Who Competes for Whom?
Monopsony in Segmented Labor Markets

Luca Lorenzini*
UCLA, Anderson School of Management

First version: February 2024. This version: December 2025
Link to the latest version
[Preliminary: not for circulation]

Abstract

This paper develops and tests a quantitative general-equilibrium model to study how sorting and segmentation in the labor market shape firms' monopsony power, aggregate efficiency, and the distribution of welfare across heterogeneous workers. Firms (workers) are heterogeneous in productivity (ability), and intrafirm spillovers imply that firm productivity depends on the average quality of the workforce. This mechanism creates a trade-off between firm size and composition: hiring lower-ability workers reduces a firm's average productivity, endogenizing labor market segmentation by worker ability. These forces generate localized competition, with firms primarily competing against others targeting similar workers. Less (more) productive firms exert greater labor-market power over lower- (higher-) ability workers. Using matched employer–employee data for Italy and Germany, I document that (i) high-paying firms impose higher hiring thresholds, (ii) low- (high-) paid workers are disproportionately employed in low- (high-) paying, smaller (larger) firms, (iii) concentration indices are non-monotonic across the worker-pay distribution, and (iv) exogenous increases (decreases) in workforce average quality causally raise (lower) firm-level productivity. Calibrated to the data, the model reproduces these moments and shows that, relative to a homogeneous-labor benchmark, segmentation reduces aggregate misallocation by weakening the link between firm size and market power. Competition is weakest at the tails of the ability distribution, where workers face more concentrated labor markets and larger welfare losses.

*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, Michael Rubens, Gianluca Violante, Brian Wheaton, and many others, as well as participants at UCLA seminars, the Midwest Macro Meeting, SED, RIDGE, and other 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. It also uses the Sample of Integrated Employer–Employee Data (SIEED 7518) from the German Institute for Employment Research (IAB); data access was provided via remote data access under project number `fdz2701`. This paper received the Consultaccount Award for Best Paper presented by a PhD student at the 17th PEJ Annual Meeting.

Contents

1 Introduction

The study of labor market power has attracted growing attention in research and policy for its potential implications for wage suppression, aggregate efficiency, and welfare. The U.S. Treasury's *State of Labor Market Competition Report* (2022) argues that employer concentration and anticompetitive practices can depress productivity, reduce wages and exacerbate income disparities. A complementary and growing body of research links rising wage inequality to patterns of worker–firm sorting and ability segmentation,[1] documenting that high-ability workers tend to match with high-wage firms while low-ability workers cluster in low-wage firms (e.g., Song et al., 2019). Yet we know far less about how sorting and segmentation determine *who competes for whom* in the labor market—and, in turn, how these forces govern firms' labor-market power. This paper asks: how do sorting and segmentation influence competition for workers across firms? What are the aggregate efficiency costs of labor-market power, and which workers—low, middle, or high ability—bear them?

To answer these questions, I develop and test a quantitative general-equilibrium model of monopsony in segmented labor markets, where firms compete for workers who differ in ability and where competition is shaped by labor market sorting and segmentation. The model is deliberately parsimonious: with a small number of interpretable parameters, it reproduces key empirical regularities and isolates a clear mechanism through which sorting generates segmentation, and segmentation shapes monopsony distortions.

Sorting arises from production complementarities between worker ability and firm productivity, while intrafirm spillovers generate segmentation. Workers not only contribute to output but also occupy a scarce, firm-specific resource subject to decreasing returns—for example, office workspace. Given this production structure, hiring an additional worker affects output through two distinct channels. The first is a size effect: expanding employment directly raises output. The second operates through firm-level productivity—or, equivalently, the quality of the firm's output—which now depends on the average output of its workers.

As a result, a worker's marginal product depends on the ability mix of coworkers: adding a low-ability worker is particularly costly for high-productivity firms employing many high-ability workers. This interaction gives rise to a size–quality trade-off: high-productivity firms exclude low-ability workers, narrowing their employment options—i.e., their choice sets—and segmenting the labor market into distinct ability clusters. Low-ability workers become crowded into low-productivity firms—although these firms are smaller—while high-ability workers cluster in high-productivity firms. From a competition standpoint, segmentation localizes rivalry: firms primarily compete with others of similar productivity that target the same range of worker abilities.

I bring the model to administrative linked employer–employee data from Italy (including also balance-sheet information) and Germany. Four empirical patterns emerge. (i) High-productivity

[1] Sorting refers to high-ability workers disproportionately employed by high-wage (or high-productivity) firms, whereas segmentation describes high- and low-ability workers clustering in separate firms.

firms exhibit higher ability thresholds for hiring. (ii) low- (high-) ability workers are disproportionately employed in low- (high-) productivity, smaller (larger) firms. (iii) Ability-specific concentration indices—the sufficient statistic for competition over each ability type—are non-monotonic across the ability distribution. (iv) Exogenous shocks to workforce composition, using unexpected worker deaths, causally affect firm-level revenue productivity: positive (negative) changes in the average ability of the workforce raise (lower) productivity.

A calibrated version of the model, estimated on these moments, replicates the empirical patterns and quantifies the efficiency and distributional consequences of monopsony. Two main messages arise. *First*, distortions may be disconnected from firm size. Because firms compete over ability-specific choice sets, even small, low-productivity firms can exert substantial monopsony power over the workers they employ. Segmentation therefore weakens the link between distortions and productivity, reducing aggregate misallocation and narrowing the efficiency losses. *Second*, competition is weakest at both tails of the ability distribution. High-ability workers receive disproportionately many offers from a small set of top firms, generating highly concentrated markets and amplifying welfare losses. Low-ability workers face severely restricted choice sets, which likewise concentrate their labor markets among lower-quality, smaller firms and raise their welfare losses. Crucially, left-tail workers have limited consumption bundles to begin with—reflecting their lower marginal product of labor—and these are further eroded by imperfect competition.

Model Overview. The model features three types of agents: workers, firms, and entrepreneurs. There is a continuum of workers who differ in latent ability, and a continuum of local labor markets, each containing a heterogeneous number of firms. Within each market, firms are granular and differ in their baseline productivity exogenous type, which is drawn from a distribution. All firms are owned by a representative entrepreneur, who receives income from firms' profits and capital rents, and determines aggregate capital accumulation. Since product market power is not central to the analysis, final goods are assumed to be perfect substitutes, implying no markups.[2]

Worker supply is modeled through a *preference heterogeneity* framework, building on D. Berger et al. (2022, hereafter BHM). Workers first choose a local labor market and then select a firm from their available choice set within that market. These choices maximize indirect utility, which depends on wages and idiosyncratic taste shocks that capture factors such as moving costs, commuting distance, or preferences for firm culture. From the worker's perspective, preference heterogeneity implies that jobs are differentiated. Consequently, firms face upward-sloping labor supply curves for each worker type, which they internalize when setting employment levels.

Firms engage in strategic interactions within each local labor market, competing à la Cournot by choosing employment schedules across worker ability types. Heterogeneity in markdowns is modeled following Atkeson and Burstein (2008). Larger labor market power arises when a firm

[2]Production is allowed to exhibit decreasing returns to scale. Hence, monopolistic competition with a constant markup could be incorporated without loss of generality. In that case, firms would choose prices that apply a constant markup over marginal cost and optimize a decreasing-return revenue function instead of a decreasing-return production function. All results would remain unchanged under this alternative formulation.

commands a substantial market share for a given worker type as firms compete over workers' *choice sets*. Intuitively, when a firm is large from the perspective of a worker—relative to the worker's available alternatives—it internalizes the lower elasticity of substitution across markets and can therefore exert greater market power.

Optimal wages take the form of firm–worker-specific markdowns relative to the efficient wage, defined as the worker's marginal product of labor.[3] In each market, concentration indices are directly linked to the average markdown by worker ability and therefore capture the intensity of competition faced by different worker types. Throughout the paper, *aggregate production inefficiency* refers to the loss in total output arising from the misallocation of labor across firms due to market power. In contrast, *welfare losses* are evaluated at the level of worker ability and entrepreneurs, comparing their respective outcomes to those under the efficient allocation that would prevail in the absence of labor market power.

Production and Segmentation. The key innovation of the model lies in the firm's production function, which generates labor market segmentation and endogenizes workers' choice sets as a function of their abilities. A firm's realized productivity[4] is endogenous, determined by the *average* output produced by its workforce, where each worker's output depends on their ability and the firm's exogenous type. Employing higher-ability (lower-ability) workers increases (reduces) a firm's productivity by raising (lowering) the average output of its workforce. Under this production structure, hiring an additional worker affects output through two channels. The first is a *size effect*: expanding employment directly increases output. The second operates through *firm-level productivity*, which depends on the average output of the existing workforce. This second effect can offset the first: if the new worker's output falls below the firm's current average, overall productivity declines, and the firm may optimally choose not to hire even at a one-cent wage.

The model nests three transparent benchmark environments with distinct empirical signatures, obtained by taking limiting cases of key parameters. First, when worker output depends solely on firm productivity rather than worker ability, the model nests the BHM benchmark, which assumes homogeneous labor in production. In this case, market shares and concentration indices are constant across ability: more productive firms are uniformly larger for every worker type, and labor market power distortions are entirely driven by firm heterogeneity. Second, when worker output is supermodular but not log-supermodular—yielding a production function similar to that in Helpman et al. (2010)—the equilibrium allocation of workers to firms, and therefore firms' labor market power, coincides with the homogeneous-labor case up to a renormalization. *Third*, when

[3] Throughout the paper, I refer to $w = MPL$ as the *efficient wage*, where MPL denotes the marginal revenue product of labor. These wages are efficient in the sense that they would implement the Planner's allocation in a decentralized economy. Under labor market power, the actual wage becomes a markdown over the MPL, expressed as $w = \mu MPL$, where μ is the firm's endogenous markdown. A value of μ closer to one implies a smaller markdown and thus a wage nearer to the competitive benchmark. Conversely, a lower μ indicates stronger market power and a larger markdown. In general, μ depends on the firm-level labor supply elasticity ϵ, as $w = \frac{\epsilon}{\epsilon+1} MPL$. Hence, a lower elasticity leads to a smaller μ and a greater markdown. Larger markdowns do not necessarily imply lower wages, since wages depend on both μ and MPL: a firm with a sufficiently high MPL may still offer high wages despite a sizable markdown.

[4] Or product quality, as the two concepts are isomorphic in the context of this model.

the production function exhibits no decreasing returns to labor, within-firm spillovers vanish and the production function collapses to the specification in Costinot and Vogel (2010)[5]. If worker output is log-supermodular, the allocation of workers to firms is characterized by sorting but not segmentation: high-productivity firms hire relatively more high-ability workers, but also more of every worker type. Consequently, concentration indices rise with ability, welfare losses increase with ability, and labor-market distortions remain closely linked to firm productivity.

When within-firm spillovers are present and worker output is log-supermodular, the resulting size–quality trade-off induces segmentation in the labor market. In equilibrium, depending on the calibration, high-ability workers cluster in high-wage firms (*sorting*), while low-ability workers face limited employment opportunities and are concentrated in low-type firms (*segmentation*). From the perspective of these workers, such firms may appear large and exert considerable market power, even if they are small relative to the aggregate economy, as competition occurs within their ability-specific choice sets. This segmentation localizes competition: firms primarily compete with others of similar exogenous productivity that target the same range of worker abilities.

Empirical Evidence. I present empirical evidence consistent with the model's key mechanisms. These patterns, drawn from matched employer–employee data for Italy and Germany, can be viewed either as stand-alone empirical facts or through the lens of the model, and they provide moments for calibration and validation.

The main dataset is the matched employer–employee panel for Italy (INPS, 1974–2024), merged with CERVED balance-sheet data (1996–2018). As a complementary source, I use the German Sample of Integrated Employer–Employee Data (SIEED) from the Institute for Employment Research (IAB). The Italian analysis focuses on non-managerial workers and defines firms at the location–commodity–sector level. Results are reported under two alternative definitions of local labor markets—(i) 3-digit industry × commuting zone and (ii) 3-digit occupation × commuting zone—while the German analysis uses occupation–location markets.[6]

Worker and firm unobservable heterogeneity are inferred from wage variation using a two-way fixed-effects decomposition à la Abowd et al. (1999, AKM), extended with a K-means clustering of firms. The empirical analysis can be viewed through two complementary lenses. From a model-free perspective, the estimated worker and firm-cluster effects capture persistent heterogeneity in pay outcomes, summarizing stable differences in wages across workers and firms. From a model-based perspective, these fixed effects are interpreted as empirical signals of the underlying structural types. I show that the within–local labor market rankings based on AKM fixed effects closely reproduce the ordering of model-implied worker and firm types, supporting their use as valid *indirect inference* measures of rank heterogeneity.

[5]This does not necessarily correspond to the same revenue function in Costinot and Vogel (2010), particularly when monopolistic competition in product markets is assumed.

[6]In the version of the SIEED data to which I have access, industry codes are available only at the 2-digit level, motivating the use of occupation–location markets.

Fact 1: Hiring Thresholds. Higher-ranked firms hire workers characterized by higher minimum worker fixed effects. The minimum fixed effect rises almost linearly with firm rank: top-decile firms hire workers about 0.6–0.8 standard deviations in the worker fixed effect distribution above those in the bottom decile. The relationship is stronger when local markets are defined by occupation and remains robust to alternative specifications and controls for hiring intensity. The same pattern is observed in the German data.

Fact 2: Segmentation. Employment-share matrices computed *within local labor markets* show that low-rank workers concentrate in low-rank, smaller firms, while high-rank workers cluster in high-rank, larger firms. Bottom-decile workers allocate about three times more employment to bottom-decile firms than to top-decile firms, and the reverse holds for top-decile workers. The pattern displays a gradual and ordered shift: as worker rank increases, employment moves steadily from lower- to higher-rank firms. Segmentation is sharper under the occupation-based market definition and persists across alternative measures of firm rank—such as revenue productivity, average wage, or co-worker composition. The same pattern is observed in the German data, indicating that segmentation is a pervasive and robust feature of labor markets.

Fact 3: Concentration and Worker Rank. Labor market concentration varies non-monotonically along the worker fixed-effect distribution. Defining local markets by occupation, wage-bill Herfindahl–Hirschman indices (HHIs), computed by worker-rank decile within markets, fall from 0.17 among low-rank workers to 0.13 in the middle deciles before rising again to about 0.16 for top-rank workers, displaying a clear U-shape. Under the industry definition, concentration is flatter but follows the same qualitative pattern, declining from roughly 0.28 to 0.26 in mid-deciles and increasing to about 0.32 at the top. Further disaggregation of industry markets by occupational group (white- versus blue-collar) shows higher average concentration among white-collar jobs (0.36 versus 0.30) and a similarly U-shaped profile. Overall, concentration is strongest for workers at both tails of the rank distribution and weakest for those in the middle.

Fact 4: Productivity Effects of Worker Deaths. Finally, I test causally whether firm revenue productivity[7] responds to changes in the firm's average worker fixed effect. Using unexpected deaths of above- versus below-average workers as quasi-random shocks, I estimate a stacked difference-in-differences design comparing treated firms to matched controls. Following the death of a below- (above-) average worker, the firm's average worker fixed effect increases (decreases), and revenue productivity rises (falls) by about 3–5% over the subsequent two to three years. The positive productivity response to the exit of a below-average worker is consistent with congestion effects: the exit of a low-type worker raises total factor revenue productivity. Taken together, these results provide direct causal evidence that firm-level productivity is shaped by the firm's average worker fixed effect, in line with the model's within-firm spillover mechanism.

[7] Firm revenue productivity is computed using standard production function estimation following Ackerberg et al. (2015) and the available balance sheet data.

Model calibration. I calibrate the model in three steps, drawing on the Italian matched employer–employee and firm balance-sheet data and defining local labor markets as industry–commuting-zone cells, a definition that enables the use of detailed firm-level balance-sheet information[8].

First, a small set of macro and technology parameters is fixed externally, and production-function parameters are estimated directly from firm-level data.

Second, I estimate the two labor-supply elasticities governing worker mobility within and across markets. The within-market elasticity is identified through a quasi-experimental design exploiting unexpected worker deaths as exogenous firm-level labor-demand shocks affecting new hires. The across-market elasticity is calibrated to match the cross-sectional relationship between firm-level labor wedges and market shares. In the model, the firm's average wage equals the average marginal product times a markdown wedge—combining the mean markdown with a covariance term between markdowns and worker type—and equals one in the absence of markdowns.[9] Firms in smaller markets have larger market shares and thus face wedges farther from one, as weaker competition allows them to impose larger markdowns. The across-market elasticity is therefore disciplined by the empirical slope of labor wedges with respect to market shares for a given estimate of the within market elasticity.[10] [11] Finally, the remaining parameters—governing worker and firm heterogeneity and the degree of complementarities in production—are calibrated by indirect inference. I simulate a synthetic employer–employee panel from the model and jointly match key moments: (i) the standard deviation of worker fixed effects, (ii) the employment-weighted wage-bill HHI, (iii) the share of top-quartile workers employed by top-quartile firms within local labor markets, (iv) the share of bottom-quartile workers employed by bottom-quartile firms within local labor markets. The calibrated model provides a close quantitative match to both targeted and untargeted moments. Beyond reproducing the moments used for calibration, it successfully replicates untargeted features such as the observed dispersion of firm size, the relationship between hiring thresholds and measures of firm quality, heterogeneity in HHI indeces by worker AKM ranks, and the joint distribution of worker and firm types. As a benchmark, I also re-calibrate two versions of a homogeneous-labor model that shut down worker heterogeneity and revert to the BHM benchmark. Calibration 1 matches the aggregate HHI index but implies too little firm-level heterogeneity, whereas Calibration 2 matches firm-level heterogeneity more closely but generates substantially higher concentration indices.

[8] If local labor markets—and thus firms—were instead defined at the occupation level, the same balance-sheet observations would be mechanically split across multiple "firms," generating ambiguous and internally inconsistent measures of productivity, input use, and revenues.

[9] A stronger covariance implies that firms underpay workers with higher marginal products, lowering the labor share and increasing the wedge's distance from one.

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

[11] I rely on the correlation between labor wedges and market shares rather than their level because balance-sheet data report revenues but not quantities. This precludes direct estimation of firm-specific output elasticities, although they can be *controlled for* via high-dimensional industry and size fixed effects (see Ridder et al. (2025)).

Results. The calibrated model serves as a measurement device for three closely related objects: markdowns by worker type, firm-level wedges in the labor share, and the associated efficiency and welfare losses. In the data, average markdowns by ability can only be approximated indirectly using HHI-based formulas and AKM worker fixed effects, which are noisy proxies and typically grouped into coarse bins; the model instead recovers markdowns at the underlying ability level. Similarly, firm-level wedges (i.e., the distortion that labor market power imposes on the firm-level labor share) could potentially be identified in the data using cost shares, but caution is required to interpret those estimated wedges as actual wedges, since both their level and dispersion are confounded by measurement error and other distortions; the structural model isolates the labor-market-power component. Finally, mapping these wedges into aggregate efficiency and welfare losses for different worker types requires the full equilibrium structure.

Measurement. The model delivers rich implications for the distribution of markdowns across workers and firms. On average, workers take home roughly 72.5% of their marginal product of labor, with markdowns varying systematically in worker ability: high-ability workers receive about 70.5% of their marginal product, and the take-home share is not monotonic in worker ability, and maximized for workers whose log ability lies roughly one standard deviation below the mean. At the firm level, the mean and median firm-level wedge are very similar across the baseline and both BHM calibrations, but the revenue-weighted mean is markedly lower in BHM Calibration 2 (0.68 versus 0.73 in the baseline), indicating that larger-revenue firms in that benchmark operate with lower wedges and hence generate a stronger impact of labor market power on the aggregate labor share. Across productivity deciles, the wedges in the BHM specifications are monotonically decreasing in firm productivity, while in the baseline calibration they display a mildly non-monotonic profile, reflecting the fact that labor market segmentation alters the correlation between firm productivity and distortions.

Misallocation. I compare the decentralized equilibrium with labor market power to the Pareto-efficient allocation corresponding to the planner's solution. Under the baseline heterogeneous-worker calibration, aggregate misallocation reduces output by 3.63% relative to the efficient allocation. The first BHM calibration generates welfare losses similar in magnitude to the baseline (3.58%), while the second BHM calibration implies a substantially larger output loss of 5.33%. Decomposing these losses reveals that once within-market misallocation is eliminated, the remaining GDP losses across the baseline and BHM Calibration 1 and 2 are nearly identical (3.33%, 3.31%, and 3.52%, respectively). This pattern indicates that inefficiencies decline more sharply under BHM Calibration 2 and that the additional inefficiency in BHM Calibration 2 is driven by stronger within-market misallocation. In contrast, labor market segmentation in the baseline model dampens the extent to which firm heterogeneity translates into within-market distortions. The baseline calibration also attributes a sizeable fraction of the remaining loss to distortions in aggregate and cross-market labor supply: when aggregate labor supply is set to its efficient level, output is only 1.87% below the efficient benchmark, and when the cross-market misallocation is removed, aggregate inefficiency falls to 2.10% of efficient output.

Welfare distribution. Welfare effects are large and heterogeneous. Entrepreneurs benefit substantially from labor market power: in the baseline calibration, they would need to reduce their consumption by roughly 66.5% to attain the same utility level they would enjoy if workers were paid their marginal product. Workers, by contrast, experience welfare losses ranging from about 42% at the tails of the ability distribution to 37% for middle-ability workers. High-ability workers face highly concentrated markets dominated by a small set of large, high-paying firms. Low-ability workers, in turn, are excluded from most employers and obtain jobs only in a limited set of low-productivity, smaller firms, so their effective labor market is likewise highly concentrated. Greater concentration steepens markdowns and thereby increases welfare losses. For low-ability workers, these losses arise entirely from imperfect competition rather than redistributive motives, highlighting that imperfect competition further compresses an already depressed consumption level.

Overall, the analysis shows that labor market segmentation not only dampens within market misallocation but also reshapes the distribution of welfare losses across workers, with the largest welfare costs borne by workers at the extremes of the ability distribution.

The remainder of the paper is organized as follows. I first review the related literature. Section 3 develops the theoretical framework, characterizing firm behavior, worker sorting, and equilibrium with labor market power. Section 4 describes the data and presents the key empirical evidence that motivates and disciplines the model. Section 5 outlines the calibration, estimation, and validation of the model parameters. Section 6 presents the main quantification exercises, evaluating the aggregate efficiency and welfare implications of labor market power under the baseline calibration and relevant counterfactuals.

2 Related Literature

Monopsony and oligopsonistic labor markets. This paper contributes to the literature on monopsony and oligopsonistic labor markets, building on classical treatments of monopsony power (Robinson, 1933) and equilibrium search models with frictions (Burdett and Judd, 1983; Manning, 2003), as well as recent quantitative and empirical work using matched employer–employee data.

A growing body of research employs the *preference heterogeneity* approach to model worker labor supply, an approach consistent with extensive empirical evidence (Card, 2022). Recent contributions estimate firm-specific labor supply elasticities, markdowns, and concentration effects (Card, Cardoso, et al., 2018; Lamadon et al., 2022; Azar et al., 2022; Sharma, 2023; Haanwinckel, 2023; D. W. Berger et al., 2023). Bils et al. (2025) develops a model combining Roy-style selection with monopsony power. Related work examines the interaction of monopsony with labor market institutions such as minimum wages (D. Berger et al., 2025), documents implications for gender wage differences (Sharma, 2023), and demonstrates how trade exposure causally affects concentration indices (Felix, 2021).

Methodologically, this paper builds on the preference heterogeneity framework developed by Card, Cardoso, et al. (2018) and D. Berger et al. (2022), which provide tractable quantitative models of imperfect competition across firms and local labor markets. The present paper departs from this literature in three fundamental ways.

First, I introduce *endogenous choice sets* determined by worker ability and firm production decisions. Existing models typically assume workers can access any firm within their market, a limitation acknowledged in recent reviews (Card, 2022; Manning, 2021). In contrast, this paper recognizes that high-productivity firms actively exclude low-ability workers through a microfounded mechanism of intrafirm spillovers, while low-ability workers face severely restricted employment opportunities. This fundamentally alters rivalry across firms: competition occurs not over the entire labor market but within ability-specific segments, reshaping the distribution of market power across the ability distribution.

Second, I shift the analytical focus from aggregate or firm-level measures of monopsony to the question of *who competes for whom*. By explicitly modeling heterogeneity in both worker ability and firm productivity—combined with log-supermodular complementarities and within-firm spillovers—the model generates equilibrium segmentation where workers face different sets of competing employers. This allows me to characterize ability-specific concentration, document that labor market power varies non-monotonically across the worker distribution, and show that welfare losses are largest at the tails rather than being monotonic in ability.

Third, I provide novel empirical evidence from matched employer–employee data for Italy and Germany documenting hiring thresholds, ability-specific segmentation patterns, and non-monotonic concentration profiles. Critically, I establish a causal link between workforce composition and firm productivity using unexpected worker deaths as quasi-experimental shocks, validating the model's intrafirm spillover mechanism.

