTE)	6.07	2.19
SD(log employment)	1.50	1.24
SD(residual log employment)	1.34	1.18
Observations (firm- / occ–firm-years)	7,312,822	15,758,453

Notes: Employment is measured in FTE workers, defined as total weeks of contribution over the calendar year divided by 52. The first column uses the location–commodity sector identifier as the firm ID; the second uses the intersection of this firm ID with the 3-digit occupation code as the occupation–firm unit. For each year from 2015 to 2019, employment is aggregated to the relevant unit and the cross-sectional distribution is computed across units. Log employment is the natural logarithm of FTE employment. Residual log employment is obtained by regressing log employment on a flexible polynomial in firm age, year fixed effects, and detailed sector fixed effects, with sector-specific age profiles in the firm-level specification and occupation-by-sector age profiles in the occupation–firm specification. Reported means and standard deviations pool all years in 2015–2019.

Worker age and education (2015–2019). Table B.7 summarizes the age distribution and prevalence of tertiary education in the Italian matched employer–employee data over 2015–2019. At the worker level, the average age is about 42 and the distribution is fairly dispersed, with an interquartile range of roughly 34 to 50 years. About 21.7% of workers are college educated when worker–spell observations are averaged using FTE worked weeks as weights. From the firm perspective, the average workforce has a mean age of about 41, and the average share of college-educated workers among those with non-missing schooling is approximately 39%, with substantial cross-firm heterogeneity.

Table B.7: Worker age and education (2015–2019)

	Worker-level	**Firm-level**
Age		
Mean age (years)	42.29	40.68
Median age (years)	42	40.87
Observations	2,308,527,727	7,312,822
College education		
Mean college share	0.217	0.390
Median college share	0.120	0.267
Observations	1,444,174,396	1,195,743

Notes: Worker-level statistics are computed on the cleaned worker–firm spell panel for 2015–2019. Each observation is an employment spell, and moments are weighted by FTE worked weeks. Age refers to the worker's age in the reference year.
Firm-level statistics are computed on firm-year observations after collapsing spells to the employer (location–commodity sector) level. The mean and median college share report the firm-level share of college graduates among workers with non-missing education, measured in FTE weeks.

B.5.1 AKM fixed effects and covariance structure

For the industry-based AKM specification, the estimation sample comprises 17,228,756 observations and 2,130,582 firms. The share of observations with missing fixed effects is 1.58%. For the occupation–firm AKM specification, the corresponding sample comprises 15,003,751 workers and 4,985,173 occupation–firm units, with a missing-fixed-effect share of 1.62%.

In the industry-based AKM, $\mathrm{Var}(\mathrm{WFE}) = 0.0682$, $\mathrm{Var}(\mathrm{FFE}) = 0.0341$, and $\mathrm{Cov}(\mathrm{WFE}, \mathrm{FFE}) = 0.0182$. In the occupation–firm AKM, $\mathrm{Var}(\mathrm{WFE}) = 0.0428$, $\mathrm{Var}(\mathrm{FFE}) = 0.0557$, and $\mathrm{Cov}(\mathrm{WFE}, \mathrm{FFE}) = 0.0202$. These statistics are computed using employment weights.

B.6 German SIEED

This section elaborates on the Sample of Integrated Employer–Employee Data (SIEED) and the methodology applied to process this data. Data access was provided via remote access at the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB). A comprehensive description is available in Schmidtlein et al., 2020; Lochner et al., 2024. I base my cleaning procedure closely on the publicly available code from Card et al., 2013, who use a dataset from the same source.

The individual data points originate from labor administration records and social security data processing. The SIEED dataset encompasses every worker at a randomly selected sample of establishments, along with their complete employment histories, including periods in which they are employed outside the sample establishments. To ensure robust coverage of the labor market, I do not restrict the dataset to the panel establishments alone. The dataset provides variables

such as the worker's establishment, average daily wage, and an extensive set of characteristics including employment status, age, gender, tenure, occupation, and education. Throughout, I employ the 3-digit occupational classification according to the "Classification of Occupations 2010" (Klassifikation der Berufe 2010, KldB2010, Bundesagentur für Arbeit, 2011). The occupational title of the job performed by the employee during the notification period is part of the employment details submitted by the employer. If more than one job title with different classification codes applies for one employee, the employer is required to select the job title that best defines the main activity performed. Employment notifications with an end date earlier than 30 November 2011 are reported using the old occupation code 1988 (KldB 1988); the less detailed occupational subgroup is recorded by the first four digits of the code. The skill level required for a job, which is recorded in the fifth digit of the KldB2010 codes, is made available separately in the variable "level of requirement."

The employment biographies are provided in spell format, which I transform into an annual panel following the data processing described in Card et al., 2013. For individuals with multiple jobs within the same year, I select the job with the highest daily wage as the main episode. All nominal values are adjusted for inflation using the Consumer Price Index (2015 = 100). My sample selection criteria align with other studies that utilize this dataset or examine similar research topics. I focus on employees aged 20–60, employed in full-time positions in West Germany, who are liable to social security contributions. Part-time and marginal employment cases are excluded. In addition, I drop jobs with real daily earnings below 10 euros.

A well-known limitation of the German matched employer–employee data is the top-coding of the earnings variable at the social security system's contribution assessment limit ("Beitragsbemessungsgrenze"). To address this right-censoring issue, I follow the approach in Card et al., 2013. Specifically, I fit a set of Tobit models to log daily wages and impute uncensored values for censored observations using the estimated parameters and random draws from the corresponding conditional distribution. I estimate 16 Tobit models (across four age and four education groups) after applying the sample restrictions described above. Following Card et al., 2013, I include controls for age, firm size and its square, a dummy for firms with more than ten employees, the mean log wage of co-workers, and the share of co-workers with censored wages.

Crucially for the AKM-based analysis, the SIEED dataset also provides pre-estimated worker and firm fixed effects constructed by the IAB following the methodology in Card et al., 2013 on the full administrative sample. These AKM fixed effects are directly linked to the establishment identifiers in the data and are used as my empirical measures of worker and firm types for this dataset. For robustness, I also re-estimate an AKM model on my selected SIEED sample. However, as documented below, this exercise yields a substantial share of missing worker and firm effects due to limited mobility within the restricted sample. By contrast, the IAB-provided AKM effects are estimated on the full administrative universe, display essentially no missingness at the firm level, and therefore provide a more reliable and comparable measure of worker and firm types for my purposes.

B.6.1 German SIEED: Descriptive Statistics

This section summarizes key features of the processed SIEED dataset used in the empirical analysis over 2010–2017. I report worker- and firm-level characteristics, coverage and dispersion of fixed-effect measures, cross-sectional wage and employment dispersion, and the most common low- and high-skill occupations.

Table B.8 reports basic worker- and firm-level characteristics. Table B.10 summarizes the coverage of fixed-effect measures and the cross-sectional dispersion of worker and firm effects. Table B.9 reports wage dispersion, the variance decomposition of residual wages, and the dispersion of firm size. Table B.11 lists the five most frequent low-skill and high-skill occupations in the SIEED sample.

Overall, the descriptive statistics document a relatively mature and experienced workforce, substantial cross-sectional wage dispersion, and a large between-firm component in residual wage variance, motivating the focus on firm heterogeneity and sorting in the main analysis. These patterns are very similar to those documented for Italy, reinforcing the interpretation that firm-driven pay policies and worker sorting are central features of wage determination in both settings.

Table B.8: Worker- and firm-level characteristics (2010–2017)

	Firm-level	**Worker-level**
Mean log wage	4.5170	–
Mean years of schooling	12.0601	12.0396
Mean age	–	42.4551
Mean experience	–	23.2875

Notes: Statistics are computed on the processed SIEED sample for 2010–2017. Firm-level means are obtained by first averaging variables within establishments and then averaging across establishments. Worker-level means are computed on worker-year observations. Experience is measured as potential labor-market experience in years.

Table B.9: Wage and employment dispersion (2010–2017)

Statistic	**Value**
Wage dispersion	
SD (Mincer residual)	0.4150
SD (Mincer–occupation residual)	0.3677
SD (raw log wage)	0.4837
Variance decomposition of residual earnings	
Within-firm variance	0.0232
Between-firm variance	0.1122
Total variance	0.1354
Firm-size dispersion	
SD (log employment)	0.9926
SD (log employment residual)	0.9550

Notes: "Mincer residual" refers to log wages residualized on a flexible polynomial in age and tenure interacted with education and year fixed effects. "Mincer–occupation residual" additionally includes detailed occupation fixed effects. The variance decomposition is based on Mincer residuals: the within-firm component is the average squared deviation of residual wages from firm means, the between-firm component is the variance of firm means, and total variance is their sum. Employment is measured at the establishment level. Log-employment residuals are obtained by regressing log employment on a flexible polynomial in firm age and absorbing year and detailed sector fixed effects, including sector-specific age profiles. All statistics are computed for 2010–2017.

Table B.10: Coverage and dispersion of AKM fixed effects in Germany (2010–2017)

	FFE$_o$	**FFE**	**WFE$_o$**	**WFE**
Coverage				
Total units	3,473,220	3,473,220	2,525,434	2,525,434
Share missing fixed effects	1.68%	28.74%	3.43%	6.77%
Cross-sectional dispersion				
SD of fixed effects	0.2453	0.3609	0.3870	0.1994

Notes: FFE$_o$ and WFE$_o$ denote the firm and worker fixed effects provided by the IAB and estimated on the full German administrative universe following Card et al. (2013). FFE and WFE denote the occupation–establishment AKM fixed effects re-estimated on the SIEED sample. Coverage is reported as the total number of distinct firms or workers and the share with missing fixed effects over 2010–2017. Dispersion is measured as the standard deviation across non-missing fixed effects.

Table B.11: Top 5 low-skill and high-skill occupations by frequency (SIEED)

Low-skill occupations		**High-skill occupations**	
Occupation	Frequency	Occupation	Frequency
Machine-building and operating	1,635,278	Electrical engineering	592,561
Building construction	1,594,797	Technical research	741,231
Warehousing and logistics	2,627,747	Computer science	574,599
Drivers in road traffic	2,642,567	Purchasing and sales	594,187
Office clerks and secretaries	1,917,099	Business organization	922,970

Notes: Occupations are classified using KldB 2010. Low-skill and high-skill status is determined by the requirement level in the fifth digit of the KldB 2010 code. Frequencies are counts of worker-year observations in the SIEED sample over the full analysis period.

C Numerical Implementation and Validation

C.1 Numerical algorithm to solve the general equilibrium

This appendix summarizes the numerical algorithm used to compute the decentralized equilibrium and the efficient benchmark. The solver has three nested steps that mirror the model: (i) a *within-market* fixed point over firm–worker allocations for given type–market masses, (ii) *across-market* aggregation for given aggregate labor supply by ability, and (iii) an *outer* fixed point that updates aggregate participation by ability using the inverse labor-supply system.

Objects, inputs, and notation

Workers belong to discrete ability groups $a \in \mathcal{A}$ with population masses $f_a(a)$. Aggregate labor supply is summarized by a participation vector $S(a)$, so the economy-wide supply of type-a labor is

$$N(a) = S(a)\, f_a(a),$$

where $N(a)$ is the CES labor-supply index entering the inverse labor-supply system (2) from the main text. To economize on notation, I use $S(a)$, $N(a)$, $n_j(a)$, and $s_{ij}(a)$ both for scalar components and for the corresponding collections over abilities, markets, and firms; when needed, I make this explicit by writing, for example, $\{S(a)\}_{a\in\mathcal{A}}$ or $\{s_{ij}(a)\}_{i,j}$.

Given $N(a)$, workers choose across local labor markets according to the across-market CES structure with elasticity θ. Let S_{aj} denote the across-market *wage-bill share* of market j in type-a labor income. Equilibrium choices satisfy

$$S_{aj} = \left(\frac{w_j(a)}{W(a)}\right)^{1+\theta}, \qquad W(a) = \left(\int_0^1 w_j(a)^{1+\theta}\, dj\right)^{1/(1+\theta)},$$

where $w_j(a)$ is the market wage index and $W(a)$ the aggregate index for type a. The implied market-level CES index of type-a labor is

$$n_j(a) = N(a)\, S_{aj}^{\theta/(\theta+1)},$$

so that $N(a)$ and $\{n_j(a)\}_j$ satisfy the across-market CES aggregator.[2]

Inside each market j, there are m_j firms indexed by i, each with baseline productivity z_{ij}. Worker–firm productivity is given by $\phi(a, z_{ij})$. Let $s_{ij}(a)$ denote the *within-market* wage-bill share: the share of type-a wage income in market j paid by firm (i, j). For the numerical implementation, I augment the within-market share matrix with a final column capturing *non-employment within the market*. Thus, for each ability type a and market j, the vector $\{s_{ij}(a)\}_i$ includes all firms $i = 1, \ldots, m_j$ plus an additional non-employment entry. In equilibrium this share is zero, but along the fixed-point iterations it may be positive if, for a given market-level labor supply, all firms in that market deliver non-positive marginal products for a given type.

Model primitives—$\alpha, \gamma, \eta, \theta, \sigma, \varphi, R$—govern production, labor supply, and firm behavior. Given candidate matrices collecting $n_j(a)$ and $s_{ij}(a)$, the inner step solves the market-level fixed point; the middle step updates the across-market shares S_{aj}; and the outer step updates participation $S(a)$.

Inner step — within-market equilibrium

For each local labor market j, the inner step solves for firm hiring, wages, markdowns, and within-market wage-bill shares $s_{ij}(a)$, taking as given the market-level CES masses $n_j(a)$.

Starting from a candidate matrix of within-market shares $s_{ij}(a)$, the algorithm iterates on the following mapping:

1. **Within-market labor allocation.** For each type a and firm i in market j, the within-market CES structure implies

   $$n_{ij}(a) = n_j(a)\, s_{ij}(a)^{\eta/(\eta+1)},$$

   and total firm employment is

   $$h_{ij} = \sum_{a \in \mathcal{A}} n_{ij}(a).$$

   This is consistent with the within-market CES aggregator

   $$n_j(a) = \left[\sum_i n_{ij}(a)^{\frac{\eta+1}{\eta}} \right]^{\frac{\eta}{\eta+1}}.$$

[2] In the numerical implementation, the continuum of markets $j \in [0, 1]$ is approximated by a finite number M of markets, so integrals are replaced by sums.

2. **Realized productivity.** Given worker composition, realized productivity at firm (i, j) is

$$\Phi_{ij} = \sum_{a \in \mathcal{A}} \frac{n_{ij}(a)}{h_{ij}} \phi(a, z_{ij}).$$

3. **Marginal products.** Using the main text equation (11), the algorithm computes $MPL_{ij}(a)$ for each type at each firm. This schedule depends on both scale, through h_{ij}, and composition, through Φ_{ij}. The implementation imposes

$$MPL_{ij}(a) \geq 0$$

by truncating negative values at zero.

4. **Wage-setting via markdowns.** Wages satisfy

$$w_{ij}(a) = \begin{cases} \mu_{ij}(a)\, MPL_{ij}(a), & MPL_{ij}(a) > 0, \\ 0, & MPL_{ij}(a) \leq 0, \end{cases}$$

where

$$\epsilon_{ij}(a) = \left[\tfrac{1}{\eta} + \left(\tfrac{1}{\theta} - \tfrac{1}{\eta}\right) s_{ij}(a)\right]^{-1}, \qquad \mu_{ij}(a) = \frac{\epsilon_{ij}(a)}{\epsilon_{ij}(a) + 1}.$$

Here

$$s_{ij}(a) = \frac{w_{ij}(a) n_{ij}(a)}{\sum_k w_{kj}(a) n_{kj}(a)}$$

is the within-market wage-bill share. In the efficient benchmark, the algorithm imposes $\mu_{ij}(a) \equiv 1$ for all (i, j, a).

5. **Update wage-bill shares.** Given the updated wage schedule, within-market shares are updated as

$$s_{ij}^{\text{new}}(a) = \frac{\left(\mu_{ij}(a)\, MPL_{ij}(a)\right)^{1+\eta}}{\sum_k \left(\mu_{kj}(a)\, MPL_{kj}(a)\right)^{1+\eta}},$$

for all firms $i = 1, \ldots, m_j$. The residual non-employment share is

$$s_{0j}^{\text{new}}(a) = 1 - \sum_{i=1}^{m_j} s_{ij}^{\text{new}}(a).$$

Along the fixed point, if $MPL_{ij}(a) \leq 0$ for all firms in market j, the corresponding mass is assigned to the non-employment column; the across-market step then reallocates it toward markets with strictly positive wages.

6. **Iterate.** The algorithm updates shares using under-relaxation,

$$s_{ij}(a) \leftarrow \upsilon\, s_{ij}^{\text{new}}(a) + (1 - \upsilon)\, s_{ij}(a),$$

for $\upsilon \in (0,1)$, and repeats until

$$\max_{i,j,a} \left| s_{ij}^{\text{new}}(a) - s_{ij}(a) \right| < \text{tol}_S.$$

Given convergence, the algorithm also computes the wage schedule $w_{ij}(a)$, market wage indices

$$w_j(a) = \Big(\sum_i w_{ij}(a)^{1+\eta} \Big)^{1/(1+\eta)},$$

and firm profits π_{ij}.

Middle step — aggregation across markets

For a given participation vector $S(a)$ and implied aggregate indices $N(a) = S(a) f_a(a)$, workers choose across the M local markets according to the CES structure with elasticity θ.

1. **Across-market CES indices.** A candidate matrix of across-market shares S_{aj} implies market-level CES indices

   $$n_j(a) = N(a)\, S_{aj}^{\theta/(\theta+1)}.$$

   These are the inputs into the inner step.

2. **Solve each market.** For each j, the algorithm feeds the vector $n_j(a)$ into the inner step. This yields within-market allocations $n_{ij}(a)$, wage schedules $w_{ij}(a)$, market wage indices $w_j(a)$, and firm profits π_{ij}.

3. **Update across-market shares.** Given the market wage indices $w_j(a)$ and the aggregate index

   $$W(a) = \Big(\int_0^1 w_j(a)^{1+\theta}\, dj \Big)^{1/(1+\theta)},$$

   across-market shares are updated as

   $$S_{aj}^{\text{new}} = \left(\frac{w_j(a)}{W(a)} \right)^{1+\theta}, \qquad \sum_j S_{aj}^{\text{new}} = 1,$$

   for each ability type a. As in the inner step, the numerical implementation keeps a residual non-employment column, which is driven to zero in equilibrium.

4. **Iterate.** The algorithm updates

   $$S_{aj} \leftarrow \upsilon\, S_{aj}^{\text{new}} + (1-\upsilon)\, S_{aj},$$

   until

   $$\max_{a,j} \left| S_{aj}^{\text{new}} - S_{aj} \right| < \text{tol}_{S,A}.$$

This middle step maps firm-level outcomes into type–market allocations and market wage indices consistently with the across-market nested-CES structure.

Outer step — aggregate labor supply (general equilibrium)

The outer step ensures that aggregate labor supply by ability is consistent with the wage indices generated by the two inner steps.

1. **Aggregate supply.** Given a trial participation vector $S(a)$, compute

$$N(a) = S(a)\, f_a(a).$$

2. **Solve the inner and middle steps.** With $N(a)$ as input, run the across-market and within-market steps to obtain the equilibrium wage indices $W(a)$ for each ability type.

3. **Update participation.** The steady-state inverse labor-supply condition

$$\left(\frac{N(a)}{f_a(a)}\right)^{\frac{1}{\varphi}+\sigma} = W(a)^{\,1-\sigma}$$

implies

$$S(a) = \frac{N(a)}{f_a(a)} = W(a)^{\frac{(1-\sigma)\varphi}{1+\sigma\varphi}}.$$

The algorithm therefore updates

$$S^{\text{new}}(a) = W(a)^{\frac{(1-\sigma)\varphi}{1+\sigma\varphi}},$$

and applies under-relaxation:

$$S(a) \leftarrow \upsilon\, S^{\text{new}}(a) + (1-\upsilon)\, S(a).$$

4. **Iterate.** Repeat until

$$\max_a \left| S^{\text{new}}(a) - S(a) \right| < \text{tol}.$$

At convergence, the allocation satisfies firm optimality, markdown-based wage-setting, workers' nested-CES choices across firms and markets, and aggregate labor-supply behavior.

Outputs and numerical safeguards

The algorithm returns within-market wage-bill shares $s_{ij}(a)$, across-market shares S_{aj}, employment allocations $n_{ij}(a)$, firm employment h_{ij}, wages $w_{ij}(a)$, market wage indices $w_j(a)$, and aggregate indices $W(a)$ together with their market-level and aggregate counterparts.

Safeguards enforce non-negativity of marginal products, remove indeterminate values, and stabilize convergence through under-relaxation and small positive lower bounds on shares. The efficient benchmark is obtained by imposing $\mu_{ij}(a) \equiv 1$ throughout; the fixed-point architecture is otherwise unchanged.

C.2 Numerical Verification via the Dual Firm Problem in $(\mathbb{E}_{g(a)}[\phi(a,z)], h)$

This appendix explains how I reduce each firm's high-dimensional employment problem to a two-dimensional dual problem in $(\mathbb{E}_{g(a)}[\phi(a,z)], h)$ and how this dual structure is used to implement a global-deviation check. The exact dual problem is written in terms of the conditional cost function $C(\Phi, h)$, whose minimizer is characterized by KKT conditions because the inner cost-minimization problem is convex. For each candidate pair (Φ, h), the KKT system implies an employment vector; the candidate is retained only if that vector reproduces the same (Φ, h). For points that fail this composition-consistency condition, the allocation generated by the KKT mapping does not deliver the candidate pair (Φ, h). The associated profit is therefore not an evaluation of the dual objective $R(\Phi, h) - C(\Phi, h)$ at that candidate pair. They are recorded only as auxiliary diagnostics for the fixed-point search.

Technology, Aggregates, and the Dual Firm Problem

For the numerical implementation, ability is discretized into a finite grid $\mathcal{A}$, and a firm of type z chooses employment

$$\{n(a)\}_{a\in\mathcal{A}}, \qquad n(a) \geq 0.$$

Define total employment, composition, and the composition-dependent productivity index as

$$h \equiv \sum_{a\in\mathcal{A}} n(a), \qquad g(a) \equiv \frac{n(a)}{h}, \qquad \Phi \equiv \sum_{a\in\mathcal{A}} \phi(a,z)\, g(a) = \frac{1}{h}\sum_{a\in\mathcal{A}} \phi(a,z)\, n(a).$$

After eliminating capital using the firm's FOC, net revenue from production can be written as

$$R(\Phi, h) = \Xi\, \Phi^{\frac{1}{1-\alpha(1-\gamma)}}\, h^{\frac{\alpha\gamma}{1-\alpha(1-\gamma)}}, \qquad 0 < \alpha < 1,\ 0 < \gamma < 1,\ \Xi > 0. \tag{C.1}$$

Given $\{n(a)\}$, the induced aggregates are $(\Phi(\{n\}), h(\{n\}))$. The original static firm problem can therefore be written as

$$\max_{\{n(a)\geq 0\}} \left\{ R(\Phi(\{n\}), h(\{n\})) - \sum_{a\in\mathcal{A}} c_a(n(a)) \right\}, \tag{C.2}$$

where

$$c_a(n) := w_a(n)\, n$$

is the type-a wage-bill schedule implied by the nested-CES labor-supply system, holding the en-

vironment fixed. Because $R(\Phi, h)$ is generally not concave in (Φ, h), the objective in (C.2) is not concave in $\{n(a)\}$, so first-order conditions are necessary but not sufficient. With hundreds of types, directly checking global optimality in $\{n(a)\}$ is infeasible.