Misallocation and aggregate costs of micro distortions. This paper also relates to the literature quantifying aggregate output and welfare losses from micro-level distortions and misallocation (Harberger, 1954; H. Hopenhayn and Rogerson, 1993; Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009; H. A. Hopenhayn, 2014; Baqaee and Farhi, 2020; Edmond et al., 2023). A large empirical tradition uses production function estimation to recover firm-level wedges from observable data—such as markups from cost shares and revenue—and map these into aggregate productivity losses (e.g., De Loecker et al., 2020; Bond et al., 2021; Wu et al., 2025; Ridder et al., 2025).

I bring this perspective to labor market power by using the calibrated model as a measurement device to recover ability-specific markdowns and firm-level wedges in the labor share, and to trace how these distortions translate into aggregate misallocation and heterogeneous welfare losses. A further contribution is to embed a tractable model of worker heterogeneity, sorting, and segmented labor markets into a classical framework of firm size and productivity, thereby bridging the two literatures. A key contribution is extending the cost-share approach to environments with heterogeneous markdowns by worker ability, showing that the observed firm-level

wedge equals the average markdown contaminated by a covariance term between markdowns and worker types. This decomposition reveals that firms may exhibit large wedges not only because they impose high average markdowns, but also because they disproportionately underpay high-marginal-product workers.

Moreover, I demonstrate that once sorting and segmentation are introduced, the tight link between firm size, productivity, and distortions implicit in homogeneous-labor models is broken. Segmentation dampens within-market misallocation relative to homogeneous-labor benchmarks, yet substantial welfare losses persist—particularly for workers at the tails of the ability distribution. These losses arise purely from unequal exposure to imperfect competition rather than from redistributive concerns, and cannot be captured by models that abstract from worker heterogeneity or assume uniform market access.

Intrafirm spillovers, segmented labor markets, and assortative matching. This paper is related to empirical work using matched employer–employee data to study firm and worker heterogeneity, co-worker interactions, peer effects, and firm-specific wage components (Abowd et al., 1999; Moretti, 2004; Mas and Moretti, 2009; Card, Heining, et al., 2013; Song et al., 2019; Bender et al., 2018; Lamadon et al., 2022), and to theoretical contributions on sorting, segmentation, and assortative matching between workers and firms (Becker, 1973; Teulings, 1995; Shimer and Smith, 2000; Saint-Paul, 2001; Yeaple, 2005; Helpman et al., 2010; Costinot and Vogel, 2010; Eeckhout and Kircher, 2018; Freund, 2022; Haanwinckel, 2023; Bils et al., 2025).

Relative to this literature, I make both empirical and theoretical contributions. Empirically, I document novel facts on assortative matching, labor market segmentation, firm-level hiring thresholds, and ability-specific concentration patterns using matched employer–employee data from Italy and Germany. Theoretically, I develop a deliberately parsimonious general-equilibrium model that rationalizes these empirical regularities within a unified framework with a transparent mechanism: intrafirm spillovers in production generate endogenous segmentation, which in turn shapes the distribution of monopsony power across worker types.

The model is intentionally tractable. With a small number of interpretable parameters, it isolates how sorting generates segmentation, and how segmentation shapes monopsony distortions, while nesting several benchmark environments—including Helpman et al. (2010) and Costinot and Vogel (2010)—as special cases. This parsimony enables the model to flexibly match empirically disciplined segmentation patterns that distinguish it from existing frameworks, without sacrificing analytical clarity.

I validate the core spillover mechanism using an event-study design based on unexpected worker deaths, providing causal evidence that changes in workforce composition affect firm-level productivity. The calibrated model reproduces a wide range of targeted and untargeted empirical moments, demonstrating that the parsimonious structure captures essential features of segmented labor markets. The framework then serves as a measurement device to characterize competition across ability-specific segments, recover ability-specific markdowns that vary non-monotonically

with worker type, and quantify both the aggregate efficiency costs of monopsony and their heterogeneous incidence across the worker-ability distribution.

3 The Model

This section develops a general-equilibrium model of sorting and segmentation in the labor market and characterizes the resulting steady-state equilibrium. Time is dynamic, but I suppress time subscripts except when describing agents' problems. All analytical derivations and proofs appear in Appendix A.

3.1 Environment

Agents — The economy consists of households (workers) whose abilities take discrete values $a \in \mathcal{A}$, where $\mathcal{A}$ is a countably infinite set. Abilities are distributed according to a cumulative distribution function F_a with probability mass function $f_a(a)$. The economy also contains a collection of firms and a representative entrepreneur.

Local labor markets are indexed by $j \in [0, 1]$. Each market draws a random number of firms $m_j \sim F_m$, which is the only *ex ante* source of heterogeneity across markets. Firms differ in baseline productivity z_{ij}, where i indexes firms within market j, and productivities are drawn from F_z.

Income and rents — Firms are owned by the representative entrepreneur, who receives profits as lump-sum payments and accumulates capital that is subsequently rented to firms. Workers supply labor and earn wage income. The entrepreneur earns profits and capital rents.[12]

3.2 Workers

Each worker derives utility from consumption and disutility from supplying labor. Preferences are concave, and labor disutility follows a nested-CES specification—building on D. Berger et al. (2022) (hereafter BHM)—described in detail below. Utility is expressed in per-capita terms and scaled by the population density of ability type a, denoted $f_a(a)$. Since product-market power is not central to the analysis, final goods are assumed to be perfect substitutes, implying no markups.[13] The consumption good is the numéraire.

Each household of type a receives a set of job offers that depends on its ability. Let $\mathcal{S}_j(a)$ denote the set of firms offering jobs to ability type a in local labor market j; this choice set is 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

[12] For tractability, I impose a sharp distinction between workers—who earn almost exclusively from labor income—and the entrepreneur, who receives profits. This assumption is consistent with evidence from the *Survey of Consumer Finances*, which shows that most households rely primarily on labor income, whereas a small subset derives the bulk of its income from capital. For example, the capital-to-labor income ratio is roughly 0.05 for college-educated workers and below 0.02 for high-school–educated workers D. Berger et al., 2025.

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

$\{w_{ijt}(a)\}$ and choice sets $\{\mathcal{S}_j(a)\}$ as given:

$$\begin{aligned} U_0(a) &= \max_{\{n_{ijt}(a),\, c_{ijt}(a)\}_{t\geq 0}} \sum_{t=0}^{\infty} \beta^t U\left(\frac{C_t(a)}{f_a(a)},\, \frac{N_t(a)}{f_a(a)}\right) \\ &= \sum_{t=0}^{\infty} \beta^t \left[\left(\frac{C_t(a)}{f_a(a)(1-\sigma)^{1/(1-\sigma)}}\right)^{1-\sigma} - \left(\frac{N_t(a)}{f_a(a)(1+1/\varphi)^{1/(1+\varphi^{-1})}}\right)^{1+1/\varphi}\right]. \end{aligned} \tag{1}$$

Aggregate consumption and labor supply are given by

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

where market-level labor input is

$$n_{jt}(a) \coloneqq \left[\sum_{i\in\mathcal{S}_j(a)} n_{ijt}(a)^{\frac{\eta+1}{\eta}}\right]^{\frac{\eta}{\eta+1}}, \qquad \eta > \theta > 0.$$

The household's period budget constraint is

$$C_t(a) = \int_0^1 \sum_{i\in\mathcal{S}_j(a)} w_{ijt}(a)\, n_{ijt}(a)\, dj.$$

Discussion. The nested-CES labor-supply system embeds the broader class of *preference heterogeneity* models microfounded by BHM[14] and surveyed in Manning (2021) and Card (2022).[15] Workers consider both wages and idiosyncratic taste shocks when choosing employers. These unobserved components capture differences in commuting costs, firm culture, work environment, and relocation costs, generating imperfect substitutability across firms. Thus, workers do not necessarily supply labor to the highest-wage firm but to the employer that maximizes indirect utility.

The parameter η governs the elasticity of substitution among firms within a market, while θ governs substitution across markets. When $\eta > \theta$, intra-market mobility is easier than inter-market mobility. As $\eta \to \infty$, firms within a market become perfect substitutes, and the worker allocates all labor to the highest-wage firm in $\mathcal{S}_j(a)$. As $\theta \to \infty$, markets become perfect substitutes. The neoclassical monopsony model arises as the special case $\eta = \theta$.

[14] Appendix A.1 extends the BHM microfoundation showing that it's unaltered by the endogenous choice set $\mathcal{S}_j(a)$.

[15] Following BHM, the labor-supply structure arises from workers' discrete choices along three margins: (i) employment, (ii) selection across markets, and (iii) selection across firms within a market. Assuming correlated Gumbel taste shocks yields the nested-CES structure, where θ governs dispersion across markets and η governs dispersion across firms within a market.

Optimality conditions. In steady state, the first-order conditions deliver the inverse labor-supply system:

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

Given aggregate labor supply, firm-level labor supply depends on two wage indices: the market index $w_j(a)$ and the aggregate index $W(a)$, defined by

$$w_j(a)n_j(a) = \sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)n_{ij}(a), \qquad W(a)N(a) = \int_0^1 w_j(a)n_j(a)\,dj.$$

Using these definitions and (2) yields

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

I focus on a steady state in which employment by ability type is constant over time. One interpretation is that, in each period, a random fraction of workers separates and is instantly replaced by new hires of the same type. The steady-state allocation is thus a continual reshuffling of workers across firms, under which the inverse labor-supply relationships in (2)—and their microfoundation in Appendix A.1—apply isomorphically to the flow of new hires.

3.3 Representative Entrepreneur

The representative entrepreneur e has monotonic 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

$$U_0(e) = \max_{\{K_{t+1},\, c_{ijt}(e)\}_{t\geq 0}} \sum_{t=0}^{\infty} \beta^t U(C_t(e)) = \sum_{t=0}^{\infty} \beta^t \left(\frac{C_t(e)}{1-\sigma}\right)^{1-\sigma}, \tag{3}$$

subject to the budget constraint

$$C_t(e) + K_{t+1} - (1-\delta)K_t = \Pi_t + R_t K_t. \tag{4}$$

Aggregate entrepreneurial consumption, capital services, and profits are

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

Optimality condition. In steady state, the entrepreneur's Euler equation for capital accumulation is

$$1 = \beta(R + 1 - \delta). \tag{5}$$

3.4 Firms

Firm (i, j) is endowed with an exogenous type z_{ij}, drawn from F_z. A worker of ability $a \in \mathcal{A}$ employed at firm ij produces $\phi(a, z_{ij})$ units of output when combined with one unit of capital.

Let k_{ijt} denote capital. Total employment and the induced ability distribution are $h_{ijt} = \sum_{a \in \mathcal{A}} n_{ijt}(a)$ and $g_{ijt}(a) = n_{ijt}(a)/h_{ijt}$.

Output is produced according to

$$y_{ijt} = \mathbb{E}_{g_{ijt}}[\phi(a, z_{ij})] \left(k_{ijt}^{1-\gamma} h_{ijt}^{\gamma}\right)^{\alpha}, \qquad \mathbb{E}_{g_{ijt}}[\phi(a, z_{ij})] \equiv \sum_{a \in \mathcal{A}} \phi(a, z_{ij})\, g_{ijt}(a). \tag{6}$$

Firms thus have an exogenous technological component z_{ij}, but realized productivity is endogenous and reflects the ability composition of their workforce.

Quantitative specification. In the quantitative analysis, the worker–firm productivity function $\phi(a, z)$ is assumed to follow a constant-elasticity-of-substitution (CES) form:

$$\phi(a, z) = \left[(1 - \omega_a)\, z^{\frac{\rho-1}{\rho}} + \omega_a\, a^{\frac{\rho-1}{\rho}}\right]^{\frac{\rho}{\rho-1}}, \qquad \rho \leq 1,\ \omega_a \in [0, 1]. \tag{7}$$

This specification delivers a flexible degree of complementarity between firm productivity z and worker ability a. For $\rho < 1$, $\phi(a, z)$ is log-supermodular.

Special cases. The CES structure in (7) nests several benchmark production functions:

1. **Cobb–Douglas benchmark.** When $\omega_a = 0$, worker output depends solely on firm productivity and $\phi(a, z) = z$. Firm-level output reduces to

   $$y = z\,(k^{1-\gamma} h^{\gamma})^{\alpha}.$$

   In this limit, the production structure of BHM is nested.[16]

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

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

[16] While the production function coincides with BHM, the income distribution differs. In D. Berger et al., 2022, a representative household receives wages, profits, and capital rents. In the present framework, entrepreneurs own capital and collect profits, whereas workers earn only wages; equilibrium allocations may therefore differ through rent sharing and, with income effects in labor supply, through aggregate employment responses.

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

3. **Additive aggregation.** When $\alpha = \gamma = 1$, output exhibits constant returns to the number of workers, and labor enters through efficiency units $\phi(a, z)$. The production function reduces to the additive form in Costinot and Vogel (2010):

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

Discussion. Appendix A.2 provides a microfoundation for the production structure in which workers share a common firm-level resource (e.g., capital equipment or workspace) that is allocated uniformly across employees.[17] Under this interpretation, $\phi(a, z)$ represents the output produced by a type-a worker at a firm with type z when combined with a standardized amount of common resources, and $\mathbb{E}_{g(a)}[\phi(a, z)]$ is realized firm productivity as the average of these individual outputs. This mechanism builds upon theories proposed by Kremer, 1993, Saint-Paul, 2001, and Helpman et al., 2010, and is closely related to theories of sorting and productivity interactions such as Shimer and Smith, 2000 and Eeckhout and Kircher, 2018.

In later sections, I provide direct empirical evidence consistent with this mechanism. Using plausibly exogenous worker exits due to unexpected deaths, I show that the loss of a lower-ability worker *increases* firm-level TFPR, whereas the loss of a higher-ability worker *reduces* it. These asymmetric responses point to within-firm congestion or dilution effects: employing workers whose ability lies below the firm's average depresses the productivity-relevant aggregate and thereby imposes a negative spillover on firm-level efficiency.

More broadly, the specification offers a parsimonious way to capture a variety of channels through which the composition of a firm's workforce influences measured productivity. These include within-firm spillovers documented in coworker-composition studies Moretti, 2004; Gennaioli et al., 2013; Bender et al., 2018 as well as peer and social-pressure effects operating within production units Ichino and Maggi, 2000; Falk and Ichino, 2006; Bandiera et al., 2010; Mas and Moretti, 2009. While the model does not explicitly incorporate the full structure of these mechanisms, it is constructed to capture their central implications in a stylized and parsimonious form. As developed in later sections, the framework reproduces rich patterns of sorting and segmentation with a small number of interpretable parameters, while also encompassing standard benchmark environments as limiting cases.

Firm's Problem. Given the specified production function, firms are infinitesimal with respect to the aggregate economy but granular within local labor markets. They treat aggregate quantities

[17] For intuition, consider a division where workers share office space, equipment, or support staff that cannot be tailored to individual ability levels. Even when workers differ in productivity, management cannot differentially allocate these shared inputs, so each worker effectively receives an identical bundle.

$N_t(a)$ and $W_t(a)$ as exogenous, while internalizing the impact of their own hiring decisions on market-level variables $n_{jt}(a)$ and $w_{jt}(a)$. The equilibrium features Cournot competition: each firm takes competitors' employment choices n^*_{-ijt} as given. Each firm maximizes profits taking R_t and the labor supply as given, choosing capital k_{ijt} and the number of workers of each ability type $n_{ijt}(a)$. An equivalent interpretation is that the firm chooses the number of job slots to open, h_{ijt}, and a workforce composition rule $g_{ijt}(a) \in [0,1]$ specifying how these slots are allocated across abilities.[18]

Given this representation, the firm's problem is

$$\pi_{ijt} = \max_{n_{ijt}(a),\, k_{ijt}} \left\{ \mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})] \left(k_{ijt}^{1-\gamma} h_{ijt}^{\gamma}\right)^{\alpha} - R_t k_{ijt} - h_{ijt}\, \mathbb{E}_{g_{ijt}(a)}[w_{ijt}(a)] \right\}, \tag{8}$$

subject to $g_{ijt}(a) = \frac{n_{ijt}(a)}{h_{ijt}} \geq 0$, $\sum_a n_{ijt}(a) = h_{ijt}$, and the inverse labor-supply condition (2).

The first-order condition for capital is

$$\left(\mathbb{E}_{g_{ij}(a)}[\phi(a, z_{ij})]\right)^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\frac{\gamma\alpha}{1-\alpha(1-\gamma)}} = \frac{R k_{ij}}{(1-\gamma)\alpha}. \tag{9}$$

Let $Z = \left(\frac{\alpha(1-\gamma)}{R}\right)^{\frac{(1-\gamma)\alpha}{1-(1-\gamma)\alpha}}$. Substituting the optimal k_{ij}, hiring an additional worker of type a affects output through two channels.

First, total employment h rises, generating a positive *size effect*: $\frac{\partial y}{\partial h}\frac{dh}{dn(a)} = \frac{\partial y}{\partial h}$.[19] This term coincides with the marginal product if all workers are homogeneous with productivity $\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]$.[20]

Second, hiring affects the firm's endogenous productivity through $\frac{\partial y}{\partial \mathbb{E}_{g_{ijt}(a)}[\phi(a,z_{ijt})]} \frac{d\mathbb{E}_{g_{ijt}(a)}[\phi(a,z_{ijt})]}{dn(a)}$,[21] which is negative whenever the worker's productivity is below the firm average. Combining both channels gives a compact expression for the marginal product of labor:

$$\begin{aligned} MPL_{ij}(a \mid \mathbb{E}_{g_{ijt}(a)}[\phi(\cdot)], h) &= \overline{MPL}_{ij} \left[1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\phi(z_{ij}, a)}{\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]}\right)\right], \\ \overline{MPL}_{ij} &\equiv Z\alpha\gamma \left(\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]\right)^{\frac{1}{1-\alpha(1-\gamma)}} h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}}. \end{aligned} \tag{10}$$

[18] For notational simplicity I suppress firm and competitor indices below. When writing $w_{ijt}(n_{ijt}(a))$, it should be understood as a function of own hiring, competitors' hiring, and aggregate quantities: $w_{ijt}(a, n_{ijt}(a), n^*_{-ijt}, N_t(a), W_t(a))$.

[19] Differentiating output with respect to total employment gives

$$\frac{\partial y}{\partial h} = Z\alpha\gamma \left(\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]\right)^{\frac{1}{1-\alpha(1-\gamma)}} h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}}.$$

[20] Under homogeneous labor, $y = z(k^{1-\gamma}h^{\gamma})^{\alpha}$, implying $MPL = Z\alpha\gamma z^{\frac{1}{1-\alpha(1-\gamma)}} h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}}$.

[21]

$$\frac{\partial y}{\partial \mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]} \frac{d\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]}{dn(a)} = Z \left(\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]\right)^{\frac{1}{1-\alpha(1-\gamma)}} h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}} \left(1 - \frac{\phi(z_{ij}, a)}{\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]}\right).$$

For expositional simplicity, I will refer to this expression as $MPL_{ij}(a)$.

Under decreasing returns to labor ($\alpha\gamma < 1$), the productivity term in (10) may outweigh the positive size effect, so that hiring additional low-ability workers—whose productivity $\phi(z_{ij}, a)$ falls sufficiently below the firm average—reduces the marginal product of labor; if the decline in endogenous productivity dominates the scale gain, the marginal product can even turn negative. Microfoundationally, each worker not only contributes to output but also occupies a *spot*—an equal share of the firm's resources (e.g., capital). When returns are decreasing, occupying a spot entails a shadow cost: additional hires dilute resources per worker, lowering average productivity. Under constant returns, this dilution effect vanishes—firms can scale up proportionally, and a worker's marginal product reverts to $\phi(a, z)$.

Intra-Firm Spillovers. The production structure generates intra-firm spillovers because each worker's marginal product depends on the firm-level productivity term $\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]$. Holding total employment h fixed, the sensitivity of $MPL_{ij}(a)$ to changes in firm productivity is

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

Because this expression is linear in $\phi(a, z_{ij})$, there exists a threshold ability a^* such that firm-level productivity improvements reduce the marginal product of workers with $a < a^*$ and increase it for those with $a > a^*$. Linearity further implies that spillover intensity rises with worker ability: productivity gains amplify intra-firm heterogeneity by disproportionately raising the marginal products of high-ability workers.

Under decreasing returns to labor, expanding employment imposes a shadow cost by diluting the firm's effective resources. As firm efficiency rises, this shadow cost strengthens and depresses the marginal products of all worker types, especially those with low ability. At the same time, the firm's capital first-order condition implies that higher efficiency induces greater capital accumulation, which increases marginal products across the board. When this capital–productivity channel dominates the shadow-cost effect, high-ability workers experience net productivity gains, generating positive intra-firm spillovers.

Under constant returns to labor, capital is absent and the resource-dilution mechanism disappears. In this case, each worker's marginal product is independent of the firm's composition, and intra-firm spillovers vanish.

Firm Wages. If the marginal product of a worker is positive, the standard rearrangement of the firm's first-order condition with respect to employment $n_{ij}(a)$ yields a Lerner-type condition: the equilibrium wage equals an endogenous markdown $\mu_{ij}(a) \leq 1$ applied to the marginal product of labor. When $MPL_{ij}(a) \leq 0$, the firm does not hire the worker, implying $w_{ij}(a) = 0$. Proposition 1 characterizes the resulting wage schedule.

Proposition 1. *Let the marginal product of a worker of type a be given by equation* (10)*. A firm optimum exists, and the maximizing wage schedule satisfies*

$$w_{ij}(a) = \begin{cases} \mu_{ij}(a) \cdot MPL_{ij}(a) & \text{if } MPL_{ij}(a) > 0, \\ 0 & \text{if } MPL_{ij}(a) \leq 0, \end{cases} \tag{11}$$

where

$$\mu_{ij}(a) = \frac{\epsilon_{ij}(a)}{\epsilon_{ij}(a)+1}, \qquad \epsilon_{ij}(a) \coloneqq \left[\frac{\partial \log w_{ij}(a)}{\partial \log n_{ij}(a)} \bigg|_{n^*_{-ij}(a)} \right]^{-1}, \tag{12}$$

and $\epsilon_{ij}(a)$ denotes the inverse elasticity of firm-specific labor supply. Under the assumed structure,

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

Lemma 1. *Under no capital in production ($\gamma = 1$), the wage structure in equation* (11) *is also sufficient for firm maximization.*[22]

Proposition 1 implies that the firm's problem can be expressed as a fixed-point system. Given initial employment choices $\{n_{ij}(a)\}$ and competitors' allocations $n^*_{-ij}(a)$, one can recover $\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]$ and h_{ij}, update marginal products and wages, and iterate until convergence.

Firm Labor Market Power. In the model, firm labor market power is captured by firm–worker–specific markdowns, which measure the wedge between a worker's marginal product of labor and the wage paid.[23] The markdown $\mu_{ij}(a)$ is determined by the firm–worker employment elasticity $\epsilon_{ij}(a)$, which reflects worker ability, firm characteristics, and the competitive environment. In equilibrium, $\epsilon_{ij}(a)$ depends on the firm's wage share among all employers of ability-a workers in market j, following D. Berger et al., 2022 and Atkeson and Burstein, 2008.

Firms compete for each worker type against type–specific wage indexes. Importantly, competition occurs within a worker's relevant choice set. Consequently, a firm may be macroeconomically small yet large relative to the subset of employers available to a particular worker type. When a firm represents 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. Conversely, when the firm is small

[22] The sufficiency result is established by first proving the existence of a maximum through a coercivity argument and then showing that any candidate satisfying the first-order conditions is unique. Uniqueness follows from partitioning the domain into two exhaustive regions: in the first, profits are strictly concave; in the second, a single-crossing property rules out multiple optima. This ensures both necessity and sufficiency of the wage schedule when capital is fixed or absent.

When capital is freely adjustable ($\gamma \neq 1$), the single-crossing argument cannot be derived analytically because capital accumulation generates cross-type positive spillovers. In that case, sufficiency is verified numerically.

[23] The standard interpretation is that increasing employment forces the firm to raise wages for all inframarginal workers. In this environment—where firms set wages by worker type—an equivalent probabilistic interpretation applies: raising the wage increases the probability of hiring but also raises the wage in all states in which the worker would have accepted anyway. The smaller the elasticity of labor supply to the firm, the larger this inframarginal cost.

within the choice set ($s_{ij}(a) \approx 0$), the markdown is governed primarily by the within-market elasticity η, as in standard oligopsonistic models.

The following proposition relates the average wage and average MPL to profits and markdowns under oligopsonistic behavior, and shows how markdowns shape firm profits and the labor share.