Dual cost function. For any fixed aggregate pair (Φ, h), define the dual cost function

$$\begin{aligned} C(\Phi, h) \equiv \min_{\{n(a)\geq 0\}} \quad & \sum_{a\in\mathcal{A}} c_a(n(a)) \\ \text{s.t.} \quad & \sum_{a\in\mathcal{A}} n(a) = h, \\ & \sum_{a\in\mathcal{A}} \phi(a, z)\, n(a) = \Phi\, h. \end{aligned} \tag{C.3}$$

The objective is strictly convex in $\{n(a)\}$ (Lemma C.5 below), and the constraints are linear. Whenever (Φ, h) is technologically feasible, this inner problem has:

- a *unique* minimizing vector $\{n^*(a \mid \Phi, h)\}$;
- a well-defined value $C(\Phi, h)$ that is continuous in (Φ, h) on the feasible set.

Let

$$\mathcal{F} \equiv \left\{(\Phi, h) :\ \exists\{n(a) \geq 0\} \text{ with } \sum_a n(a) = h,\ \sum_a \phi(a, z)n(a) = \Phi h\right\}$$

denote the set of feasible aggregate pairs. By construction of $C(\Phi, h)$, for any feasible $\{n(a)\}$ with induced (Φ, h),

$$C(\Phi, h) \leq \sum_a c_a(n(a)).$$

Therefore,

$$R(\Phi, h) - \sum_a c_a(n(a)) \leq R(\Phi, h) - C(\Phi, h).$$

Equality is attained when $\{n(a)\}$ solves the inner problem (C.3) at (Φ, h). It follows that the original firm problem (C.2) is exactly equivalent, in value, to the two-dimensional dual problem

$$\max_{(\Phi,h)\in\mathcal{F}} \{R(\Phi, h) - C(\Phi, h)\}. \tag{C.4}$$

Given an optimal $(\Phi^\star, h^\star)$ for (C.4), the associated employment vector is uniquely recovered as $\{n^*(a \mid \Phi^\star, h^\star)\}$.

KKT characterization at fixed (Φ, h). For fixed (Φ, h), revenue is constant, so the firm's problem reduces to choosing the least-cost composition of workers that delivers those aggregates. This is exactly the conditional cost problem (C.3). Its objective is strictly convex and its constraints are linear, so it has a unique minimizer.

The key point is that, at a stationary point of the dual problem, the first-order conditions of the conditional cost problem coincide with the firm's full KKT conditions. The connection between the conditional cost problem and the firm's KKT conditions is as follows. Let $\phi_a \equiv \phi(a, z)$. Since

$$h = \sum_a n(a), \qquad \Phi = \frac{\sum_a \phi_a n(a)}{h},$$

the derivative of revenue with respect to employment of type a can be written as

$$\frac{\partial R(\Phi, h)}{\partial n(a)} = R_h(\Phi, h) + R_\Phi(\Phi, h)\frac{\phi_a - \Phi}{h}.$$

Thus the full-firm KKT condition for an interior type is

$$c_a'(n(a)) = R_h(\Phi, h) + R_\Phi(\Phi, h)\frac{\phi_a - \Phi}{h},$$

with the usual complementary-slackness inequalities for types with $n(a) = 0$. This condition is affine in ϕ_a, exactly as in the KKT conditions of the conditional cost problem (C.3), whose interior conditions are

$$c_a'(n(a)) = \lambda_h + \lambda_\Phi \phi_a.$$

At a stationary point of the dual problem, the cost-minimization multipliers satisfy

$$\lambda_\Phi = \frac{R_\Phi(\Phi, h)}{h}, \qquad \lambda_h = R_h(\Phi, h) - \frac{\Phi}{h} R_\Phi(\Phi, h).$$

Hence a composition-consistent solution of the firm's KKT system is exactly the cost-minimizing composition associated with that aggregate pair, and it is a stationary candidate of the original high-dimensional firm problem.

Strict Convexity of the Wage Bill

The key property needed for the dual formulation is that the wage bill is strictly convex in the firm's own employment vector.

Lemma C.5 (Strict convexity of the wage bill)**.** *Fix other firms' wages and quantities and consider a single firm ij. Let*

$$C_{ij}(n_{ij}) = \sum_{a \in \mathcal{A}} w_{ij}(a; n_{ij}(a))\, n_{ij}(a)$$

denote its wage bill as a function of its own employment vector $n_{ij}(\cdot)$, holding the environment fixed. Under the nested-CES structure with $\eta > \theta$, $C_{ij}(n_{ij})$ is strictly convex in $n_{ij}(\cdot)$.

Proof. Fix an ability type a and suppress indices ij, a for clarity. Consider the one-dimensional function

$$f(n) := w(n)\, n,$$

where $w(n) \equiv w_{ij}(a;n)$ is the firm's wage for type a as a function of its own employment $n := n_{ij}(a)$, holding the environment fixed. Let

$$\varepsilon(n) := \left(\frac{\partial \log w(n)}{\partial \log n} \right)^{-1}$$

denote the firm-specific elasticity of labor supply; by construction $\varepsilon(n) > 0$.

From the definition of $\varepsilon(n)$,

$$\frac{\partial \log w}{\partial \log n} = \frac{n}{w}\frac{\partial w}{\partial n} = \frac{1}{\varepsilon(n)} \quad \Rightarrow \quad \frac{\partial w}{\partial n} = \frac{w}{n} \cdot \frac{1}{\varepsilon(n)}.$$

The first derivative of $f(n)$ is

$$f'(n) = \frac{\partial}{\partial n}[w(n)n] = w(n) + \frac{\partial w}{\partial n} n = w(n)\Big(1 + \frac{1}{\varepsilon(n)}\Big) > 0.$$

Differentiating again,

$$f''(n) = w'(n)\Big(1 + \frac{1}{\varepsilon(n)}\Big) + w(n)\frac{\partial}{\partial n}\Big(\frac{1}{\varepsilon(n)}\Big).$$

Using $w'(n) = [w/(n\varepsilon(n))]$ and

$$\frac{\partial}{\partial n}\Big(\frac{1}{\varepsilon(n)}\Big) = -\frac{1}{\varepsilon(n)^2}\,\frac{\partial \varepsilon(n)}{\partial n},$$

we obtain

$$f''(n) = \underbrace{\frac{w(n)}{n\varepsilon(n)}\Big(1 + \frac{1}{\varepsilon(n)}\Big)}_{>0} \; - \; \underbrace{w(n)\,\frac{1}{\varepsilon(n)^2}\,\frac{\partial \varepsilon(n)}{\partial n}}_{\text{sign depends on } \partial\varepsilon/\partial n}.$$

Under the nested-CES structure,

$$\varepsilon(n) = \left[\frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big)s(n)\right]^{-1},$$

with $s(n)$ weakly increasing in own employment n. Since $\eta > \theta$, the bracket is increasing in $s(n)$, so $\varepsilon(n)$ is weakly decreasing in n, i.e. $\partial\varepsilon(n)/\partial n \le 0$. It follows that

$$-\,w(n)\,\frac{1}{\varepsilon(n)^2}\,\frac{\partial \varepsilon(n)}{\partial n} \ge 0.$$

Hence the first term in $f''(n)$ is strictly positive, the second weakly, implying $f''(n) > 0$ for all $n > 0$. Thus $f(n) = w(n)n$ is strictly convex in own employment n for each type a.

Finally, the wage bill is a finite sum of strictly convex, separable functions,

$$C_{ij}(n_{ij}) = \sum_{a \in \mathcal{A}} f_a(n_{ij}(a)),$$

so $C_{ij}(n_{ij})$ is strictly convex in the vector $n_{ij}(\cdot)$. □

Lemma C.5 implies that the inner problem (C.3) has a unique minimizer for each feasible (Φ, h). In what follows, I exploit this uniqueness and the dual parameterization to organize a search over KKT-consistent deviations.

Monopsony Benchmark: Analytical KKT Mapping and Composition Consistency

I start with the monopsony benchmark, to build intuition in a simpler environment. In a monopsony benchmark, residual supply is isoelastic $w(a) = k(a)n(a)^{1/\eta}$ and the markdown is constant $w(a) = \mu\,\text{MPL}(a)$. In that case, the firm's first-order conditions imply a closed-form mapping from any candidate (Φ, h) to a KKT-consistent employment vector $\{n(a \mid \Phi, h)\}$.

Given a candidate (Φ, h):

1. Compute $\text{MPL}(a \mid \Phi, h, z)$ as in main text equation (11).

2. Use the constant-markdown Lerner condition to obtain wages:

$$w(a \mid \Phi, h) = \mu\,\text{MPL}(a \mid \Phi, h, z).$$

3. Use inverse labor supply to recover a unique KKT-consistent employment vector $n(a \mid \Phi, h)$:

$$n(a \mid \Phi, h) = \left(\frac{w(a \mid \Phi, h)}{k(a)}\right)^{\eta}.$$

4. Compute the implied aggregates

$$h'(\Phi, h) = \sum_a n(a \mid \Phi, h), \qquad \Phi'(\Phi, h) = \frac{1}{h'(\Phi, h)} \sum_a \phi(a, z)\, n(a \mid \Phi, h).$$

5. If $(h'(\Phi, h), \Phi'(\Phi, h)) \approx (h, \Phi)$ (composition-consistent), then $\{n(a \mid \Phi, h)\}$ induces the candidate aggregates and satisfies the firm's KKT conditions at (Φ, h). Since the wage-bill objective is strictly convex and the aggregate constraints are linear, this vector is also the unique solution to the conditional cost problem (C.3) for that pair. Otherwise, the candidate pair is not feasible as a stationary point of the original firm problem and is discarded.

For each composition-consistent (Φ, h), I compute profits as

$$\pi(\Phi, h) = R(\Phi, h) - \sum_a w(a \mid \Phi, h)\, n(a \mid \Phi, h).$$

A global deviation check then consists of searching over a grid of candidate pairs (Φ, h), retaining composition-consistent points as the main admissible deviations, and comparing their implied

profits $\pi(\Phi, h)$ to the equilibrium value. Any interior global maximizer of the original firm problem must satisfy the firm's KKT conditions. Therefore, the global-deviation check searches over composition-consistent KKT candidates in (Φ, h) space and compares their profits to the equilibrium value. Composition-inconsistent points are excluded from the admissible set. Their implied profits are recorded as an auxiliary diagnostic, and to diagnose whether the grid or fixed-point procedure is missing suspicious regions.

Oligopsony: Numerical Inner Problem via Wage Fixed Point

In the full nested-CES oligopsony environment, wages satisfy

$$w_{ij}(a) = \mu_{ij}(a)\,\mathrm{MPL}_{ij}(a), \qquad \mu_{ij}(a) = \frac{\epsilon_{ij}(a)}{\epsilon_{ij}(a)+1},$$

with

$$\epsilon_{ij}(a) = \left[\frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta}\right) s_{ij}(a)\right]^{-1}, \qquad s_{ij}(a) = \frac{w_{ij}(a) n_{ij}(a)}{\sum_{i' \in \mathcal{S}_j(a)} w_{i'j}(a) n_{i'j}(a)}.$$

At the same time, nested-CES labor supply relates wages and employment via

$$w_{ij}(a) = \left(\frac{n_{ij}(a)}{n_j(a)}\right)^{1/\eta} \left(\frac{n_j(a)}{N(a)}\right)^{1/\theta} W(a),$$

where $N(a)$ and $W(a)$ are fixed since the firm is infinitesimal with respect to the aggregate economy, and $n_j(a)$ is determined by the environment and the deviating firm's own $n_{ij}(a)$.

In this case, the inner problem (C.3) cannot be solved in closed form. Instead, I compute, for each candidate (Φ_{ij}, h_{ij}), a KKT-consistent pair $\{n_{ij}(a), w_{ij}(a)\}$ by solving the firm's wage and employment FOCs via a fixed-point algorithm.

Inner problem for given (Φ_{ij}, h_{ij}). Fix a candidate deviation (Φ_{ij}, h_{ij}) for firm ij and hold competitors' wage schedules at their equilibrium values. Given the resulting wage system, reconstruct employment using the nested-CES labor-supply equations. Proceed as follows:

1. Compute $\mathrm{MPL}_{ij}(a \mid \Phi_{ij}, h_{ij}, z_{ij})$ using main text equation (11); this depends on the candidate pair only through (Φ_{ij}, h_{ij}).
2. Initialize the deviating firm's wages at their equilibrium values,

$$w_{ij}^{(0)}(a) = w_{ij}^{\mathrm{eq}}(a), \qquad \forall a.$$

3. For each iteration t:
   - Form

$$w_{-i,j}(a) = \sum_{i' \neq i} w_{i'j}(a)^{1+\eta}, \qquad s_{ij}^{(t)}(a) = \frac{\left(w_{ij}^{(t)}(a)\right)^{1+\eta}}{w_{-i,j}(a) + \left(w_{ij}^{(t)}(a)\right)^{1+\eta}}.$$

- Compute elasticities and markdowns

$$\epsilon_{ij}^{(t)}(a) = \left[\frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta}\right) s_{ij}^{(t)}(a)\right]^{-1}, \qquad \mu_{ij}^{(t)}(a) = \frac{\epsilon_{ij}^{(t)}(a)}{1 + \epsilon_{ij}^{(t)}(a)}.$$

- Update wages using the Lerner condition,

$$\widetilde{w}_{ij}^{(t)}(a) = \mu_{ij}^{(t)}(a)\,\mathrm{MPL}_{ij}(a \mid \Phi_{ij}, h_{ij}, z_{ij}).$$

- Apply relaxation:

$$w_{ij}^{(t+1)}(a) = \lambda\,\widetilde{w}_{ij}^{(t)}(a) + (1-\lambda)\,w_{ij}^{(t)}(a), \qquad \lambda \in (0,1].$$

4. Iterate until convergence,

$$\max_a \left|w_{ij}^{(t+1)}(a) - w_{ij}^{(t)}(a)\right| < \mathrm{tol}_{\mathrm{FP}}.$$

5. Given the converged wages $w_{ij}(a)$, reconstruct employment in a way that mirrors the equilibrium solver. For each ability type a, first compute the market wage index

$$w_j(a) = \Big(\sum_{i'} w_{i'j}(a)^{1+\eta}\Big)^{1/(1+\eta)},$$

and then recover market employment

$$n_j(a) = N(a)\Big(\frac{w_j(a)}{W(a)}\Big)^{\theta}.$$

Next, construct within-market shares

$$S_{ij}(a) = \frac{w_{ij}(a)^{1+\eta}}{\sum_{i'} w_{i'j}(a)^{1+\eta}}.$$

To maintain consistency with the equilibrium solver, I apply the same numerical sparsity rule used in the fixed-point algorithm, setting sufficiently small shares to zero before mapping them into employment. Firm-level employment is then recovered as

$$n_{ij}(a) = S_{ij}(a)^{\eta/(1+\eta)}\, n_j(a).$$

6. Compute implied aggregates

$$h'_{ij} = \sum_a n_{ij}(a), \qquad \Phi'_{ij} = \frac{1}{h'_{ij}} \sum_a \phi(a, z_{ij})\, n_{ij}(a).$$

7. If (h'_{ij}, Φ'_{ij}) is not sufficiently close to (h_{ij}, Φ_{ij}), as measured by the numerical tolerances used in the algorithm, I classify the candidate (Φ_{ij}, h_{ij}) as composition-inconsistent and exclude it from the main admissible-deviation set. I nevertheless record the profit implied by the associated allocation as an auxiliary diagnostic.

8. Otherwise, the converged $\{n_{ij}(a), w_{ij}(a)\}$ are treated as the KKT-consistent allocation associated with this candidate pair, and the implied allocation is used to evaluate profits using the same profit formula as in the equilibrium solver.

Global deviation check. The global deviation check then proceeds as in the monopsony case, but with the inner problem solved numerically:

1. For each firm, compute the equilibrium employment vector $\{n_{ij}^{\text{eq}}(a)\}$ implied by main text Proposition 1, the induced aggregates $(\Phi_{ij}^{\text{eq}}, h_{ij}^{\text{eq}})$, and equilibrium profits π_{ij}^{eq}.

2. Choose a grid of candidate pairs (Φ_{ij}, h_{ij}) covering the feasible range around and beyond the equilibrium point, including boundary regions. In practice, this search is implemented adaptively: I begin with a coarse grid around the equilibrium point, identify the most suspicious cases based on implied gains and boundary solutions, and then refine the grid locally around those points.

3. At each grid point, solve the inner problem as described above: use the wage fixed point and the composition-consistency filter to recover the implied allocation $(\{n_{ij}(a)\}, \{w_{ij}(a)\})$ whenever the algorithm converges. For each converged candidate, compute the implied aggregates
$$h'_{ij} = \sum_a n_{ij}(a), \qquad \Phi'_{ij} = \frac{1}{h'_{ij}} \sum_a \phi(a, z_{ij})\, n_{ij}(a),$$
as well as implied profits.

4. If (h'_{ij}, Φ'_{ij}) is sufficiently close to (h_{ij}, Φ_{ij}), the candidate is classified as composition-consistent and is included in the main admissible-deviation set. If not, it is classified as composition-inconsistent and excluded from the main set, although the associated implied profit is recorded as an auxiliary diagnostic.

5. If, after the adaptive grid-refinement procedure, no composition-consistent candidate yields profits above π_{ij}^{eq} up to the profit tolerance used in the numerical procedure, I treat the equilibrium allocation as numerically verified against unilateral deviations in $\{n_{ij}(a)\}$.

In other words, the dual problem (C.4) provides the conceptual foundation, and the wage fixed-point plus composition-consistency algorithm delivers a numerical search over candidate unilateral deviations parameterized by (Φ, h). The main admissible-deviation test is conducted on composition-consistent points, while implied profits at composition-inconsistent points are used only as an auxiliary robustness check.

Implementation. The numerical sufficiency test at the baseline calibration yields no evidence of economically meaningful profitable unilateral deviations. In the coarse pass, only 1 out of 13,911 firms has no composition-consistent candidate, and no firm exhibits a profitable consistent or inconsistent-implied deviation above the chosen profit tolerance. I then refine the search adaptively around the 300 most suspicious firms using denser local grids. In the final refinement, all consistent gains remain below 10^{-6}, while even the inconsistent-implied gains remain below 10^{-5}, and the candidate points associated with the largest consistent gains remain extremely close to the corresponding equilibrium values of (Φ, h). Thus, both the implied profit gains and the associated deviations in allocation are numerically negligible, providing no evidence against sufficiency of the computed equilibrium.

D Empirical Facts: Robustness

This appendix replicates, where possible, the main empirical facts documented in Section 2 in the main text under a set of robustness exercises. I proceed in two steps. First, as an external-validity exercise, I replicate the facts on market shares and hiring thresholds for Germany. Second, I replicate the analysis including full-time workers under their main job spell in a given year.

Recap of empirical strategy. Across all three empirical facts, I use a common empirical strategy. I first rank workers and firms into *within-market* deciles of their AKM fixed effects. Using these decile rankings, I construct employment-share matrices (Fact 1) by taking a worker-side perspective and computing, for each worker decile, the distribution of employment across firm deciles. If high-paying firms simply employed more of every worker type, each worker decile would allocate more employment to higher firm ranks. Instead, main text Section 2.2.1 shows strong within-market segmentation, with lower-ranked workers disproportionately employed at lower-ranked firms; this pattern remains robust throughout this appendix.

I then construct firm hiring thresholds as the minimum worker fixed effect among new hires and relate these thresholds to alternative measures of firm rank, controlling for log new hires and market and year fixed effects (Fact 2). Higher-ranked firms systematically exhibit higher hiring thresholds, and this relationship is also robust. Finally, I compute wage-bill HHI indices by worker decile within local labor markets and aggregate them using employment weights to obtain an economy-wide concentration profile across the worker fixed-effect distribution (Fact 3). The resulting HHI profile is U-shaped and highest for top-ranked workers, a pattern likewise confirmed in the robustness exercises.

D.1 Market shares: robustness to ranking firms by average log wages

This appendix verifies that the segregation pattern documented in the main text does not depend on ranking firms by their AKM firm fixed effect. Instead, I rank firms within each local labor market and year by their average log wage and recompute the worker–firm employment matrix

exactly as in Section 2.2.1 in the main text. Workers continue to be ranked by their AKM worker fixed effect within local labor market and year, and for each worker-decile bin I compute the share of employment allocated to each firm-decile bin. As in the baseline specification, deciles are assigned separately within each local labor market and year, the resulting matrices are averaged over 2015–2019 using employment weights, and the sample is restricted to local labor markets with at least 100 workers and 100 firms.

Figure D.1 shows that the resulting allocation is very similar to the baseline AKM-based classification of Figure 3a in the main text. High-ranked workers remain strongly concentrated in high-ranked firms: top-decile workers allocate about 40% of their employment to top-decile firms and only about 4% to bottom-decile firms. By contrast, the employment distribution of lower-ranked workers remains more diffuse across firm ranks, although it still tilts toward lower-ranked firms. Bottom-decile workers allocate about 15% of their employment to bottom-decile firms and about 10 to 13% to top-decile firms. Relative to the baseline AKM-based ranking, the average-log-wage classification therefore delivers a quantitatively very similar segregation pattern, with somewhat stronger concentration in the upper tail.

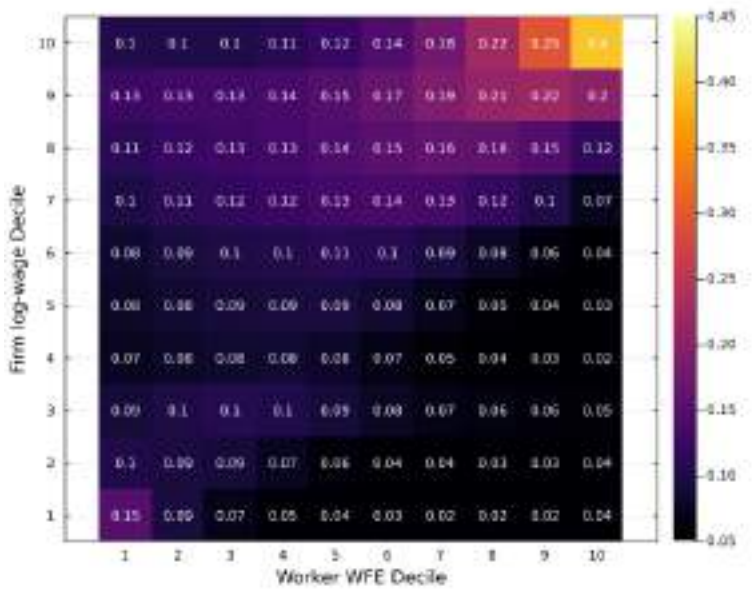


Figure D.1: Employment shares across worker fixed-effect deciles when firms are ranked by average log wages

Notes: The figure reports, for Italy, the distribution of employment shares across firm deciles for each worker fixed-effect decile. Workers are ranked by their AKM worker fixed effect within local labor market and year. Local labor markets are defined by industry–commuting zones. Firms are ranked by average log wage within local labor market and year. Decile assignments are performed separately within each local labor market and year, and the resulting matrix is averaged over 2015–2019 using employment weights. The sample is restricted to local labor markets with at least 100 workers and 100 firms.

D.2 Germany

As an external-validity exercise, I replicate the analysis using German administrative data from the IAB, described in the Data Appendix. The IAB provides AKM worker and firm fixed effects constructed on the full administrative sample following Card et al. (2013). These firm effects are defined at the establishment level and therefore do not require any additional firm-clustering procedure. Because the German data available to me are a subsample rather than the full population, I take these AKM fixed effects as given and do not re-estimate them. When I define firm types at the firm–occupation level, I assign the establishment-level AKM firm fixed effect to all occupations within the same establishment.