Proposition 2 (Firm-Level Quantities)**.** *Define the firm-level labor market power wedge:*

$$\widetilde{\psi}_{ij} = \overline{\mu}_{ij} + cov_{g_{ij}}\left(\mu, \frac{\phi}{\mathbb{E}_{g_{ij}(a)}[\phi(a, z_{ijt})]}\right) \leq 1,$$

where the covariance is taken with respect to g_{ij}. Let $MPL_{ij}(a)$ denote the marginal product of type-a labor, $\overline{MPL}_{ij}$ the average marginal product, $\overline{w}_{ij}$ the average wage, and ls_{ij} the labor share. Then:

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

$$\overline{w}_{ij} = \overline{MPL}_{ij} \cdot \widetilde{\psi}_{ij}, \qquad ls_{ij} = \alpha\gamma \cdot \widetilde{\psi}_{ij}.$$

Absent markdowns, $\widetilde{\psi}_{ij} = 1$, and the average wage equals the average marginal product. In this benchmark, the firm's labor share and profits coincide with those implied by a Cobb–Douglas technology with homogeneous labor: labor receives a constant share $\alpha\gamma$ of revenue, while profits account for the remaining share $1 - \alpha$, despite the presence of worker heterogeneity.

The labor-market-power wedge $\widetilde{\psi}_{ij}$ determines the gap between the firm's average marginal product and its average wage. It has two components. The first reflects the firm's average markdown, so $\widetilde{\psi}_{ij}$ (and thus the labor share) falls as the firm exercises greater monopsony power. The second component—the covariance term—captures a second-order effect: markdowns reduce the labor share more strongly when their covariance with workers' relative output, $\frac{\phi}{\mathbb{E}_{g_{ij}(a)}[\phi(a,z_{ijt})]}$, is more negative. This covariance is taken with respect to the within-firm distribution of worker abilities.

Intuitively, a more negative covariance between markdowns μ and worker output ϕ means the firm underpays workers with higher relative $MPL_{ij}(a)$, who account for a disproportionate share of total wage payments. Shifting earnings away from these high-MPL workers compresses the right tail of the within-firm earnings distribution, lowers the labor share, and raises profits. In an economy with sorting, more productive firms exhibit a stronger covariance between worker output and market shares, further contributing to a second-order reduction in their labor share.

Proposition 2 also informs the empirical measurement of markdowns. The production-function approach recovers firm-level markdowns by comparing a plant's marginal revenue product of labor with its wage (e.g., Yeh et al., 2022; Kirov and Traina, 2021). The proposition shows that, in the presence of unobservable workforce heterogeneity, this empirical measure corresponds to the model-consistent labor-market-power wedge $\widetilde{\psi}_{ij}$, which differs from the average markdown because it incorporates the covariance term. Importantly, $\widetilde{\psi}_{ij}$ remains the relevant sufficient statistic

for assessing how labor market power shapes the firm-level labor share.

Definition 1 (Steady-State General Equilibrium)**.** Given primitives $\{F_a, F_m, F_z\}$, parameters $(\sigma, \varphi, \alpha, \gamma)$, and technology $\phi(a, z)$, a steady-state general equilibrium is a collection of allocations and prices

$$\{n_{ij}(a),\, k_{ij},\, \pi_{ij},\, w_{ij}(a),\, K,\, C(a),\, C(e),\, R\}_{i,j,a}$$

such that:

(i) **Households.** For each ability $a \in \mathcal{A}$, households satisfy the optimality conditions in (2), determining equilibrium labor supply $N(a)$ and firm-level allocations $\{n_{ij}(a)\}_{i,j}$.

(ii) **Capital markets.** The steady-state return to capital satisfies the Euler equation (5), and aggregate capital equals the sum of firm-level capital,

$$K = \int_0^1 \sum_{i=1}^{m_j} k_{ij}\, dj.$$

(iii) **Firms.** Each firm (i, j) maximizes profits taking R and labor supply as given, choosing $\{n_{ij}(a)\}_a$ and k_{ij} according to the first-order conditions (9)–(10) and the wage condition (11).

(iv) **Feasibility.** Aggregate goods-market clearing holds:

$$\int_0^1 \sum_{i=1}^{m_j} y_{ij}\, dj = \sum_{a \in \mathcal{A}} C(a) f_a(a) + C(e).$$

Individual consumption equals individual income:

$$C(a) = \int_0^1 \sum_{i \in \mathcal{S}_j(a)} w_{ij}(a) n_{ij}(a)\, dj, \qquad C(e) = \Pi + (R - \delta)K,$$

where total profits are

$$\Pi := \int_0^1 \sum_{i=1}^{m_j} \pi_{ij}\, dj.$$

Lemma 2. *A steady-state equilibrium exists.*

Definition 2 (Planner's Problem)**.** The social planner chooses allocations $\{C(a),\, C(e),\, K,\, n_{ij}(a),\, k_{ij}\}_{i,j,a}$ to maximize weighted social welfare:

$$\mathcal{W} = \sum_{a \in \mathcal{A}} \psi(a)\, U(C(a), N(a))\, f_a(a) \;+\; \psi(e)\, U(C(e))\,,$$

subject to the resource constraint

$$\sum_{a\in\mathcal{A}} C(a)f_a(a) + C(e) + K = \int_0^1 \sum_{i=1}^{m_j} y_{ij} + (1-\delta)K,$$

and the feasibility conditions defining $n_{ij}(a)$, h_{ij}, $g_{ij}(a)$, and y_{ij}.

Proposition 3 (Efficiency Without Markdown)**.** *If all firms pay no markdowns ($\mu_{ij}(a) = 1$ for all i, j, a), so that*

$$w_{ij}(a) = MPL_{ij}(a),$$

then the decentralized equilibrium coincides with the planner's allocation.

3.5 Market Equilibrium: Sorting, Segmentation, and Labor Market Power

I now examine how different parameterizations of the model generate distinct patterns of sorting and segmentation in the labor market. These mechanisms determine the intensity of competition across firms for different worker types and, in turn, shape the allocation distortions and welfare effects induced by labor-market power. Whenever a firm pays a worker less than their marginal product, it attracts too few workers of that type relative to the efficient allocation, who are misallocated either toward other firms or toward leisure. As established in Proposition 3, such misallocation reduces aggregate welfare and redistributes income from workers to the entrepreneur. To illustrate the underlying mechanisms, I analyze a single labor market with a random realization of firm productivities and a fixed aggregate labor supply. The market contains one hundred firms whose productivities are drawn from a lognormal distribution, while worker abilities follow a discretized lognormal distribution. The baseline calibration for this exercise is reported in Table 1. I then vary one parameter at a time to trace how sorting, segmentation, and markdown heterogeneity emerge in equilibrium under alternative parameter configurations. The results in this subsection are qualitative in nature: they are designed to elucidate the model's mechanisms rather than deliver quantitative predictions.

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

Table 1: Baseline calibration parameters for the illustrative labor-market exercise.

Definition 3 (Market Equilibrium)**.** Fix the labor-supply density in market j as $f_a(a)$. A *market equilibrium* is a collection of employment shares $\{q_{ij}(a)\}$ satisfying

$$q_{ij}(a) = \frac{w_{ij}(a)^\eta}{\sum_{i'\in\mathcal{S}_j(a)} w_{i'j}(a)^\eta}, \qquad \mathcal{S}_j(a) = \{i : MPL_{ij}(a) > 0\},$$

with wages given by $w_{ij}(a) = \mu_{ij}(a)MPL_{ij}(a)$. The efficient equilibrium satisfies the same system with $\mu_{ij}(a) = 1$, i.e. $w_{ij}(a) = MPL_{ij}(a)$.

Linking markdowns and concentration. Before examining the equilibrium benchmarks, it is useful to establish two preliminary results.

Lemma 3 (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 and let $s_{ij}(a)$ be firm i's wage bill share for that type. Then*

$$\sum_i s_{ij}(a)\,\mu_{ij}(a)^{-1} = \frac{\overline{\mathrm{MPL}}_j(a)}{\overline{W}_j(a)} = 1 + \frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big) HHI_j(a),$$

where $HHI_j(a) = \sum_i s_{ij}(a)^2$. Thus the wage-bill–weighted, ability-specific average inverse markdown depends directly on market concentration.

Lemma 4 (Finite–sample decomposition of concentration)**.** *Let $m_j(a)$ be the number of firms hiring ability a in market j, and let $\mathrm{CV}_j(a)$ be 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 market concentration reflects both the size of a worker's choice set and the dispersion in effective wage offers across firms.

Together, Lemma 3 and Lemma 4 show how average markdowns by ability type are shaped by market concentration, the number of available employers, and the heterogeneity of effective wage offers within a market.

Equilibrium under Homogeneous Labor and Supermodular Technology. I begin with the benchmark case of homogeneous labor, as in BHM, obtained by setting $\omega_a = 0$. Because all workers contribute identically to production, the economy exhibits neither sorting nor segmentation. Higher firm productivity z_{ij} raises the marginal product of every worker uniformly. More productive firms therefore pay higher wages and hire the same share of each ability type, expanding proportionally across the entire workforce. These firms capture a larger market share for every worker type and exert greater market power symmetrically across the ability distribution. Since more productive firms apply larger markdowns, they pay wages that fall further below the efficient benchmark and operate at inefficiently small scale relative to the planner's allocation. I refer to this as the *firm-size distortion*, to distinguish it from the additional distortions that arise once workers differ in productivity.

Next, consider the case where $\omega_a > 0$ but let $\rho \to 1$, so that worker-level output is supermodular but not log-supermodular. Under the $\rho \to 1$ specification, after defining the firm-level productivity term $\tilde{z}_{ij} := z_{ij}^{1-\omega_a}\mathbb{E}_{g_{ij}(a)}[a^{\omega_a}]/\bar{a}_{ij}$, firm output and worker's marginal product can be written as

$$y_{ij} = \tilde{z}_{ij}\,\bar{a}_{ij}\,\Big(k_{ij}^{(1-\gamma)}h_{ij}^{\gamma}\Big)^{\alpha}, \qquad MPL_{ij}(a) = \tilde{z}_{ij}\,k_{ij}^{\alpha(1-\gamma)}h_{ij}^{\gamma\alpha-1}\big[a - (1-\alpha)\bar{a}_{ij}\big].$$

Let $\bar{a}$ be the average ability in the market, computed using the exogenous aggregate labor supply density. In the limit $\rho \to 1$, when firms hire the same density of workers equal to the market density of labor supply, they share the same average ability $\bar{a}_{ij} = \bar{a}$, and the dependence of the marginal product on ability a enters multiplicatively through a term common to all firms for a given a.

Lemma 5 (Equilibrium invariance from $\omega_a = 0$ to $\rho \to 1$)**.** *Let $\{q_{ij}(a)\}$ denote the equilibrium assignment under $\omega_a = 0$. Let $\bar{a}$ be the average ability in the market, given the exogenous aggregate labor supply, and assume that $a_l > (1-\alpha)\bar{a}$*[24]*. Define the renormalized firm productivities $\tilde{z}_{ij} = z_{ij}^{1-\omega_a} \frac{\mathbb{E}_{g(a)}[a^{\omega_a}]}{\bar{a}}$. Then the same assignment $\{q_{ij}(a)\}$ constitutes a market equilibrium under the $\rho \to 1$ technology $y_{ij} = \tilde{z}_{ij}\, \bar{a}_{ij} \left(k_{ij}^{(1-\gamma)} h_{ij}^{\gamma}\right)^{\alpha}$.*

When $\rho \to 1$, worker ability scales firm productivity proportionally across all firms, preserving relative productivities and thereby leaving equilibrium allocations unchanged. The two equilibria—under homogeneous labor ($\omega_a = 0$) and under the multiplicative-in-ability technology ($\rho \to 1$)—are *isomorphic* up to a renormalization: they feature identical market shares, markdowns, welfare allocations, and distributions of distortions across firms. Ability heterogeneity affects all firms symmetrically and therefore does not alter the equilibrium mapping between productivity, market power, and welfare. Panel (a) of Figure 1 reports the employment market shares for three selected firms in terms of their z, the least, the median, and the highest z within the market distribution. As predicted by the lemma, higher quality firms employ more of each labor type, in a uniform manner. The market shares with $\omega_a = 0$ are identical up to a renormalization and are therefore omitted.

Equilibrium under Sorting and No Spillovers. Consider the benchmark with no decreasing returns to labor ($\alpha = \gamma = 1$), so that within-firm spillovers are absent and the marginal product reduces to the worker–firm production term, $MPL_{ij}(a) = \phi(a, z_{ij})$. Let $\omega_a > 0$ and $\rho < 1$ so that $\phi(a, z)$ is log-supermodular and strictly increasing in both worker ability a and firm productivity z. For any given ability, more productive firms therefore offer higher marginal products. Panel (b) and Panel (c) of Figure 1 report the resulting market shares and concentration indices for this benchmark economy. More generally, the core results below hold whenever $MPL_{ij}(a)$ is log-supermodular and increasing in (a, z), regardless of whether $\alpha = \gamma = 1$.

Lemma 6 (Monotonicity of employment shares in firm productivity)**.** *Fix an ability type a. If $MPL_{ij}(a)$ is (weakly) increasing in z_{ij} and $\mu_{ij}(a)$ is (weakly) decreasing in the wage-bill share, then for any $z_{i'j} > z_{ij}$,*

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

with strict inequality whenever either $MPL_{ij}(a)$ increases strictly in z_{ij} or $\mu_{ij}(a)$ is strictly decreasing. Hence employment shares $q_{ij}(a)$ are (weakly) increasing in firm productivity.

[24] If this condition does not hold, the only difference is that worker types with negative marginal product are not hired by any firm under $\rho \to 1$, whereas they would be hired under $\omega_a = 0$.

Lemma 7 (Positive assortative matching)**.** *Let $z_{i'j} > z_{ij}$. If $MPL_{ij}(a)$ is log-supermodular, 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 more productive firms.

Proposition 4. *Under the same assumptions, the concentration index $HHI_j(a) = \sum_i q_{ij}(a)^2$ is strictly increasing in ability: for any $a' > a$, $HHI_j(a') > HHI_j(a)$. Higher-ability workers are therefore employed in more concentrated segments of the market and face greater firm-level labor market power.*

With complementarities in production, high-z firms have a comparative advantage in high-a workers and, in the absence of spillovers, an absolute advantage for every worker type. Consequently, high-z firms employ more workers of all types, and disproportionately more of the high-ability ones, as illustrated in Panel (b) of Figure 1. This pattern generates an $HHI_j(a)$—and hence welfare losses—that rise with ability. Higher-ability workers are therefore subject to stronger market power and experience larger welfare losses relative to the efficient benchmark.
Regarding firm-size distortions, high-z firms apply larger markdowns across all worker types and are distorted for each of them. In the efficient allocation, these firms would employ more workers of every type—particularly of higher-ability types, for which their market shares are largest. Firm-size distortions are thus positively correlated with firm productivity, as illustrated in Panel (c) of Figure 1.

Market Equilibrium under Within-Firm Spillovers. Figure 2 reports market shares, size distortions, HHI indices, and firm ability thresholds (i.e., the minimum ability level employed by each firm) under the production structure specified in this model, using the calibration in Table 1, for labor market supply fixed and from a discretized log-normal distribution. The equilibrium is now characterized by a rationing of workers' choice sets: access to the most productive firms is precluded for workers at the lower end of the ability distribution. The following lemma establishes a sufficient condition for this outcome.

Lemma 8. *If the equilibrium marginal product of labor $MPL_{ij}(a)$ is log-supermodular in (a, z_{ij}), then (i) realized productivity $E_{g_{ij}(a)}[\phi(a, z_{ij})]$ is weakly increasing in firm productivity, and (ii) the screening threshold $\widetilde{a}_{ij}$, defined as the minimum worker ability among employees, is weakly increasing in z_{ij}.*

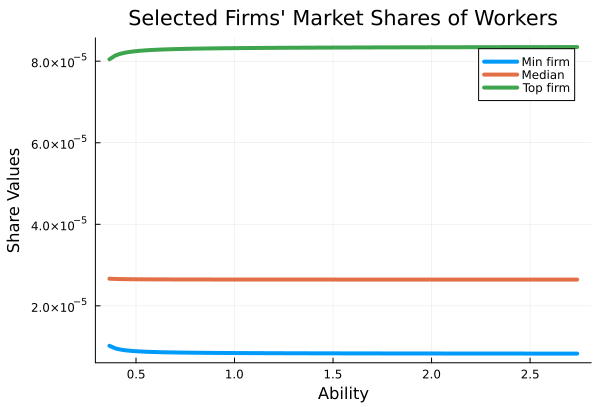
Spillovers within firms alter the allocation of workers across productivity levels. The marginal product of low-ability workers may be higher in low-productivity firms than in high-productivity ones, when the negative externality they impose on productivity in the latter outweighs the positive scale effect from firm size. As a result, low-ability workers are segregated into low-type firms, which have limited ability to pay high wages. The labor market thus exhibits both *sorting* and *segmentation*: high-ability workers are more likely to be employed by high-type firms, while high- and low-ability workers cluster respectively within high- and low-type firms. Competition

among firms becomes *localized*, as each firm specializes in a specific segment of the ability distribution and competes primarily with nearby firms in productivity space targeting similar worker types. Firms at the lower end of the productivity distribution therefore hire mainly low-ability workers facing restricted choice sets. Following Lemma 4, strong segmentation implies that when low-ability workers are concentrated in low-productivity firms—and when such firms are relatively few—competition for these workers weakens. Each low-productivity firm internalizes that it constitutes a large share of the choice set of low-ability workers, even if it is small in the aggregate economy, and therefore exercises greater monopsony power, paying wages below the efficient level. Conversely, more productive firms shift their market-share mass toward higher-ability workers. As these firms extend offers to highly skilled workers, competition again diminishes: although these workers do not face restricted choice sets, offers from high-productivity firms dominate their outside options—an outcome analogous to Proposition 4.

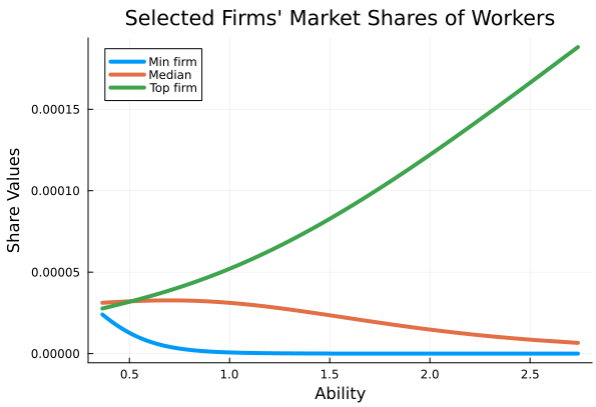
Turning to production distortions, segmentation weakens the correlation between firm productivity and the magnitude of distortions. Panel (c) of Figure 2, relative to Panel (b) of Figure 1, illustrates how the firm-size distortion varies under a baseline parametrization that induces labor market segmentation. As segmentation intensifies, less productive firms acquire greater labor market power, and the firm size distortion flattens. Because aggregate output losses from misallocation are particularly severe when distortions rise with firm productivity (Restuccia and Rogerson, 2008; H. A. Hopenhayn, 2014), and present in market equilibrium only if there is dispersion in market power, breaking this positive correlation reduces the aggregate efficiency cost of labor market power.[25]

Taking stocks. The different model environments discussed above generate distinct equilibrium predictions for realized market shares and concentration patterns across the ability distribution. In the next section, I turn to the data to examine how market shares and the $HHI_j(a)$ index vary empirically with worker ability. I then use these empirical moments to benchmark and quantify the welfare and production-efficiency effects of labor market power, comparing them to the homogeneous-worker BHM economy.

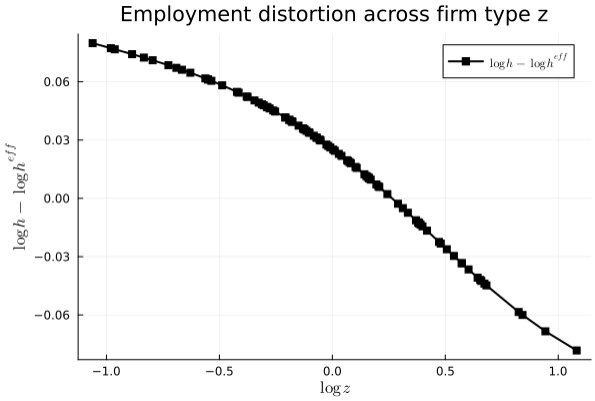
[25] If aggregate output is $Y = \sum_i y_{ij}$, then $\frac{Y}{Y_{\text{eff}}} = \sum_i \frac{y_{ij}}{y_{ij,\text{eff}}} \frac{y_{ij,\text{eff}}}{Y_{\text{eff}}}$, which is a weighted average of firm-level output losses. Hence, total output losses are larger when firm-level distortions are concentrated among the most productive firms, which have greater revenue shares.



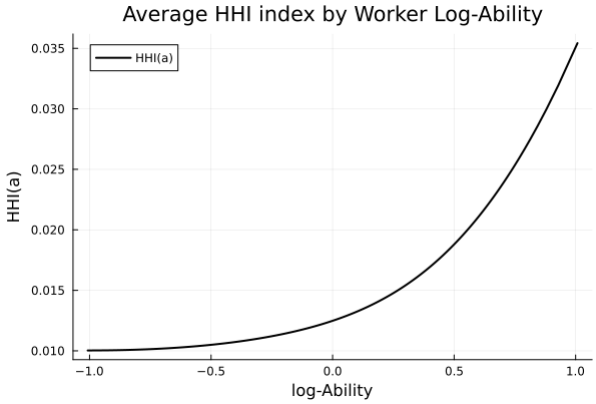
(a) Employment market shares



(b) Employment market shares



(c) Size distortion by firm type

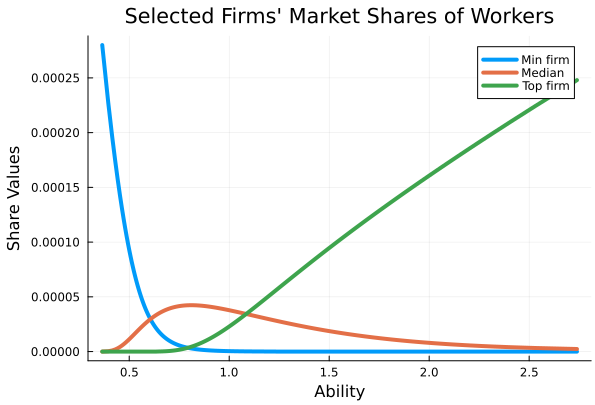


(d) HHI index by worker log-ability

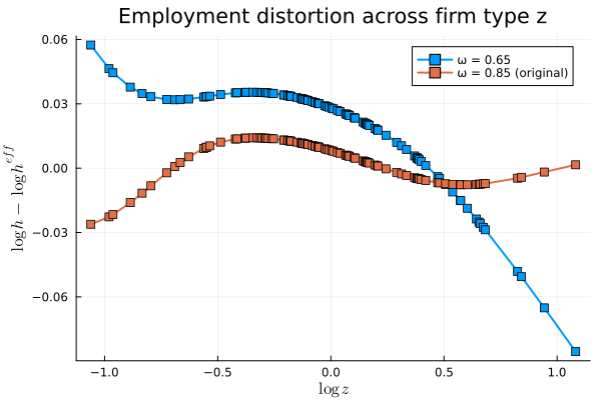
Figure 1: Equilibrium Market Shares, Size Distortions, and HHI Index Across Benchmark Parameterizations

Notes:

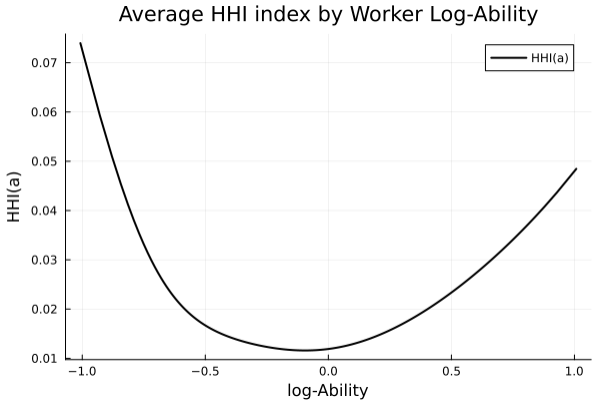
This figure is divided into four panels, labeled a–d, and reports equilibrium outcomes for alternative benchmark parameterizations of the model. Panels (a) and (b) display employment market shares for three selected firms—the bottom, median, and top firm in the productivity distribution—by worker log-ability, under the supermodular worker-level output ($\rho \to 1$) and the sorting case without spillovers ($\alpha = \gamma = 1$), respectively. Panel (c) depicts the size distortion, measured as the log deviation of firm employment from the efficient level, in the sorting case without spillovers ($\alpha = \gamma = 1$). Panel (d) reports the corresponding Herfindahl–Hirschman index ($HHI_j(a)$) across the ability distribution. Markdowns by worker ability are not reported, as they are a mirror image of the market-share distribution across abilities, and the average markdown by worker log-ability is similarly a mirror image of the $HHI_j(a)$ pattern. When $\omega_a \to 0$, market shares and distortions coincide—up to a rescaling—with those in the $\rho \to 1$ case and are therefore omitted. In both the $\omega_a = 0$ and $\rho \to 1$ environments, the $HHI_j(a)$ index is flat across the ability distribution and thus not reported. Likewise, the firm-size distortion for $\omega_a = 0$ exhibits the same shape as in the $\alpha = \gamma = 1$ economy and is omitted for brevity. All simulations use the calibration reported in Table 1, with one parameter tilted at a time relative to the baseline to isolate each theoretical mechanism.



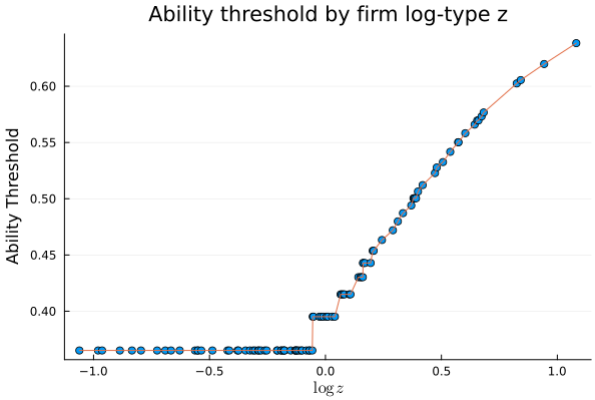
(a) Employment market shares



(b) Size distortion by firm type for different ω_a



(c) HHI index by worker log-ability



(d) Ability threshold by firm type

Figure 2: Equilibrium Employment Market Shares, Firm-Size Distortions, Concentration, and Ability Thresholds at the Baseline Calibration

Notes:

This figure summarizes equilibrium outcomes at the baseline calibration reported in Table 1. Panel (a) shows equilibrium employment market shares for three representative firms—the bottom, median, and top firm in the productivity distribution—by worker log-ability. Panel (b) reports the corresponding firm-size distortions, measured as the log deviation of equilibrium employment from the efficient allocation, for the baseline ω_a and a lower one $\omega_a = 0.5$. Panel (c) displays the Herfindahl–Hirschman index ($HHI_j(a)$) across the worker log-ability distribution, summarizing market concentration by ability. Panel (d) plots the ability threshold $\tilde{a}_j$, defined as the minimum worker ability employed in firm j, as a function of firm type z. Markdowns by worker ability are omitted because they mirror the market-share distribution across abilities, and the average markdown by worker log-ability is similarly a mirror image of the $HHI_j(a)$ pattern. All results are computed at the baseline parameter values from Table 1.

4 Empirical Evidence

4.1 Data Sources

This section introduces the empirical evidence that informs the analysis. The evidence can be interpreted in two complementary ways. First, it provides a set of model-free empirical facts about sorting, segmentation, and wage-setting in labor markets. Second, these facts can be viewed through the lens of the theoretical framework developed in this paper.

In the subsequent calibration, I use a subset of these empirical moments as targets to discipline the model quantitatively, while leaving others deliberately untargeted to serve as validation moments.

4.1.1 Italian Data

Social Security Data (INPS). The main dataset used in the empirical analysis is the matched employer–employee microdata for Italy obtained through the VisitINPS Scholars access to the administrative archives of the Istituto Nazionale della Previdenza Sociale (INPS). The VisitINPS panel covers the universe of private-sector dependent employees whose employers make social-security contributions over the period 1983–2024. To account for major changes in labor-market institutions over this horizon, I divide the data into rolling sample windows aligned with the timing of key institutional breaks[26] . The data exclude the self-employed, public employees, agricultural workers, and contractors. Each administrative spell record contains a harmonized employer identifier, spell start and end dates, gross labor income (including bonuses and overtime), contributory weeks, contract type, and qualification codes distinguishing blue- and white-collar workers from managers. Both employers and employees are observed at the municipality level, which is aggregated into commuting zones following the official crosswalk provided by the institute. The worker-level data are merged with an anagraphic registry providing demographic information such as year of birth and, when applicable, year of death. A corresponding firm registry supplies legal form, industry classification, geographic location, and year of foundation.

The dataset includes two distinct firm identifiers: (i) an *enterprise* identifier that aggregates all establishments owned by the same legal entity, and (ii) a *location–commodity sector* identifier that captures the specific activity–location unit within the enterprise. Each location–commodity sector identifier is uniquely associated with one industry and one commuting zone, though it does not necessarily coincide with an establishment identifier, as multiple establishments within the same location–sector combination appear under a single code. The activity–location identifier provides a more appropriate definition of the relevant local labor market in which firms compete for workers. Accordingly, I adopt the location–commodity sector identifier as the firm unit throughout the

[26] I divide the sample into six rolling windows corresponding to major phases of the Italian labor market: (i) 1983–1990 (pre-Euro, pre-Treu reforms), (ii) 1991–1997 (early 1990s crisis and social pacts), (iii) 1998–2007 (Eurozone entry and the Biagi reform), (iv) 2008–2013 (global financial crisis and sovereign-debt crisis), (v) 2014–2019 (Jobs Act and subsequent recovery), and (vi) 2020–2024 (COVID-19 and the subsequent inflation shock).

analysis; henceforth, the term *firm* refers to this unit.

I apply a set of sample restrictions to construct an annual panel suitable for analysis. The core sample includes full-time individuals aged 20–65 and excludes managers[27], apprentices, and special categories (e.g., aviation workers), retaining only white- and blue-collar employees.

I further restrict the sample to private, for-profit employers—corporations, partnerships, profit-oriented cooperatives, and sole proprietorships—excluding public entities, non-profit organizations, religious or educational institutions, and other residual legal forms. These excluded categories operate under distinct objectives, wage-setting mechanisms, and regulatory frameworks that are not comparable to those governing market-based employment relationships. Each observation contains information on total compensation and weeks worked during a job spell. Because hours of work are not observed, restricting attention to full-time jobs minimizes the confounding effect of hours dispersion that could otherwise bias measures of hourly pay. Wages for each spell are computed as the sum of regular and excess taxable income plus figurative events,[28] and are deflated to 2022 euros. Each worker is assigned to a single job per year, corresponding to the spell with the highest annual earnings. The resulting measure of the worker's annual wage is the real weekly wage from that spell. Observations with implausibly low weekly wages (below €50) are excluded. Firm-level employment is measured as the total number of headcounts derived from these selected job spells. Unless otherwise noted, cross-sectional results are reported for 2014–2019, the last period for which balance-sheet information is available.

Occupation and Education Extract (ISCO Codes). A separate extract from the INPS archives provides detailed information on workers' occupation and education, coded according to the International Standard Classification of Occupations (ISCO). These variables are derived from mandatory employer notifications filed when contracts are initiated or modified. Since 2010, employers have been required to report ISCO codes for all new contracts and subsequent contract changes, so coverage increases steadily from that year onward. The data are therefore available primarily for *movers*—workers who change employer or contract—and are used to enrich worker-level education and occupation information whenever present. Because these variables are not universally observed, analyses relying on occupation or education data are based on the subset of movers.

CERVED Balance-Sheet Data (Firm-Level Analysis). The analysis is complemented with firm-level financial information from the CERVED database, merged to the INPS data via the enterprise identifier. CERVED provides annual balance-sheet variables—including total assets, turnover,

[27] Managers are excluded because managerial quality may primarily reflect into firm-level productivity and organizational decisions (Lucas Jr, 1978), rather than the production-side labor market mechanisms modeled here. Moreover, managers' labor supply decisions are unlikely to follow the behavioral mechanisms assumed in the model, and empirical evidence on managerial labor-supply elasticities remains limited.

[28] Figurative events (*eventi figurativi*) are administrative imputed earnings recorded by INPS for periods in which workers receive social insurance benefits—such as short-time work, sickness, or parental leave—during which contributions are credited even though no direct wage payment occurs.

value added, revenue, wage bill, legal form, year of foundation, and firm status (active, suspended, or closed)—for the universe of incorporated businesses in Italy over the period 1996–2018. Because CERVED financials are reported at the enterprise level rather than by location–commodity sector, each enterprise is linked to the location–commodity sector with the highest cumulative employment across years, thereby assigning it to a unique local labor market. Intermediate inputs are measured using the accounting definition of value added (revenues minus intermediate consumption), and the capital stock is constructed as the sum of tangible and intangible assets.

Summary statistics. For the period 2014–2019, the final analysis sample comprises 1,566,564 unique firms and 12,951,361 unique workers. The mean worker age is 42 years. The median weekly wage is €427 and the mean weekly wage is €477, with a standard deviation of log wages equal to 0.47. At the firm (location–commodity sector) level, the median firm employs two workers while the mean firm employs nine; the standard deviation of log employment is 1.12. The average firm age is 18 years.

4.1.2 German Data

To complement the Italian analysis, I use the German Sample of Integrated Employer–Employee Data (SIEED) provided by the Institute for Employment Research (IAB). Using a second country is informative both empirically—by highlighting institutional and market differences—and theoretically, by showing whether the model's mechanisms are specific to Italy or general across settings.

The SIEED contains a representative 1.5% sample of all German establishments, covering complete employment biographies of workers, including periods when they are not employed by sampled establishments. Establishments are categorized by ownership type, 2-digit industry, and geographic location (141 labor markets). Worker information includes total earnings, days worked annually, and detailed characteristics such as education, occupation, employment status (part- or full-time), age, nationality, and sex. The dataset also includes establishment and worker AKM fixed effects estimated on the full administrative sample following Card, Heining, et al., 2013; see below for more details.[29]

Following the original construction, I divide the data into five overlapping periods—1985–1992, 1993–1999, 1998–2004, 2003–2010, and 2010–2017. From the spell-level data, I build an annual panel following a similar cleaning procedures as for Italy. The sample is restricted to individuals aged 20–60 who are fully employed in establishments located in West Germany, with real daily wages above €10. I exclude vocational training, freelance, and part-time jobs to mitigate the confounding effect of hours dispersion. When multiple employment spells overlap, I retain the highest-paying spell as the main job for that year. Wages, recorded on a daily basis, are deflated to 2015 euros, and top-coded observations are adjusted using standard imputation methods.

The final dataset spans 1985–2017 and, in the last period (2010–2017), includes 887,497 unique establishments and 2,575,727 workers. The mean establishment employs 5.5 workers, with stan-

[29] See Lochner et al., n.d. for further documentation.

dard deviation of log employment distribution equal to 0.95. The mean log wage is 4.5—approximately €90 per day—with a standard deviation of 0.48. The average worker age is 42 years.

4.1.3 Local Markets, Worker Types, and Firm Heterogeneity

Local Labor Market Definition. Each firm is assigned to a local labor market. Because this assignment is inherently discretionary, I employ two alternative definitions. The first defines a local labor market as the intersection of a 3-digit industry and a commuting zone, as in D. Berger et al., 2022 and Yeh et al., 2022. The second defines it as the intersection of a 3-digit occupation and a commuting zone, as in Felix, 2021. Since the boundary between these categories is partly arbitrary, I report results under both definitions whenever possible to limit the influence of classification-driven differences. When the local labor market is defined by occupation, I redefine the firm identifier at the occupation level—that is, each firm–occupation pair is treated as a distinct unit—so that the number of firm identifiers mechanically increases relative to the industry-based definition. For analyses requiring firm financial variables—which are not available in disaggregated form at the firm-occupation level—I use the 3-digit industry–commuting-zone definition as the baseline. There are 271 three-digit industries, 145 three-digit occupations, and 724 commuting zones, yielding a total of 77,078 distinct markets under the industry–commuting-zone definition and 66,810 under the occupation–commuting-zone definition. For the German data, I use a definition that combines three-digit occupation and 141 local labor markets for a total of 14550 local labor markets.

Indirect Inference Identification of Unobservable Types. A central challenge in studying labor-market sorting and wage determination is that much of the relevant heterogeneity is unobservable to the researcher. Because this heterogeneity cannot be directly measured in administrative data, it must instead be inferred from the structure of wage variation across workers and firms. To do so, I estimate a two-way fixed-effects decomposition of workers' log wages, separating individual heterogeneity from firm-specific wage components (Abowd et al., 1999; Card, Heining, et al., 2013; Song et al., 2019; Bonhomme et al., 2022).

To operationalize this idea, I estimate an additive wage equation of the form[30]

$$\log w_{a,ij,t} = \alpha_a + \psi_{J(a,t)} + x'_{at}\beta + \epsilon_{a,ij,t}, \tag{14}$$

where $\log w_{a,ij,t}$ denotes the log real daily wage of worker a in year t in firm ij, α_a is a worker-specific component capturing time-invariant ability, $\psi_{J(a,t)}$ is a firm-specific (or firm-type–specific) pay premium associated with the employer $J(a,t)$, x_{at} is a vector of time varying observable characteristics, and $\epsilon_{a,ij,t}$ is an idiosyncratic residual. The covariate vector x_{at} includes year and profession (white- vs. blue-collar) fixed effects, along with a flexible polynomial in age and experience fully interacted with profession. This specification allows for distinct life-cycle and experience

[30] Appendix B.5 contains additional information on how the empirical estimation is implemented.

profiles across occupational groups, ensuring that the estimated fixed effects capture persistent worker and firm heterogeneity rather than systematic differences in age–earnings profiles.

For the Italian data, $J(a,t)$ indexes the *firm cluster* identified through the discretization procedure of Bonhomme et al., 2022, rather than a specific firm. I estimate worker and firm effects using this clustering approach, which extends the standard Abowd et al., 1999 framework to mitigate the well-known limited mobility bias. Specifically, I apply a pre-estimation K-means clustering algorithm to the distribution of residualized log wages at the firm level. This procedure groups firms into a finite number of latent "types" that approximate the underlying firm heterogeneity relevant for pay-setting and sorting. While firms are discretized into K-means clusters, worker effects are estimated individually, as in Lamadon et al., 2022.

For the German data, $J(a,t)$ refers to the actual establishment employing worker a in year t. Because the full administrative universe is not accessible, I rely on the AKM-style worker and establishment fixed effects pre-estimated by the Institute for Employment Research (IAB) on the complete German administrative sample following Card, Heining, et al., 2013.

The usual interpretation of the worker and firm fixed effects is as latent components of productivity. The worker effect α_a captures persistent, portable productivity, whereas the firm effect ψ reflects persistent pay premia associated with firm-specific productivity, rents, or labor market power. The economic framework developed in this paper does not impose the log-additive wage equation of the AKM form in (14), but instead allows for heterogeneous markdowns, complementarities, and spillover effects across workers within firms. In Appendix B.5, I show using model-simulated data that the within–local labor market rankings of worker and firm types obtained from the AKM fixed effects closely align with the corresponding rankings of the structural types, z and a[31]. This correspondence supports the use of AKM-based rankings within local labor markets as *indirect inference* valid proxies for latent heterogeneity.

Alternative Measures of Firm Heterogeneity. While worker heterogeneity is inferred from estimated fixed effects, firm heterogeneity can be measured in several complementary ways that are consistent with the model. In the main analysis, I use the firm pay-premium type—derived from the clustering—as the baseline measure of firm heterogeneity. For robustness, I also consider alternative indicators, including (i) firm revenue productivity from balance-sheet data, (ii) the firm's average wage, and (iii) the average characteristics of co-workers, measured as the mean co-worker wage or mean co-worker fixed effect.

Firm-level revenue productivity is estimated from the CERVED balance-sheet data using the semiparametric production-function estimator of Levinsohn and Petrin, 2003, with the correction proposed by Ackerberg et al., 2015 (details in Appendix B.6). Specifically, I estimate Cobb–Douglas value-added production functions separately by three-digit industry and obtain firm-level log productivity ω_{ij} as the residual term.[32]

[31] Ranking firms and workers within a local labor market ensures that most of the variation in their pay fixed effects reflects differences in underlying heterogeneity rather than differences in competitive environments.

[32] The estimation controls for year fixed effects, treating value added as output, employment as the freely variable

Following Foster et al., 2008, $\omega_{ij,t}$ represents firm revenue productivity. Although this measure is based on value-added data and thus reflects revenue rather than physical productivity, the two concepts are tightly linked, and their within–local labor market rankings are effectively interchangeable. Accordingly, I use the ranking in firm revenue productivity as an empirically valid proxy for the underlying firm type.

4.2 Hiring Thresholds

A key implication of the model developed in Section 3 is that, in the presence of within-firm spillovers, higher-quality firms set higher ability thresholds—defined as the minimum worker ability among their hires.

I begin the empirical analysis by examining how firms' hiring standards correlate with measures of firm quality. The analysis is restricted to firms and workers transitioning between firms within markets containing at least ten such firms or workers to ensure meaningful decile rankings. This restriction provides comparable measures of worker and firm quality and limits confounding from employment frictions such as firing costs. Between 2014 and 2019, the data include an average of 1,828,121 worker transitions, corresponding to 22% of all workers. In the German data, the share of switchers is similar, with an annual average of 251,412 worker transitions.

For each firm i located in labor market j and year t, I indirectly infer its hiring threshold as the minimum worker fixed effect among new hires from the job market:

$$\widetilde{a}_{ij,t} = \min\{\alpha \mid \alpha \in \text{new hires of firm } i \text{ in market } j \text{ at time } t\},$$

where α denotes the worker fixed effect.[33] To ensure comparability across markets, I standardize each firm's threshold so that results are expressed in standard deviations of the local labor market worker fixed-effect distribution.[34] I then estimate the following empirical relationship between firms' hiring thresholds and their position in the local firm-type distribution:

$$\widetilde{a}_{ij,t} = \beta_0 + f(\text{Firm Decile}_{ij,t}) + \beta \log(\text{New Hires}_{ij,t}) + \gamma_m + \gamma_t + \varepsilon_{ij,t}, \tag{15}$$

where $\widetilde{a}_{ij,t}$ is the standardized minimum worker fixed effect among new hires at firm i in market j and year t. Firm decile is again measured by its position in the within local labor distribution, and $f(\cdot)$ denotes a fully nonparametric specification based on firm-decile dummies.

The only control included in the preferred specification is $\log(\text{New Hires})$ included for two main reasons. First, it absorbs temporary hiring shocks or firm expansions that mechanically

input, capital as the state variable, and material inputs as the proxy for unobserved productivity shocks. The procedure follows the approach of Levinsohn and Petrin, 2003, originally pioneered by Olley and Pakes, 1996, and incorporates the correction proposed by Ackerberg et al., 2015. A third-degree polynomial approximation to the control function is used to obtain firm-specific estimates of log productivity $\omega_{ij,t}$ as the residual.

[33] As an alternative, I also use the average fixed effect among new hires (as in Carrillo-Tudela et al., 2023). Results are nearly identical, but the minimum-based measure is preferred as it more directly reflects the firm's hiring threshold.

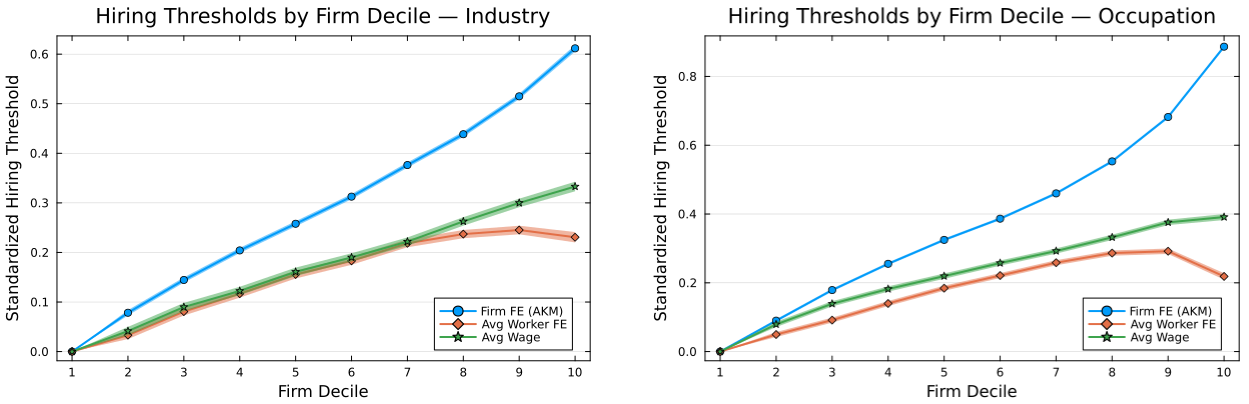
[34] Market and year fixed effects already ensure that the relationship is estimated within local labor markets.

lower the minimum observed hire as firms become less selective (Carrillo-Tudela et al., 2023). Second, because the dependent variable is a minimum order statistic, larger hiring cohorts mechanically generate lower minima even absent true changes in selectivity. Controlling for hiring intensity therefore isolates the structural link between firm quality and screening behavior. Year (γ_t) and market (γ_m) fixed effects absorb aggregate and local conditions. Alternative control sets—including worker composition, firm age, and additional firm observables—are presented in Appendix C.2 and yield highly similar results.

Figure 3 plots the nonparametric relationship implied by equation (15). Each panel shows estimates obtained under the preferred specification for Italy, using three alternative firm-type measures: the AKM firm fixed effect, the average worker fixed effect among incumbent employees, and the average incumbent log wage. Across all definitions, the relationship is nearly perfectly linear and strongly positive: higher-ranked firms hire strictly higher-minimum-ability workers.

Quantitatively, when defining local labor markets by *industry–commuting zones*, firms in the top decile set hiring thresholds approximately 0.64 standard deviations above those of bottom-decile firms. When local markets are defined by *occupation–commuting zones*, segmentation is stronger, with an implied difference of roughly 0.84 standard deviations between the top and bottom deciles. Results for Germany exhibit the same qualitative patterns and are presented in Appendix C.2.

Fact 1. *Within each local labor market, higher-ranked firms impose higher screening thresholds: they hire workers with systematically higher minimum fixed effects. This relationship is nearly linear, robust across alternative firm-type measures, and strongest when local labor markets are defined by occupations.*



(a) Industry–Commuting Zone LLM Classification

(b) Occupation–Commuting Zone LLM Classification

Figure 3: Nonparametric Estimates of Hiring Thresholds by Firm Decile

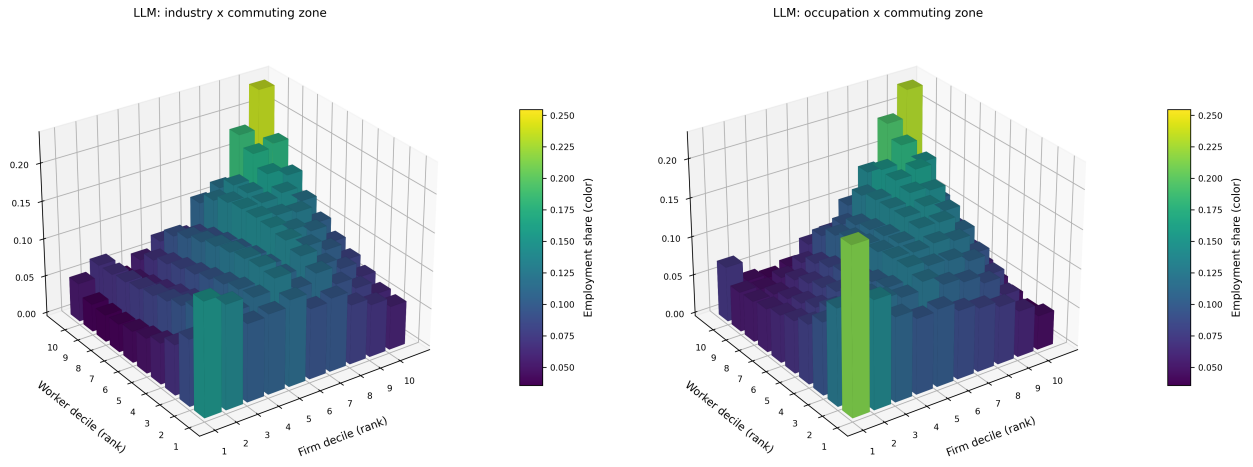
Notes: Each panel reports nonparametric point estimates and 95% confidence intervals from indicator regressions of the standardized hiring threshold on firm–decile dummies, as specified in equation 15. The dependent variable, $\widetilde{a}_{ij,t}$, is the market–year standardized minimum worker fixed effect among new hires. For each local labor market, firms are ranked according to three alternative measures of firm type: (i) the AKM firm fixed effect, (ii) the average worker fixed effect of incumbent employees, and (iii) the average log wage of incumbent employees. Decile 1 is the omitted category, so coefficients are interpreted relative to the lowest-type firms under each ranking. All specifications include year and local-market fixed effects and control for $\log(\text{New Hires})$. Standard errors are heteroskedasticity-robust and clustered at the firm level (industry sample) or at the firm$\times$occupation level (occupation sample).