Because the German data are a subsample, I also relax the sample-size restriction on local labor markets. Instead of requiring markets to contain more than 100 workers and firms, I retain markets with more than 10, so that decile classifications remain meaningful while preserving a sufficiently large number of markets. I define local labor markets at the three-digit occupation level and replicate, for the 2010–2017 sample window, the two core empirical exercises from the Italian analysis: (i) construction of market-share matrices documenting within-market segmentation between worker and firm types, and (ii) estimation of hiring-threshold regressions relating firms' minimum worker fixed effects among new hires to alternative measures of firm rank. Because the data do not cover the full firm population, I do not construct or report wage-bill concentration (HHI) indices for Germany.

D.2.1 Market Shares

I first replicate the market-share analysis for Germany using the establishment-level AKM fixed effects provided by the IAB. Local labor markets are defined at the three-digit occupation level, and firm types at the establishment–occupation level, as described above. Within each local labor market and year, I rank workers and firms into deciles of their AKM fixed effects and, from the worker side, compute for each worker decile the distribution of employment across firm deciles. To preserve a sufficiently large number of markets in the German subsample, I restrict attention to markets with at least 10 workers and 10 firms.

Figure D.2 reports the resulting employment-share matrix. Segmentation is pronounced and closely parallels the Italian results. Workers in the bottom AKM decile allocate about 18.2% of their employment to bottom-decile firms and only 7.6% to top-decile firms, whereas workers in the top decile allocate about 17.4% of their employment to top-decile firms and only 7.4% to bottom-decile firms. Middle-ranked workers display a more diffuse allocation. Overall, the pattern is one of strong top–top and bottom–bottom clustering, indicating that labor-market segmentation is also a robust feature of the German data.

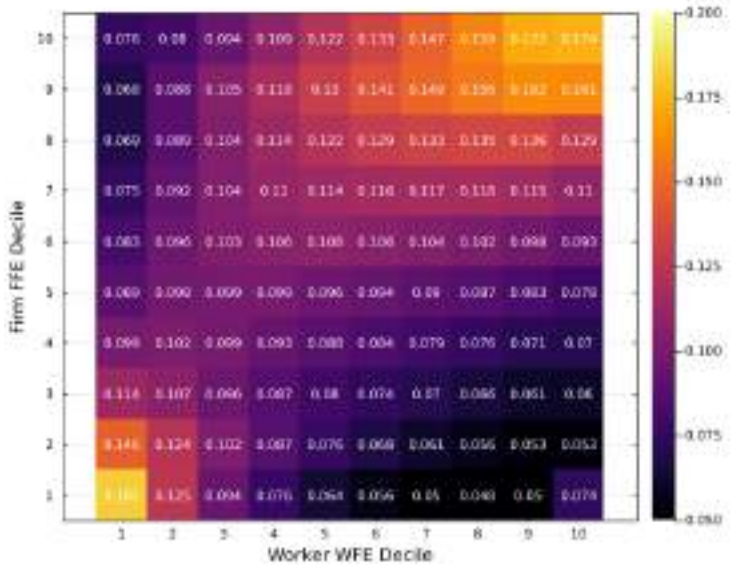


Figure D.2: Employment market shares across worker and firm AKM deciles (occupation-based local labor markets, Germany)

Notes: The figure reports, for each worker AKM fixed-effect decile, the distribution of employment shares across firm AKM fixed-effect deciles in Germany. Workers and firms are ranked into deciles within each local labor market and year, where local labor markets are defined by three-digit occupation. Firm types are defined at the establishment–occupation level. The matrix is averaged over 2010–2017 using employment weights. The sample is restricted to local labor markets with at least 10 workers and 10 firms.

D.2.2 Hiring Thresholds

I next replicate the hiring-threshold analysis for Germany using the occupation-based definition of local labor markets. As in main text Section 2.2.2, I define a firm's hiring threshold as the minimum worker AKM fixed effect among its new hires and standardize this variable within markets and years. I then regress the standardized hiring threshold on firm-decile indicators, controlling for $\log(\text{New Hires})$ and market and year fixed effects, and estimate separate specifications in which firms are ranked by (i) their AKM fixed effects, (ii) the average AKM fixed effect of their workforce (coworker quality), and (iii) their average log wage. Local labor markets are defined by three-digit occupation, and firm types at the establishment–occupation level.

Figure D.3 reports the resulting hiring-threshold profiles by firm decile. In all three specifications, hiring thresholds rise steeply and approximately monotonically with firm rank. When firms are ranked by their AKM fixed effects, the standardized hiring threshold rises from zero for bottom-decile firms to more than 0.5 standard deviations for top-decile firms. The gradients are even steeper when firms are ranked by coworker quality or average log wage: relative to the bottom decile, top-decile firms exhibit hiring thresholds roughly 1.1 standard deviations higher under the average-AKM-worker-fixed-effect ranking and about 1.0 standard deviations higher

under the average-log-wage ranking. Standard errors are small throughout, and the confidence bands for high- and low-ranked firms do not overlap. Overall, the German evidence confirms the Italian pattern: higher-type firms systematically screen more aggressively on worker quality, and this conclusion is robust to alternative measures of firm rank.

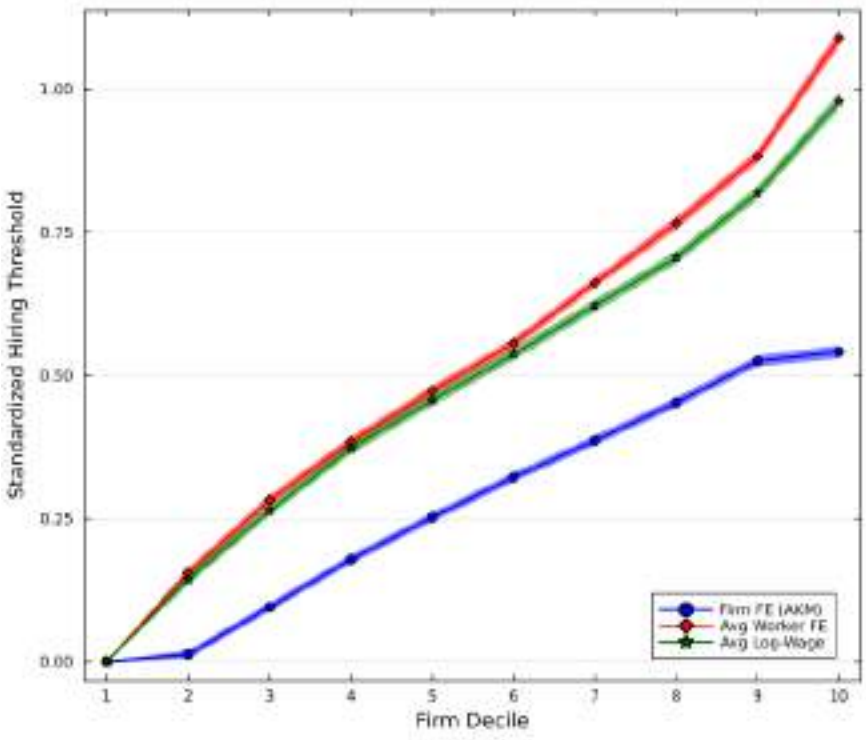


Figure D.3: Hiring thresholds by firm decile (occupation-based local labor markets, Germany)

Notes: The figure plots standardized hiring thresholds by firm decile in Germany. The dependent variable is the minimum worker AKM fixed effect among new hires, standardized within local labor markets and years. Local labor markets are defined by three-digit occupation, and firm types at the establishment–occupation level. Each line corresponds to a separate regression in which firms are ranked by (i) their AKM fixed effects, (ii) the average AKM fixed effect of their workforce (coworker quality), or (iii) their average log wage. In each specification, I regress the standardized hiring threshold on firm-decile indicators, controlling for $\log(\text{New Hires})$ and market and year fixed effects. The plotted points report the estimated coefficients by firm decile, normalized to zero for bottom-decile firms; shaded areas indicate 95% confidence intervals based on clustered standard errors. The sample covers 2010–2017 and is restricted to local labor markets with at least 10 workers and 10 firms.

D.3 Replication: Occupation-Based Local Labor Markets

D.3.1 Market shares: robustness to occupation-based local labor markets

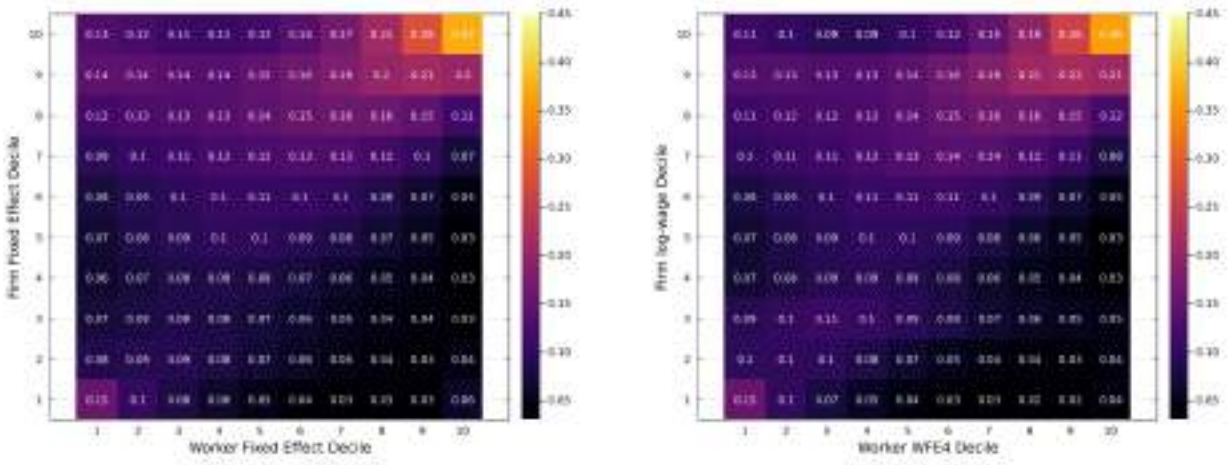
This appendix verifies that the segregation results are robust to defining local labor markets by occupation–commuting zones rather than industry–commuting zones. I repeat the exercise in main text Section 2.2.1, but now construct all worker and firm deciles within occupation-based local labor markets. This alternative definition is useful because workers may compete most di-

rectly with others performing similar tasks rather than with all workers employed in the same industry–commuting zone.

Figure D.4 reports the resulting worker–firm employment matrices when firms are ranked by their AKM firm fixed effect and by their average log wage, respectively. In both cases, the pattern is very similar to the baseline industry-based specification. Bottom-decile workers remain disproportionately concentrated in bottom-decile firms, while top-decile workers remain disproportionately concentrated in top-decile firms. Under the AKM-based firm ranking, bottom-decile workers allocate about 15% of their employment to bottom-decile firms and about 11% to top-decile firms, whereas top-decile workers allocate about 37% to top-decile firms and about 6% to bottom-decile firms. When firms are ranked by average log wages, the same pattern obtains, with somewhat stronger concentration in the upper tail.

Figure D.5 reports the analogous coworker-type matrix under the occupation-based definition. The figure is visually very similar to its industry-based counterpart in main text Figure 3b. Bottom-decile workers are disproportionately surrounded by low-ability coworkers, while top-decile workers are disproportionately surrounded by high-ability coworkers. Thus, the segregation pattern is robust not only in the joint distribution of worker and firm types, but also in coworker composition within firms.

Taken together, these results show that the evidence on worker segregation is not sensitive to whether local labor markets are defined by industry or occupation. If anything, the occupation-based definition delivers a slightly sharper pattern, but the quantitative differences are small relative to the overall strength of top–top and bottom–bottom clustering.



(a) Firms ranked by AKM firm fixed effect

(b) Firms ranked by average log wage

Figure D.4: Employment shares across worker fixed-effect deciles in occupation-based local labor markets

Notes: The figures report, for Italy, the distribution of employment shares across firm deciles for each worker fixed-effect decile. Workers are ranked by their AKM worker fixed effect within local labor market and year. Local labor markets are defined by occupation–commuting zones. In panel (a), firms are ranked by their AKM firm fixed effect; in panel (b), by their average log wage. Decile assignments are performed separately within each local labor market and year, and the resulting matrices are averaged over 2015–2019 using employment weights. The sample is restricted to local labor markets with at least 100 workers and 100 firms.

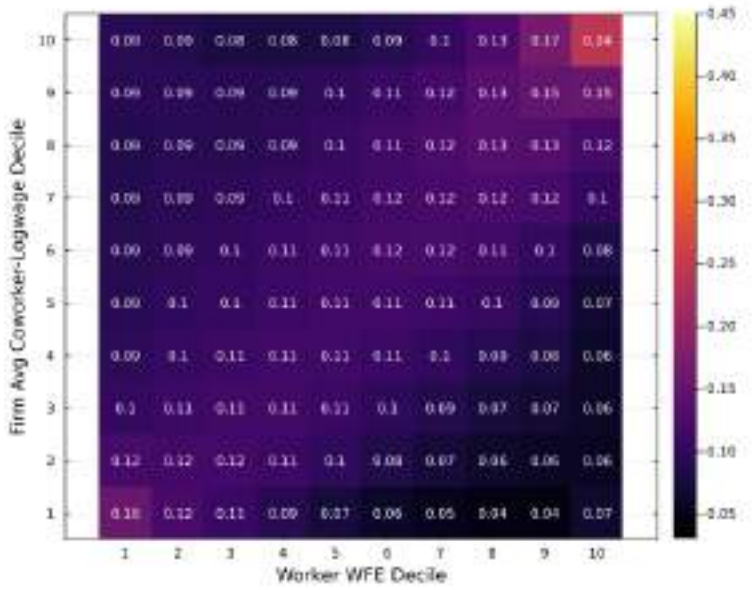


Figure D.5: Employment shares by worker fixed-effect decile and coworker-type decile in occupation-based local labor markets

Notes: The figure reports, for Italy, the distribution of employment shares across coworker-type deciles for each worker fixed-effect decile. Workers are ranked by their AKM worker fixed effect within local labor market and year. Local labor markets are defined by occupation–commuting zones. Firms are ranked into deciles based on the within-firm average AKM worker fixed effect among coworkers, excluding the worker herself. Decile assignments are performed separately within each local labor market and year, and the resulting matrix is averaged over 2015–2019 using employment weights. The sample is restricted to local labor markets with at least 100 workers and 100 firms.

D.3.2 Hiring thresholds in occupation-based local labor markets

This appendix repeats the hiring-threshold analysis of main text Section 2.2.2 using occupation–commuting zones rather than industry–commuting zones to define local labor markets. All worker and firm deciles are constructed within occupation-based local labor markets, and the estimating equation remains identical to main text Equation (16).

Figure D.6 reports the estimated relationship between hiring thresholds and firm deciles under the three firm-quality measures used in the main text. The results are qualitatively similar to those obtained under the industry-based definition. When firms are ranked by their AKM firm fixed effect, the relationship is strongly positive and approximately linear: the hiring threshold rises by about 0.80 standard deviations between the bottom and top deciles. When firms are ranked by average incumbent log wages, the corresponding increase is about 0.50 standard deviations. By contrast, when firms are ranked by the average worker fixed effect among incumbents, the relationship is again nearly flat and slightly negative in the upper tail.

Thus, redefining local labor markets by occupation rather than industry leaves the main con-

clusion unchanged: firms with higher wage premia or higher wages set systematically higher hiring thresholds. At the same time, the weaker relationship obtained when firms are ranked by incumbent worker composition is also robust.

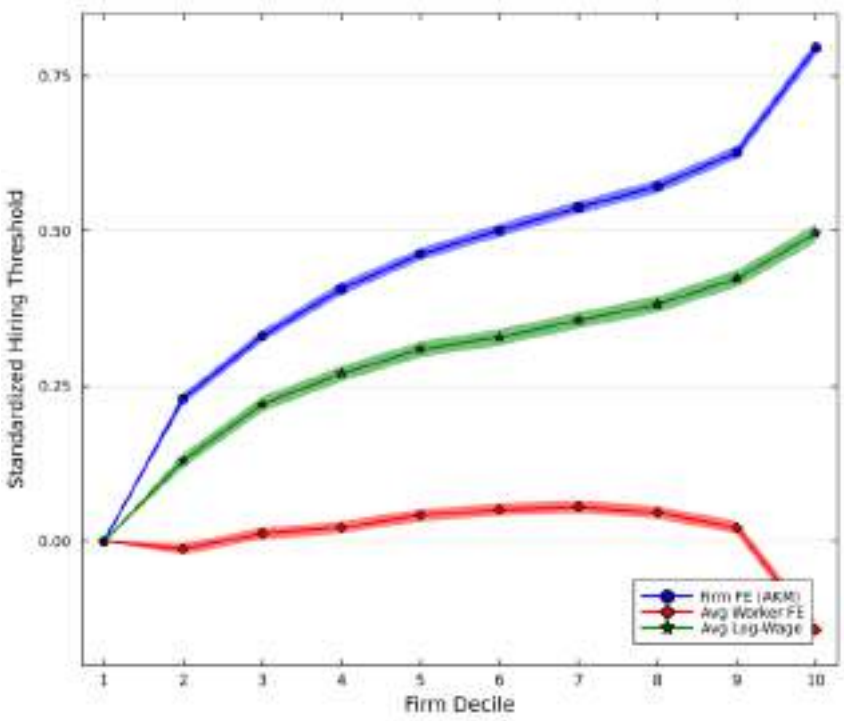


Figure D.6: Hiring thresholds by firm decile in occupation-based local labor markets

Notes: The figure plots the estimated relationship between firms' hiring thresholds and their position in the local firm-type distribution, as in main text Equation (16), for occupation–commuting-zone local labor markets. The dependent variable is the standardized minimum worker fixed effect among new hires at the firm, expressed in units of the local worker fixed-effect standard deviation. Firms are ranked into deciles using three alternative measures of firm quality: the AKM firm fixed effect, the average worker fixed effect among incumbent employees, and the average incumbent log wage; each line corresponds to one ranking measure. Estimates control for $\log(\text{New Hires}_{ij,t})$ and include market and year fixed effects; ribbons depict 95% confidence intervals based on standard errors two-way clustered by firm and market–year.

D.3.3 Concentration indices in occupation-based local labor markets

This appendix repeats the HHI analysis of main text Section 2.2.3 using occupation–commuting zones rather than industry–commuting zones to define local labor markets. As in the baseline specification, workers are assigned to ten deciles based on the within-market distribution of AKM worker fixed effects, and for each worker-decile group I compute a local wage-bill HHI over firms. These local HHIs are then aggregated to the national level using employment weights based on FTE worked weeks. I also report the corresponding HHI profiles separately for white-collar and blue-collar occupations.

Figure D.7 shows that the qualitative pattern is similar to the one reported in the main text.

Under the occupation-based definition, the overall mean HHI is about 0.13, compared with 0.16 for blue-collar workers and 0.15 for white-collar workers. Across worker fixed-effect deciles, the overall HHI is about 0.17 at the bottom of the distribution, declines to roughly 0.12 around the fifth decile, and rises again to about 0.14 in the top decile. Hence, the relationship between worker ability and market concentration remains U-shaped, although levels are lower overall than under the industry-based definition.

Within occupation-based local labor markets, blue-collar jobs are somewhat more concentrated than white-collar jobs on average. Within both groups, however, concentration is again lowest for middle-ranked workers and higher toward the tails of the fixed-effect distribution. Thus, redefining local labor markets by occupation rather than industry changes the level of measured concentration, but not the basic non-monotonic relationship between worker rank and concentration.

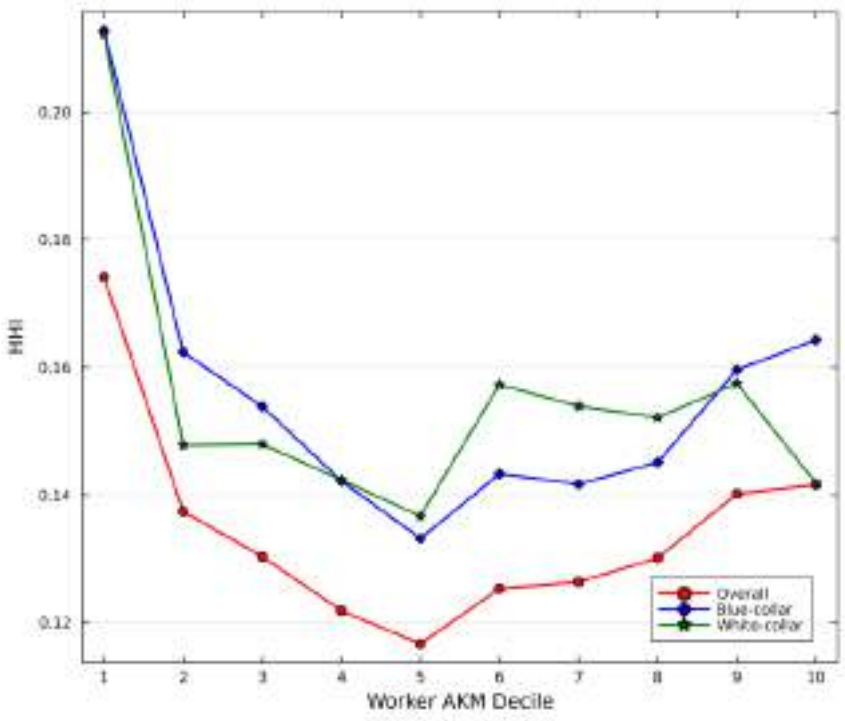


Figure D.7: Wage-bill Herfindahl indices by worker fixed-effect decile in occupation-based local labor markets

Notes: The figure plots the wage-bill Herfindahl–Hirschman Index (HHI) across worker fixed-effect deciles in Italy when local labor markets are defined by occupation–commuting zones. For each local labor market and year, workers are assigned to ten deciles based on the within-market distribution of AKM worker fixed effects. For each worker-decile group, a local wage-bill HHI is computed as the sum of squared firm wage-bill shares within that group. The figure reports employment-weighted averages of these local HHIs across all markets, where weights are total employment in FTE worked weeks for the corresponding worker-decile group. In addition to the overall HHI, each local labor market is further split into broad occupational segments (white-collar and blue-collar jobs), and the same HHI construction is repeated within these finer market definitions; the three lines correspond to the overall sample, blue-collar workers, and white-collar workers.

D.4 Empirical facts: full-time main jobs

I now replicate the AKM-based moments on a subsample that focuses on the extensive margin for prime-age workers in relatively stable employment relationships. Specifically, I restrict the sample to full-time employees aged 20–65.

I assign each worker to a single firm in each calendar year, namely the firm at which the worker earns the highest annual wage. This "main job" definition follows standard practice in the AKM literature (e.g., Card et al., 2013). On this restricted sample, I estimate the main text AKM regression (15) and use the resulting worker and firm fixed effects to construct the empirical moments used in the calibration.

Relative to the baseline sample, which includes part-time work and multiple job spells, this robustness sample is selected into larger firms and exhibits more dispersed wages. The mean firm now employs roughly 9 workers rather than 6, and the standard deviation of log wages is about 0.50 rather than 0.40 in the full population.

I first replicate the market-share analysis in main text Section 2.2.1, then the hiring-threshold analysis in main text Section 2.2.2, and finally the HHI profiles in main text Section 2.2.3.

D.4.1 Market shares: robustness to full-time main jobs

I now revisit the market-share analysis using the full-time main-job sample described in Appendix D.4. As in main text Section 2.2.1, I take a worker-side perspective and compute, for each decile of the within-year worker fixed-effect distribution, the employment share in each decile of the firm fixed-effect distribution. Decile assignments are performed separately within each local labor market and year, and the resulting matrices are averaged over 2015–2019 using employment (week) weights. To avoid distortions from very small markets, I restrict the sample to local labor markets with at least 100 workers and 100 firms.

Figure D.8 reports the employment matrix in this robustness sample, where both workers and firms are ranked by their AKM fixed effects. Figure D.9 replicates the exercise ranking firms instead by their average log wage while continuing to rank workers by their AKM fixed effects. In both cases, segmentation remains pronounced: low-ranked workers are disproportionately employed in low-ranked firms, while high-ranked workers are clustered in high-ranked firms. The gradient across worker deciles remains smooth, with higher-decile workers progressively more likely to be employed in higher-decile firms.

Quantitatively, when local labor markets are defined by industry–commuting zones and firms are ranked by their AKM fixed effects, bottom-decile workers allocate about 21.2% of their employment to bottom-decile firms but only 5.7% to top-decile firms. Conversely, top-decile workers allocate 27.1% of their employment to top-decile firms and only 7.3% to bottom-decile firms. Very similar magnitudes obtain when firm deciles are defined by average log wages rather than AKM fixed effects. Relative to the baseline results in main text Section 2.2.1, the bottom–bottom concentration is somewhat larger, indicating that ability segregation is, if anything, stronger once

attention is restricted to full-time main jobs.

A natural interpretation is that high-paying firms also employ a nontrivial number of part-time and secondary low-wage jobs. When these jobs are included, as in the baseline FTE-weighted analysis, they dilute the observed degree of sorting between high-AKM workers and high-AKM firms. By contrast, the full-time main-job sample strips out much of this intensive-margin and secondary-job variation, revealing more clearly the association between worker and firm types. This robustness exercise therefore reinforces the main conclusion that segmentation is a salient feature of the worker–firm allocation rather than an artifact of part-time or multi-job employment.

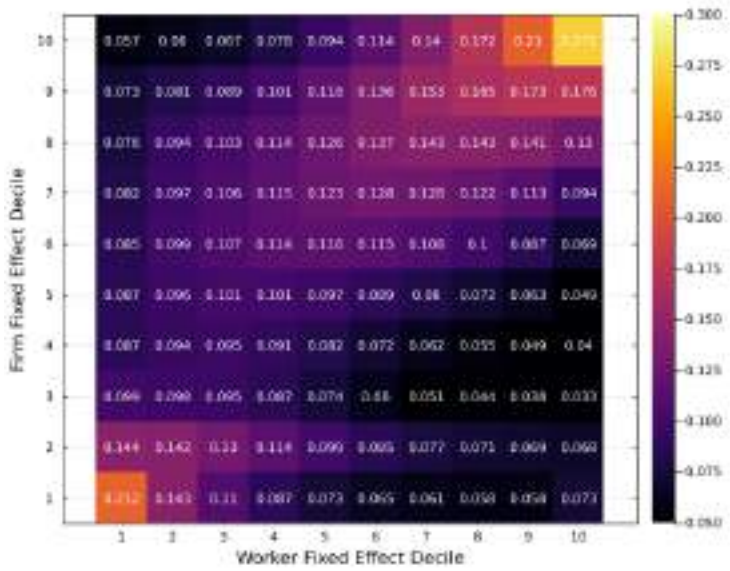


Figure D.8: Employment market shares across worker and firm AKM deciles (full-time main jobs, Italy)

Notes: The figure reports, for each worker AKM fixed-effect decile, the distribution of employment shares across firm AKM fixed-effect deciles in Italy. The sample is restricted to full-time employees aged 20–65, and each worker is assigned to a single "main" employer in each calendar year, defined as the firm at which the worker earns the highest annual wage. Job spells are weighted by the number of weeks worked in the spell. Workers and firms are ranked into deciles within each local labor market and year, where local labor markets are defined by industry–commuting zone and firms are identified at the activity–location level. The matrix is averaged over 2015–2019 using employment (week) weights. The sample is restricted to local labor markets with at least 100 workers and 100 firms.

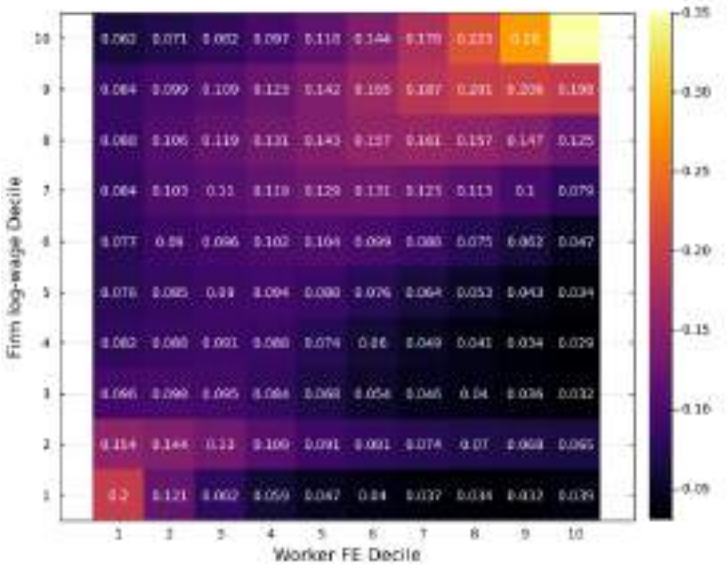


Figure D.9: Employment market shares across worker AKM deciles and firm average-log-wage deciles (full-time main jobs, Italy)

Notes: The figure reports, for each worker AKM fixed-effect decile, the distribution of employment shares across firm deciles in Italy, where firms are ranked by their average log wage. The sample is restricted to full-time employees aged 20–65, and each worker is assigned to a single "main" employer in each calendar year, defined as the firm at which the worker earns the highest annual wage. Job spells are weighted by the number of weeks worked in the spell. Workers are ranked by their AKM fixed effects; firms are ranked by their average log wage. Rankings are constructed within each local labor market and year, where local labor markets are defined by industry–commuting zone and firms are identified at the activity–location level. The matrix is averaged over 2015–2019 using employment (week) weights. The sample is restricted to local labor markets with at least 100 workers and 100 firms.

D.4.2 Hiring thresholds: robustness to full-time main jobs

I now revisit the hiring-threshold analysis using the full-time main-job sample described in Appendix D.4. I re-estimate main text Regressions (16) on full-time employees aged 20–65, assigning each worker to a single "main" employer in each calendar year, defined as the firm at which the worker earns the highest annual wage. As in the main text, the analysis is restricted to within-market transitions in local labor markets with sufficient numbers of switching firms and workers to construct meaningful firm and worker deciles.

Figure D.10 plots the resulting nonparametric relationship between firms' hiring thresholds and their decile in the local firm-type distribution for the three ranking schemes used in the main analysis: the AKM firm fixed effect, the average worker fixed effect among incumbents, and the average incumbent log wage. In all cases, the patterns remain strongly positive and approximately linear: higher-ranked firms continue to hire workers with strictly higher minimum AKM fixed effects.

Quantitatively, firms in the top decile set hiring thresholds about 0.8 standard deviations above those of bottom-decile firms in the full-time main-job sample, very similar to the magnitudes in the baseline results. Overall, the implied slopes are close to those in the baseline, confirming that the strong positive link between firm quality and hiring thresholds documented in Fact 2 is not driven by part-time work or multiple simultaneous jobs.

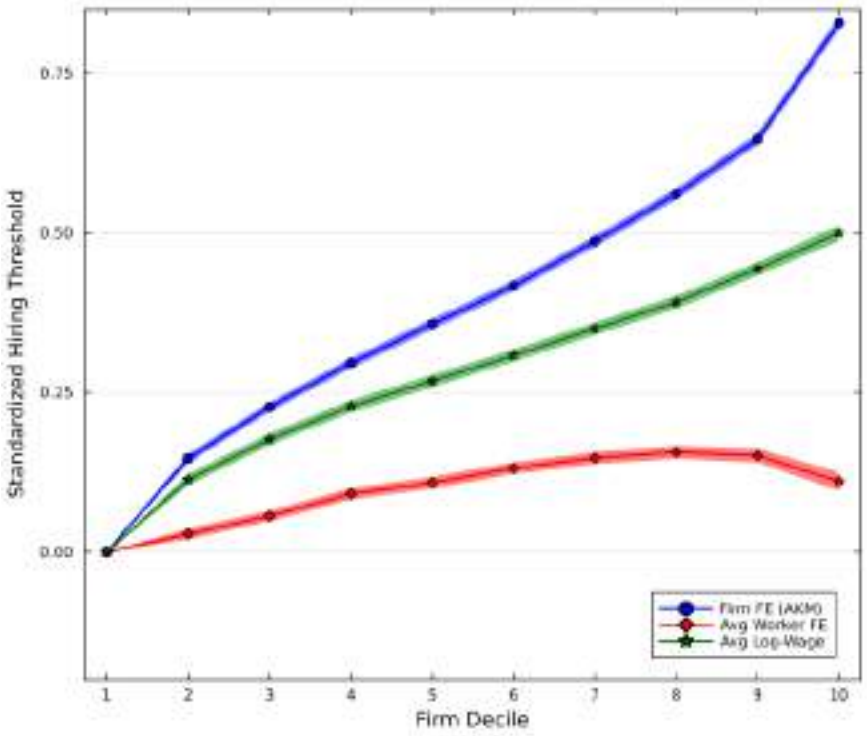


Figure D.10: Hiring thresholds by firm decile (full-time only, Italy)

Notes: Each panel plots estimated hiring thresholds by firm decile within local labor markets, comparing the data (solid blue line with 95% confidence bands) to the model-implied relationship (dashed orange line). The dependent variable is the standardized minimum worker fixed effect among new hires at firm i in market j and year t, as in main text Equation (16). Firm deciles are defined within local labor markets based on (a) the AKM firm fixed effect, (b) the average worker fixed effect among incumbent employees, and (c) the average incumbent log wage. The sample is restricted to full-time employees aged 20–65, with each worker assigned to a single main employer per year, defined as the firm paying the highest annual wage. All regressions control for $\log(\text{New Hires}_{ij,t})$ and include market and year fixed effects.

D.4.3 HHI indices by broad occupation group

Figure D.11 plots wage-bill Herfindahl–Hirschman indices (HHI) by worker AKM fixed-effect decile for the overall workforce and separately for blue- and white-collar jobs, using the full-time main-job sample described in Appendix D.4. For the overall workforce, the HHI profile retains the familiar U-shape documented in the main text, with somewhat steeper variation across worker fixed-effect deciles once attention is restricted to full-time main jobs. When disaggregating by broad occupation group, concentration is uniformly higher and more sharply U-shaped for white-

collar workers than for blue-collar workers, with especially elevated HHI levels in the tails of the white-collar distribution. These patterns confirm that concentration—and hence employer market power—is especially pronounced for low- and high-ranked white-collar workers, and they show that the main-text results are robust to excluding part-time work and secondary jobs.

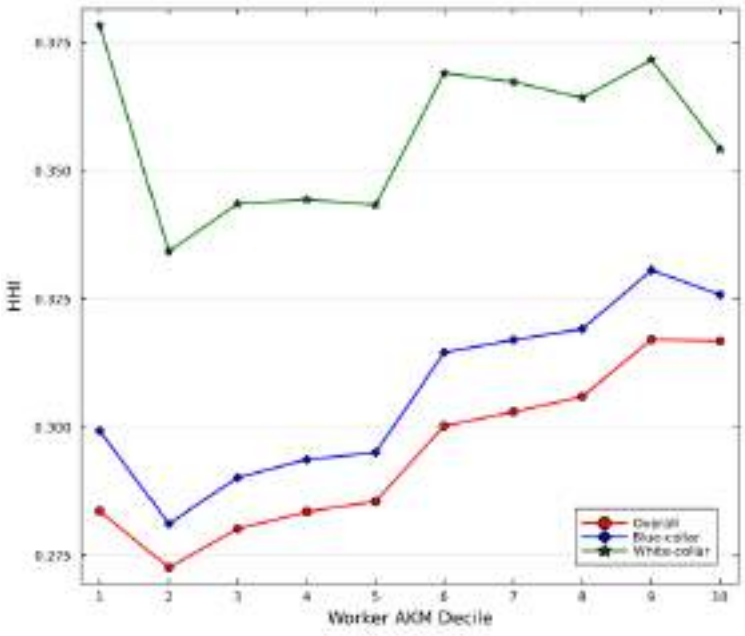


Figure D.11: Wage-bill HHI by worker fixed-effect decile (full-time main jobs, Italy)

Notes: For each local labor market (industry–commuting zone) and each worker AKM fixed-effect decile, I compute the wage-bill Herfindahl–Hirschman index (HHI) as the sum of squared firm wage-bill shares among workers in that decile. The sample is restricted to full-time employees aged 20–65, and each worker is assigned to a single main employer per year, defined as the firm at which the worker earns the highest annual wage. The figure reports the resulting HHI profiles over 2015–2019 for the overall workforce and separately for blue- and white-collar workers. Local HHI values are aggregated to the national level by weighting each market–decile observation by total employment in the corresponding group.

D.5 Taking stock

Taken together, the robustness exercises deliver a remarkably consistent picture. First, the segmentation patterns documented in the main text are not sensitive to how firms are ranked or how local labor markets are defined. Whether firms are ranked by AKM fixed effects or by average log wages, and whether markets are defined by industry or by occupation, low-ranked workers remain disproportionately employed in low-ranked firms and high-ranked workers in high-ranked firms. The same qualitative pattern also appears in the German data, indicating that strong worker–firm segmentation is not specific to the Italian setting.

Second, the positive relationship between firm rank and hiring thresholds is equally robust. In Italy, higher-ranked firms systematically hire workers with higher minimum AKM fixed effects under both industry- and occupation-based market definitions and under alternative measures of

firm quality. The same qualitative pattern appears in Germany. Restricting attention to full-time main jobs leaves this conclusion essentially unchanged, showing that the result is not driven by part-time work or multiple simultaneous jobs.

Third, the concentration results are also stable. Across alternative market definitions and sample restrictions, wage-bill HHI profiles remain U-shaped in worker rank, with relatively high concentration for workers in the tails of the fixed-effect distribution. The level of HHI changes somewhat—for example, occupation-based markets yield lower measured concentration than industry-based markets—but the basic non-monotonic pattern is unchanged. In the full-time main-job sample, the profile is, if anything, somewhat steeper, especially for white-collar workers.

Overall, these exercises show that the three core empirical facts emphasized in the main text—strong within-market segmentation, selective hiring by higher-type firms, and elevated concentration for workers at the tails of the ability distribution—are not artifacts of a particular market definition, ranking measure, country, or treatment of part-time and multiple-job employment.

E Taking the Model to the Data

This appendix provides additional details for main-text Section 3.

E.1 Calibration of the Distribution of Firms Across Local Labor Markets

To discipline the distribution of firms across local labor markets, I simulate an economy with $M = 1000$ markets, each hosting up to 200 potential firms. In the data, a local labor market is defined as the intersection of a three-digit industry and a commuting zone, consistent with the balance-sheet data used in the calibration.

The number of active firms per local labor market is assumed to follow a Pareto distribution with upper bound 200. I also include a mass point at one-firm markets to capture the large share of such markets in the data. The location, tail, and scale parameters of this distribution are estimated by maximum likelihood using the empirical distribution of firms across markets. The calibrated distribution closely matches the data, as shown in Figure E.1.

Table E.1: Calibration of the Firm-Size Distribution Across Local Labor Markets

Parameter	**Value**
Location parameter of Pareto distribution	1.000
Tail parameter (shape)	1.0673
Scale parameter	4.101
Mass of markets with one firm	0.190
Maximum number of firms per market	200

Notes: The table reports the calibrated parameters governing the distribution of firms across local labor markets. The distribution is estimated by maximum likelihood using the empirical distribution of firms per market. A mass point at one-firm markets is included to capture the substantial share of one-firm markets in the data.

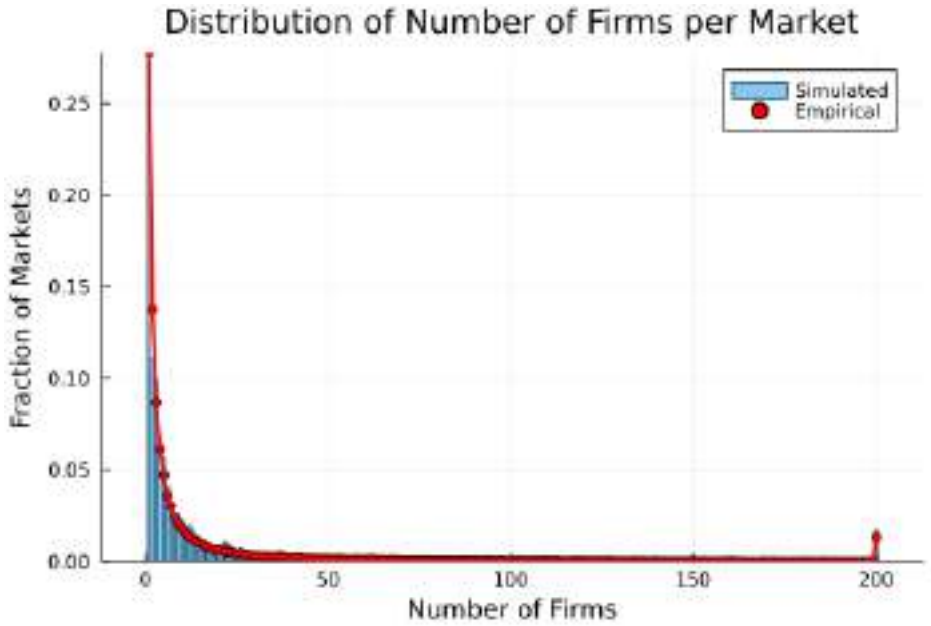


Figure E.1: Distribution of Firms Across Local Labor Markets: Data vs. Model

E.2 Production Function Estimation: Parameters (α, γ)

This section describes how I estimate firm-level revenue productivity using the merged INPS–CERVED panel for Italy. The goal is to use the estimated output elasticities to discipline the calibration of the production parameters α and γ, that is, the decreasing-returns parameter and the share parameters in the capital–labor composite.

E.2.1 Sample and Variable Definition

The estimation sample is the merged INPS–CERVED firm-year panel for 2015–2018, restricted to private for-profit firms as described in Appendix B.3. For each firm i in local labor market j and year t, I construct:

- **Value added:** VA_{ijt}, defined from CERVED balance sheets as revenues minus intermediate inputs (materials and services). The dependent variable in the production function is log value added,

  $$y_{ijt} \equiv \log(VA_{ijt})$$

  Since VA_{ijt} is measured in current prices, the production function is estimated in revenue terms.

- **Capital:** K_{ijt}, measured as the sum of tangible and intangible fixed assets. The corresponding regressor is $\log K_{ijt}$.

- **Labor:** L_{ijt}, measured as the firm's full-time-equivalent (FTE) worked weeks derived from the INPS employment panel; the corresponding regressor is $\log L_{ijt}$.[3]
- **Intermediate inputs:** M_{ijt}, proxied by material expenditures. This input is used as a *proxy* for unobserved productivity in the control-function estimation, but does not enter the final value-added specification directly.

The goal is to recover sector-level production-function elasticities. To limit the influence of outliers in balance-sheet and employment data, I winsorize log value added, log capital, log employment, log materials, and the log wage bill symmetrically at the 5th and 95th percentiles. Observations with missing values after winsorization are dropped. The resulting sample contains about 1.3 million firm-year observations over 2015–2018.

Firms are assigned to sectors using the three-digit ATECO industry code. I estimate separate production functions by three-digit sector, retaining only sectors with at least 200 observations. Sector-specific year fixed effects absorb sector–year-specific price levels and common shocks.

E.2.2 Empirical Specification and Identification

Within each three-digit sector s, I estimate a Cobb–Douglas value-added production function of the form

$$y_{ijt} = \beta_{\ell s} \log L_{ijt} + \beta_{ks} \log K_{ijt} + \delta_t^{(s)} + \omega_{ijt} + \varepsilon_{ijt} \tag{E.1}$$

where $\delta_t^{(s)}$ are sector-specific year fixed effects, ω_{ijt} is log firm revenue productivity, and ε_{ijt} is an idiosyncratic error term. The coefficients $\beta_{\ell s}$ and β_{ks} are sector-specific output elasticities with respect to labor and capital.

Input choices $(L_{ijt}, K_{ijt}, M_{ijt})$ may be correlated with ω_{ijt}. I address this endogeneity using the control-function approach of Levinsohn et al. (2003), with the identification refinements of Ackerberg et al. (2015). The key assumptions are that (i) intermediate inputs respond monotonically to productivity, conditional on capital, and (ii) productivity follows a first-order Markov process. Intermediate materials then serve as a proxy: conditional on K_{ijt}, the demand for M_{ijt} can be inverted to recover ω_{ijt} up to a nonparametric transformation. In practice, I approximate this control function with a low-order polynomial in $\log K_{ijt}$ and $\log M_{ijt}$.

Labor is treated as a freely adjustable input, capital as a predetermined state variable, and intermediate inputs as the productivity proxy. The ACF moment conditions exploit the Markov structure of ω_{ijt} and lagged inputs to disentangle the contemporaneous correlation between inputs and productivity. Estimation is carried out separately within each sector s, and sectors with very small samples are excluded to avoid poorly identified parameters.

Because the production function is estimated in value-added (revenue) terms, ω_{ijt} should be interpreted as revenue productivity (TFPR). Sector-specific year fixed effects $\delta_t^{(s)}$ absorb sector-level price and demand shocks.

[3] To make employment comparable to balance-sheet information, employment for this production-function estimation is computed at the enterprise level.

E.2.3 Elasticities and Productivity Residuals

For each sector s, I recover the estimated elasticities $\hat{\beta}_{\ell s}$ and $\hat{\beta}_{ks}$ together with the sector-specific year fixed effects $\hat{\delta}_t^{(s)}$. The fitted value from (E.1) is

$$\hat{y}_{ijt} = \hat{\beta}_{\ell s} \log L_{ijt} + \hat{\beta}_{ks} \log K_{ijt} + \hat{\delta}_t^{(s)},$$

and I define the firm's log revenue productivity as the residual

$$\hat{\omega}_{ijt} = y_{ijt} - \hat{y}_{ijt} \tag{E.2}$$

Across all sectors in 2015–2018, the employment-weighted mean of the estimated labor elasticity $\hat{\beta}_{\ell s}$ is approximately 0.729, while the corresponding mean capital elasticity $\hat{\beta}_{ks}$ is about 0.10. Only $\hat{\beta}_{\ell s}$ is used to discipline the calibration of (α, γ). I do not use $\hat{\beta}_{ks}$ because capital elasticities estimated from balance-sheet data are subject to substantial attenuation bias from measurement error in the capital stock (Collard-Wexler et al., 2016); in these data, the implied micro-level capital elasticity is also substantially below the aggregate capital share. This overidentification failure is consistent with Collard-Wexler et al. (2016) and motivates calibrating the capital side from macro data rather than from the estimated $\hat{\beta}_{ks}$.

Accordingly, I calibrate $(1-\gamma)\alpha$ to match the aggregate value-added share of capital computed from firm-level balance-sheet data (CERVED). The aggregate user cost of capital is taken from the Penn World Table (Feenstra et al., 2015) and applied to the stock of tangible capital reported in balance sheets. This yields an aggregate capital share of 0.20777.[4] I then recover the effective labor output elasticity, $\alpha\gamma$, using the sector-level production-function estimates: sector-specific labor elasticities $\hat{\beta}_{\ell s}$ are aggregated across three-digit sectors using employment weights to obtain an economy-wide target for $\alpha\gamma$.