4.3 Market Shares

I examine how workers of a given type are allocated across firms of different types, providing the empirical counterpart to the theoretical market shares implied by the model. For each year, I compute the employment share of workers in a given decile of the within-year worker fixed-effect distribution who are employed in firms belonging to a specific decile of the firm fixed-effect distribution. Decile assignments are performed separately for each local labor market and each year, and the resulting matrices are averaged over the 2014–2019 period to obtain a representative allocation of worker types across firm types. To ensure balanced decile partitions and to avoid distortions arising in very small markets, I restrict the Italian sample to local labor markets with at least 100 workers and 100 firms.

Figure 4 summarizes the Italian employment-share matrices in visual form. For each worker fixed-effect decile, the figure reports the distribution of employment shares across firm deciles under two alternative definitions of local labor markets: industry–commuting zones (panel a) and occupation–commuting zones (panel b). The full employment-share tables underlying the figure are provided in Appendix C.1[35].

[35]Closely related matrices appear in other papers such as Card, Heining, et al. (2013). The main difference is to



(a) Industry–Commuting Zone LLM Classification

(b) Occupation–Commuting Zone LLM Classification

Figure 4: Employment Market Shares Across Worker Fixed-Effect Deciles

Notes: The figure reports, for each worker fixed-effect decile, the distribution of employment shares across firm fixed-effect deciles in Italy. Both firms and workers are ranked by their AKM fixed effect. Workers and firms are ranked within their local labor market, defined either by industry–commuting zones (panel a) or occupation–commuting zones (panel b). The full underlying matrices are provided in Appendix C.1.

In the absence of segmentation, workers would be evenly distributed across firm types, with larger shares in higher ranked firms. Instead, the data reveal substantial *segmentation*: low-ranked workers concentrate disproportionately in low-ranked firms, and high-ranked workers cluster in high-ranked firms. The gradient is smooth across the worker distribution, with higher-decile workers progressively more likely to be employed in higher-decile firms. High-type firms also employ a larger overall share of the workforce.

Quantitatively, when local labor markets are defined by *industry–commuting zones*, bottom-decile workers allocate 15.0% of their employment to bottom-decile firms but only 5.8% to top-decile firms, even though bottom-decile firms collectively employ just 6.1% of all workers while top-decile firms employ 11.4%. Conversely, top-decile workers allocate 23.6% of their employment to top-decile firms and only 5.1% to bottom-decile firms.

Segregation intensifies when local labor markets are defined by *occupation–commuting zones*: the bottom–bottom share increases to 21.3%, and the top–top share slightly decreases to 22.9%. These magnitudes imply pronounced clustering of low-ability workers in low-type firms within each labor market.

Comparable results are obtained using administrative data from Germany—reported in Appendix C.1—. There, the bottom–bottom employment share is 18.2% (versus 7.6% in top-decile firms) and the top–top share is 17.4% (versus 7.4% in bottom-decile firms). The persistence of strong top–top and bottom–bottom clustering across institutional contexts confirms that segmentation is a general feature of labor markets rather than specific to Italy. As in the Italian case,

classify worker and firm types *within* each local labor market—the relevant competitive environment in the model.

low-quality workers are disproportionately employed in low-quality firms, even though these firms account for a smaller share of total employment.

Additional robustness exercises—reported in Appendix C.1—replace the firm fixed effect with alternative measures of firm quality, such as realized revenue productivity or mean co-worker wages. These specifications deliver closely aligned patterns: bottom-type workers continue to match disproportionately with bottom-type firms, and top-type workers remain concentrated in top-type firms. The magnitudes vary modestly across definitions but consistently indicate strong segmentation.

Fact 2. *Both the Italian and German labor markets exhibit strong segmentation: within each local labor market, low-ranked workers are concentrated in low-ranked firms, whereas high-ranked workers are clustered in high-ranked, larger firms.*

4.4 Concentration Indices by Worker AKM

From Lemma 3, the average markdown faced by workers of a given ability within a local labor market depends on the wage-bill Herfindahl–Hirschman Index (HHI) by worker ability and the corresponding labor supply elasticities. Hence, HHI indices capture the degree of labor market power exerted over each worker-ability type and are tightly linked to the associated welfare losses. Moreover, a key implication of the model developed in Section 3 is that when labor markets are segmented and low-ability workers are disproportionately employed by low-quality firms, the relationship between worker ability and market concentration need not be monotonic. This pattern arises not only from the extent of segmentation itself but also from the dispersion of wage offers within each segment of the labor market, as characterized in Proposition 4[36].

I therefore empirically examine wage-bill HHI indices across the worker ability distribution. This analysis provides an indirect measure of the intensity of labor market power faced by different types of workers and signals how segmentation and wage dispersion shape the concentration of employment across firms within each local labor market.

In the data, I approximate this distribution by discretizing each local labor market into ten ability deciles based on the within-market worker fixed-effect distribution and computing the local wage-bill HHI for each decile. I then aggregate these local HHI measures to the national level by weighting each market's HHI by its total employment in the corresponding ability group. This procedure yields an empirical counterpart to the theoretical relationship between worker ability and market concentration, albeit in an approximate form. The approximation arises for two reasons. First, AKM worker fixed effects capture heterogeneous worker quality only indirectly. Second, the use of ten ability deciles necessarily averages over heterogeneity within each group, smoothing finer variation in worker types.

[36]For instance, if markets were segmented but the underlying distribution of firm productivity were Pareto rather than log-normal, HHI indices for low-ability workers would not necessarily exceed those for high-ability workers. In that case, the thick left tail of the productivity distribution would imply a large number of low-productivity firms hiring low-ability workers, thereby limiting concentration and market power at the bottom of the ability distribution.

Figure 5 displays how aggregate market concentration varies with worker fixed-effect deciles. There is a marked level difference between the two definitions of local labor markets: concentration is systematically higher when markets are defined by industry.[37]

When local labor markets are defined by *industry*, the wage-bill HHI index is relatively stable across the worker-ability distribution, averaging around 0.281. The index reaches a minimum of 0.266 in the second decile and rises gradually to about 0.315 for the top decile. By contrast, when markets are defined by *occupation*, concentration levels are substantially lower on average (mean HHI of 0.16) and display a more pronounced nonmonotonicity: the index declines sharply from 0.17 in the first decile to 0.134 around the fifth decile, before increasing modestly to 0.16 in the upper part of the ability distribution. This pattern indirectly signals that labor market power—and the associated markdowns and welfare losses—may be non monotonic across the worker ability distribution.

I further examine whether these patterns differ systematically between high-skill and low-skill segments of the labor market. Understanding this distinction is important because one might argue that segmentation has limited macroeconomic relevance if it does not extend to the broader workforce but rather is something specific of high-skill workers. The distinction is also relevant for policy: high-skill labor markets may be differentially affected by interventions such as minimum wage policies or collective bargaining institutions. To assess the scope of these effects, I separate the sample into white- and blue-collar occupations, as classified in the dataset, and replicate the analysis for each group separately.[38]

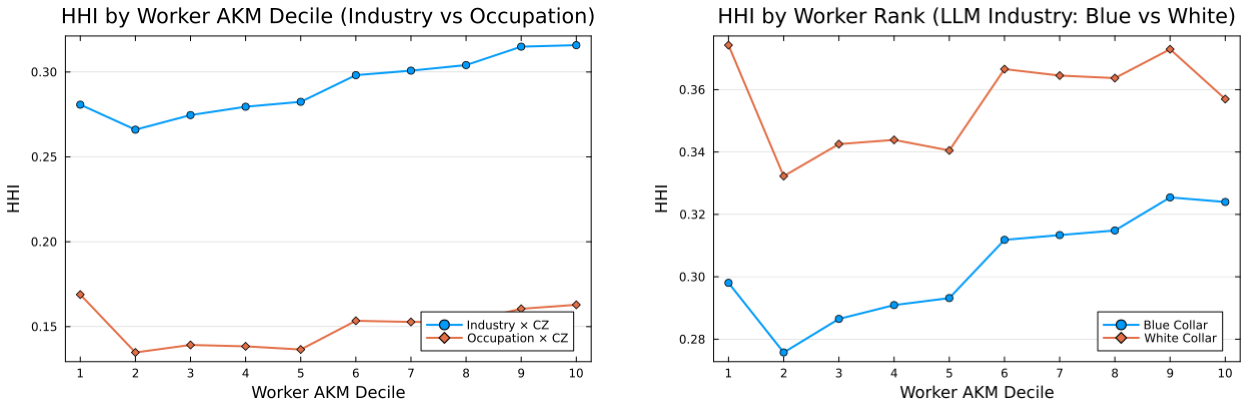
As shown in Panel (b) of Figure 5, concentration is higher overall among white-collar workers, with an average HHI of 0.36 compared to 0.30 among blue-collar workers. Both groups exhibit a U-shaped relationship between HHI and worker fixed effects: concentration is lowest for middle-ranked workers (around the second and third deciles, with HHI values near 0.28 for blue-collar and 0.33 for white-collar jobs) and higher at both tails of the distribution. At the bottom decile, HHI equals 0.30 for blue-collar and 0.37 for white-collar workers, while at the top decile it rises again to 0.32 and 0.36, respectively. The same heterogeneity is also reported for occupation in Appendix C.3. There, the mean HHI under the occupation-based market definition is approximately 0.18. The index displays a non-monotonic pattern across the ability distribution: it falls from about 0.21 in the lowest decile to roughly 0.5 near the second decile, then rises back to about 0.18 among the highest deciles.

Fact 3. *The relationship between labor market concentration and worker AKM rank is U-shaped: wage-bill HHI indices—computed within local labor markets and aggregated using employment weights—are lowest for middle-ranked workers and higher at both ends of the rank distribution. Concentration levels are*

[37] This difference partly reflects the underlying distribution of firms per labor market. The number of firms across Italian local labor markets is more left-skewed under the industry definition, with over 30% of markets containing a single firm, compared to roughly 20% under the occupation-based definition.

[38] 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*.

systematically larger for white-collar relative to blue-collar jobs.



(a) HHI: Industry vs. Occupation LLM classification

(b) HHI: Blue- vs. White-Collar Jobs

Figure 5: Employment weighted HHI Indices by Worker Fixed-Effect Decile

Notes: The figure plots employment-weighted wage-bill Herfindahl–Hirschman indices (HHI) by worker fixed-effect decile, computed within each local labor market as the sum of squared firm wage-bill shares for workers in a given ability decile. Panel (a) reports results separately for labor markets defined by industry (blue line) and by occupation (red line). Panel (b) reports results separately for blue-collar (blue line) and white-collar (red line) jobs, where local labor markets are defined by industry–commuting zones.

4.5 Event Study: Productivity Effects of Unexpected Worker Deaths

I now test the model's central mechanism—that firm-level productivity depends on the average quality of the workforce, and that increases (decreases) in average worker quality lead to corresponding increases (decreases) in firm-level productivity. To identify this relationship, I exploit a quasi-random natural experiment based on *unexpected non-managerial worker deaths*, which generate plausibly exogenous shocks to a firm's production workforce ability composition[39] .

The identifying intuition is directly tied to the theory: if a firm's realized productivity depends on the average quality of its workforce, the unanticipated exit of a worker whose quality lies *below* the firm average should raise the firm's average ability and thereby increase its realized revenue productivity. Conversely, the loss of an above-average worker should lower average ability and reduce productivity. Worker unexpected deaths provide a powerful identification strategy because they induce sudden and plausibly exogenous separations, unrelated to contemporaneous firm-level conditions—such as hiring waves or demand shocks—that would otherwise confound identification.

In addition to shifting the average quality of the workforce, worker deaths generate an exogenous demand for replacement hires. This feature allows me to use the same events to estimate labor-supply elasticities at the firm level, as explained in detailed in Section 5. I next de-

[39]The use of worker or manager deaths as a source of exogenous variation builds on earlier studies, including Jaravel et al. (2018), Azoulay et al. (2019), Bennedsen et al. (2020), Sauvagnat and Schivardi (2024), and Jäger et al. (2024).

scribe the sample construction and empirical strategy used for all death events. The sample employed to study firm-level productivity is necessarily smaller, as it is limited to firms with available balance-sheet data over roughly twenty years rather than the full forty-year horizon covered by the worker-level panel.

4.5.1 Sample construction and matched comparison group

Identifying death events. Death events are identified using the anagraphic registry, assigning the event year d to the calendar year reported in the registry's date-of-death record.[40] To avoid spurious records, I drop cases with post–death employment records, which plausibly reflect reporting errors.

To enhance the plausibility of exogeneity, I restrict the sample to deaths that are sudden and unlikely to reflect long-term health deterioration. The sample is limited to workers younger than 60 years of age who were employed full time in the year of death and during 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 without actual work activity, corresponding to periods of sickness, maternity leave, unemployment benefits, or other legally protected absences. To exclude deaths preceded by protracted illness or labor-market detachment, I restrict the sample to workers with zero figurative weeks in the event year and in each of the four years prior. Finally, I retain only firm–year observations with a single worker death to ensure that results are not confounded by larger accidents or disasters that could have independent effects on firm outcomes.

Firm-level and sample-size restrictions. To ensure that worker deaths represent sharp, idiosyncratic shocks rather than marginal fluctuations, the analysis focuses on firms of stable, small-to-medium size. I retain firms whose average employment in the four years preceding the event lies within the interval $[3,30]$, so that the exit of a single worker meaningfully alters the firm's workforce composition. Firms exhibiting inconsistent post-death employment records—such as continued employment activity more than 30 days after the reported death date—are excluded to remove spurious or misreported cases.

Treatment definition and position indicator. For each death, I construct an indicator $L = \mathbf{1}\{\hat{\alpha}_a < \overline{\hat{\alpha}}_{ij,d-1}\}$, which equals one if the deceased worker's fixed effect from the AKM decomposition, $\hat{\alpha}_a$, is lower than the firm's pre-event mean $\overline{\hat{\alpha}}_{ij,d-1}$ of its workforce. Hence, $L=1$ ("left") denotes the loss of a below-average (low-quality) worker, while $L=0$ ("right") indicates the loss of an above-average (high-quality) worker. The firm-level average is measured in the year preceding the death ($d-1$) to avoid using a contemporaneous value that may already be affected by the shock.

Comparison pool and matched sampling procedure. To construct a control group of placebo death events, I draw from a pool of worker–firm pairs that never experienced a valid death in

[40] Employer notifications reporting deaths are available from 2005 onward and are used as a robustness check.

any event year. This restriction prevents contamination of the control group by treated units and thus avoids the "forbidden comparisons" emphasized in recent event-study research (Goodman-Bacon, 2021; Sun and Abraham, 2021).

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 never experienced a death and satisfies the same pre-event sample restrictions. The matching is performed exactly on gender, two-year age group, the side relative to the firm's pre-period mean (L, left or right), and the decile of the worker's distance from the firm's average fixed effect $(\hat{\alpha}_a - \overline{\hat{\alpha}}_{ij,d-1})$. At the firm level, I match on the pre-period employment bin and on the decile of the firm's average worker fixed effect, both computed as the mean values over the four years preceding the event ($d-4$ to $d-1$). Matching is conducted separately for each event year to ensure temporal comparability. When multiple candidate controls satisfy all matching conditions, I select the control with the smallest absolute difference in worker fixed effects $|\hat{\alpha}_a - \overline{\hat{\alpha}}_{ij,d-1}|$.

The resulting design yields a 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 the analysis, $T = 1$ denotes an actual worker death, while $T = 0$ indicates a placebo (control) event.

Balance, acceptance rate, and final sample. The matching procedure achieves a very high matching rate: 98.9% of treated worker–firm pairs are successfully matched to a placebo event, yielding a final sample of 34,030 matched events (17,015 treated and 17,015 controls). Observations for which no valid match is found are excluded. Table B.21 reports pre-event summary statistics showing that the matching produces a well-balanced comparison group across all key characteristics. Among below-average ($L = 1$) worker deaths, treated and control units are nearly identical in average worker fixed effects (-0.140 vs. -0.137), age (45.8 vs. 45.7 years), experience (25.1 vs. 24.4 years), and log wages (-0.134 vs. -0.161). For above-average ($L = 0$) worker deaths, the treated and control samples are likewise balanced in worker fixed effects (0.178 vs. 0.174), age (50.1 vs. 50.0 years), experience (28.3 vs. 27.5 years), and log wages (0.282 vs. 0.076). Occupation shares are also virtually identical across groups: 20.7% white-collar among low-side events, and 27.2% white-collar among high-side events. Regarding gender composition, deaths are predominantly male dominated, with a share of femal of 15.4% in the low-side events and 11.2% in the high-side events[41]. These similarities confirm that the matched sampling procedure effectively balances observed worker and firm characteristics prior to the event.

Empirical model. I implement a stacked *dynamic event-study* design that flexibly traces the evolution of outcomes around the event. The unit of observation is the firm-event ij observed in relative event time $t = \tau - d_{ij}$, where d_{ij} denotes the year of the death (or its matched placebo).

[41] Gender is mechanically identical across groups, as exact matching was performed on this variable.

The outcome variable is expressed as a deviation from its pre-event mean, $y_{ij,t} \equiv Y_{ij,t} - \overline{Y}_{ij,\text{pre}}$, so that coefficients can be interpreted relative to the pre-death level. This transformation also absorbs any time-invariant heterogeneity across firms, ensuring that identification relies solely on within-firm variation over event time. The estimating equation is

$$y_{ij,t} = \sum_{\tau \neq -1} \left[\gamma_\tau \mathbf{1}\{t = \tau\} + \gamma_\tau^L \mathbf{1}\{t = \tau\} L_{ij} + \beta_\tau \mathbf{1}\{t = \tau\} T_{ij} + \beta_\tau^L \mathbf{1}\{t = \tau\} T_{ij} L_{ij} \right] + \varepsilon_{ij,t}. \quad (16)$$

where $T_{ij} = 1$ if the firm experienced an actual death and $T_{ij} = 0$ for the matched placebo, and $L_{ij} = 1$ if the deceased worker's fixed effect is below the firm's pre-event average (a "low-side" death). The coefficients β_τ trace the dynamic response of firms that lose above-average workers ($L = 0$), while β_τ^L measures the additional differential effect for firms losing below-average workers ($L = 1$). I omit the relative year $\tau = -1$ as the reference period.

Standard errors are clustered at the death-event level to account for serial correlation in firm outcomes and for shared shocks within matched pairs (Abadie and Spiess, 2022). The event window is restricted to $t \in [-3, +5]$ years around the death, ensuring balanced coverage across firms. The plotted coefficients correspond to $\widehat{\beta}_\tau$ for high-quality ($L = 0$) worker losses and $\widehat{\beta}_\tau + \widehat{\beta}_\tau^L$ for low-quality ($L = 1$) losses.

Worker quality is conceptually a continuous treatment, since the impact of a worker's departure depends on the exact magnitude of the departing worker's fixed effect relative to the firm mean. As Callaway et al. (2024) note, causal interpretation of continuous treatments requires stronger forms of the parallel-trends assumption. To avoid these additional requirements, I discretize the treatment into two groups—$L = 1$ versus $L = 0$—which identifies the average treatment effect for firms losing below- versus above-average workers under standard difference-in-differences assumptions.[42]

Identification assumptions. Identification relies on the assumption that worker deaths are as good as random shocks to firms' labor supply. Under this assumption, treated and control firms would have followed parallel outcome trends in the absence of a worker death. The dynamic specification directly tests the plausibility of this assumption by allowing for pre-event coefficients that trace potential differences in trends before the death.

The design embeds two layers of identification. The first difference-in-differences, captured by the coefficients β_τ, compares outcomes of treated and matched control firms that lose workers of similar quality ($L = 0$) before and after the event. The second layer, represented by β_τ^L, adds a triple-difference dimension by comparing how post-event differences between treated and control firms vary across losses of high- and low-ability workers ($L = 1$ versus $L = 0$). This triple-difference identifies the causal effect of a shift in the average quality of the workforce—holding constant the overall shock of a worker exit—under the assumption that, absent the death, high-

[42]The treatment could be discretized more finely to reflect variation in shock size, but the available sample limits statistical power as the number of bins increases.

and low-L firms would have experienced parallel changes in outcomes. Together, these assumptions imply that the estimated effects can be interpreted as the causal impact of an exogenous exit of a worker—below or above the firm's average quality, as measured by the worker's AKM fixed effect—induced by an unexpected death.

4.5.2 Results: effect on Firm-Level Revenue Productivity

The analysis of firm-level revenue productivity is restricted to the period 1996–2018, for which CERVED balance-sheet data are available. The outcome variable is the log of firm revenue productivity, $\omega_{ij,t}$, estimated from value-added production functions following Ackerberg et al. (2015) as described in Section 4.1.3. The resulting panel includes approximately 10,000 worker–firm event pairs per event time.

Table B.22 presents pre-event ($t = -1$) summary statistics—reporting means, medians, and standard deviations—for treated and control firms separately by L group. The matching procedure yields highly comparable treated and control firms across all dimensions. Firm size, age, and productivity are closely aligned: the average firm employs about ten workers, is roughly 15–16 years old, and has an average log revenue productivity around 1.25. The distributions of firm pay premia, employment shares, and worker fixed effects are also nearly identical across treatment status.

Dynamic specification. Figure 6 visualizes the estimated event-time coefficients from the dynamic difference-in-differences specification in equation (16). Each coefficient measures the deviation in firm revenue productivity, $y_{ij,t}$, relative to the pre-event period ($t = -1$). Panel (a) plots the estimated coefficients $\widehat{\beta}_\tau$ for treated firms that lost above-average workers ($L = 0$), while Panel (b) shows the corresponding linear combination $\widehat{\beta}_\tau + \widehat{\beta}^L_\tau$, which captures the total effect for firms that lost below-average workers ($L = 1$). Shaded areas denote 95% confidence intervals, and standard errors are clustered at the firm-death event level.

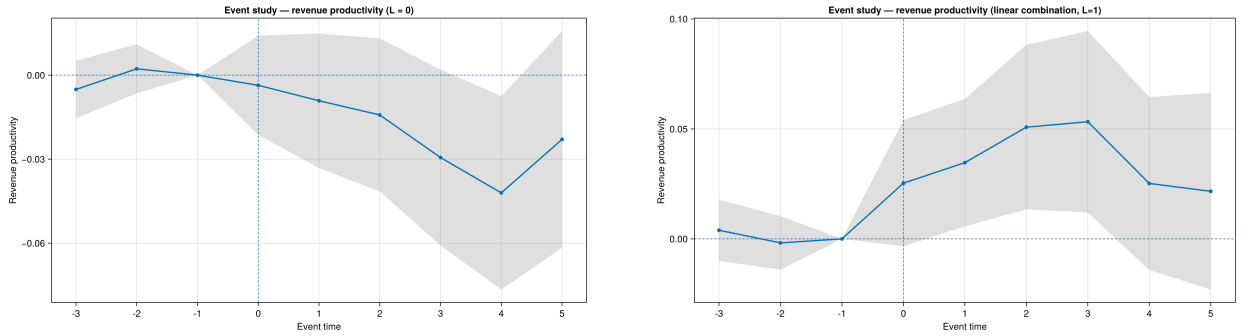
The coefficients for the two pre-event years ($t = -3$ and $t = -2$) are close to zero and statistically insignificant, indicating no systematic differences in pre-treatment trends between treated and control firms. This absence of pre-trends lends credibility to the parallel-trends assumption underlying the identification strategy and supports the interpretation of the post-event dynamics as causal effects of the worker death shocks.

Panel (a) of Figure 6 shows that following the death of an above-average worker ($L = 0$), firm productivity tends to decline. The point estimates are negative in all post-event periods, statistically significant only at year four, with a maximum decrease of about 4.2%. This pattern is consistent with the theoretical prediction that the loss of a high-ability worker lowers firm productivity.

Panel (b) displays the corresponding linear combination of coefficients ($\widehat{\beta}_\tau + \widehat{\beta}^L_\tau$), which captures the total effect of a below-average worker's death ($L = 1$). In this case, firm productivity rises

steadily over the first three years after the event, peaking at roughly 5.0–5.3% around $t = 2$–$t = 3$ ($p < 0.01$), before mean-reverting.

Overall, the dynamic specification delivers a consistent pattern: the unexpected death of a below-average worker increases firm revenue productivity, whereas the loss of an above-average worker lowers it. This evidence is consistent with the model's mechanism whereby firm productivity depends on the average quality of its workforce—an exogenous improvement in worker composition, here induced by an unexpected death, raises realized productivity.



(a) Above-average worker deaths ($L = 0$)

(b) Below-average worker deaths ($L = 1$)

Figure 6: Dynamic Event-Study Estimates of Firm Revenue Productivity

Notes: The figure reports event-time coefficients from the dynamic difference-in-differences specification in equation (16). Each coefficient measures the deviation in firm revenue productivity, $y_{ij,t}$, relative to the pre-event period ($t = -1$). Panel (a) plots the coefficients $\widehat{\beta}_\tau$ for treated firms that lost above-average workers ($L = 0$), comparing them to matched controls. Panel (b) plots the corresponding linear combination $\widehat{\beta}_\tau + \widehat{\beta}^L_\tau$, which captures the total effect for firms that lost below-average workers ($L = 1$). Standard errors are clustered at the event level, and shaded areas represent 95% confidence intervals.

Taking stock. The analysis of unexpected worker deaths provides a clean empirical test of the model's central mechanism linking workforce composition and firm productivity. I find a clear and internally consistent pattern: firms that lose below-average workers experience a statistically significant increase in revenue productivity, whereas firms that lose above-average workers experience a moderate decline. The dynamic estimates reveal no differential pre-trends, supporting the credibility of the identification strategy, and the effects materialize gradually within two to three years.

These findings speak directly to the model's congestion mechanism. In the Cobb–Douglas production formulation, firm output depends on headcount and on a TFP term given by the average quality of employees. Low-ability workers therefore impose a negative externality on firm efficiency: holding employment fixed, replacing a low-ability worker with a higher-ability one raises TFP and improves production efficiency. An unexpected death removes one worker; when the lost worker is below average, the shock simultaneously improves the composition of the workforce and alleviates the *congestion* created by low-ability workers, leading to a measurable increase in revenue productivity. Conversely, the loss of an above-average worker lowers average ability

and reduces productivity. The empirical patterns align tightly with this mechanism.

Fact 4. *An exogenous worker exit induced by a plausibly unexpected death raises firm revenue productivity when the departing worker's fixed effect lies below the firm's average, and lowers it when the worker's fixed effect lies above the firm's average.*

5 Model Calibration and Quantification

In this section, I take the model to the data and discipline its key parameters through calibration. I consider two versions of the model: a benchmark specification with homogeneous workers and a baseline specification with heterogeneous workers. The calibration serves three main purposes. First, it quantifies the production inefficiencies generated by labor market power and how these change when labor market segmentation is introduced. Second, it evaluates the markdowns and welfare losses across workers with different levels of ability, thereby identifying who bears the largest costs of labor market power. Third, it allows me to assess the extent and nature of sorting in the data—specifically, how worker and firm types interact in equilibrium. For calibration purposes, a local labor market is defined as the intersection between a three-digit industry and a commuting zone.[43] This industry–commuting-zone definition delivers a conservative calibration: segmentation is even stronger when local labor markets are defined by occupations, so any benchmark-model deviations I uncover are, if anything, understated.

The distribution of welfare gains and losses across both workers and entrepreneurs is measured by the percentage change in per capita consumption required to equalize steady-state utilities relative to the decentralized Pareto-efficient equilibrium, in which workers are paid their marginal product of labor, as characterized in Proposition 3.[44]

5.1 Taking the Model to the Data

Externally Calibrated Parameters and Local Labor-Market Structure. A subset of parameters is fixed externally at standard values. I set the real interest rate to $R = 0.10$ and the parameter governing the Frisch elasticity of labor supply to $\varphi = 0.5$. The curvature of utility from consumption is calibrated so that the Marshallian elasticity of labor supply equals 0.06, in line with Keane (2011), which implies $\sigma = 0.83$. The depreciation rate is set to $\delta = 0.08$.

The model is simulated over $M = 2000$ local labor markets, each containing up to 200 firms. The number of firms across local labor markets is assumed to follow a Pareto distribution. The shape and scale parameters of this distribution are estimated by maximum likelihood using the

[43] The calibration relies on firm-level balance-sheet data, which are available only at this level of aggregation. Defining local labor markets by occupation would fragment firm identifiers across multiple categories, preventing the use of balance-sheet information.

[44] Formally, for workers, $\lambda(a)$ satisfies $U\Big((1+\lambda(a))\frac{C(a)}{f_a(a)}, \frac{N(a)}{f_a(a)}\Big) = U\Big(\frac{C_{\text{eff}}(a)}{f_a(a)}, \frac{N_{\text{eff}}(a)}{f_a(a)}\Big)$, and for entrepreneurs, $\lambda(e)$ satisfies $U((1+\lambda(e))C(e)) = U(C_{\text{eff}}(e))$.

observed empirical distribution of firms per market. Figure B.3 compares the empirical distribution to its model-implied counterpart.

Production Parameters. From firms' first-order condition for capital (Equation 9), each firm chooses a capital share of value added equal to $(1-\gamma)\alpha$. 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 procedure yields an aggregate capital share of 0.2117.[45]

I then recover the effective labor output elasticity, $\alpha\gamma$, using the sector-level production function estimates described in Section 4.1.3 and aggregate these across three-digit sectors using employment weights. Together, these two calibration steps jointly pin down the pair (α, γ) used in the quantitative analysis. The resulting parameter values are $\gamma = 0.775$ and $\alpha = 0.940$.[46]

Labor Supply Elasticities. To discipline labor-supply behavior, I estimate two elasticities that govern worker adjustment within and across local labor markets. The within-market elasticity η measures how workers substitute across firms within a local labor market, while θ captures how they substitute across markets. I identify η using a quasi-experimental design that exploits workers' unexpected deaths as exogenous firm-level labor demand shocks affecting new hires. I then calibrate θ by indirect inference, targeting the cross-sectional relationship between firm-level wedges $\tilde{\psi}$ and market shares. The wedge $\tilde{\psi}$ represents the firm-specific distortion arising from labor market power (see Proposition 2) and can be measured analogously to Morlacco (2019) and Yeh et al. (2022). This strategy is similar in spirit to that of Edmond et al. (2023), who calibrate the model-implied relationship between markups and market shares in an oligopolistic framework.

The calibration explained below yields the following estimated pair of labor-supply elasticities. The estimated within-market elasticity is $\widehat{\eta} = 7.69$, indicating a relatively elastic supply of new hires within local labor markets. The calibrated across-market elasticity is $\widehat{\theta} = 0.84$, consistent with substantially lower substitutability across markets than within them. Together, these values discipline the strength of worker reallocation margins that govern the model's heterogeneity in firm-level labor-market power.

Within-Market Elasticity (η). To estimate η, I exploit firm-level variation in hiring responses to plausibly exogenous shifts in labor demand induced by unexpected worker deaths. The empirical specification builds on the isomorphism between employment and new-hire labor supply established above: in steady state, separations are offset by inflows of new hires, so the same nested-CES structure governs both stocks and flows. Formally, the flow of type-a new hires to firm ij follows the inverse labor-supply relationship in equation (2), where $n_{ij,t}(a)$ and $w_{ij,t}(a)$ are

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

[46] D. Berger et al. (2022) calibrate $\alpha = 0.940$ and $\gamma = 0.808$ for the U.S. economy.

interpreted as, respectively, the flow of new hires $n^{\text{nh}}_{ij,t}(a)$ and their wage $w^{\text{nh}}_{ij,t}(a)$. Aggregating over ability types yields total new hires $h^{\text{nh}}_{ij,t}$ and the composition density $g^{\text{nh}}_{ij,t}(a) \equiv n^{\text{nh}}_{ij,t}(a)/h^{\text{nh}}_{ij,t}$. A first-order log-linearization with respect to a small, firm-specific wage change yields the discrete-time empirical analogue

$$\begin{aligned}\Delta \log h^{\text{nh}}_{ij,t+1} = \eta \underbrace{\left(\mathbb{E}_{g^{\text{nh}}_{ij,t+1}}\left[\log w^{\text{nh}}_{ij,t+1}(a)\right] - \mathbb{E}_{g^{\text{nh}}_{ij,t}}\left[\log w^{\text{nh}}_{ij,t}(a)\right]\right)}_{\text{Change in average log wage of new hires}} \\ + \eta \underbrace{\sum_a \Delta g^{\text{nh}}_{ij,t}(a)\, w^{\text{nh}}_{ij,t+1}(a)}_{\text{Composition effect}} + (\theta - \eta) \underbrace{\mathbb{E}_{g^{\text{nh}}_{ij,t}}\left[\lambda_{ij,t}(a)\, \Delta \log w^{\text{nh}}_{ij,t+1}(a)\right]}_{\text{Oligopsonistic effect}} + \varepsilon,\end{aligned} \tag{17}$$

where $\Delta x_{t+1} \equiv x_{t+1} - x_t$, and

$$\lambda_{ij,t}(a) \equiv \left.\frac{\mathrm{d} \log w_j(a)}{\mathrm{d} \log w_{ij}(a)}\right|_t$$

denotes the total effect on the local-market wage index $w_j(a)$ of a marginal, idiosyncratic change in the wage posted by firm ij at time t[47] [48] [49].

This approximation decomposes the elasticity of new hires with respect to wages into three components: a direct wage response, a composition effect capturing changes in the ability distribution of new hires, and an oligopsonistic adjustment reflecting firms' wage-bill shares.

Estimating equation (17) poses two primary challenges. First, a naive IV regression of changes in the log of total new hires on changes in the average log wage—instrumented with a demand shifter to address amenity-related omitted-variable bias—is subject to the classical oligopsonistic omitted-variable problem identified by BHM. In particular, the market CES wage index enter the first-order approximation and bias the estimate of η. Second, the naive specification neglects that a demand shock—here, an unexpected worker death—may alter the composition of incoming hires, generating a composition bias in the estimated wage response. For example, if the shock induces a shift toward higher-ability workers, the average wage of new hires would rise for a given number of hires, leading the naive regression to understate the true elasticity.

[47] Under homogeneous-worker assumptions equation (17) holds exactly rather than as an approximation, and one can difference out the oligopsonistic term using market–time fixed effects, as in Felix (2021), to identify η. In the present setting this strategy is not available for two reasons. First, the composition term depends on firm-specific changes in the ability distribution of new hires, $\Delta g^{\text{nh}}_{ij,t}(a)$, and therefore does not load on a common market–time component that could be removed by fixed effects. Second, the CES index $w_{j,t}(a)$ is defined at the worker-type level, with infinitely many types a, so the oligopsonistic term involves an infinite-dimensional set of market–time objects, which is infeasible to saturate with market–time fixed effects.

[48] The continuous-time first-order approximation is

$$\mathrm{d} \log h^{\text{nh}}_{ij,t} = \eta\, \mathbb{E}_{g^{\text{nh}}_{ij,t}}\left[\mathrm{d} \log w^{\text{nh}}_{ij,t}(a)\right] + (\theta - \eta)\, \mathbb{E}_{g^{\text{nh}}_{ij,t}}\left[\lambda_{ij,t}(a)\, \mathrm{d} \log w^{\text{nh}}_{ij,t}(a)\right],$$

where $\lambda_{ij,t}(a)$ is defined as above and the total derivative of the CES price index with respect to firm ij's wage is evaluated at time t.

[49] Equation (17) should be interpreted as a local approximation around the observed equilibrium path, tracing the response of new hires to small, firm-specific wage changes induced by idiosyncratic shocks to firm ij, while treating the evolution of market-wide wage indices as given.

I address these concerns in two complementary ways. First, I allow for heterogeneous elasticities by interacting the key wage-change term with an indicator for whether the firm is located in a large local labor market. In large markets, individual firms have negligible wage-bill shares, so strategic (oligopsonistic) responses are muted. Comparing estimates across market-size bins both reduces oligopsony bias in large markets and provides a direct test for strategic wage-setting behavior. Second, I control explicitly for composition changes induced by the shock. In the model of Section 3, a firm-level demand shock from an unexpected worker exit affects the composition of new hires by shifting the firm's average worker type. Accordingly, the empirical specification is saturated with fixed effects that absorb the realized magnitude of the implied compositional shock—that is, the change in the firm-level average worker type implied by the shock—so that identification of η comes from within-composition variation in wages rather than from mechanically driven reweighting of hire types.

Empirical model, estimation, and results. I start with a *naive* specification that corrects for the standard amenity bias by instrumenting employment changes with a labor-demand shifter—the exogenous worker death[50]—but does not account for compositional changes in the quality of new hires or for potential oligopsonistic wage responses. To estimate the within-market elasticity of labor supply, η, I exploit firm-level variation in hiring responses to these shocks using a difference-in-differences design where all variables are expressed relative to firm-specific pre-event means.

The outcome variable is the log deviation of average new-hire wages for firm ij from its pre-event mean, $\Delta\overline{\log w}^{\mathrm{nh}}_{ij,t} \equiv \overline{\log w}^{\mathrm{nh}}_{ij,t} - \overline{\log w}^{\mathrm{nh}}_{ij,\mathrm{pre}}$, and the key regressor is the corresponding deviation of log new hires $\Delta \log h^{\mathrm{nh}}_{ij,t} \equiv \log h^{\mathrm{nh}}_{ij,t} - \overline{\log h}^{\mathrm{nh}}_{ij,\mathrm{pre}}$. Pre-event means are computed over three pre-event periods $t = -3, -2, -1$ and are fixed for each firm–event. Because wages are endogenously determined with respect to employment, I instrument the post-event change in new hires with the interaction between the worker-death indicator and the post-event dummy, $\mathrm{Post}_t \times T_{ij}$, which captures the exogenous labor-demand shock generated by the unexpected worker exit. The empirical model in the naive specification consists of the following two stages:

$$\Delta \log h^{\mathrm{nh}}_{ij,t} = \pi_0 + \pi_1 \big(\mathrm{Post}_t \times T_{ij}\big) + \pi_2 \mathrm{Post}_t + \nu_{ij,t}, \qquad \text{[First stage]}$$

$$\Delta\overline{\log w}^{\mathrm{nh}}_{ij,t} = \beta_0 + \beta_1 \, \widehat{\Delta \log h^{\mathrm{nh}}_{ij,t}} + \beta_2 \mathrm{Post}_t + \varepsilon_{ij,t}. \qquad \text{[Second stage]}$$

where $\Delta x_{ij,t}$ denotes the deviation of $x_{ij,t}$ from its firm-specific pre-event average. $\overline{\log w}^{\mathrm{nh}}_{ij,t}$ denotes the firm-level average log wage of new hires in period t. Standard errors are clustered at the event level (death–firm ij identifier).

This specification is referred to as the *naive regression*. It corrects for the classical amenity bias by exploiting exogenous labor-demand shocks but remains potentially biased due to (i) oligop-

[50]The estimation sample is larger than in the TFPR event study because it exploits the full matched employer–employee panel rather than being restricted to firms with balance-sheet data. The new-hire event study contains approximately 47,000 firm–event observations per event year.

sonistic wage-setting behavior, which flattens the observed wage–employment elasticity, and (ii) composition effects that arise when worker deaths alter the ability distribution of new hires. I therefore extend the empirical specification in two steps. First, I control explicitly for composition changes induced by the worker-death shock. I construct an indicator L_{ij} equal to one if the deceased worker's ability lies below the firm's pre-event average. The specification adds L_{ij} and the triple interaction $(\text{Post}_t \times T_{ij} \times L_{ij})$ as controls, allowing post-event wage changes to differ systematically when the exiting worker is of below-average ability. This adjustment isolates the within-composition wage response by controlling for mechanical shifts in new-hire ability specific to low-type exits.

Second, I allow for heterogeneous elasticities across local markets by interacting the wage-change term with an indicator for firms that account for an above-median share of the local wage bill, denoted SmallMarket_{ij}. The median share is approximately 1.5%, so firms with a wage-bill share below this threshold operate in relatively large, less concentrated markets. These "high-share" firms operate in more concentrated or smaller labor markets and are therefore more likely to possess oligopsony power. In addition to the interaction $\text{SmallMarket}_{ij} \times \Delta\widehat{\log h^{\text{nh}}_{ij,t}}$, I include $\text{Post}_t \times \text{SmallMarket}_{ij}$ to absorb differential level shifts in wages across market types. Let $H_{ij} \equiv \text{SmallMarket}_{ij}$ for brevity. The extended empirical specification can be written as:

$$\begin{aligned}\Delta\overline{\log w}^{\text{nh}}_{ij,t} = \beta_0 + \beta_1\,\Delta\widehat{\log h}^{\text{nh}}_{ij,t} + \beta_2\left(H_{ij} \times \Delta\widehat{\log h}^{\text{nh}}_{ij,t}\right) \\ + \beta_3\,L_{ij} + \beta_4\left(\text{Post}_t \times T_{ij} \times L_{ij}\right) + \beta_5\,\text{Post}_t + \beta_6\left(\text{Post}_t \times H_{ij}\right) + \varepsilon_{ij,t}.\end{aligned} \tag{18}$$

As in the naive specification, estimation proceeds by two-stage least squares (2SLS). In the first stage, the deviation in firm ij's new hires, $\Delta \log h^{\text{nh}}_{ij,t}$, is regressed on the interaction between the death indicator and the post-event dummy, $\text{Post}_t \times T_{ij}$, together with all interaction terms that enter the second stage (including $\text{Post}_t \times H_{ij}$ and $\text{Post}_t \times T_{ij} \times L_{ij}$). In the full specification with heterogeneous elasticities across markets, I also treat the interaction $H_{ij} \times \Delta \log h^{\text{nh}}_{ij,t}$ as endogenous and instrument it with $\text{Post}_t \times T_{ij} \times H_{ij}$. The second stage then estimates the effect of these fitted changes in hiring on average new-hire wages.

Because all variables are expressed relative to pre-event averages, the specification relies exclusively on within-firm variation and absorbs time-invariant heterogeneity by construction. Standard errors are clustered at the death–firm ij level to account for serial correlation within events. The coefficients of interest are β_1 and β_2. The former measures the wage–employment elasticity among firms in low-share (more competitive) markets, while β_2 captures how this elasticity differs in high-share, more concentrated markets. A positive and statistically significant β_2 implies that labor supply is less elastic in small, concentrated markets, consistent with stronger oligopsonistic wage-setting power.

Table 2 reports the estimates of the wage–employment elasticity among new hires. Column (1) corresponds to the naive IV specification, column (2) adds the composition controls based on L_{ij}, and column (3) further introduces heterogeneity across markets via the interaction with SmallMarket_{ij}. In the naive regression, which corrects only for amenity bias by instrumenting employment with

Table 2: Wage–employment elasticity among new hires and instrument relevance

	(1) Naive IV	(2) + Low-type	(3) + Low-type & market het.
Panel A: Second-stage coefficients			
$\Delta \log h_{ij,t}^{\mathrm{nh}}$	0.113^{**}	0.211^{***}	0.134^{**}
	(0.045)	(0.067)	(0.060)
$\Delta \log h_{ij,t}^{\mathrm{nh}} \times H_{ij}$			0.211
			(0.145)
Controls (additional regressors)			
Composition Controls		✓	✓
Panel B: First-stage diagnostic statistics			
SW F for $\Delta \log h_{ij,t}^{\mathrm{nh}}$	41.5	25.8	30.1
SW F for $\Delta \log h_{ij,t}^{\mathrm{nh}} \times H_{ij}$			28.7
Observations	244,570	244,570	244,570
Clusters (events)	43,023	43,023	43,023

Notes: Outcome is the deviation of average log new-hire wages from the firm-specific pre-event mean. All regressors are defined analogously. Column (1) reports the naive 2SLS specification that instruments $\Delta \log h_{ij,t}^{\mathrm{nh}}$ with $\mathrm{Post}_t \times T_{ij}$ and includes Post_t as a control (not shown). Column (2) adds controls for composition effects based on the low-ability indicator L_{ij} and its interaction with the post-treatment period and death. Column (3) further includes interactions with the high wage-bill share indicator H_{ij} and treats both $\Delta \log h_{ij,t}^{\mathrm{nh}}$ and $H_{ij} \times \Delta \log h_{ij,t}^{\mathrm{nh}}$ as endogenous, instrumented by $\mathrm{Post}_t \times T_{ij}$ and $\mathrm{Post}_t \times T_{ij} \times H_{ij}$. All specifications are estimated by 2SLS with standard errors clustered at the death–firm ij level. Robust standard errors in parentheses. Sanderson–Windmeijer (SW) F-statistics report the conditional first-stage F for each endogenous regressor. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

the death shock, the coefficient on the log change in new hires is positive and statistically significant ($\widehat{\beta}_1 = 0.113$, $p < 0.05$). This estimate implies that a 1% increase in hiring is associated with a roughly 0.11% increase in the average wage of new hires, consistent with a relatively elastic within-market labor supply. When the specification is augmented to control for compositional shifts in workforce quality, the estimated elasticity rises to $\widehat{\beta}_1 = 0.211$ ($p < 0.01$). In the full specification with both composition controls and market-heterogeneity interactions, the coefficient on the hiring term remains positive and statistically significant, $\widehat{\beta}_1 = 0.134$ ($p < 0.05$), corresponding to an implied elasticity of $\widehat{\eta} \approx 7.5$, comparable to the value ($\eta = 10$) reported by BHM for the United States. The interaction term for high-share markets is sizable, $\widehat{\beta}_2 = 0.211$ ($p < 0.15$). Although very noisy, the point estimates indicates substantially lower elasticities in less competitive labor markets, in line with stronger oligopsonistic power among firms operating in smaller or more concentrated markets. Instrument diagnostics are reported in the lower panel of Table 2. Sanderson–Windmeijer conditional F-statistics are 30.1 and 28.7 for the two endogenous regressors, respectively, indicating that the instruments are strong for both coefficients of interest. Taken together, these diagnostics suggest that weak-instrument concerns are limited and that the results provide a reliable characterization of within-market labor-supply elasticities.

Across-Market Elasticity (θ). Having estimated the within-market elasticity η, I now turn to the identification of the across-market elasticity θ. To recover θ, I employ an indirect-inference calibration that matches the empirical relationship between firm markdowns and their market positions. The guiding thought experiment is straightforward: a firm operating in a less competitive labor market should exhibit a larger distortion—i.e., a lower $\widetilde{\psi}_{ij}$—than an otherwise identical firm located in a larger, more competitive market.

In the baseline model of Section 3, I abstract from several features emphasized in the empirical literature—such as variable markups, time-varying output elasticities, and intermediate inputs—and show that the firm-level labor wedge $\widetilde{\psi}_{ij}$ can be expressed in terms of wages and marginal products (Proposition 2). For the empirical implementation, I allow for these additional layers of realism: firms may exercise product-market power, use a flexible and undistorted material input, and feature firm-specific output elasticities with respect to labor and materials. Under the same structure for heterogeneous worker productivity as in Section 3, the composite wage wedge $\widetilde{\psi}_{ij,t} = \mathbb{E}_{g_{ij,t}(a)}[\mu_{ij,t}(a)\psi_{ij,t}(a)]$ remains identified from observable cost shares and the ratio of output elasticities.

Lemma 9 (Recovering the wage wedge from cost shares)**.** *Let firm ij produce output $y_{ij,t}$ with labor headcount $h_{ij,t}$ 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. As in the model of Section 3, suppose that the marginal product of a worker of type a can be written as*

$$MPL_{ij,t}(a) = \frac{\partial y_{ij,t}}{\partial n_{ij,t}(a)} = \frac{\partial y_{ij,t}}{\partial h_{ij,t}}\,\psi_{ij,t}(a),$$

where $\psi_{ij,t}(a)$ captures heterogeneous marginal products across worker types. If the firm cost-minimizes and the input $x_{ij,t}$ has a marginal price p_x, then the firm-level wedge $\widetilde{\psi}_{ij,t}$ satisfies

$$\frac{1}{\widetilde{\psi}_{ij,t}} = \frac{p_x x_{ij,t}}{\overline{w}_{ij,t}\,h_{ij,t}}\,\frac{\alpha_{l,ij}}{\alpha_{m,ij}}, \tag{19}$$

so that $\widetilde{\psi}_{ij,t}$ is pinned down by the ratio of the material bill to the wage bill and the ratio of output elasticities.

In the simpler environment of Proposition 2, the wedge $\widetilde{\psi}_{ij}$ shows up directly in the labor share, $ls_{ij} = \alpha\gamma\,\widetilde{\psi}_{ij}$. Lemma 9 extends this insight to a richer setting with product-market power and more general production technologies: even when labor's share is not mechanically equal to an output elasticity, the average markdown $\widetilde{\psi}_{ij}$ remains point-identified from the ratio of the flexible input expenditure to the wage bill and the ratio of output elasticities.

In the data, I exploit intermediate expenditure as flexible input expenditure. Following Ridder et al. (2025), one can estimate how this wedge correlates with firm market shares by taking logarithms of (19) and controlling sufficiently for heterogeneity in production elasticities. In the benchmark case of a Cobb–Douglas, time-invariant production function with elasticities varying at the three-digit industry level—and under a competitive material market—this amounts to including three-digit industry fixed effects.

Thus, to calibrate θ, I target using indirect inference the empirical coefficient β in the panel regression

$$\log\left(\frac{m_{ij,t}}{\overline{w}_{ij,t}h_{ij,t}}\right) = \beta\, S_{ij,t}(\text{Wage bill}) + \Psi_{ij,t} + \varepsilon_{ij,t}, \tag{20}$$

where the dependent variable is the logarithm of the material-to-labor expenditure ratio, $S_{ij,t}(\text{Wage bill})$ denotes the firm's wage-bill share within the local labor market, and $\varepsilon_{ij,t}$ is an idiosyncratic error term.