These two calibration steps jointly pin down the pair (α, γ) used in the quantitative analysis. The resulting parameter values are $\gamma = 0.7784$ and $\alpha = 0.937$.[5] These values imply an effective labor elasticity $\alpha\gamma \approx 0.729$ and a capital elasticity $(1-\gamma)\alpha \approx 0.208$, consistent with both the estimated employment-weighted labor elasticity from the production-function regressions and the aggregate capital share in the Italian data.

E.3 Labor Supply Elasticities

In this appendix, I first describe the empirical approach used to estimate η. I then explain the construction of the death shock, the sample restrictions and pretrend checks that support identification, and the robustness exercises. Finally, I discuss how I recover $\widetilde{\psi}_{ij}$ from observables and the empirical specification used to calibrate θ.

[4] For comparison, the corresponding aggregate capital share in the United States is approximately 18% (Barkai, 2020).

[5] Berger et al. (2022) calibrate $\alpha = 0.940$ and $\gamma = 0.808$ for the U.S. economy.

E.3.1 Within-Market Elasticity η

Derivation of the inverse labor-supply approximation in Equation (17). In the empirical design, the shock is the death of a single worker at firm (i, j) in period t. Let $\tilde{n}_{ij,t}(a)$ denote the labor input (in FTE worked weeks) of the deceased worker of ability a in that firm–year, and define employment of *all remaining workers* ("survivors") as

$$n^{\text{surv}}_{ij,t}(a) \equiv n_{ij,t}(a) - \tilde{n}_{ij,t}(a), \qquad h^{\text{surv}}_{ij,t} \equiv \sum_a n^{\text{surv}}_{ij,t}(a),$$

with corresponding within-firm type shares

$$g^{\text{surv}}_{ij,t}(a) \equiv \frac{n^{\text{surv}}_{ij,t}(a)}{h^{\text{surv}}_{ij,t}}.$$

Survivor-level inverse labor supply. The discrete-choice microfoundation in Appendix A.1 is formulated for a continuum of type-a workers sharing a common taste-shock distribution. Any subset of that continuum that retains the same taste-shock distribution and faces the same wage offers satisfies the same nested-logit choice probabilities, and aggregation continues to deliver the same nested-CES labor-supply system with the same parameters (η, θ). Removing the deceased worker and, symmetrically, a matched placebo worker in control firms therefore redefines the relevant population as the set of survivors but leaves the functional form and the parameters (η, θ) unchanged.

Under a firm-specific shock in a large economy, the aggregate wage index $W_t(a)$ can be treated as fixed. The type-a employment of surviving workers at firm (i, j) then satisfies the same type-level relationship as in the full population:

$$\mathrm{d}\log n^{\text{surv}}_{ij,t}(a) = \eta\,\mathrm{d}\log w^{\text{surv}}_{ij,t}(a) + (\theta - \eta)\,\mathrm{d}\log w^{\text{surv}}_{j,t}(a)$$

where $w^{\text{surv}}_{ij,t}(a)$ is the wage paid to surviving type-a workers and $w^{\text{surv}}_{j,t}(a)$ is the survivor-based local-market wage index for type a. In response to a firm-specific shock,

$$\mathrm{d}\log n^{\text{surv}}_{ij,t}(a) = \big[\eta + (\theta - \eta)\lambda_{ij,t}(a)\big]\,\mathrm{d}\log w^{\text{surv}}_{ij,t}(a), \qquad \lambda_{ij,t}(a) \equiv \frac{\mathrm{d}\log w^{\text{surv}}_{j,t}(a)}{\mathrm{d}\log w^{\text{surv}}_{ij,t}(a)}\Big|_t$$

Aggregating over ability types and using $h^{\text{surv}}_{ij,t} = \sum_a n^{\text{surv}}_{ij,t}(a)$ and $\mathrm{d}\log h^{\text{surv}}_{ij,t} = \sum_a g^{\text{surv}}_{ij,t}(a)\,\mathrm{d}\log n^{\text{surv}}_{ij,t}(a) = \mathbb{E}_{g^{\text{surv}}_{ij,t}}\big[\mathrm{d}\log n^{\text{surv}}_{ij,t}(a)\big]$ yields the continuous-time direct supply relation for surviving workers:

$$\mathrm{d}\log h^{\text{surv}}_{ij,t} = \eta\,\mathbb{E}_{g^{\text{surv}}_{ij,t}}\big[\mathrm{d}\log w^{\text{surv}}_{ij,t}(a)\big] + (\theta - \eta)\,\mathbb{E}_{g^{\text{surv}}_{ij,t}}\big[\lambda_{ij,t}(a)\,\mathrm{d}\log w^{\text{surv}}_{ij,t}(a)\big].$$

In discrete time, I replace differentials with first differences over $[t, t+1]$ and interpret these as

first-order local approximations. To link the direct relation to observables, note that

$$\Delta\overline{\log w}^{\text{surv}}_{ij,t+1} = \mathbb{E}_{g^{\text{surv}}_{ij,t}}\big[\Delta \log w^{\text{surv}}_{ij,t+1}(a)\big] + \sum_a \Delta g^{\text{surv}}_{ij,t}(a)\ \log w^{\text{surv}}_{ij,t+1}(a)$$

where, for any variable x, $\Delta x_{t+1} := x_{t+1} - x_t$, and

$$\overline{\log w}^{\text{surv}}_{ij,t} \equiv \sum_a g^{\text{surv}}_{ij,t}(a) \log w^{\text{surv}}_{ij,t}(a).$$

Solving for $\mathbb{E}_{g^{\text{surv}}_{ij,t}}[\Delta \log w^{\text{surv}}_{ij,t+1}(a)]$ and substituting into the direct relation yields the discrete-time first-order inverse labor-supply approximation for survivors:

$$\Delta\overline{\log w}^{\text{surv}}_{ij,t+1} \approx \frac{1}{\eta}\,\Delta \log h^{\text{surv}}_{ij,t+1} + \underbrace{\sum_a \Delta g^{\text{surv}}_{ij,t}(a)\ \log w^{\text{surv}}_{ij,t+1}(a)}_{\Delta\text{Comp}^{\text{surv}}_{ij,t+1}} - \frac{\theta-\eta}{\eta}\ \underbrace{\mathbb{E}_{g^{\text{surv}}_{ij,t}}\big[\lambda_{ij,t}(a)\,\Delta \log w^{\text{surv}}_{ij,t+1}(a)\big]}_{\Delta\text{Olig}^{\text{surv}}_{ij,t+1}} + \varepsilon^{\text{surv}}_{ij,t+1} \tag{E.3}$$

Equation (E.3) is the survivor-based counterpart of main text Equation (17). It provides the log-linear inverse labor-supply relation underlying the estimation of η: the structural inverse elasticity is $1/\eta$, while composition changes and oligopsony enter as additional terms.

Estimating Equation (E.3) raises three identification challenges. First, the object of interest is the *local* slope of inverse labor supply around the equilibrium prevailing at time t. As in standard labor-supply estimation, this requires a labor-demand shifter, but here the demand shifter must also be *small*, exogenous, and firm-specific, so that it traces out the local slope rather than inducing large nonlinear adjustments. Second, a naive IV regression of changes in average log wages on changes in log employment suffers from the omitted-variable problem emphasized by BHM: the firm-level shock enters the market CES wage index, which enters the estimating equation through the oligopsony term and biases the estimate of $1/\eta$. Third, Equation (E.3) makes clear that a firm-level demand shock may change the ability composition of the workforce, so that $\Delta\overline{\log w}^{\text{surv}}_{ij,t+1}$ partly reflects reweighting across types rather than movement along a fixed inverse labor-supply curve.

I address these concerns in three complementary ways. First, as the demand shifter, I use the *unexpected death* of a worker, in the spirit of Jäger et al. (2024), as a quasi-random shock to labor demand for surviving workers and replacement hires[6]. A single worker death is a small perturbation to firm employment, making the log-linear approximation in Equation (17) empirically plausible.[7] I focus on the impact response between event times $\tau = -1$ and $\tau = 1$,[8] comparing the

[6]The use of worker or manager deaths as a source of exogenous variation is common in many studies, including Jaravel et al. (2018), Azoulay et al. (2019), Bennedsen et al. (2020), Sauvagnat et al. (2024), and Jäger et al. (2024).

[7]The average firm in the estimation sample employs about 13 FTE workers, and a single worker can supply at most 1 FTE. Thus, the largest relevant employment change from a single death is $1/13$. Any second-order term in a smooth approximation is of order $(1/13)^2$, and is therefore quantitatively negligible at the scale of the shocks studied here.

[8]Here, τ denotes event time in the stacked difference-in-differences design, with $\tau = 0$ the calendar year in which the death occurs. I do not use the change from $\tau = -1$ to $\tau = 0$, because the data are annual, whereas deaths occur at different dates within the event year. When a death occurs late in calendar year $\tau = 0$, part or all of the adjustment

response of survivor employment to the response of survivor wages. I also restrict the sample to events in firms with preshock employment below 50 FTE workers. This cap ensures that the death has a nonnegligible effect on the firm while excluding very large employers, where oligopsonistic forces are more likely to matter. Restricting the sample in this way further reduces concerns that the estimates are contaminated by market-level wage-index responses driven by large employers.

Second, I address potential oligopsony bias by allowing labor-supply elasticities to vary with the firm's baseline wage-bill share. Specifically, I interact the change in survivor employment with the firm's preshock wage-bill share in the local labor market. This specification allows the slope of inverse labor supply to vary with the firm's competitive environment and recovers the atomistic inverse elasticity at an approximately zero wage-bill share. As an additional robustness check, I reestimate the specification on the subset of firms whose preshock market wage-bill share lies below the median in the shocked-control estimation sample, equal to 1.07%.

Third, I explicitly control for composition changes induced by the death shock, so that identification of $1/\eta$ relies on within-composition wage variation rather than mechanically induced changes in the composition of employment. Specifically, I control for the deceased worker's position in the preshock firm wage distribution (below or above the mean), and I include nonparametric controls for changes in the average AKM worker fixed effect among survivors and among new hires. To avoid contamination from postevent adjustment, these fixed effects are estimated using a rolling window ending before the event year. As discussed in Appendix B.4, AKM fixed effects provide an informative indirect proxy for latent worker ability in this setting, so controlling for changes in their distribution also helps absorb changes in latent worker composition.

The remainder of this subsection describes the construction of the death shock, the baseline characteristics of treated and matched control firms, the empirical specification and pretrend checks, and the main estimates of η together with a set of robustness exercises.

Identifying death events. I identify worker deaths using the population registry and assign the event year d to the calendar year of the recorded date of death.[9] To remove spurious cases, I drop observations with postdeath employment records, which likely reflect reporting errors (0.02% of all deaths).

To strengthen the plausibility of exogeneity, I restrict attention to deaths that are sudden and unlikely to be preceded by prolonged health deterioration. Specifically, I retain workers younger than 60 who were employed full time both in the year of death and in each of the four preceding years ($d-4,\dots,d$). The data record the total number of *figurative weeks*—weeks in which social security contributions are credited despite the absence of actual work activity, including sickness, maternity leave, unemployment benefits, and other legally protected absences. To exclude deaths preceded by prolonged illness or labor-market detachment, I keep only workers with zero figurative weeks in the event year and in each of the previous four years. Finally, I retain only firm–year

may be realized only in $\tau = 1$. The change from $\tau = -1$ to $\tau = 1$ therefore provides a cleaner measure of the impact response.

[9]Employer notifications reporting deaths are available from 2005 onward and are used as a robustness check.

observations with a single worker death, so that the estimated effects are not confounded by larger accidents or disasters that may independently affect firm outcomes.

Firm-level and sample-size restrictions. To ensure that a worker death has a meaningful effect on firm outcomes, the analysis focuses on stable small- and medium-sized firms. I retain firms whose preshock employment lies in the interval $[1, 50]$. I also exclude firms with inconsistent postdeath employment records to remove spurious or misreported cases.

Treatment definition and position indicator. For each death event, I construct the indicator

$$L_{ij} = \mathbf{1}\{\log w_{a,ij,d-1} < \overline{\log w}_{ij,d-1}\},$$

which equals one if the deceased worker's pre-event log wage $\log w_{a,ij,d-1}$ lies below the firm's pre-event average log wage $\overline{\log w}_{ij,d-1}$. Thus, $L_{ij} = 1$ ("left") denotes the loss of a below-average-wage worker, whereas $L_{ij} = 0$ ("right") denotes the loss of an above-average-wage worker. The classification uses lagged wages to avoid contamination by contemporaneous responses.

Comparison pool and matched sampling procedure. To construct a control group of placebo death events, I draw from a pool of worker–firm–year triples that did not experience a death.

Each treated worker–firm pair—corresponding to an actual death in year d—is matched to a placebo worker–firm pair observed in the same year that did not experience a death and satisfies the same pre-event restrictions. Matching uses a rich set of worker and firm characteristics. On the worker side, I match exactly on gender, a fine two-year age group, the indicator for being below or above the firm's pre-event mean wage (L_{ij}), and the decile of the worker's log-wage deviation from the firm's pre-event average,

$$\log w_{a,ij,d-1} - \overline{\log w}_{ij,d-1}.$$

On the firm side, I match exactly on a detailed firm-size bin based on employment in $d-4$, the $d-4$ decile of the firm's average log wage, the event year, and, in the most stringent pass, a three-way firm-age group (0–5, 6–15, and > 15 years).

Matching proceeds in up to three increasingly coarse rounds. In the first round, I require exact agreement on all worker and firm variables, including the firm-age group and the wage-distance group. In the second round, I relax the firm-age requirement and match exactly on gender, wage-position indicator L_{ij}, age group, firm-size bin, event year, firm wage decile, and wage-distance group. In the third round, I further relax the wage-distance requirement and match only on gender, wage position, age group, firm-size bin, event year, and firm wage decile. Within each round, if multiple candidate controls satisfy the exact-matching criteria, I select the control that minimizes first the absolute difference in employment relative to the treated firm and then the absolute difference in the worker's distance from the firm's average wage. Most treated observations are

matched in the first and third rounds, while the second round contributes relatively few additional pairs.

Overall, the three-round procedure yields a match rate above 99%, and treated units that remain unmatched after all rounds are excluded. I also keep only treated–control pairs in which the control firm has not yet been treated, to avoid the forbidden comparisons emphasized in the recent event-study literature (e.g., Goodman-Bacon, 2021; Callaway et al., 2021). After imposing all sample-selection criteria, I retain only balanced treated–control pairs for which both the treated firm and its matched control are active and observed throughout the event-study window $t \in [-3, +1]$. The main estimation sample therefore comprises 27,769 treated worker–firm death events and an equal number of matched placebo events.

The resulting design yields 1:1 matching without replacement within each event year: every treated pair is matched to a single control pair in the same year d, and each control unit is used at most once per year. Placebo firms and workers may, however, serve as controls for different treated events in other years. Throughout, $T_{ij} = 1$ denotes an actual worker death, while $T_{ij} = 0$ denotes a placebo event.

Balance and final sample. Table E.2 reports pre-event summary statistics by treatment status and position in the firm wage distribution.

Among below-average ($L_{ij} = 1$) worker deaths, treated and control units are broadly similar in worker characteristics. Mean worker fixed effects are -0.107 for controls and -0.089 for deceased workers; mean age is 46.7 in both groups; experience is 15.2 versus 16.1 years; tenure at the firm is 8.2 versus 9.4 years; and mean log wages are 6.131 versus 6.180. Among above-average ($L_{ij} = 0$) worker deaths, treated and control samples are likewise close on age, experience, and tenure, though deceased workers have somewhat higher pre-event wages: mean worker fixed effects are 0.098 for controls and 0.103 for deceased workers, age is 48.0 versus 48.1 years, experience is 18.7 versus 18.5 years, tenure is 11.3 years in both groups, and mean log wages are 6.397 versus 6.524. Overall, observable differences between treated and control workers are modest relative to within-group dispersion, although the wage gap is somewhat more pronounced for above-average deaths.

Occupational composition and gender are similarly balanced across treated and control units within each wage-position group. For below-average ($L_{ij} = 1$) worker deaths, the female share is 14.1% in both groups, while the mean occupation measure—defined as 1 for blue-collar and 2 for white-collar jobs—is 1.172 among controls and 1.177 among deceased workers. For above-average ($L_{ij} = 0$) worker deaths, the female share is again 11.4% in both groups, and the corresponding occupation measure is 1.331 for controls and 1.350 for deceased workers.

Relative to the full estimation sample, both deceased and placebo workers are older, more experienced, longer-tenured, and less likely to be female than the average worker. Overall, the matched sampling procedure delivers a control group that resembles the treated workers closely along the key observed dimensions used in the design.

Table E.3 reports firm-level characteristics in the year before the event ($t = -1$) for treated and control firms, separately for above- and below-firm-average worker deaths. Along most dimensions, treated and control firms are very similar within each group. For above-firm-average worker deaths ($L_{ij} = 0$), mean firm fixed effects are -0.008 for controls and -0.011 for treated firms, average firm age is 19.2 versus 19.5 years, and employment is essentially identical at 15.1 workers in both groups. The number of new hires, incumbent employment, and employment market shares are also very similar. For below-firm-average worker deaths ($L_{ij} = 1$), treated and control firms are likewise comparable: mean FFE is -0.004 in the control group and -0.003 in the treated group, firm age is 19.2 versus 19.7 years, and employment is 16.6 versus 16.5 workers. Incumbent employment, employment shares, and new hires are again close across groups.

Relative to the full firm population, firms in the matched sample are older and larger, but treated and control firms are well balanced on pre-event observables. The sector composition of the matched sample also broadly resembles that of the overall economy: several of the most common 2-digit ATECO sectors among event firms also rank among the most prevalent sectors in the full firm population, indicating that events are not concentrated in a narrow or atypical set of industries.[10]

Identification assumptions and structural equation. Identification relies on the assumption that, conditional on observables and the matching procedure, worker deaths are as-good-as-random shocks to firms' labor demand for remaining workers and replacement hires. Under this assumption, treated and control firms would have followed parallel trends absent a worker death. I assess the plausibility of this assumption using a stacked event-study difference-in-differences design that allows explicit pretrend tests.

Specifically, I estimate the model over the relative-time window $\tau \in [-3, 1]$, where $\tau = 0$ denotes the death year. For each event i and relative year τ, define event-time indicators

$$D_{i\tau}^{k} = \mathbf{1}\{\tau = k\}, \qquad k \in \{-3, -2, 0, 1\},$$

omitting $k = -1$ as the reference period. Let $Y_{i\tau}$ denote the outcome for event i in relative year τ. I consider two outcomes: (i) log FTE employment excluding the deceased worker, and (ii) the average log wage of all remaining workers, including both incumbents and new hires. In the estimating equation, I express each outcome relative to its value in the year immediately preceding the death, $Y_{i\tau} - Y_{i,-1}$, thereby differencing out time-invariant event-level heterogeneity. These are also the variables that later enter the IV specification, with log survivor employment as the

[10] Among event firms, the five most frequent 2-digit ATECO sectors are repair, maintenance, and installation of machinery and equipment; construction of buildings; wholesale trade (excluding motor vehicles and motorcycles); retail trade (excluding motor vehicles and motorcycles); and specialized construction activities. In the full economy, the five most frequent sectors are retail trade (excluding motor vehicles and motorcycles); food and beverage service activities; repair, maintenance, and installation of machinery and equipment; construction of buildings; and wholesale trade (excluding motor vehicles and motorcycles). Thus, four of the five most common sectors in the event sample also appear in the top five of the full economy, indicating that the event sample remains broadly representative in sectoral terms even if the exact ranking differs.

Table E.2: Pre-Event Characteristics of Deceased and Placebo Workers

	WFE	Age	Experience	Tenure at firm	Log wage
Below-firm-average wage ($L_{ij} = 1$)					
Control ($T_{ij} = 0$)	-0.107	46.72	15.18	8.23	6.131
Death ($T_{ij} = 1$)	-0.089	46.75	16.13	9.41	6.180
Above-firm-average wage ($L_{ij} = 0$)					
Control ($T_{ij} = 0$)	0.098	48.04	18.72	11.31	6.397
Death ($T_{ij} = 1$)	0.103	48.07	18.54	11.33	6.524

	Profession measure	Female (%)
Below-firm-average wage ($L_{ij} = 1$)		
Control ($T_{ij} = 0$)	1.172	14.1
Death ($T_{ij} = 1$)	1.177	14.1
Above-firm-average wage ($L_{ij} = 0$)		
Control ($T_{ij} = 0$)	1.331	11.4
Death ($T_{ij} = 1$)	1.350	11.4

Notes: The table reports pre-event characteristics of treated worker–firm pairs (actual deaths, $T_{ij} = 1$) and matched placebo pairs ($T_{ij} = 0$), separately for workers below and above the firm's pre-event average wage ($L_{ij} = 1$ and $L_{ij} = 0$, respectively). All worker characteristics are measured in the year prior to the event. Female is reported as a percentage.

endogenous quantity variable and average log survivor wages as the price variable.

The stacked event-study specification is

$$Y_{i\tau} - Y_{i,-1} \;=\; \sum_{k \neq -1} \Big[\gamma_k \, D^k_{i\tau} + \beta_k \, D^k_{i\tau} T_i\Big] + X'_{i\tau}\theta + \varepsilon_{i\tau} \tag{E.4}$$

where T_i is an indicator for an actual death event, $X_{i\tau}$ collects time-varying controls for changes in workforce composition, including flexible controls for changes in the average preshock worker fixed effects of survivors and new hires, and standard errors are clustered at the death-event level. Because $k = -1$ is omitted, the coefficients β_k measure treatment effects relative to the year immediately preceding the death. Testing for pretrends therefore amounts to assessing whether β_k for $k < 0$ are statistically and economically close to zero.

Figure E.2 plots the estimated coefficients for survivor employment and survivor wages. For log FTE employment of remaining workers, the pre-event coefficients are very small and statistically insignificant: -0.003 at $k = -3$ and 0.001 at $k = -2$. By contrast, employment rises sharply after the shock, by about 3.3 log points in the event year and 6.9 log points in the following year, both relative to $k = -1$. This increase in FTE labor input among remaining workers is the empirical counterpart of the firm's labor-demand shock after a death, and is absorbed through a combination of replacement hiring and higher labor input supplied by incumbent workers. For average log wages, the pre-event coefficients are again close to zero and statistically insignificant, while postevent responses are more muted: the estimate at $k = 0$ is positive but small and not statisti-

Table E.3: Firm Covariates Before the Shock ($t = -1$)

	Control ($T_{ij} = 0$)	Death ($T_{ij} = 1$)
Panel A: Above-Firm-Average Worker Deaths ($L_{ij} = 0$)		
FFE	-0.008	-0.011
Firm age	19.15	19.46
Employment	15.08	15.06
New hires	1.25	1.25
Incumbents	13.83	13.81
Employment share	0.099	0.102
Panel B: Below-Firm-Average Worker Deaths ($L_{ij} = 1$)		
FFE	-0.004	-0.003
Firm age	19.15	19.73
Employment	16.61	16.52
New hires	1.42	1.21
Incumbents	15.19	15.31
Employment share	0.108	0.105

Notes: The table reports mean firm characteristics in the year before the event ($t = -1$) for treated firms (actual deaths, $T_{ij} = 1$) and matched control firms ($T_{ij} = 0$), separately for deaths of workers above and below the firm's pre-event average wage. FFE denotes the firm fixed effect from the wage regression; firm age is measured in years; employment, new hires, and incumbents are measured in FTE workers. The estimation sample is restricted to firms with employment below 50.

cally distinguishable from zero, whereas by $k = 1$ average log wages rise by about 0.4 log points relative to $k = -1$, and the effect is statistically significant. Taken together, these results support the parallel-trends assumption and suggest that worker deaths induce a sizeable increase in firms' demand for labor input, while the corresponding wage adjustment is comparatively modest, consistent with firms facing a relatively elastic labor supply.