The vector $\Psi_{ij,t}$ includes a comprehensive set of controls that absorb cross-sectional heterogeneity in production technology, potential distortions in intermediate input markets, and aggregate time effects. Specifically, I include five-digit industry fixed effects, allowing production elasticities to vary nonparametrically across narrowly defined sectors; calendar-year fixed effects to capture aggregate shocks; and nonparametric controls for firm size, combined with employment-weighted observations, to alleviate potential biases arising from quantity discounts in intermediate input markets, as documented by Lorenzini and Martner (2025)[51]. In addition, I include nonparametric controls for firm age to capture life-cycle effects.

The wage-bill share is instrumented with the three-year lag of the revenue share for two main reasons. First, measurement error in the firm wage bill would induce both attenuation bias and omitted-variable bias, since the wage bill appears in the denominator of the dependent variable and in the numerator of the regressor. Second, instrumenting helps alleviate concerns that unobserved labor-augmenting productivity not fully absorbed by $\Psi_{ij,t}$ may bias the coefficient downward: firms with higher labor-augmenting productivity both command larger wage-bill shares and exhibit lower measured $\widetilde{\psi}_{ij}$, since a given degree of labor market power translates into a lower material-to-labor ratio (Wu et al., 2025)[52].

Table 3 reports the instrumental-variables estimates of equation (20). In the preferred specification—which includes five-digit industry, age-group, and employment-bin fixed effects, together with a flexible cubic polynomial in log employment—the IV coefficient on the firm's wage-bill share is 0.66 (s.e. 0.03). This estimate implies a strong positive association between firms' labor-market positions and their material-to-labor expenditure ratios, consistent with the model's prediction that weaker local competition translates into larger markdown wedges.

The alternative specifications, which vary the covariates used to control for treatment of production-technology heterogeneity, deliver point estimates of similar magnitude, reinforcing the stability of the relationship. Across all specifications, the first-stage Kleibergen–Paap Wald F-statistics exceed 10^4, far above conventional weak-instrument thresholds, confirming that lagged revenue share provides extremely strong identifying variation.

[51] If quantity discounts take the form of a two-part tariff, the relationship in equation (19) would hold only at marginal prices, which coincide with unit prices only for sufficiently large firms.

[52] Notice that even with variable product market power, only wagebill share affects labor market power while revenue share affects market power in the product market (Gutiérrez, 2023).

Table 3: Relationship Between Material-to-Labor Expenditures and Wage-Bill Share

	(1) Baseline	(2) +5d/age FE	(3) Preferred
Dependent variable:		$\log\left(\frac{m_{ij,t}}{\overline{w}_{ij,t}h_{ij,t}}\right)$	
Wage-bill share	0.564***	0.467***	0.663***
	(0.0288)	(0.0280)	(0.0276)
First stage: Kleibergen–Paap F	1.4×10^4	1.2×10^4	1.4×10^4
Fixed effects and controls:			
3-digit industry FE	✓		
5-digit industry FE		✓	✓
Calendar-year FE	✓	✓	✓
Age-group FE		✓	✓
Employment-bin FE			✓
Log employment polynomial			✓
Observations	461,797	461,797	461,797
Clusters (markets)	42,195	42,195	42,195

Notes: The table reports IV estimates of equation (20). All regressions are weighted by firm employment, and standard errors are clustered at the market level. Column (1) includes three-digit industry and calendar-year fixed effects. Column (2) replaces these with five-digit industry and age-group fixed effects. Column (3), the preferred specification, adds a cubic polynomial in log employment and employment-group fixed effects. The excluded instrument is the lagged (three-year) revenue share. *** $p < 0.01$.

Heterogeneity and Complementarity Parameters. The remaining parameters govern worker and firm heterogeneity (σ_a, σ_z) and the structure of worker–firm complementarities in production (ρ, ω_a).[53] The baseline calibration jointly identifies all remaining parameters using an indirect inference strategy. As discussed in Section 3.5, different combinations of $(\rho, \omega_a, \sigma_a, \sigma_z)$ generate distinct implications for sorting and segmentation in the labor market. Accordingly, I simulate a worker–firm synthetic panel dataset from the model[54] and form moments on this artificial dataset, and compare them to their empirical counterparts. The targeted moments include: (i) the aver-

[53]The across-market elasticity θ is also calibrated by indirect inference. The regression coefficient on firms' wage bill shares almost exactly identifies this parameter. In practice, I initialize θ to match the empirical coefficient, simulate the model, recalibrate the remaining parameters conditional on that value, and iterate until convergence.

[54]The synthetic panel is generated by repeatedly assigning workers to firms according to the equilibrium allocation probabilities implied by the model. A number of workers in proportion to firms are drawn from the ability distribution—mirroring the empirical ratio of workers to firms in the data. Each worker is matched to a firm by first drawing a market conditional on ability, and then a firm within that market using normalized employment shares. The simulation runs for periods corresponding to the sample period in the targeted data, and in each new period, a worker switches to a new firm with probability λ, set to match the empirical job-switching rate. For every matched pair, the simulated dataset records wages, firm and worker identifiers, and other model-implied variables, allowing for a standard AKM wage decomposition on the simulated panel. Details are reported in AppendixD.3.

age employment-weighted average of the worker-rank-specific wage-bill HHI within each worker fixed-effect decile; (ii) the standard deviation of AKM worker fixed effects; (iii) the employment share of workers in the top quartile of the worker AKM fixed-effect distribution employed by firms in the top quartile of the firm AKM distribution, computed within markets with more than 100 firms; and (iv) the employment share of workers in the bottom quartile of the worker AKM fixed-effect distribution employed by firms in the bottom quartile of the firm AKM distribution, computed within markets with more than 100 firms. Heuristically, the standard deviation of worker fixed effects disciplines σ_a; the wage-bill HHI is informative about firm heterogeneity σ_z; the share of top-quartile workers employed by top-quartile firms is informative about the heterogeneity of wages—for a given labor supply elasticity—across top and lower-ranked firms, thereby identifying the strength of firm weight and thus heterogeneity in worker-level production function $1-\omega_a$; and the fraction of low ranked worker in low ranked firms identifies segmentation, thus regulating the remaining parameter ρ, as together with ω_a, it reflects the degree of segmentation implied by the technology. All parameters are jointly estimated to match these moments. Further details on the calibration algorithm, simulation design, and convergence criteria are provided in Appendix D.3.

5.2 Calibration Results and Validation

Calibration Results. The indirect inference procedure delivers a close match between the simulated and empirical moments. Allowing for complementarities in production ($\omega > 0$), the estimated parameters are $\rho = 0.35$, $\omega_a = 0.80$, $\sigma_a = 0.29$, and $\sigma_z = 0.513$. Table 4 shows that the model reproduces well the targeted moments: the dispersion of worker fixed effects, the degree of assortative matching between high-type workers and high-paying firms, the degree of segmentation as measured by the employment share of low ranked workers to low ranked firms, and the average employment-weighted HHI index by worker decile rank. Table 5 reports the full set of parameters under the baseline calibration.

Table 4: Targeted Moments: Data vs. Model

Moment	**Data**	**Model**
Std. dev. of worker fixed effects (WFE)	0.31	0.32
Share of top-quartile workers in top-quartile firms	0.43	0.43
Share of bottom-quartile workers in bottom-quartile firms	0.28	0.28
Employment-weighted HHI index	0.29	0.30

Notes: This table reports the moments targeted in the indirect inference calibration. The empirical moments are computed from the matched employer–employee Italian panel described in Section 4, where the local labor market is defined as a three-digit industry crossed with a commuting zone, over the period 2014–2019. The model moments are obtained from a synthetic simulated panel containing nine times as many workers as firms—as in the data—with workers assigned to firms according to the model-implied market shares. Each worker is reshuffled across firms with probability $\lambda = 0.24$, matching the empirical job-switching rate, over five simulated periods corresponding to 2014–2019. The moments capture, respectively: (i) the dispersion of worker fixed effects estimated from AKM decompositions; (ii) the degree of assortative matching, measured as the share of top-quartile workers employed by top-quartile firms, ranked within local labor markets; (iii) the degree of segmentation, measured as the share of bottom-quartile workers employed by bottom-quartile firms, ranked within local labor markets; and (iv) the employment-weighted wage-bill Herfindahl–Hirschman Index (HHI), computed by worker decile rank, and averaged to reflect market concentration in wage payments.

Table 5: Summary of Calibrated Model Parameters

Parameter	Interpretation	Value
γ	Output elasticity of labor	0.775
α	Returns to scale (labor–capital composite)	0.940
ρ	Complementarity parameter	0.350
ξ	Worker-output scale parameter	1.000
ω_a	Weight on worker ability in production	0.800
η	Within-type labor supply elasticity	7.692
θ	Across-type labor supply elasticity	0.843
μ_a	Mean of worker ability distribution	0.000
σ_a	Std. dev. of worker ability distribution	0.290
μ_z	Mean of firm productivity distribution	0.000
σ_z	Std. dev. of firm productivity distribution	0.513
R	Interest rate	0.100
σ	Utility curvature (CRRA)	0.830
φ	Labor disutility parameter	0.500
δ	Capital depreciation rate	0.080
S	Number of worker types	500
M	Number of local labor markets	1000
Pareto location	Firm size distribution location parameter	1.000
Pareto tail	Firm size distribution tail parameter	0.966
Pareto scale	Firm size distribution scale parameter	3.519
Mass of single-firm markets	Share of LLMs with a single firm	0.212
Max firms per market	Market size cap	200

Notes: This table reports the full set of parameters used in the final calibration. The structural parameters governing heterogeneity and complementarities, $(\rho, \omega_a, \sigma_a, \sigma_z)$, are jointly estimated by indirect inference. The within-market labor supply elasticity η is identified using variation from unexpected worker deaths, and the across-market elasticity θ is chosen to match the empirical relationship between markdown wedges and firms' revenue shares. Parameters γ and α are calibrated to match aggregate moments related to the labor share and returns to scale. The simulated economy features $S = 500$ worker types and $M = 1000$ local labor markets, with firm and worker heterogeneity drawn from log-normal and Pareto distributions. All parameters refer to the steady-state allocation of the calibrated model.

Model Validation. To assess external validity, I compare the calibrated model's implications for a set of non-targeted moments and joint-distributional patterns between workers and firms. Table 6 reports these validation statistics. Overall, the model replicates well the empirical dispersion of wages (0.46 in the model vs. 0.48 in the data) and the share of wage variance originating within firms (46% vs. 41%), indicating that the model generates realistic within- and between-firm wage structure. The distribution of firm sizes is also well matched: the standard deviation of log em-

ployment is 1.19 in the model compared to 1.12 in the data. The model slightly overstates the dispersion in firm fixed effects (0.29 vs. 0.18) and slightly understates the empirical covariance between worker and firm fixed effects (0.010 vs. 0.0138). Finally, the aggregate labor share generated by the model (0.535) aligns well with its empirical counterpart (0.51), obtained from FRED data (Feenstra et al., 2015) and averaged over 2014–2019. Taken together, these non-targeted moments show that the calibrated model successfully reproduces key moments of the joint worker–firm distribution.

Table 6: Untargeted Validation Moments: Data vs. Model

Moment	**Data**	**Model**
Std. dev. of log wages	0.48	0.46
Share of wage variance within firms (%)	41	46
Std. dev. of log employment	1.12	1.19
Std. dev. of firm fixed effects (FFE)	0.18	0.29
Covariance of worker and firm fixed effects	0.0138	0.010
Labor share	0.51	0.535

Notes: This table reports a set of *untargeted* moments used to evaluate the external validity of the calibrated model. These statistics do not enter the indirect-inference objective; instead, they provide an out-of-sample assessment of how well the model reproduces salient empirical patterns in the joint worker–firm distribution. The first two moments summarize wage dispersion and the share of total wage variance attributable to within-firm differences, jointly testing whether the model captures the heterogeneity embedded in worker and firm types. The standard deviation of log employment assesses the model's ability to generate realistic dispersion in firm size. The dispersion of firm fixed effects (FFE) and the covariance between worker and firm fixed effects measure the strength of sorting between worker ability and firm quality. Finally, the labor share compares the aggregate wage bill generated by the model to its empirical counterpart. Overall, the moments reveal that the model reproduces well the main empirical features of wages, employment, and sorting patterns, despite not being directly calibrated to match them.

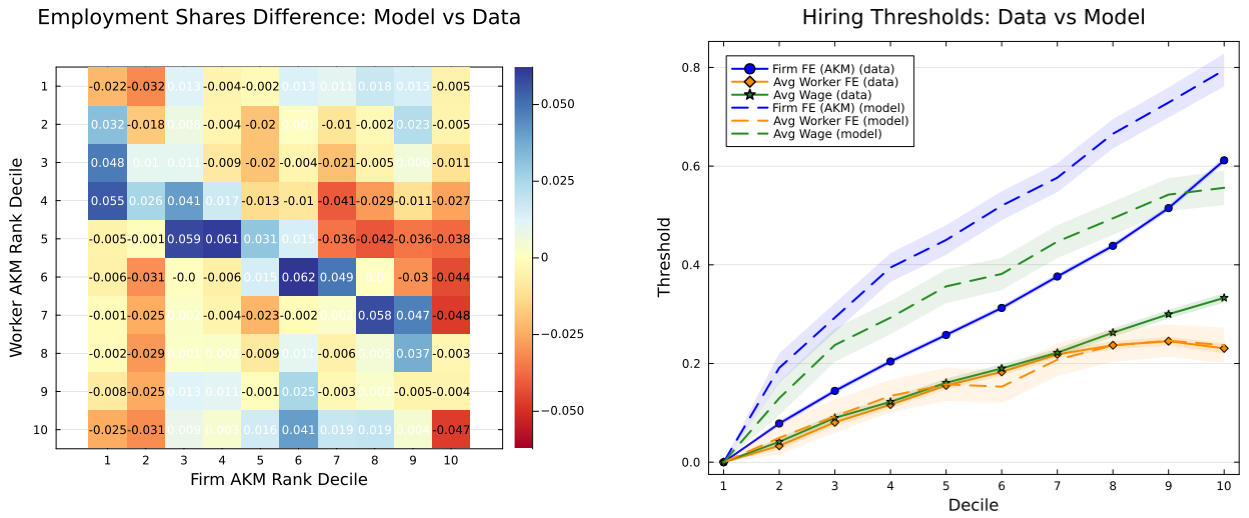
Beyond moment comparisons, I further evaluate the model's ability to reproduce the *joint distribution* of worker and firm fixed effects. Figure 7a reports the cell-by-cell differences between the model-implied and empirical decile matrices introduced in Section 4.3. Discrepancies are modest overall: the model slightly underpredicts sorting at the top of the distribution—placing too few high-ability workers in high-paying firms—and overpredicts matching in the middle deciles. These deviations are economically small: the maximum absolute difference across all worker–firm decile cells employment shares is approximately 0.06. Taken together, these results show that the calibrated model not only reproduces the targeted moments by construction but also matches the joint worker–firm allocation patterns with high fidelity.

Moreover, I compare the hiring-threshold gradients derived in Section 4.2 using model-generated data with their empirical counterparts. I conduct this comparison for all three firm-quality measures used in the data: (i) the AKM firm fixed effect, (ii) the average fixed effect of incumbent workers, and (iii) the average incumbent log wage. Figure 7b display these relationships.

The model successfully reproduces both the *ordering* and the *relative steepness* of hiring thresholds across the different ranking schemes. Consistent with the data, hiring thresholds are steepest when firms are ranked by their AKM pay premium, followed by rankings based on average incumbent wages, and are shallowest when based on average incumbent worker fixed effects. Quantitatively, the model matches exactly the slope of hiring thresholds with respect to average incumbent worker fixed effects. It moderately overpredicts the sensitivity of thresholds to AKM firm ranks (model: 0.80; data: 0.60) and to average incumbent wages (model: 0.58; data: 0.38), but the magnitudes remain in close empirical proximity.

In addition, I examine heterogeneity in wage-bill concentration across the worker AKM rank distribution by comparing the model-implied and empirical HHI profiles by worker fixed-effect decile. Only the aggregate mean HHI enters the calibration, so the decile-specific indices provide a partially untargeted validation. The model closely reproduces the empirical pattern: the maximum absolute deviation in HHI across worker deciles is approximately 0.02, and the full decile profile is reported in Appendix D.4.

Overall, these joint-distributional and hiring-threshold validations demonstrate that the calibrated model provides a quantitatively accurate representation of worker sorting, firm heterogeneity, and hiring thresholds across local labor markets. The model matches targeted and non-targeted features alike, suggesting that its structural mechanisms capture the key forces shaping worker–firm allocation in the data.



(a) Employment Shares: Model vs. Data (AKM Firm-Worker Ranks)

(b) Hiring Thresholds: Model vs. Data

Figure 7: Model vs. Data: Employment Shares and Hiring thresholds

5.3 Model Benchmark: BHM

As a benchmark, I estimate a version of the model with homogeneous labor by imposing $\omega_a = 0$, following D. Berger et al. (2022). In this environment, workers no longer differ in ability, and

sorting and segmentation mechanisms are mechanically shut down. Only firm heterogeneity, captured by σ_z, remains to be calibrated.

I consider two alternative calibrations for σ_z. The first calibration targets the aggregate employment-weighted wage-bill HHI (the "Total" statistic in Table B.25), yielding $\sigma_z = 0.095$. Although this parametrization matches the aggregate HHI by construction, it implies too little dispersion in log wages relative to the data.

For this reason, I implement a second calibration that instead targets the empirical standard deviation of log wages; this produces $\sigma_z = 0.37$. This parametrization increases wage dispersion to empirical levels but no longer matches the aggregate HHI. Because labor is homogeneous in the BHM benchmark, neither specification generates worker–firm sorting or segmentation, and such moments are not defined in this environment.

Table 7 summarizes key model moments under the two alternative calibrations.

Table 7: Comparison of BHM Calibrations: HHI-Targeted vs. Wage-Dispersion-Targeted

Moment	**Calibration 1: HHI Target**	**Calibration 2: Std. Log Wage Target**
Employment-weighted HHI	0.292	0.413
Std. dev. of log employment	1.179	2.608
Labor share	0.536	0.494
Std. dev. of log wages	0.305	0.480

Notes: This table reports key model-generated moments under two alternative calibrations of the homogeneous-labor benchmark (BHM). Calibration 1 sets σ_z to match the aggregate employment-weighted wage-bill HHI, while Calibration 2 chooses σ_z to match the empirical standard deviation of log wages. Because worker heterogeneity is shut down under $\omega_a = 0$, moments involving worker–firm sorting or segmentation are not defined in this benchmark environment.

6 Results

In this section I first use the calibrated model as a measurement device to recover the average markdown by worker type, $\mu(a)$. Empirical counterparts to this object can be constructed from HHI indices as implied by Lemma 3, but these estimates are fragile because worker ability is only indirectly proxied by AKM worker fixed effects and is typically grouped into coarse bins. This discretization potentially masks substantial within-bin heterogeneity and delivers markdowns at the level of noisy proxies rather than at the underlying ability a. The model instead provides $\mu(a)$ directly at the structural ability level.

I also rely on the model to measure moments of the firm-level wedge $\tilde{\psi}_{ij}$. In principle, these wedges can be estimated from the data using cost shares, as in Lemma 9, but such empirical estimates are difficult to interpret: the *level* of the wedge cannot be cleanly identified with firm-level balance-sheet data (Ridder et al., 2025), and the *dispersion* in wedges may conflate true markdown variation with other distortions, such as measurement error, implicit taxes, and adjustment costs.

For this reason, I use the structural model to obtain a well-defined distribution of wedges that isolates the labor market power component.

Finally, I use the model to quantify the aggregate efficiency cost of labor market power and the welfare losses by worker type. I begin by comparing aggregate output under labor market power with that of the efficient allocation, and then quantify heterogeneous welfare losses across the ability distribution.[55]

6.1 Measurement

Panel (a) of Figure 8 reports the employment-weighted average markdown by worker type. On average, workers take home roughly 72.5% of their marginal product of labor. This take-home share varies systematically with worker ability. Markdowns are largest for high-ability workers, who on average receive about 70.5% of their marginal product. The relationship is non-monotonic in worker ability: the take-home share attains its maximum for workers whose log ability lies roughly one standard deviation below the mean of the log-ability distribution. I next use the model to characterize the distribution of the firm-level wedge $\tilde{\psi}_{ij}$. For each calibration, I summarize the distribution of $\tilde{\psi}_{ij}$ across firms by its mean, revenue-weighted mean, median, and standard deviation.[56] In the baseline model, the mean of $\tilde{\psi}_{ij}$ is 0.83, the revenue-weighted mean is 0.73, the median is 0.87, and the standard deviation is 0.09. In BHM 1, the mean, revenue-weighted mean, median, and standard deviation of $\tilde{\psi}_{ij}$ are 0.83, 0.74, 0.87, and 0.09, respectively, while in BHM 2 they are 0.84, 0.68, 0.88, and 0.10. Thus, although the mean and median wedges are very close across the three calibrations, the revenue-weighted mean is substantially lower in BHM 2, indicating that larger-revenue firms are associated with lower values of $\tilde{\psi}_{ij}$ and, consequently, a stronger impact of labor market power on the aggregate labor share in that specification. Consistent with this, the aggregate profit share of GDP is 0.2535 in the baseline calibration, 0.2520 in BHM Calibration 1, and 0.2940 in BHM Calibration 2. The corresponding labor shares are 0.535 in the baseline, 0.536 in Calibration 1, and 0.494 in Calibration 2.

Table 8 provides additional detail by reporting average wedges by productivity decile. In both BHM calibrations, $\tilde{\psi}_{ij}$ is monotonically decreasing in firm productivity deciles, whereas in the baseline model the profile is mildly non-monotonic, with wedges that do not decline uniformly across the productivity distribution.

Overall, the model developed in this paper, together with the empirical evidence, substantially reshape the distribution of labor market power both across firms and across workers. It implies different patterns for wedges, markdowns, and their correlation with productivity than those em-

[55]The model is required for this calculation: even if average markdowns by worker ability are observed, they are not themselves welfare-relevant statistics and do not capture the full set of distortions affecting individual welfare.

[56]Let ls_{ij} denote the actual labor share in firm ij, and let ls^* denote the efficient labor share in the absence of labor market power, costant across firms. By definition, the firm-level wedge satisfies $ls_{ij} = \tilde{\psi}_{ij}\,\ell^*$. The aggregate labor share is then

$$LS^{\text{agg}} = \sum_{ij} \frac{y_{ij}}{\sum_{i'j'} y_{i'j'}}\, ls_{ij} = ls^* \sum_{ij} \frac{y_{ij}}{\sum_{i'j'} y_{i'j'}}\, \tilde{\psi}_{ij},$$

so the aggregate wedge in the labor share is given by the revenue-weighted mean of $\tilde{\psi}_{ij}$.

bedded in benchmark models. I next turn to study the implications of these features for aggregate efficiency and for the distribution of welfare losses by worker type.

Table 8: Average $\tilde{\psi}$ by Productivity Decile: Baseline and BHM Benchmarks

	Productivity decile (from lowest to highest)									
	1	2	3	4	5	6	7	8	9	10
Baseline	0.862	0.851	0.843	0.838	0.824	0.822	0.810	0.801	0.814	0.847
BHM 1	0.852	0.842	0.843	0.841	0.835	0.884	0.833	0.825	0.818	0.811
BHM 2	0.870	0.862	0.860	0.857	0.850	0.855	0.841	0.823	0.803	0.758

Notes: This table reports the average firm-level wedge $\tilde{\psi}_{ij}$ by firm productivity decile for the baseline calibration and the two homogeneous-labor benchmarks (BHM 1 and BHM 2). Firm productivity is measured as the expected production term $\mathbb{E}_{g_{ij}(a)}[\phi(a, z_{ij})]$ for each firm. Productivity deciles are constructed from the distribution of $\mathbb{E}_{g_{ij}(a)}[\phi(a, z_{ij})]$ across all firms in the simulated economy: firms are sorted by this measure and assigned to ten equally sized groups (deciles), from the lowest- to the highest-productivity firms. For each calibration and each decile, the table reports the unweighted mean of $\tilde{\psi}_{ij}$ over all firms belonging to that decile.

6.2 Production Efficiency

Under imperfect competition, firms exert labor market power by paying wages below the efficient level. Because markdowns are higher for firms with larger employment shares within each worker type, firms that capture a greater share of the type-a labor market face the strongest markdowns relative to the market wage index for that ability group. Consequently, firms are under-resourced precisely for the worker types they employ most intensively. This interaction between firm size and worker heterogeneity distorts multiple equilibrium objects that jointly determine aggregate production.

I identify four distinct distortions arising from labor market power. First, *Size Distortion*: markdowns distort firm size, causing some firms to become larger and others smaller than in the efficient allocation. Second, *Misallocation of Talent*: because firms face heterogeneous markdowns across worker types, they misallocate ability within the firm—some firms employ too many low-ability workers, while others attract too many high-ability workers. This heterogeneity generates deviations in endogenous productivity across firms relative to the efficient benchmark. Third, *Market Supply Distortion*: cross-market misallocation of labor. Fourth, *Aggregate Labor Supply Distortion*: markdowns alter total labor supply.