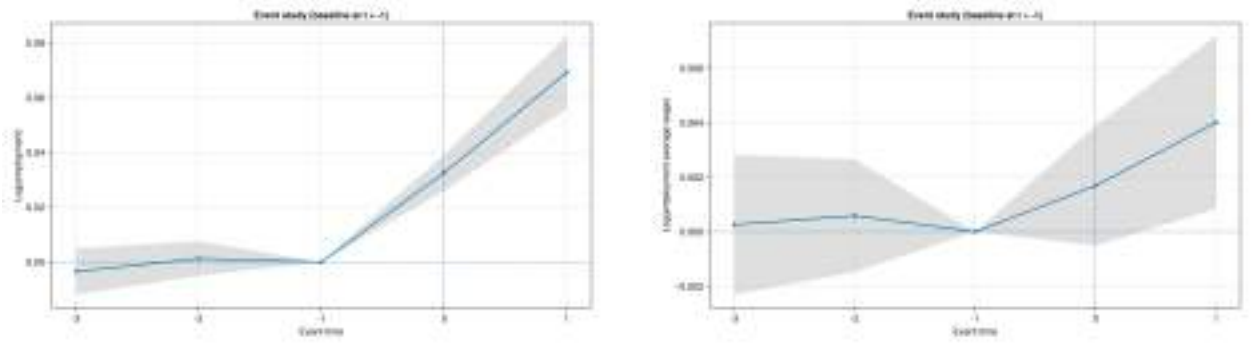
IV estimation and results. To estimate the inverse labor-supply elasticity, I collapse the panel to event-level changes in survivor outcomes between event times $\tau = -1$ and $\tau = 1$. For each matched treated–control death event, I compute

$$\Delta X_i \equiv X_{i,\tau=1} - X_{i,\tau=-1}, \qquad \Delta Y_i \equiv Y_{i,\tau=1} - Y_{i,\tau=-1},$$

where X_i denotes log FTE employment of all workers other than the deceased and Y_i denotes their average log wage. Thus, ΔX_i is the quantity response to the death-induced labor-demand shock, while ΔY_i is the corresponding wage response.

In the preferred specification, I treat both ΔX_i and its interaction with the deceased worker's baseline share in the local market wage bill,

$$xS_i \equiv \Delta X_i \times \text{wshare}_{i,-1},$$



(a) Log FTE employment of survivors (b) Average log wage of survivors

Figure E.2: Event-study estimates for survivor employment and wages around worker deaths

Notes: Each panel reports stacked event-study estimates $\hat{\beta}_k$ from Equation (E.4), using a balanced sample of treated and matched placebo firms observed at all relative times $k \in [-3, 1]$. Outcomes are (a) the log of FTE employment of all workers other than the deceased and (b) their average log wage. These correspond, respectively, to the quantity and price variables used in the IV specification. The omitted period is $k = -1$, so each coefficient is interpreted relative to the year immediately preceding the death. The employment response captures changes in total labor input among remaining workers, including both replacement hiring and changes in labor supplied by incumbent workers. All specifications include flexible controls for changes in workforce composition, and standard errors are clustered at the death-event level.

as endogenous. The second-stage equation is

$$\Delta Y_i = \beta_1 \Delta X_i + \beta_2 xS_i + W_i'\theta + \varepsilon_i \tag{E.5}$$

where β_1 is the inverse labor-supply elasticity of interest and W_i collects controls. The first-stage system for the two endogenous regressors is

$$\Delta X_i = \pi_1 T_i + \pi_2 T_i \times \text{wshare}_{i,-1} + W_i'\delta + u_i, \tag{E.6}$$

$$xS_i = \phi_1 T_i + \phi_2 T_i \times \text{wshare}_{i,-1} + W_i'\kappa + v_i, \tag{E.7}$$

where T_i is an indicator for an actual death in the matched pair and $T_i \times \text{wshare}_{i,-1}$ its interaction with the deceased worker's baseline market wage-bill share. Standard errors are clustered at the death-event level. The control vector W_i always includes a rich set of quantile-bin fixed effects for the change in the average worker fixed effect among survivors,[11] which flexibly controls for composition shifts in the postshock workforce. In the preferred specification, W_i also includes the indicator for a "left-side" death (L_i) and the deceased worker's baseline market wage-bill share, $\text{wshare}_{i,-1}$.

Table E.4 reports the IV estimates from a sequence of specifications. Column (1) presents the baseline 2SLS regression of ΔY_i on ΔX_i, instrumenting ΔX_i with the death indicator T_i and including no further controls or heterogeneous response. The implied inverse labor-supply elasticity

[11] In the implementation, I partition $\Delta\overline{\text{WFE}}_i$ into 220 bins using sample quantiles and absorb the resulting fixed effects. Some bins are empty after sample restrictions, so the number of absorbed categories is slightly smaller in the estimation output.

is $\hat{\beta}_1 = 0.054$ (s.e. 0.017). Column (2) adds flexible controls for composition through the binned change in average worker fixed effects, $\Delta\overline{\text{WFE}}_i$; the point estimate rises to $\hat{\beta}_1 = 0.067$ (s.e. 0.018), consistent with the idea that death shocks induce composition changes that would otherwise attenuate the estimated wage response.

Column (3) adds the left-side indicator L_i and flexible controls for the size of the implied composition shock induced by the death.[12] The estimate remains very similar, $\hat{\beta}_1 = 0.069$ (s.e. 0.018), indicating that the baseline result is not driven by coarse differences in the type of death event or by variation in the mechanical effect of the death on the firm's average worker quality.

Column (4) allows the inverse labor-supply slope to vary with the firm's competitive environment. Specifically, I treat both ΔX_i and its interaction with baseline market wage-bill share, $xS_i = \Delta X_i \times \text{wshare}_{i,-1}$, as endogenous and instrument them jointly with T_i and $T_i \times \text{wshare}_{i,-1}$. This specification also includes L_i and the baseline wage-bill share as controls. The estimated coefficient on ΔX_i is $\hat{\beta}_1 = 0.076$ (s.e. 0.019), while the interaction term is imprecisely estimated, $\hat{\beta}_2 = -0.128$ (s.e. 0.091). I take this as the preferred specification: it controls explicitly for composition, allows the inverse labor-supply slope to vary with the firm's local market power, and remains well identified.

Finally, Column (5) provides a robustness check restricting the sample to firms whose baseline market wage-bill share lies below the median, equal to 1.07%. The specification reduces to a just-identified IV regression of ΔY_i on ΔX_i, instrumented by T_i, while controlling for L_i and the composition fixed effects. The resulting estimate, $\hat{\beta}_1 = 0.068$ (s.e. 0.025), is very close to the corresponding estimates in the full sample.

Weak identification is not a concern. In the exactly identified one-endogenous-regressor specifications, the Kleibergen–Paap rk Wald F-statistics are around 205–215. In the preferred specification with two endogenous regressors and two excluded instruments, the corresponding statistic is 94.95, well above the conventional Stock–Yogo benchmark of 7.03 for the two-by-two case. Substantively, the preferred specification implies a within-market labor-supply elasticity of approximately $\widehat{\eta} = 13.20$. This points to a fairly elastic labor supply facing the firm and is broadly in line with the benchmark estimate reported by Berger et al. (2022) (BHM) for the United States.

E.3.2 Across-Market Elasticity θ

Having estimated the within-market elasticity η, I next calibrate the across-market elasticity θ. Conditional on η, cross-firm heterogeneity in markdowns is governed by the gap $\eta - \theta$, so θ is identified from how labor-market distortions vary with firms' competitive environment. I discipline θ by indirect inference, targeting the cross-sectional relationship between the firm-level labor-market wedge $\widetilde{\psi}_{ij}$ from main text Proposition 2 and either the market HHI or the firm wage-

[12] The implied composition shock is measured as $\text{shock}_i = \text{dist\_logwage}_i/(\text{emp\_baseline}_i - 1)$, where dist_logwage_i denotes the deceased worker's deviation from the firm's pre-event average log wage. Dividing by baseline employment net of the deceased worker rescales this deviation into the counterfactual change in the firm's average log wage that would result mechanically from the worker's removal, absent any behavioral adjustment. I then include flexible bins of this variable as controls.

Table E.4: Estimation Results for Equation (E.5)

	(1) Basic IV	(2) + comp. FE	(3) + shock controls	(4) Preferred	(5) Low-share firms
ΔX (log FTE employment change)	0.0539***	0.0674***	0.0690***	0.0757***	0.0676***
	(0.0173)	(0.0176)	(0.0179)	(0.0193)	(0.0252)
$\Delta X \times$ wagebill share				-0.1276	
				(0.0907)	
Binned $\Delta\overline{\text{WFE}}$ controls	No	Yes	Yes	Yes	Yes
Controls for L	No	No	Yes	Yes	Yes
Controls for baseline wage-bill share	No	No	No	Yes	No
Shock controls (shock$_d$)	No	No	Yes	No	No
Low-share subsample ($<$ median)	No	No	No	No	Yes
K–P rk Wald F	215.0	210.5	205.2	95.0	97.9
Observations	55,538	55,253	55,253	55,253	27,589
Clusters (events)	55,538	55,253	55,253	55,253	27,589

Notes: The dependent variable is the change in the average log wage of all workers other than the deceased between event times $\tau = -1$ and $\tau = 1$, ΔY_i. The main regressor is the corresponding change in log FTE employment of the remaining workforce, ΔX_i. All specifications are estimated by 2SLS without weighting, and standard errors are clustered at the death-event (`event_id`) level. Column (1) instruments ΔX_i with an indicator for an actual death in the matched pair, T_i. Column (2) adds flexible controls for composition through binned values of the change in the average worker fixed effect, $\Delta\overline{\text{WFE}}_i$. Column (3) additionally includes the left-side indicator L_i and flexible controls for the size of the implied composition shock induced by the death. Column (4) is the preferred specification: it instruments both ΔX_i and its interaction with the deceased worker's baseline market wage-bill share using T_i and $T_i \times \text{wshare}_{i,-1}$, and includes controls for $\Delta\overline{\text{WFE}}_i$, L_i, and the baseline wage-bill share. Column (5) restricts the sample to firms with baseline market wage-bill share below the median (1.07%). Kleibergen–Paap rk Wald F-statistics are reported as weak-identification diagnostics. Robust standard errors are in parentheses. *** $p < 0.01$.

bill share within a market. Different values of θ generate different model-implied mappings from local market share to $\widetilde{\psi}_{ij}$, making this relationship highly informative about θ. When firms operate in large markets, competition is strong and $\widetilde{\psi}_{ij}$ is primarily governed by η; in the atomistic limit, $\widetilde{\psi}_{ij} = \eta/(\eta+1)$. By contrast, when firms operate in small markets with limited competition, $\widetilde{\psi}_{ij}$ becomes increasingly governed by θ; in the one-firm-market limit, $\widetilde{\psi}_{ij} = \theta/(\theta+1)$. This strategy is analogous in spirit to Edmond et al. (2023), who infer demand elasticities by matching the relationship between markups and market shares in an oligopolistic setting.

In the simpler environment of main text Proposition 2, the wedge $\widetilde{\psi}_{ij}$ appears directly in the firm's labor share, $ls_{ij} = \alpha\gamma\,\widetilde{\psi}_{ij}$. Lemma E.6 shows how to recover $\widetilde{\psi}_{ij}$ in a richer environment with product-market power and more general production technologies: even when variation in the labor share is not driven solely by labor-market power, the labor-market wedge $\widetilde{\psi}_{ij}$ remains point-identified from the ratio of the flexible-input bill to the wage bill and the ratio of output elasticities. The next lemma formalizes this result and shows how to recover $\widetilde{\psi}_{ij,t}$ from cost shares up to production-function elasticities.

Lemma E.6 (Recovering the wage wedge from cost shares)**.** *Let firm ij in period t produce output $y_{ij,t}$ using labor input $h_{ij,t}$ and a flexible input $x_{ij,t}$. Let $\alpha_{l,ij}$ and $\alpha_{m,ij}$ denote the output elasticities of $y_{ij,t}$ with respect to $h_{ij,t}$ and $x_{ij,t}$, respectively.*

Suppose that the marginal product of a worker of type a can be written as

$$MPL_{ij,t}(a) = \frac{\partial y_{ij,t}}{\partial n_{ij,t}(a)} = \frac{\partial y_{ij,t}}{\partial h_{ij,t}}\,\psi_{ij,t}(a).$$

Assume $\psi_{ij,t}(a) > 0$ on the employment support. Assume that the following normalization holds, as in the model developed in the main text:[13]

$$\frac{1}{h_{ij,t}} \sum_{a \in \mathcal{A}} \psi_{ij,t}(a)\, n_{ij,t}(a) = 1.$$

Let the firm face inverse labor supply schedules $w_{ij,t}(a, n_{ij,t}(a))$, and define

$$\epsilon_{ij,t}(a) := \left[\frac{\partial \log w_{ij,t}(a)}{\partial \log n_{ij,t}(a)}\right]^{-1}, \qquad \mu_{ij,t}(a) := \frac{\epsilon_{ij,t}(a)}{\epsilon_{ij,t}(a)+1}.$$

Define

$$\widetilde{\psi}_{ij,t} := \frac{1}{h_{ij,t}} \sum_{a \in \mathcal{A}} \mu_{ij,t}(a)\, \psi_{ij,t}(a)\, n_{ij,t}(a), \qquad \overline{w}_{ij,t} := \frac{1}{h_{ij,t}} \sum_{a \in \mathcal{A}} w_{ij,t}(a)\, n_{ij,t}(a).$$

If the flexible input $x_{ij,t}$ has marginal price p_x, then

$$\frac{1}{\widetilde{\psi}_{ij,t}} = \frac{p_x x_{ij,t}}{\overline{w}_{ij,t}\, h_{ij,t}} \frac{\alpha_{l,ij}}{\alpha_{m,ij}} \tag{E.8}$$

Thus, under the maintained assumptions, cost shares recover the markdown-adjusted firm-level wage wedge $\widetilde{\psi}_{ij,t}$. In the special case with no markdowns, $\mu_{ij,t}(a) = 1$ for all a, so $\widetilde{\psi}_{ij,t} = 1$. More generally $0 \leq \widetilde{\psi}_{ij,t} \leq 1$. Moreover, the following decomposition holds

Under the maintained assumptions of the lemma,

$$\widetilde{\psi}_{ij,t} = \mathbb{E}_{g_{ij,t}}\big[\mu_{ij,t}(a)\psi_{ij,t}(a)\big] = \overline{\mu}_{ij,t} + \mathrm{Cov}_{g_{ij,t}}\big(\mu_{ij,t}(a), \psi_{ij,t}(a)\big),$$

where

$$\overline{\mu}_{ij,t} := \frac{1}{h_{ij,t}} \sum_{a \in \mathcal{A}} \mu_{ij,t}(a)\, n_{ij,t}(a), \qquad g_{ij,t}(a) := \frac{n_{ij,t}(a)}{h_{ij,t}}.$$

Proof of Lemma E.6. Fix a firm ij and period t, and suppress indices.

The firm faces inverse labor supply schedules $w(a, n(a))$ and chooses $\{n(a)\}_{a \in \mathcal{A}}$ and x to produce a target output Q at minimum variable expenditure:

$$\min_{\{n(a)\}, x} \quad p_x x + \sum_{a \in \mathcal{A}} w(a, n(a))\, n(a) \quad \text{s.t.} \quad y(\{n(a)\}, x) \geq Q.$$

[13] This normalization also holds when labor is measured in efficiency units, $h_{ij,t} = \sum_{a \in \mathcal{A}} \phi(a)\, n_{ij,t}(a)$, since then $MPL_{ij,t}(a) = \frac{\partial y_{ij,t}}{\partial h_{ij,t}} \phi(a)$, so $\psi_{ij,t}(a) = \phi(a)$ and $\frac{1}{h_{ij,t}} \sum_a \psi_{ij,t}(a) n_{ij,t}(a) = 1$.

Let $\lambda > 0$ be the multiplier on the output constraint. The first-order condition for x is

$$p_x = \lambda \frac{\partial y}{\partial x}.$$

For each type a, the first-order condition for $n(a)$ is

$$w(a) + n(a)\frac{\partial w(a)}{\partial n(a)} = \lambda \frac{\partial y}{\partial n(a)}.$$

Using

$$\frac{n(a)}{w(a)}\frac{\partial w(a)}{\partial n(a)} = \frac{1}{\epsilon(a)}, \qquad \mu(a) = \frac{\epsilon(a)}{\epsilon(a)+1},$$

this becomes

$$w(a)\Big(1 + \frac{1}{\epsilon(a)}\Big) = \lambda \frac{\partial y}{\partial n(a)} \quad \Longleftrightarrow \quad w(a) = \lambda\, \mu(a) \frac{\partial y}{\partial n(a)}.$$

Using

$$\frac{\partial y}{\partial n(a)} = \frac{\partial y}{\partial h}\, \psi(a),$$

we obtain

$$w(a) = \lambda\, \mu(a) \frac{\partial y}{\partial h}\, \psi(a).$$

Multiplying by $n(a)$ and summing over a gives

$$\overline{w}\, h = \lambda \frac{\partial y}{\partial h}\, \widetilde{\psi}\, h,$$

hence

$$\frac{\partial y}{\partial h} = \frac{\overline{w}}{\lambda\, \widetilde{\psi}}.$$

Therefore

$$\alpha_l = \frac{\partial y}{\partial h}\frac{h}{y} = \frac{\overline{w}}{\lambda\, \widetilde{\psi}}\frac{h}{y}, \qquad \alpha_m = \frac{\partial y}{\partial x}\frac{x}{y} = \frac{p_x}{\lambda}\frac{x}{y}.$$

Taking the ratio yields

$$\frac{\alpha_l}{\alpha_m} = \frac{\overline{w}\, h}{p_x x}\frac{1}{\widetilde{\psi}},$$

which rearranges to (E.8).

If $\mu(a) = 1$ for all a, then

$$\widetilde{\psi} = \frac{1}{h}\sum_{a \in \mathcal{A}} \psi(a) n(a) = 1$$

by the assumed normalization. More generally, if $\psi(a) \geq 0$ on the support of employment, then $0 \leq \mu(a) \leq 1$ implies

$$0 \leq \widetilde{\psi} = \sum_a g(a)\mu(a)\psi(a) \leq \sum_a g(a)\psi(a) = 1.$$

Finally, by definition,

$$\widetilde{\psi} = \mathbb{E}_g[\mu(a)\psi(a)].$$

Using the maintained normalization $\mathbb{E}_g[\psi(a)] = 1$, we obtain

$$\widetilde{\psi} = \mathbb{E}_g[\mu(a)]\,\mathbb{E}_g[\psi(a)] + \mathrm{Cov}_g\big(\mu(a), \psi(a)\big) = \mathbb{E}_g[\mu(a)] + \mathrm{Cov}_g\big(\mu(a), \psi(a)\big).$$

□

In the data, I proxy flexible-input expenditure with intermediate-input expenditure. Following De Ridder et al. (2026), I estimate how this wedge covaries with firms' competitive environment by taking logarithms of (E.8) and sufficiently controlling for heterogeneity in production elasticities. In the benchmark case of a Cobb–Douglas, time-invariant production function with elasticities that vary at the three-digit industry level—and under a competitive intermediate-input market—this amounts to including three-digit industry fixed effects.

To calibrate θ, I therefore use indirect inference to target the empirical coefficient β in the panel regression

$$\log\left(\frac{m_{ij,t}}{\overline{w}_{ij,t}h_{ij,t}}\right) = \beta\, HHI_{j,t} + \Psi_{ij,t} + \varepsilon_{ij,t} \tag{E.9}$$

where the dependent variable is the logarithm of the material-to-labor expenditure ratio, $HHI_{j,t}$ measures the firm's competitive environment through the local wage-bill HHI, and $\varepsilon_{ij,t}$ is an idiosyncratic error term. Standard errors are clustered at the market level, that is, commuting zone by three-digit industry.

The vector $\Psi_{ij,t}$ contains a rich set of controls absorbing cross-sectional heterogeneity in production technology, potential distortions in intermediate-input markets, and aggregate time effects. In particular, I include firm fixed effects, which absorb time-invariant firm-specific output elasticities as well as time-invariant firm wedges in the flexible input, and calendar-year fixed effects to capture aggregate shocks.

Two concerns arise in this specification. First, measurement error in the firm wage bill induces a mechanical negative bias in β, since the wage bill appears both in the denominator of the dependent variable and in the numerator of the regressor. Second, unobserved labor-augmenting productivity not fully absorbed by $\Psi_{ij,t}$ may also bias the coefficient downward: firms with higher labor-augmenting productivity both command larger wage-bill shares and exhibit lower measured $\widetilde{\psi}_{ij,t}$, because a given degree of labor-market power translates into a lower material-to-labor ratio (Rubens et al., 2026).

To address these concerns, I instrument the labor-market concentration measure with product-market concentration, namely the revenue-based HHI measured within the local labor market. This is natural for three reasons.

First, under the model's assumptions, revenue concentration is a valid instrument satisfying both relevance and exclusion. Revenue concentration is an informative signal of labor-market concentration because both measures load on the same underlying market structure; accordingly,

the revenue-based HHI is highly correlated with the wage-bill HHI and is a strong instrument in both the data and the model. Empirically, this relevance condition is confirmed by the very large first-stage Kleibergen–Paap F-statistics reported in Tables E.5 and E.6. At the same time, under the model's structure, product-market concentration affects the composite wage wedge $\widetilde{\psi}_{ij,t}$ only indirectly, through its effect on labor-market concentration. The exclusion restriction therefore holds: revenue concentration does not enter the measured wedge except through labor-market concentration.

Second, the same logic extends to richer environments with heterogeneous product-market power. As emphasized by Gutiérrez (2023), variation in product-market concentration shifts markups through the revenue share, whereas the labor wedge is governed by labor-market concentration. Hence, even with firm-specific markups and heterogeneous demand conditions, the revenue-based concentration index remains a valid instrument for labor-market concentration because it affects the labor wedge only through its correlation with the wage-bill HHI. This argument is reinforced by the broader markup literature, which stresses that the economically relevant competitive environment for product-market power need not coincide with the sectoral classification typically available in administrative data (e.g., Pellegrino, 2025).

Third, in a more general empirical environment, the revenue-based HHI helps address the mechanical attenuation bias generated by the presence of the wage bill in both the dependent variable and the regressor by relying on variation not mechanically induced by the wage bill itself. A remaining concern is that changes in revenue concentration, or in the associated firm-level revenue share, might be correlated with labor-biased technological change that directly affects $\widetilde{\psi}_{ij,t}$. I view this channel as unlikely to be quantitatively important beyond what is already absorbed by firm fixed effects and other controls. Moreover, to the extent that it remains, it would induce a negative correlation between the instrument and the residual in Equation (E.9), biasing the estimated coefficient on concentration downward. In this sense, the IV estimates can be interpreted as conservative with respect to the strength of the relationship between labor-market concentration and the material-to-labor expenditure ratio.

As a robustness check, I also estimate an analogous specification in which the regressor of interest is the firm-level wage-bill share, instrumented by the firm-level revenue share. The second-stage equation is

$$\log\left(\frac{m_{ij,t}}{\overline{w}_{ij,t}h_{ij,t}}\right) = \beta\, WBS_{ij,t} + \Psi_{ij,t} + \varepsilon_{ij,t} \tag{E.10}$$

where $WBS_{ij,t}$ denotes the firm-level wage-bill share. In the model, the two regressions imply approximately the same β. This additional specification therefore provides an extra over-identifying restriction on the calibration of θ and a useful validation check.

Table E.5 reports results for specifications in which the regressor of interest is the wage-bill HHI, while Table E.6 reports the corresponding specifications where the regressor is the wage-bill share. In Table E.5, Column (1) reports the estimate of β from a specification with market (commuting-zone-by-industry) and calendar-year fixed effects only; Column (2) adds firm fixed

effects in place of market fixed effects; and Column (3) reports the IV estimates instrumenting the wage-bill HHI with the revenue-based HHI. In Table E.6, Column (1) reports the estimate of β from a specification with three-digit industry and calendar-year fixed effects, instrumenting the wage-bill share with the revenue share; Column (2) includes firm and calendar-year fixed effects and instruments the wage-bill share with the wage-bill HHI; and Column (3) includes firm and calendar-year fixed effects and instruments the wage-bill share with the revenue-based HHI.

The pattern across specifications is consistent with the biases discussed above. A simple OLS regression with only market fixed effects (Column (1) of Table E.5) delivers a coefficient of $\beta \approx 0.15$, plausibly attenuated by the mechanical correlation between the wage bill in the dependent variable and in the regressor, and by omitted labor-augmenting productivity shocks. Introducing firm fixed effects (Column (2) of Table E.5) raises the coefficient to about 0.23, indicating that part of the bias is driven by time-invariant firm-specific components. Instrumenting the wage-bill HHI with the revenue-based HHI (Column (3) of Table E.5) yields a substantially larger coefficient, around 0.41. The three IV regressions in which the regressor is the firm-level wage-bill share—instrumented in turn by the firm-level revenue share, the wage-bill HHI, and the revenue-based HHI—deliver very similar estimates, all in the range $[0.41, 0.45]$. Taken together, these results point to a robust elasticity in this interval. I treat Column (3) of Table E.5, which yields $\beta \approx 0.41$, as the preferred specification and target in the calibration.

After the joint calibration of θ together with the other structural parameters, the implied value of θ is approximately 1.51.

Table E.5: Relationship Between Material-to-Labor Expenditures and Market Concentration

	(1) Market FE	(2) Firm FE (OLS)	(3) Firm FE (IV)
Dependent variable:		$\log\left(\frac{m_{ij,t}}{\overline{w}_{ij,t}h_{ij,t}}\right)$	
HHI index	0.147***	0.228***	0.414***
	(0.0199)	(0.0151)	(0.1378)
First stage: Kleibergen–Paap F			1.5×10^3
Fixed effects and controls:			
Market FE	✓		
Firm FE		✓	✓
Calendar-year FE	✓	✓	✓
IV:			
Instrument	×	×	Revenue HHI
Observations	1,525,628	1,407,714	1,407,714
Clusters (markets)	51,483	50,915	50,915

Notes: The table reports estimates of Equation (E.9), where the regressor of interest is the market-level wage-bill HHI index. The dependent variable is the logarithm of the material-to-labor expenditure ratio. Standard errors are clustered at the market level. Columns (1) and (2) report OLS estimates with market and firm fixed effects, respectively. Column (3) reports the corresponding IV estimate, instrumenting the wage-bill HHI with the market-level revenue-based HHI; the reported Kleibergen–Paap statistic is the rk Wald F. *** $p < 0.01$.

Table E.6: Relationship Between Material-to-Labor Expenditures and Wage-Bill Share

	(1) 3-digit industry FE	(2) Firm FE (HHI IV)	(3) Firm FE (Rev-HHI IV)
Dependent variable:		$\log\left(\frac{m_{ij,t}}{\overline{w}_{ij,t}h_{ij,t}}\right)$	
Wage-bill share	0.410***	0.419***	0.454***
	(0.0191)	(0.0288)	(0.1527)
First stage: Kleibergen–Paap F	4.7×10^4	6.5×10^3	9.8×10^2
Fixed effects and controls:			
3-digit industry FE	✓		
Firm FE		✓	✓
Calendar-year FE	✓	✓	✓
IV:			
Instrument	Revenue share	Wage-bill HHI	Revenue HHI
Observations	1,530,021	1,407,714	1,407,714
Clusters (markets)	55,876	50,915	50,915

Notes: The table reports IV (2SLS) estimates of Equation (E.10), where the regressor of interest is the firm-level wage-bill share. The dependent variable is the logarithm of the material-to-labor expenditure ratio. Standard errors are clustered at the market level. Column (1) includes three-digit industry and calendar-year fixed effects and instruments the wage-bill share with the revenue share. Column (2) includes firm and calendar-year fixed effects and instruments the wage-bill share with the wage-bill HHI index. Column (3) includes firm and calendar-year fixed effects and instruments the wage-bill share with the revenue-based HHI index. Reported Kleibergen–Paap statistics are the rk Wald F-statistics for the excluded instrument in each column. *** $p < 0.01$.

E.4 Simulation of the Model-Implied Panel Dataset

This subsection describes how I construct the synthetic worker–firm panel used in the quantitative analysis. The simulation is designed to satisfy three requirements simultaneously: (i) every model firm appears in the panel with its correct general-equilibrium characteristics; (ii) the cross-sectional distribution of workers across firms and ability types is representative of the model economy; and (iii) the discrete worker identifiers used to compute AKM and screening-threshold moments behave empirically like the worker identifiers in the administrative data, with roughly eight worker IDs and six FTE workers per firm on average.

Steady state and interpretation. As in the main text, I focus on a stationary equilibrium in which employment by ability type is constant over time. This can be interpreted as an environment in which, in each period, a random fraction of workers separates and is replaced by new hires of the same type. The equilibrium allocation

$$\{n_{ij}(a),\, w_{ij}(a),\, y_{ij},\, h_{ij},\, \widetilde{\psi}_{ij}\}_{i,j,a}$$

therefore describes both a stock allocation and the basis for flow probabilities: the main text inverse labor-supply relationships in (2) and their microfoundation in Appendix A.1 apply directly to the flow of new hires.

Challenges. Two practical issues arise when simulating a worker–firm panel from this allocation. First, the model contains no primitive notion of a finite number of workers per firm: firms employ continuous masses of workers, whereas in the administrative data firms employ discrete worker IDs, each supplying a firm-specific number of hours. Second, a purely worker-based simulation that samples workers and assigns them to firms proportionally to $n_{ij}(a)$ is representative of workers but not of firms: many small firms receive no synthetic workers and therefore disappear from the panel, which is problematic for representing the full distribution of firm types.

The simulation procedure below combines the advantages of both approaches. It keeps all firms generated by the general-equilibrium solution, allocates a firm-specific number of worker IDs proportional to equilibrium employment, and attaches worker-level sampling weights so that the discrete panel reproduces the model's continuous employment structure in expectation.

Firm-specific worker identifiers and weights. Let F be the total number of firms in the model. For convenience, I pool all firms across local labor markets and index them by a single index $i \in \{1, \dots, F\}$, so that each i corresponds to a firm–market pair (i, j) in the main text. Let H_i denote the model-implied employment (in FTE units) of firm i. Let $\bar{W}$ and $\bar{H}$ denote the target average numbers of worker IDs and FTE workers per firm in the simulated panel, calibrated from the administrative data (approximately $\bar{W} = 8$ and $\bar{H} = 6$).

The panel uses $W \equiv \bar{W}F$ synthetic worker IDs in total. These IDs are allocated across firms in two steps. First, each firm receives one guaranteed worker ID. Second, the remaining $R \equiv W - F$ IDs are allocated across firms through a multinomial draw with probabilities proportional to equilibrium employment:

$$(X_1, \dots, X_F) \sim \text{Multinomial}\left(R,\ \left(\frac{H_1}{\sum_{i'} H_{i'}}, \dots, \frac{H_F}{\sum_{i'} H_{i'}}\right)\right).$$

The resulting number of worker IDs at firm i is

$$m_i \equiv 1 + X_i, \qquad i = 1, \dots, F,$$

so that $\sum_i m_i = W$ exactly and $\mathbb{E}[m_i] \propto H_i$.

Each worker ID at firm i is interpreted as representing h_i FTE workers, where

$$h_i = \frac{\kappa H_i}{m_i}, \qquad \kappa \equiv \frac{\bar{H}F}{\sum_{i'} H_{i'}}.$$

The scale factor κ is chosen so that average employment equals $\bar{H}$. Thus, each firm has on average $\bar{W}$ worker IDs and $\bar{H}$ FTE workers in the panel, mirroring the empirical sample, while the cross-

sectional distribution of FTE employment remains proportional to the model-implied H_i.

Initial assignment and permanent abilities. Worker abilities are discretized into S types, with type-s ability denoted a_s. For each firm i, the equilibrium allocation implies a within-firm type distribution

$$p_i(s) \equiv \frac{n_i(a_s)}{\sum_{s'} n_i(a_{s'})} = \frac{n_i(a_s)}{H_i}.$$

In the first period of the panel, firm i draws m_i worker IDs independently from the categorical distribution $p_i(\cdot)$. Each draw is interpreted as a permanent ability type: worker k in firm i is assigned type index s and ability level a_s, and her wage at firm (i, j) is set to the equilibrium wage $w_{ij}(a_s)$. This step ensures that, in expectation, the within-firm composition of worker IDs by ability coincides with the model-implied composition, while preserving the desired average number of IDs per firm.

Dynamic reassignment. Starting in the second period, worker IDs may move across firms. Microfoundationally, this corresponds to workers drawing new idiosyncratic preference shocks in the discrete-choice foundation of the nested-CES labor supply in Appendix A.1. Let p^{switch} denote the empirical mean switching rate over 2015–2019. In each period $t \geq 2$, each worker ID independently switches employer with probability p^{switch}; workers who do not switch remain with their current firm.

For a given period, let S_i denote the number of stayers at firm i, so that $m_i - S_i$ is the number of vacancies. All workers who switch are placed in a common pool. Vacancies are then filled firm by firm. For each vacancy at firm i, I draw a *target* type $\tilde{s}$ from the within-firm distribution $p_i(\cdot)$ and assign to firm i a worker from the pool whose permanent type index is $\tilde{s}$. If no such worker is available, I assign the available worker whose type index is closest to $\tilde{s}$ on the discretized ability grid. Each worker ID therefore keeps her permanent type a_s over time but may move across firms. This procedure preserves, in expectation, both the firm-specific headcount m_i and the model-implied within-firm type composition in each cross section, while matching the empirical switching rate.

Panel structure and representativeness. For each year and worker ID, the simulation records the worker identifier, period, permanent ability type, assigned firm and market, wage and log wage, and the matched firm's productivity, size, output, wedge, and concentration. Each worker–period observation carries the weight h_i corresponding to its employing firm. Stacking these observations over all periods yields a worker–firm panel with T periods and $\sum_i m_i$ worker IDs.

By construction, this panel reproduces the empirical horizon, the average number of worker IDs and FTE workers per firm, and the mean switching rate. Moreover, weighted cross-sectional moments computed on the panel—including the variance and covariance structure of wages, AKM worker and firm fixed effects, employment shares by worker–firm deciles, and concentration measures by worker ability—coincide with their model-implied counterparts up to a common

scale factor. At the same time, the allocation of IDs across firms is approximately proportional to firm employment, so that deciles defined on worker identifiers, as in the administrative data, are close to population deciles of the model's worker distribution. This makes the synthetic panel informative for both worker-level and firm-level moments used in the quantitative evaluation.

Table E.7: Empirical Targets Used in the Panel Simulation

	Value	**Source**
Average worker IDs per firm ($\bar{W}$)	8.0	Derived from administrative data
Average FTE workers per firm ($\bar{H}$)	6.07	Derived from administrative data
Panel years	2015–2019 ($T = 5$)	Administrative data
Mean annual switching rate	0.294	Worker mobility statistics (2015–2019)

Notes: The simulation uses these empirical moments to discipline the size of the worker and firm populations, the panel length, and the probability that workers switch employers from one year to the next. The targets $\bar{W}$ and $\bar{H}$ pin down, respectively, the average number of worker identifiers and FTE workers per firm in the simulated panel, so that the discrete worker IDs behave like those in the administrative data. In the baseline simulations, the annual switching probability p^{switch} is fixed at the empirical mean rate 0.294 computed over 2015–2019.

E.4.1 Numerical Calibration Procedure

This subsection describes how the parameters are numerically disciplined. The calibration targets and their empirical construction are summarized in the main text Section 3.

Free parameters and fixed objects. The vector of free parameters is

$$\Theta = (\rho, \omega_a, \sigma_a, \sigma_z, \theta),$$

where ρ and ω_a govern complementarities and the weight of worker ability in production, σ_a and σ_z are the dispersions of worker and firm types, and θ is the across-market labor-supply elasticity. All other parameters—including (α, γ), the preference parameters, and the within-market elasticity η—are held fixed at the values obtained from the production-function estimation and the external calibration described in the main text Section 3.

Equilibrium computation. For any candidate Θ, I recompute the stationary competitive equilibrium of the model. Worker abilities are discretized into S points on an equally spaced grid of percentiles of a log-normal distribution with mean μ_a normalized to zero and standard deviation σ_a. Given $(\sigma_z, \rho, \omega_a)$ and the fixed distribution of firm draws, I solve for the fixed point in wages and employment using the nested-CES labor-supply system, and then recover the full set of equilibrium allocations

$$\left\{n_{ij}(a),\, w_{ij}(a),\, y_{ij},\, h_{ij},\, \widetilde{\psi}_{ij}\right\}_{i,j,a}$$

using the procedures in Appendix C. This yields the complete general-equilibrium allocation for the candidate Θ.

Some model-implied moments are computed directly from these equilibrium objects, without simulation. In particular, I construct the aggregate labor share, the dispersion of firm employment, and the pooled regression of $-\log\widetilde{\psi}_{ij}$ on the wage-bill share and the market HHI. These are the model counterparts of the concentration and wedge regressions estimated in the data.

Synthetic panel and AKM moments. Moments that depend on the joint distribution of worker and firm AKM fixed effects are computed on a synthetic worker–firm panel simulated from the equilibrium. The simulation procedure is described in Appendix E.4.

The simulated panel matches the empirical horizon (five years) and the worker-to-firm ratio used in the AKM estimation. AKM worker and firm fixed effects, the top–top employment shares, concentration measures by worker decile, and the screening-threshold regressions are then computed on the synthetic data using exactly the same estimators and code as in the empirical analysis.[14] To prevent simulation noise from contaminating the objective, each worker–period is assigned a fixed pseudo-random seed, so that the simulated panel is a deterministic function of Θ.

Objective function and weighting. Let $\widehat{m}$ denote the vector of empirical moments and $m(\Theta)$ the corresponding vector of model-implied moments. The calibration chooses Θ to minimize the quadratic distance

$$Q(\Theta) = \big[m(\Theta) - \widehat{m}\big]^{\top} W \big[m(\Theta) - \widehat{m}\big], \tag{E.11}$$

where W is a diagonal weighting matrix. In the baseline specification, targeted moments receive positive weights reflecting their relative importance for identifying the structural parameters, while the remaining moments receive zero weight and serve as untargeted validation checks. Moments related to worker heterogeneity, firm-size dispersion, sorting, and labor-market concentration receive the highest priority, whereas wage dispersion is left untargeted. Thus, the criterion is an indirect-inference/SMM objective in which a small set of key moments is targeted and the remaining empirical facts are used for validation.

Numerical implementation. Each evaluation of $Q(\Theta)$ requires solving for the full general equilibrium and, in the baseline specification, simulating a large worker–firm panel and re-estimating the AKM decomposition. I therefore adopt a two-step procedure. First, I explore a finite set of economically plausible parameter vectors by direct evaluation of $Q(\Theta)$, using different initial conditions for $(\rho, \omega_a, \sigma_a, \sigma_z, \theta)$ motivated by the production-side estimates and by simple comparative statics. This step yields a collection of candidate parameter vectors and their objective values. Second, I take the candidate with the lowest objective value as the starting point for a local refinement using a Nelder–Mead direct-search algorithm.

[14] In practice, the simulated panel is exported to a flat file and processed with the same Stata routines used for the administrative data, ensuring that empirical and model-implied moments are constructed identically.

The free parameters are constrained to lie in economically meaningful intervals, for example $\theta \in (0, \eta]$ and $\sigma_a, \sigma_z > 0$. To enforce these bounds, I reparameterize each element θ_k of Θ as

$$\theta_k = a_k + \frac{b_k - a_k}{1 + \exp(-x_k)},$$

where $[a_k, b_k]$ is the admissible domain and $x_k \in \mathbb{R}$ is unconstrained. The optimization is performed over x; for each trial value, I map back to Θ, recompute the equilibrium and the simulated moments, and evaluate $Q(\Theta)$. Because the mapping from x to Θ is smooth and strictly monotone in each coordinate, it converts the constrained problem into an unconstrained one without changing the location of local minima. In the reported calibration, I run the Nelder–Mead algorithm once from the best-performing initial condition found in the first step. The parameter vector reported in the main text Section 3.1.3 is the minimizer of (E.11) obtained by this procedure.[15]

E.5 Calibration Summary

Table E.8 reports the parametrization used in the main exercise of the paper.

Table E.8: Baseline parameter values

Symbol	Value	Description	Symbol	Value	Description
γ	0.78	Labor elasticity in $(k^{1-\gamma}h^{\gamma})^{\alpha}$	α	0.94	Returns to scale in the composite input
ρ	0.41	CES parameter in worker–firm technology $\phi(a, z)$	ω_a	0.81	Weight on worker ability in $\phi(a, z)$
η	13.2	Within-market labor supply elasticity	θ	1.51	Across-market labor supply elasticity
μ_a	0.00	Mean of worker ability (normalized)	μ_z	0.00	Mean of firm productivity (normalized)
σ_a	0.23	Std. dev. of worker ability distribution	σ_z	0.75	Std. dev. of firm productivity distribution
R	0.10	Real interest rate	σ	0.83	Curvature of utility from consumption
φ	0.50	Parameter governing the Frisch elasticity	δ	0.08	Capital depreciation rate
S	500	Number of worker types	M	1000	Number of local labor markets
$m_{\min}$	1.00	Minimum of firms-per-market Pareto distribution	κ_m	1.07	Tail parameter of firms-per-market Pareto
λ_m	4.10	Scale parameter of firms-per-market Pareto	p_1	0.19	Share of markets with a single firm
$\bar{m}$	200	Maximum number of firms per market			

Notes: (γ, α) govern the capital–labor production technology; (ρ, ω_a) parameterize the CES worker–firm technology $\phi(a, z)$; (η, θ) govern within- and across-market labor supply elasticities; (μ_a, σ_a) and (μ_z, σ_z) describe the distributions of worker ability and firm productivity, with $\mu_a = \mu_z = 0$ as normalizations; $(R, \sigma, \varphi, \delta)$ govern preferences and labor-supply behavior; S and M are simulation dimensions; $(m_{\min}, \kappa_m, \lambda_m)$ parameterize the Pareto distribution of the number of firms per local labor market; p_1 and $\bar{m}$ summarize the cross-market distribution of firm counts.

F Model Replication of Empirical Evidence

This appendix replicates, in the calibrated model, the empirical facts documented in Section 2 of the main text, using the parameterization reported in Table E.8. It concludes by comparing model-implied additional untargeted moments to their empirical counterparts. All model-based objects

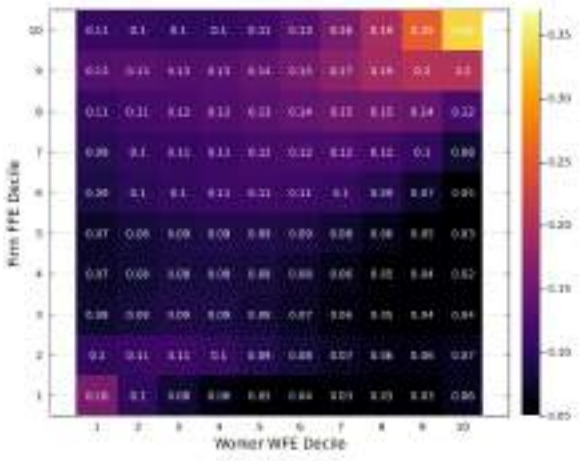
[15] All computations are implemented in Julia; the equilibrium and simulation routines are parallelized across local labor markets. The calibration is numerically costly but robust: nearby alternative starting values for the second-stage search yield essentially identical objective values and parameter vectors.

are computed from the stationary equilibrium of the calibrated economy. AKM-based statistics are obtained from the synthetic worker–firm panel described in Appendix E.4. Throughout, I use the same discretizations and ranking procedures as in the empirical analysis so that model and data objects are directly comparable.

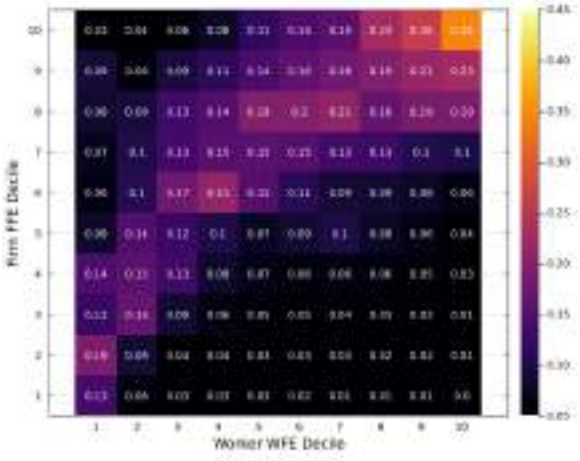
F.1 Market Shares

To assess the model's ability to reproduce the joint distribution of workers and firms, I construct worker–firm employment matrices from the simulated panel that mirror those in the main text Figure 3a. Workers are sorted into deciles of the empirical distribution of AKM worker fixed effects, and firms are ranked either by their AKM firm fixed effect, by average log wages, or by the average log wage of coworkers. For each definition of firm rank, I compute the share of total employment accounted for by each worker–firm decile pair.

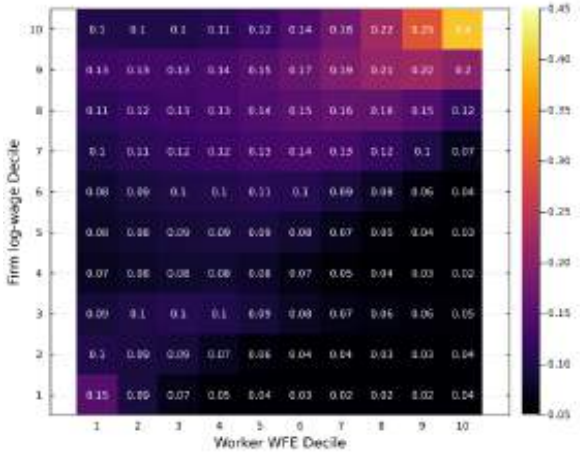
Figure F.1 compares the empirical worker-side employment matrices with their model-implied counterparts when firms are ranked by AKM fixed effects and by average log wages. The left column reports the empirical matrices, and the right column the corresponding matrices from the calibrated model. In each panel, rows correspond to worker AKM deciles and columns to firm deciles, and cell entries report employment shares. As discussed in the main text, the mean absolute error for the AKM-based matrix is about 2.9 percentage points, with a correlation of 0.82 between empirical and model cell entries. For the wage-based matrix, the mean absolute error is about 3.5 percentage points and the correlation is 0.88.



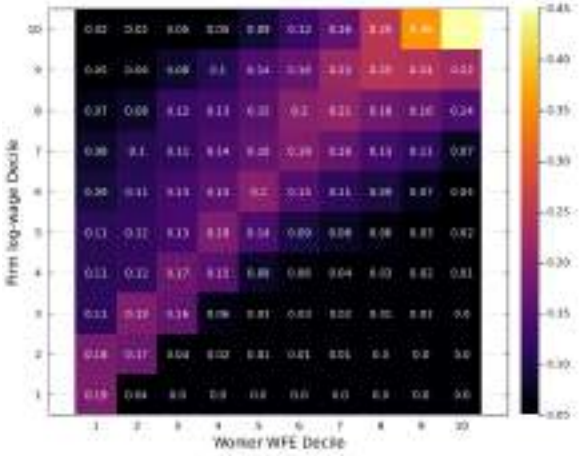
(a) Data: firms ranked by AKM FFE



(b) Model: firms ranked by AKM FFE



(c) Data: firms ranked by average log wages

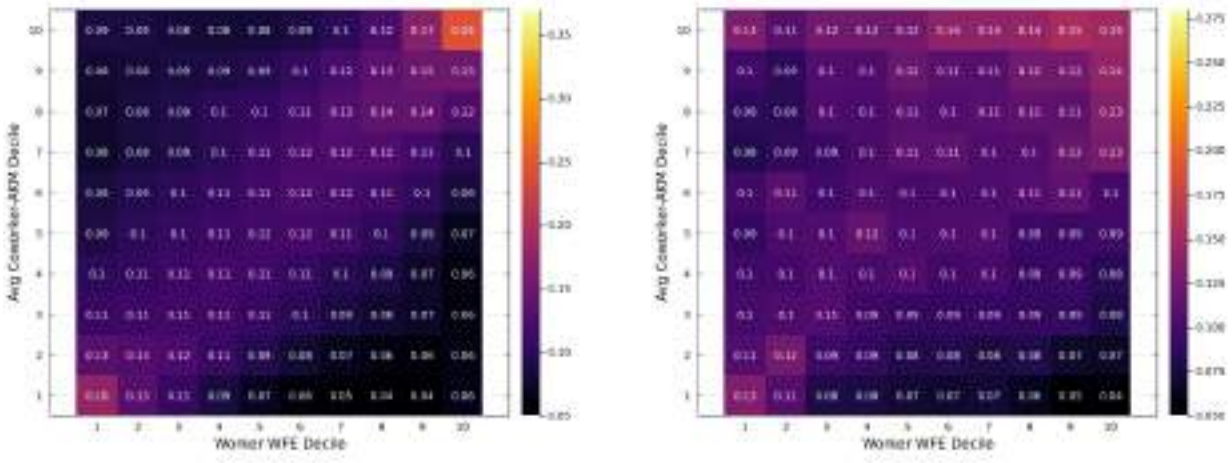


(d) Model: firms ranked by average log wages

Figure F.1: Worker–firm employment shares: data vs. model

Notes: Each panel reports the share of total employment by worker AKM decile (rows) and firm decile (columns). The left column displays the empirical matrices constructed from the Italian employer–employee data; the right column displays the corresponding matrices constructed from the simulated worker–firm panel of the calibrated model. In the top row, firms are ranked by their AKM firm fixed effect; in the bottom row, by the average log wage of their incumbent workforce. Darker cells correspond to higher employment shares.

Figure F.2 compares the empirical and model-implied coworker employment matrices, where coworkers are ranked by their average log wage. In this case, the mean absolute error between empirical and model cell entries is about 1.5 percentage points and the correlation is 0.69.



(a) Data: coworker matrix

(b) Model: coworker matrix

Figure F.2: Coworker employment shares: data vs. model

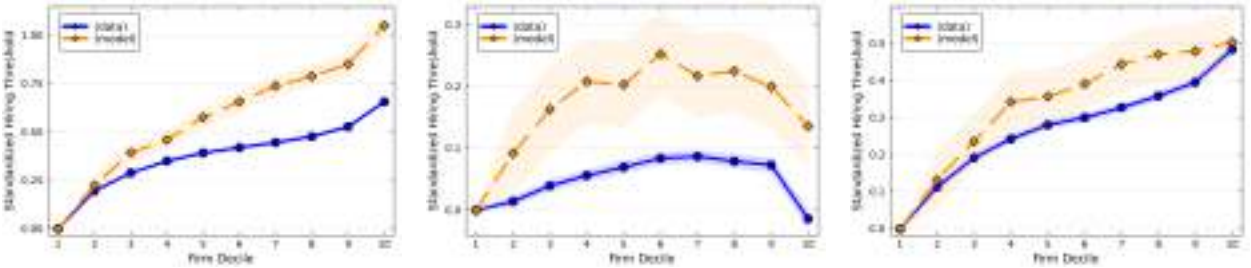
Notes: Each panel reports the share of coworker pairs by worker-decile pair, where rows correspond to the decile of the worker of interest and columns to the decile of a randomly chosen coworker from the same firm. The left panel is constructed from the empirical Italian worker–firm data; the right panel is constructed from the simulated panel of the calibrated model, using the same decile cutoffs. Darker cells indicate more frequent coworker combinations.

F.2 Hiring Thresholds

I next examine the extent to which the calibrated model reproduces the empirical hiring-threshold profiles. For each firm, I compute the minimum AKM worker fixed effect among newly hired workers and regress this measure on three notions of firm rank, as in the main text Equation (16). I rank firms by (i) their AKM firm fixed effect, (ii) the average AKM worker fixed effect of their incumbent workforce, and (iii) the average incumbent log wage. I then replicate exactly the same procedure in the simulated panel, using the same ranking definitions, sample restrictions, and regression specification.

Figure F.3 plots the empirical and model-implied hiring thresholds by firm decile for the three ranking schemes. In each panel, the solid line denotes the empirical series and the dashed line the model-implied series. As discussed in the main text, the model broadly matches both the levels and the gradients of the hiring-threshold profiles. When firms are ranked by their AKM fixed effect, the model generates a somewhat steeper slope than in the data: in the top decile, the model-implied threshold exceeds its empirical counterpart by about 0.4 local standard deviations of worker fixed effects, and the mean absolute deviation across deciles is about 0.2 standard deviations. For rankings based on average incumbent worker fixed effects and average incumbent log wages, the corresponding mean absolute deviations are about 0.12 and 0.06, while the upper-tail differences are 0.15 and 0.01, respectively. The model-implied series also feature wider confidence intervals than in the data. Overall, the calibrated economy reproduces both the level and the shape of the empirical hiring-threshold profiles, with some overstatement of selectivity in the upper tail

under AKM-based firm rankings.



(a) Firm Rank: AKM FFE deciles (b) Firm Rank: Avg. incumbent AKM deciles (c) Firm Rank: Avg. incumbent log-wage deciles

Figure F.3: Hiring-threshold profiles: data vs. model

Notes: Each panel plots the minimum AKM worker fixed effect among new hires by firm decile, as estimated from main text Equation (16). Solid lines report the empirical series constructed from the Italian worker–firm data, with 95% confidence bands; dashed lines report the corresponding series constructed from the simulated panel of the calibrated model. In panel (a), firms are ranked by their AKM firm fixed effect; in panel (b), by the average AKM worker fixed effect of their incumbent workforce; and in panel (c), by the average incumbent log wage. All thresholds are expressed in local standard deviations of worker fixed effects within the corresponding market and sample.

F.3 Concentration Indices by Worker AKM

Finally, I evaluate the model's ability to replicate concentration across workers of different AKM deciles. For each worker AKM decile, I compute a market–decile wage-bill Herfindahl–Hirschman index (HHI) across firms, both in the empirical data and in the simulated panel. The aggregate HHI is then formed using employment weights, following the same procedure as in the main text Section 2. This yields a series of ten concentration indices by worker AKM decile in both data and model.

Figure F.4 plots the empirical and model-implied HHI profiles by worker AKM decile. The solid line reports the empirical series, and the dashed line the corresponding series from the calibrated model. In levels, the model slightly overpredicts concentration: the average difference between model and data is about 0.04 HHI points, with per-decile gaps in the range 0.033–0.047. The correlation between the two series is high, around 0.98, and the model reproduces the same U-shaped pattern as in the data, with higher concentration in the tails than in the middle of the worker distribution.

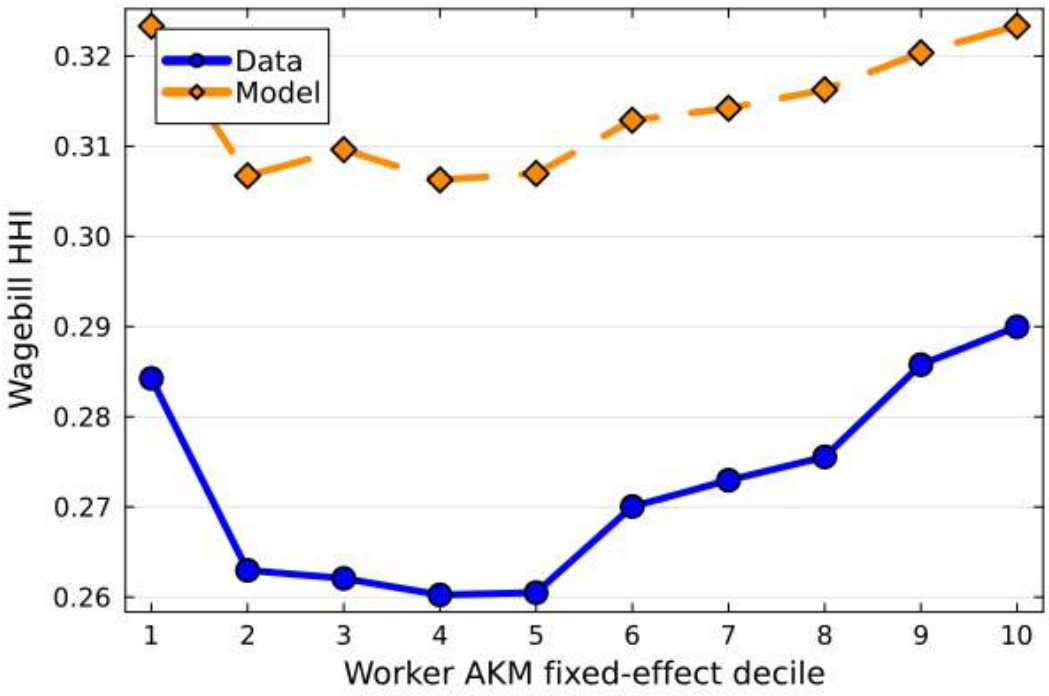


Figure F.4: Concentration by worker AKM decile: data vs. model

Notes: The figure plots the wage-bill HHI within worker deciles, averaged across LLMs using employment weights. The solid line reports the empirical series constructed from the Italian employer–employee data; the dashed line reports the corresponding series from the simulated worker–firm panel of the calibrated model. For each worker decile, the HHI is computed over firm-level wage-bill shares within that decile.

F.4 Additional Moments

Table F.1 reports the targeted moments in the data and the corresponding moments implied by the model. Together, these moments discipline dispersion in firm size and worker heterogeneity, patterns of sorting and hiring thresholds, and the link between labor wedges and concentration. Table F.2 reports a broader set of untargeted moments for the baseline calibration and the three BHM calibrations, covering firm-size dispersion, labor shares, wage dispersion, and variance decompositions.

G Additional Tables for the Quantitative Analysis

G.1 Aggregate Production

Table G.1 reports average wedges and average markdowns by firm-type decile in the baseline calibration. Both decline with firm type, albeit somewhat nonmonotonically, showing that more productive firms face lower labor wedges and wider markdown distortions on average. The average wedge $\tilde{\psi}_{ij}$ falls from 0.914 in the bottom decile to 0.880 in the top decile, while the average markdown $\overline{\mu}_{ij}$ falls from 0.913 to 0.883. The gap between the two is small in every decile, suggest-

Table F.1: Targeted Calibration Moments

Moment	Data	Model
Dispersion in wages and firm size		
Standard deviation of firm log employment	1.50	1.51
Standard deviation of AKM worker fixed effects	0.261	0.260
Sorting, segregation, and thresholds		
Share of top-decile workers employed by top-decile firms[a]	0.337	0.336
Hiring-threshold gradient w.r.t. firm rank[b]	0.045	0.045
Labor market power and concentration		
Labor-wedge gradient w.r.t. wage-bill HHI[c]	0.414	0.416

Notes: This table reports the moments targeted in the baseline calibration. All moments use industry–commuting-zone local labor markets and FTE weights unless otherwise noted.
[a] Fraction of top-decile workers employed by top-decile firms.
[b] Slope from a linear version of Equation (16), ranking firms by average incumbent log wage.
[c] Coefficient on wage-bill HHI in a regression of labor wedges on HHI.

Table F.2: Untargeted Validation Moments

Moment	Data	Baseline	BHM1	BHM2	BHM3
Std. dev. of log employment	1.50	1.51	2.56	1.53	1.50
Aggregate labor share	0.52	0.61	0.44	0.47	0.605
Variance of log wages	0.154	0.134	0.35	0.24	0.059
Within-firm component of wage variance	0.063	0.065	n/a	n/a	n/a
Between-firm within-LLM component	0.037	0.016	0.00044	0.0017	0.0026
Between-LLM component of wage variance	0.054	0.054	0.35	0.24	0.056
Std. dev. of firm fixed effects	0.185	0.228	n/a	n/a	n/a
Covariance(worker FE, firm FE)	0.018	0.006	n/a	n/a	n/a

Notes: This table reports untargeted validation moments for the baseline model and the three homogeneous-worker benchmarks. All moments use industry–commuting-zone local labor markets and FTE weights unless otherwise noted. AKM moments are not defined in the homogeneous-workers benchmarks.

ing that the covariance term embedded in $\tilde{\psi}_{ij}$ plays only a minor quantitative role overall.

Table G.1: Average Wedge and Markdown by Firm-Type Decile

	Firm-type decile (latent z_{ij}, lowest to highest)									
	1	2	3	4	5	6	7	8	9	10
Baseline: $\tilde{\psi}_{ij}$	0.914	0.907	0.907	0.903	0.898	0.900	0.896	0.888	0.883	0.880
Baseline: $\overline{\mu}_{ij}$	0.913	0.906	0.906	0.902	0.898	0.900	0.896	0.889	0.885	0.883

Notes: This table reports unweighted mean values of $\tilde{\psi}_{ij}$ and $\overline{\mu}_{ij}$ by decile of latent firm productivity z_{ij} in the simulated baseline economy.

As an accounting exercise, I distinguish three margins of misallocation: aggregate labor supply, the allocation of labor across local labor markets, and the allocation of labor across firms within markets. I then reset one margin at a time to its efficient value while holding the other two at their distorted values, so that the reported entries measure the output loss that remains after correcting that particular margin. Formally, the equilibrium allocation can be summarized by $\{S(a), s_j(a), s_{ij}(a)\}$, where $S(a)$ denotes aggregate labor supply for worker type a, $s_j(a)$ the share of type-a labor allocated to market j, and $s_{ij}(a)$ the share allocated to firm i within market j. At the firm level, distortions affect both firm size (*size misallocation*) and the within-firm allocation of worker ability (*misallocation of talent*).[16] In the baseline economy, correcting within-market allocation reduces the output loss only from 2.32% to 2.08%, whereas correcting the allocation of labor across markets or aggregate labor supply lowers the residual loss to 1.33% and 1.27%, respectively. The main inefficiencies in the heterogeneous-worker baseline therefore operate through cross-market allocation and aggregate labor supply rather than within-market misallocation.

Table G.2: Aggregate Inefficiency and Decomposition

	Baseline	BHM1	BHM2	BHM3
Total output loss	2.32	8.44	5.82	2.47
Residual output loss after fixing:				
Market-supply distortion[a]	1.33	6.82	3.79	1.56
Within-market distortion[b]	2.08	4.66	4.56	2.03
Aggregate-supply distortion[c]	1.27	5.60	3.38	1.38

Notes: This table reports percentage output losses relative to the production-efficient allocation. BHM1 uses the original BHM parameterization, BHM2 replaces only σ_z, and BHM3 additionally replaces η and θ.

[a] $s_j(a)$ set to its efficient value, holding $S(a)$ and $s_{ij}(a)$ at distorted values.

[b] $s_{ij}(a)$ set to its efficient value, holding $S(a)$ and $s_j(a)$ at distorted values.

[c] $S(a)$ set to its efficient value, holding $s_j(a)$ and $s_{ij}(a)$ at distorted values.

[16] Because these objects interact nonlinearly in equilibrium, this is not an exact additive decomposition of total misallocation; the entries should therefore be interpreted as heuristic partial contributions.

G.2 Wage Inequality

Table G.3 compares wage dispersion in the decentralized equilibrium and in the efficient allocation without labor market power. Overall wage inequality falls when markdowns are removed: the standard deviation of log wages declines from 0.366 to 0.322. This decline is driven primarily by a sharp compression in between-firm wage dispersion, which falls from 0.069 to 0.040. In particular, the between-local-labor-market component of wage variance drops from 0.054 to 0.019, while the within-local-labor-market component rises modestly from 0.016 to 0.021. Consistent with markdowns being especially large on top wages, the AKM variance of worker effects increases slightly from 0.068 to 0.070. Consistent with stronger assortative matching in the efficient allocation, the covariance between worker and firm effects rises from 0.006 to 0.008. Thus, removing markdowns steepens wage differences within markets but compresses average wages across local labor markets by enough to reduce overall wage inequality.

Table G.3: Wage Inequality in the Decentralized and Efficient Allocations

	Decentralized	Efficient
Std. dev. of log wages	0.366	0.322
Within-firm component	0.065	0.064
Between-firm component	0.069	0.040
Between local labor markets	0.054	0.019
Within local labor markets	0.016	0.021
AKM variance of worker effects $\mathrm{Var}(\alpha)$	0.068	0.070
AKM variance of firm effects $\mathrm{Var}(\psi)$	0.052	0.016
AKM covariance $\mathrm{Cov}(\alpha, \psi)$	0.006	0.008

Notes: This table compares model-implied wage dispersion in the decentralized equilibrium and the efficient allocation. The middle block decomposes wage variance into within-firm and between-firm components, with the latter split into between- and within-local-labor-market variation.

Appendix References

Abowd, John M, Francis Kramarz, and David N Margolis (1999). "High Wage Workers and High Wage Firms". In: *Econometrica* 67.2, pp. 251–334.

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer (2015). "Identification Properties of Recent Production Function Estimators". In: *Econometrica* 83.6, pp. 2411–2451.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S Graff Zivin (2019). "Does Science Advance One Funeral at a Time?" In: *American Economic Review* 109.8, pp. 2889–2920.

Barkai, Simcha (2020). "Declining Labor and Capital Shares". In: *The Journal of Finance* 75.5, pp. 2421–2463.

Bennedsen, Morten, Francisco Pérez-González, and Daniel Wolfenzon (2020). "Do CEOs Matter? Evidence from Hospitalization Events". In: *The Journal of Finance* 75.4, pp. 1877–1911.

Berger, David, Kyle Herkenhoff, and Simon Mongey (2022). "Labor Market Power". In: *American Economic Review* 112.4, pp. 1147–93.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2022). "Discretizing Unobserved Heterogeneity". In: *Econometrica* 90.2, pp. 625–643.

Callaway, Brantly and Pedro HC Sant'Anna (2021). "Difference-in-Differences with Multiple Time Periods". In: *Journal of Econometrics* 225.2, pp. 200–230.

Card, David, Jörg Heining, and Patrick Kline (2013). "Workplace Heterogeneity and the Rise of West German Wage Inequality". In: *The Quarterly Journal of Economics* 128.3, pp. 967–1015.

Collard-Wexler, Allan and Jan De Loecker (2016). *Production Function Estimation and Capital Measurement Error*. NBER Working Paper 22437. National Bureau of Economic Research.

Costinot, Arnaud and Jonathan Vogel (2010). "Matching and Inequality in the World Economy". In: *Journal of Political Economy* 118.4, pp. 747–786.

De Ridder, Maarten, Basile Grassi, and Giovanni Morzenti (2026). "The Hitchhiker's Guide to Markup Estimation: Assessing Estimates from Financial Data". In: *Econometrica* 94.1, pp. 137–168.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu (2023). "How Costly Are Markups?" In: *Journal of Political Economy* 131.7, pp. 1619–1675.

Eeckhout, Jan and Philipp Kircher (2018). "Assortative Matching With Large Firms". In: *Econometrica* 86.1, pp. 85–132.

Feenstra, Robert C, Robert Inklaar, and Marcel P Timmer (2015). "The Next Generation of the Penn World Table". In: *American Economic Review* 105.10, pp. 3150–82.

Felix, Mayara (Mar. 2026). *Trade, Labor Market Concentration, and Wages*. NBER Working Paper 35018. National Bureau of Economic Research.

Goodman-Bacon, Andrew (2021). "Difference-in-Differences with Variation in Treatment Timing". In: *Journal of Econometrics* 225.2, pp. 254–277.

Gutiérrez, Agustín (2023). *Labor Market Power and the Pro-competitive Gains from Trade*. Working paper, April 22, 2023 version.

Helpman, Elhanan, Oleg Itskhoki, and Stephen Redding (2010). "Inequality and Unemployment in a Global Economy". In: *Econometrica* 78.4, pp. 1239–1283.

Jäger, Simon, Jörg Heining, and Nathan Lazarus (2024). "How Substitutable Are Workers? Evidence from Worker Deaths". American Economic Review, conditionally accepted.

Jaravel, Xavier, Neviana Petkova, and Alex Bell (2018). "Team-Specific Capital and Innovation". In: *American Economic Review* 108.4-5, pp. 1034–1073.

Levinsohn, James and Amil Petrin (2003). "Estimating Production Functions using Inputs to Control for Unobservables". In: *The Review of Economic Studies* 70.2, pp. 317–341.

Lochner, Benjamin, Stefanie Wolter, and Stefan Seth (2024). "AKM Effects for German Labour Market Data from 1985 to 2021". In: *Journal of Economics and Statistics* 244.4, pp. 425–431.

McFadden, Daniel (1974). "Conditional Logit Analysis of Qualitative Choice Behavior". In: *Frontiers in Econometrics*. Ed. by Paul Zarembka. New York: Academic Press, pp. 105–142.

Pellegrino, Bruno (2025). "Product Differentiation and Oligopoly: A Network Approach". In: *American Economic Review* 115.4, pp. 1170–1225.

Rubens, Michael, Yingjie Wu, and Mingzhi Xu (2026). "Exploiting or Augmenting Labor?" In: *American Economic Review: Insights* 8.1, pp. 72–89.

Saint-Paul, Gilles (2001). "On the Distribution of Income and Worker Assignment under Intrafirm Spillovers, with an Application to Ideas and Networks". In: *Journal of Political Economy* 109.1, pp. 1–37.

Sauvagnat, Julien and Fabiano Schivardi (2024). "Are Executives in Short Supply? Evidence from Death Events". In: *Review of Economic Studies* 91.1, pp. 519–559.

Schmidtlein, Lisa, Stefan Seth, and Philipp vom Berge (2020). *Sample of Integrated Employer Employee Data (SIEED) 1975–2018*. FDZ-Datenreport. Documentation on Labour Market Data 14/2020 (en). Institut f"ur Arbeitsmarkt- und Berufsforschung (IAB).

Song, Jae, David J Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter (2019). "Firming Up Inequality". In: *The Quarterly Journal of Economics* 134.1, pp. 1–50.