Table 9 compares aggregate inefficiency and its decomposition between the baseline calibration and the two calibrations of the homogeneous-worker benchmark (BHM). In both the efficient and distorted allocations, the equilibrium can be summarized by the collection $\{S(a), s_j(a), s_{ij}(a)\}$: the aggregate supply of each worker type $S(a)$, defined so that the total mass of workers of ability a is $N(a) = S(a)f_a(a)$; the market wage-bill shares $s_j(a)$ describing how each type is allocated across local labor markets; and the firm-level wage-bill shares $s_{ij}(a)$ within each market. To gauge the

quantitative importance of each margin, I construct counterfactuals in which, one at a time, these components are reset to their efficient values while all other components are held at their distorted (baseline) levels. In particular, I consider counterfactuals that: (i) set aggregate labor supply $S(a)$ to its efficient level, (ii) restore the efficient distribution of workers across markets $s_j(a)$, and (iii) restore firm-level wage-bill shares $s_{ij}(a)$ to their efficient values. The firm-level component is further decomposed into a productivity component (adjusting only firms' endogenous productivity) and a size component (adjusting firm employment shares), which together generate the "Firm size + firm productivity" row in the table. The corresponding changes in output measure the marginal contribution of each distortion to aggregate inefficiency.

Two main findings emerge. First, overall efficiency losses in the baseline calibration are smaller than in the second BHM calibration: aggregate output in the baseline model is 3.63% below the efficient level, compared to 5.33% in BHM Calibration 2. The rows labeled "Firm size + firm productivity" in Table 9 show that, once firm-level distortions are shut down, the remaining GDP losses in the baseline and in BHM Calibration 1 are nearly identical (3.33% and 3.31% of efficient output), while in BHM Calibration 2 they fall to 3.52%, a much larger decrease. In BHM Calibration 1, firm heterogeneity is very limited (with σ_z close to zero), so there is little scope for within-market misallocation. The similarity between the baseline and BHM Calibration 1, together with the fact that all three parametrizations exhibit very similar inefficiencies once within-market distortions are removed (between 3.31% and 3.52% of efficient output), indicates that the additional efficiency loss in BHM Calibration 2 is driven by stronger within-market misallocation. In contrast, in the baseline model labor market segmentation dampens the extent to which firm heterogeneity translates into within-market misallocation, making its efficiency properties closer to those of the low-heterogeneity BHM Calibration 1.

Second, the baseline calibration attributes a sizeable fraction of the remaining loss to distortions in aggregate labor supply and in the allocation of labor across markets. When aggregate labor supply is set to its efficient level ("Aggregate-supply distortion"), output in the baseline economy is only 1.87% below the efficient benchmark; when the market-supply distortion is removed ("Only market-supply distortion"), aggregate inefficiency falls to 2.10% of efficient output.

Overall, the comparison indicates that in the heterogeneous-worker baseline with segmentation, within-market misallocation plays a more limited role than in the homogeneous-worker benchmark, and a larger share of the efficiency cost of labor market power operates through aggregate and cross-market labor supply distortions.

6.3 Welfare Distribution

I now analyze the heterogeneous welfare effects of labor market power across workers of different log abilities. Departures from the efficient allocation generate two distinct welfare consequences. First, production inefficiencies reduce the overall size of the economy. Second, markdowns redistribute income from workers to entrepreneurs: firms pay wages below marginal products, lowering workers' consumption, while profits and entrepreneurs' consumption increase. Because

Table 9: Inefficiency Decomposition: Baseline vs. BHM Calibrations

Statistic / Decomposition	Baseline	BHM Cal. 1	BHM Cal. 2
Total inefficiency	0.0363	0.0358	0.0533
Decomposition of inefficiency (GDP loss when only the listed distortion is removed)			
Only firm-size distortion	0.0336	0.0331	0.0352
Only market-supply distortion	0.0210	0.0205	0.0413
Only firm-productivity distortion	0.0387	0.0358	0.0533
Firm size + firm productivity	0.0333	0.0331	0.0352
Aggregate-supply distortion	0.0187	0.0183	0.0311

Notes: This table reports the aggregate inefficiency and its decomposition for three model specifications: the baseline calibrated model and two homogeneous-worker benchmarks (BHM Calibration 1 and BHM Calibration 2). "Total inefficiency" is the steady-state loss in aggregate output, expressed as a share of efficient output, when all distortions implied by the model are present. The decomposition rows report counterfactual GDP losses when only the listed distortion is removed by setting the corresponding variable to its efficient level, holding all other distortions at their baseline values. Rows labeled "Only … distortion" refer to removing a single distortion; "Firm size + firm productivity" reports the loss when both the firm-size and firm-productivity distortions are removed simultaneously; "Aggregate-supply distortion" refers to removing the distortion operating through aggregate labor supply. All numbers are reported as shares of efficient output and correspond to steady-state comparative statics under the calibrations described in Table 5.

markdowns vary across ability groups—reflecting different levels of competition for each worker type—the welfare impact of labor market power is heterogeneous across the ability distribution.

The welfare gain or loss for each group is measured as the percentage change in per capita consumption required to equalize steady-state utility with that of the Pareto-efficient allocation in which workers are paid their marginal product of labor, as characterized in Proposition 3.[57]

Entrepreneurs benefit from labor market power at the expense of workers. In the baseline calibration, they would need to reduce their consumption by roughly 66.5% to attain the same utility level they would have if workers were paid their marginal product. Consistent with this, the realized profit share of GDP is 0.2535, whereas the profit share implied by the efficient allocation under decreasing returns to scale is only about 6%, highlighting the substantial redistribution generated by imperfect competition.

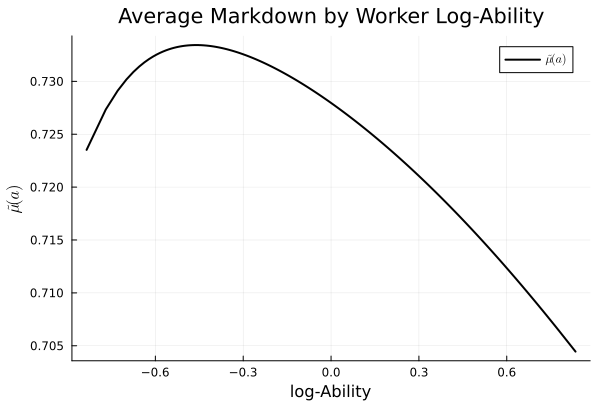
Panel (b) of Figure 8 reports heterogeneous welfare losses $\lambda(a)$ by worker log ability under the baseline calibration. Workers experience welfare losses ranging from about 42% to 37% across the ability distribution. To assess how much of this welfare loss is due to changes in leisure, Panel (b) also reports welfare losses when leisure is held fixed. Increases in leisure under labor market power reduces welfare losses by approximately 2 percentage points, a reduction that is roughly constant across worker types.

Turning to the heterogeneity across workers, Panel (b) reveals that welfare losses are largest at the tails of the ability distribution. Workers at both the bottom and top of the distribution would need to increase consumption by about 42% to achieve efficient-allocation utility, compared to 37%

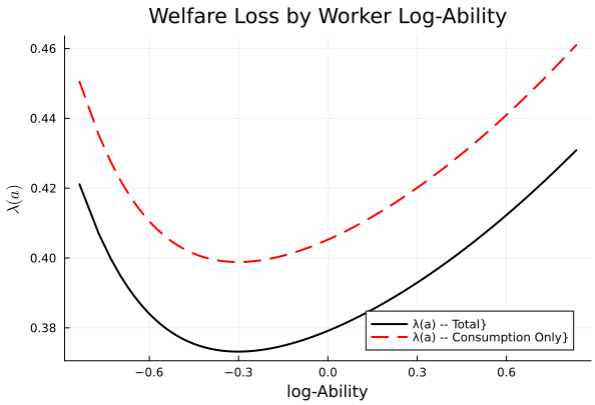
[57]Formally, for workers, $\lambda(a)$ satisfies $U\Big((1+\lambda(a))\frac{C(a)}{f_a(a)}, \frac{N(a)}{f_a(a)}\Big) = U\Big(\frac{C_{\text{eff}}(a)}{f_a(a)}, \frac{N_{\text{eff}}(a)}{f_a(a)}\Big)$, and for entrepreneurs, $\lambda(e)$ satisfies $U((1+\lambda(e))C(e)) = U(C_{\text{eff}}(e))$.

for middle-ability workers. Markdowns are shaped by the degree of competition for each type of labor. For lower-tail workers, few firms from the bottom of the productivity distribution are willing to hire them, leading to limited competition and higher markdowns. These firms internalize that, although they are small relative to the aggregate economy, they are large within the choice set of low-ability workers, who face few alternative employers. At the upper tail, high-ability workers are concentrated in the most productive firms, which compete only against a handful of similarly productive rivals; competition in that segment is likewise limited, resulting in high markdowns. Hence, both extremes of the ability distribution face weaker competition and larger welfare losses. Importantly, welfare losses at the bottom stem purely from imperfect competition rather than from redistributive considerations of the planner. If the planner also valued redistribution or inequality reduction, the efficient allocation would further increase consumption for low-ability workers beyond what efficiency alone would dictate, amplifying the contrast with the decentralized outcome.

Overall, the analysis shows that labor market segmentation not only dampens aggregate inefficiency but also reshapes the distribution of welfare losses across workers. While segmentation reduces aggregate rent extraction by entrepreneurs, failing to account for this heterogeneity would obscure important differences in how labor market power affects workers across the ability distribution. These welfare disparities are most pronounced at the extremes—particularly among low-ability workers, whose welfare is already limited by their lower marginal product and is further eroded by imperfect competition, which amplifies their losses.



(a) Employment-Weighted Markdown by Worker Log Ability



(b) Welfare Loss by Worker Log Ability

Figure 8: Markdowns and Welfare Losses by Worker Log Ability

Notes: Each panel reports model-implied objects by worker log ability. Panel (a) plots the employment-weighted average markdown $\mu(a)$ by worker log ability, where the markdown is defined as the ratio of the wage to the marginal product of labor, and employment weights are given by the steady-state allocation of workers across firms. Panel (b) plots welfare losses by worker log ability. The figure displays two series: one series reports the consumption-equivalent percentage change that makes steady-state utility under labor market power equal to steady-state utility under the Pareto-efficient allocation, allowing both consumption and labor supply to differ across allocations; the second series reports the consumption-equivalent percentage change computed by scaling consumption only while holding labor supply fixed. In both panels, worker log ability is measured on the horizontal axis and the corresponding model-implied statistic is reported on the vertical axis for each ability type.

7 Conclusion

This paper develops, tests, and quantifies a general-equilibrium model of monopsony in segmented labor markets in which sorting and segmentation jointly shape competition for workers. A key innovation is a production technology with intrafirm spillovers, whereby firm productivity depends on the average quality of the workforce, nesting standard benchmark labor models as special cases. Using matched employer–employee data for Italy and Germany, I document that (i) high-paying firms impose higher hiring thresholds, (ii) low- (high-) paid workers are disproportionately employed in low- (high-) paying, smaller (larger) firms, (iii) concentration indices are non-monotonic across the worker–pay distribution, and (iv) exogenous increases (decreases) in workforce average quality causally raise (lower) firm-level productivity.

I then calibrate the model to Italian data and validate it using these empirical moments, benchmarking against two alternative recalibrations of the homogeneous-worker model in D. Berger et al. (2022). First, I use the calibrated structure as a measurement device. On average, workers take home about 72.5% of their marginal product, with markdowns varying systematically in ability and peaking for high-ability workers. At the firm level, wedges in the labor share display similar unweighted means, medians, and dispersion in the baseline and benchmark models, but their relationship with productivity differs: in the homogeneous-labor benchmark, distortions are tightly and monotonically tied to productivity, whereas in the baseline model this link is weaker and mildly non-monotonic. As a consequence, the revenue-weighted mean wedge differs markedly across the two models, reflecting how segmentation reshapes the allocation of distortions toward high-revenue firms. Second, I quantify production efficiency. Under the baseline calibration, labor market power generates modest but nontrivial aggregate output losses that are substantially smaller than in the homogeneous-labor benchmark. A decomposition along aggregate labor supply, cross-market allocation, and within-market allocation of workers to firms shows that, in the benchmark, a large share of inefficiency operates through within-market misallocation, whereas in the baseline model segmentation weakens the link between productivity and markdowns and shifts the importance of distortions toward aggregate and cross-market labor supply. Finally, I study welfare. Entrepreneurs gain substantially from labor market power, while workers experience large and heterogeneous welfare losses. These losses are largest for low-ability workers, who are excluded from most high-productivity firms and face limited opportunities in a narrow set of small, low-productivity employers, and for high-ability workers, whose attractive job offers are concentrated among a few top-paying firms. Importantly, these welfare differences arise purely from imperfect competition, not from redistributive motives.

Taken together, the results show that sorting and segmentation fundamentally reshape monopsony power in labor markets. Segmentation weakens the tight link between firm size, productivity, and distortions that is implicit in homogeneous-labor models, attenuating within-market misallocation but leaving substantial welfare losses—especially for low-ability workers—that are driven by the unequal exposure of different worker types to imperfect competition.

References

Abadie, Alberto and Jann Spiess (2022). "Robust post-matching inference". In: *Journal of the American Statistical Association* 117.538, pp. 983–995.

Abowd, John M, Francis Kramarz, and David N Margolis (1999). "High wage workers and high wage firms". In: *Econometrica* 67.2, pp. 251–333.

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer (2015). "Identification properties of recent production function estimators". In: *Econometrica* 83.6, pp. 2411–2451.

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

Azar, José A, Steven T Berry, and Ioana Marinescu (2022). *Estimating labor market power*. Tech. rep. National Bureau of Economic Research.

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.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2010). "Social incentives in the workplace". In: *The review of economic studies* 77.2, pp. 417–458.

Baqaee, David Rezza and Emmanuel Farhi (2020). "Productivity and misallocation in general equilibrium". In: *The Quarterly Journal of Economics* 135.1, pp. 105–163.

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.

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.

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

Berger, David W, Kyle F Herkenhoff, Andreas R Kostøl, and Simon Mongey (2023). *An Anatomy of Monopsony: Search Frictions, Amenities and Bargaining in Concentrated Markets*. Tech. rep. National Bureau of Economic Research.

Bils, Mark, Bariş Kaymak, and Kai-Jie Wu (2025). *Robinson Meets Roy: Monopsony Power and Comparative Advantage*. Tech. rep. National Bureau of Economic Research.

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch (2021). "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data". In: *Journal of Monetary Economics* 121, pp. 1–14.

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

Burdett, Kenneth and Kenneth L Judd (1983). "Equilibrium price dispersion". In: *Econometrica: Journal of the Econometric Society*, pp. 955–969.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna (2024). *Difference-in-differences with a continuous treatment*. Tech. rep. National Bureau of Economic Research.

Card, David (2022). "Who set your wage?" In: *American Economic Review* 112.4, pp. 1075–90.

Card, David, Ana Rute Cardoso, Joerg 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.

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.

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. 39–57.

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. 3150–3182.

Felix, Mayara (2021). "Trade, labor market concentration, and wages". In: *Job Market Paper*.

Foster, Lucia, John Haltiwanger, and Chad Syverson (2008). "Reallocation, firm turnover, and efficiency: Selection on productivity or profitability?" In: *American Economic Review* 98.1, pp. 394–425.

Freund, Lukas (2022). "Superstar Teams: The Micro Origins and Macro Implications of Coworker Complementarities". In: *Available at SSRN 4312245*.

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.

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, version of April 22 2023. URL: https://www.freit.org/EIIT/2023/selected/gutierrez.pdf.

Haanwinckel, Daniel (2023). *Supply, demand, institutions, and firms: A theory of labor market sorting and the wage distribution*. Tech. rep. National Bureau of Economic Research.

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 and Richard Rogerson (1993). "Job turnover and policy evaluation: A general equilibrium analysis". In: *Journal of political Economy* 101.5, pp. 915–938.

Hopenhayn, Hugo A (2014). *On the measure of distortions*. Tech. rep. 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. URL: https://www.simonjaeger.com/.

Jaravel, Xavier, Neviana Petkova, and Alex Bell (2018). "Team-specific capital and innovation". In: *American Economic Review* 108.4-5, pp. 1034–1073.

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

Kirov, Ivan and James Traina (2021). *Labor market power and technological change in US manufacturing*. Tech. rep. Working paper, University of Chicago, Chicago, IL, November 11.

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.

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, Stefan Seth, and Stefanie Wolter (n.d.). "FDZ-METHODENREPORT". In: ().

Lorenzini, Luca and Antonio Martner (2025). *Aggregate Outcomes of Nonlinear Prices in Supply Chains*. Tech. rep. Version October 31, 2025. Working Paper.

Lucas Jr, Robert E (1978). "On the size distribution of business firms". In: *The Bell Journal of Economics*, pp. 508–523.

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.

McFadden, Daniel et al. (1973). "Conditional logit analysis of qualitative choice behavior". In.

Moretti, Enrico (2004). "Workers' education, spillovers, and productivity: evidence from plant-level production functions". In: *American Economic Review* 94.3, pp. 656–690.

Morlacco, Monica (2019). "Market power in input markets: Theory and evidence from french manufacturing". In: *Unpublished, March* 20, p. 2019.

Olley, G. Steven and Ariel Pakes (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry". In: *Econometrica* 64.6, pp. 1263–1297.

Restuccia, Diego and Richard Rogerson (2008). "Policy distortions and aggregate productivity with heterogeneous establishments". In: *Review of Economic dynamics* 11.4, pp. 707–720.

Ridder, Maarten De, Basile Grassi, and Giovanni Morzenti (2025). "The Hitchhiker's Guide to Markup Estimation: Assessing Estimates from Financial Data". In: *Econometrica*. Forthcoming.

Robinson, Joan (1933). *The economics of imperfect competition*. Springer.

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*. Tech. rep. Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for ...

Sharma, Garima (2023). "Monopsony and gender". In: *Unpublished Manuscript*.

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.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". In: *Journal of econometrics* 225.2, pp. 175–199.

Teulings, Coen N (1995). "The wage distribution in a model of the assignment of skills to jobs". In: *Journal of political Economy* 103.2, pp. 280–315.

Wu, Yingjie, Mingzhi Xu, and Michael Rubens (June 2025). "Exploiting or Augmenting Labor?" In: *American Economic Review: Insights*.

Yeaple, Stephen Ross (2005). "A simple model of firm heterogeneity, international trade, and wages". In: *Journal of international Economics* 65.1, pp. 1–20.

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

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 D. Berger et al. (2022) (BHM), extended to incorporate an endogenous choice set of firms for workers of ability a, denoted $\mathcal{S}_j(a)$.

Setup. For each ability type a, there exists a unit measure of ex-ante identical individuals indexed by $l \in [0,1]$. Individual l has heterogeneous preferences over firms (i,j), captured by an idiosyncratic taste shock $\zeta_{lij}(a)$. Disutility from supplying hours $h_{lij}(a)$ to firm ij is

$$\nu_{lij}(a) = e^{-\zeta_{lij}(a)h_{lij}(a)}, \qquad \log \nu_{lij}(a) = \log h_{lij}(a) - \zeta_{lij}(a).$$

The vector of shocks $\{\zeta_{lij}(a)\}_{ij}$ is drawn from a nested Gumbel distribution:

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

where $\eta > \theta > 0$ govern substitutability across firms within a market and across markets, respectively.

Each individual earns income $Y_l(a)$, drawn from distribution F_Y, and satisfies

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

conditional on choosing firm ij.

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

$$\min_{(i,j)\in\mathcal{S}(a)} \{\log h_{lij}(a) - \zeta_{lij}(a)\} \equiv \max_{(i,j)\in\mathcal{S}(a)} \{\log w_{ij}(a) - \log Y_l(a) - \zeta_{lij}(a)\}.$$

Choice probabilities. Using standard nested-logit results McFadden et al. (1973), the probability that a worker of type a chooses firm $ij \in \mathcal{S}_j(a)$ is

$$p_{ij}(a) = \frac{w_{ij}(a)^{1+\eta}}{\sum_{i'\in\mathcal{S}_j(a)} w_{i'j}(a)^{1+\eta}} \cdot \frac{\left(\sum_{i'\in\mathcal{S}_j(a)} w_{i'j}(a)^{1+\eta}\right)^{\frac{1+\theta}{1+\eta}}}{\displaystyle\int_0^1 \left(\sum_{k\in\mathcal{S}_j(a)} w_{kj}(a)^{1+\eta} \right)^{\frac{1+\theta}{1+\eta}} dj}. \tag{21}$$

This probability is identical for all individuals of the same ability type a.

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

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

Employment aggregation. Total employment of ability type a at firm ij is

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

Aggregate labor income of type a equals

$$\int Y_l(a)\, dF_Y(Y_l(a)) = W(a)N(a),$$

where

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

Substituting these expressions into the formula for $n_{ij}(a)$ yields the firm-level labor-supply function:

$$n_{ij}(a) = \left(\frac{w_{ij}(a)}{w_j(a)} \right)^{\eta} \left(\frac{w_j(a)}{W(a)} \right)^{\theta} N(a). \tag{22}$$

Representative-household formulation. The labor supply system implied by individual discrete choice is equivalent to the solution of the representative-household problem:

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

subject to

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

This establishes the nested-CES labor-supply structure used in the main text.

A.2 Production Function: Microfoundation

This subsection provides a microfoundation for the production technology used in the main text, building upon Saint-Paul, 2001, Helpman et al., 2010, and Eeckhout and Kircher, 2018.

Consider a firm producing a single output y. Production is carried out by a team of workers with heterogeneous abilities $a \in \mathcal{A}$, supervised or coordinated by a manager of type z. Each

worker produces according to

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

where χ denotes the worker's share of a common firm-level resource and ω_χ governs returns to that resource. The key assumption—unlike in Eeckhout and Kircher, 2018—is that the firm cannot allocate χ differentially across workers based on ability a. For concreteness, suppose the firm rents a stock of space or equipment k at price R and allocates it equally across all workers. Setting $\omega_\chi = 1 - \gamma$ yields the worker-level production function

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

where $h = \sum_{a \in \mathcal{A}} n(a)$ is total employment.

Aggregating output across workers gives

$$y = \sum_{a \in \mathcal{A}} \phi(a, z) \left(\frac{k}{h}\right)^{1-\gamma} n(a) = \mathbb{E}_{g(a)}[\phi(a, z)] \left(k^{1-\gamma} h^{\gamma}\right), \tag{23}$$

where

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

is realized firm productivity.

Relation to existing production functions. The production structure in (23) includes several benchmark cases as special or limiting forms. Using the CES specification for $\phi(a, z)$ in (7) and varying ρ and ω_a:

- **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 to the production function in D. 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 the structure in Helpman et al., 2010.

- **Absence of decreasing returns to scale.** If $\gamma = 1$ (i.e., no decreasing returns in employment) and $\omega_\chi = 0$, the production function becomes

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

corresponding to the additive formulation used in Costinot and Vogel, 2010.[58]

A.3 Firm Wage Structure

This subsection proves Proposition 1. We first derive the firm–level labor–supply elasticity, which enters the firm's optimality conditions. We then rewrite the firm problem after substituting out optimal capital, establish the existence of a maximizer, and characterize the implied wage schedule.

Step 1: Firm-level labor-supply elasticity. A key object in the wage characterization is the inverse elasticity of firm–specific labor supply,

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

Define the firm's market share for ability-a workers as

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

Under the inverse labor–supply system

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

the common terms cancel in the ratio and the market share simplifies to

$$s_{ij}(a) = \frac{n_{ij}(a)^{\frac{\eta+1}{\eta}}}{\sum_{i \in \mathcal{S}_j(a)} n_{ij}(a)^{\frac{\eta+1}{\eta}}}. \tag{24}$$

Differentiating $w_{ij}(a)$ with respect to $n_{ij}(a)$ gives

$$\frac{\partial \log w_{ij}(a)}{\partial \log n_{ij}(a)} = \frac{1}{\eta} + \left(\tfrac{1}{\theta} - \tfrac{1}{\eta} \right) \frac{\partial \log n_j(a)}{\partial \log n_{ij}(a)} = \frac{1}{\eta} + \left(\tfrac{1}{\theta} - \tfrac{1}{\eta} \right) s_{ij}(a),$$

where the last equality follows from (24). Hence the inverse elasticity takes the closed form

$$\epsilon_{ij}(a) = \left[\frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta} \right) s_{ij}(a) \right]^{-1}. \tag{25}$$

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 a larger markdown. In the limit $s_{ij}(a) \downarrow 0$, $\epsilon_{ij}(a) \to \eta$.

[58] This expression characterizes the production function. With CES demand, the corresponding revenue function generally differs.

Step 2: Output and the firm's objective. Using the optimal capital choice, firm output is

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

where $g_{ij}(a) = n_{ij}(a)/h_{ij}$ and $Z > 0$ is constant across firms. Define the rescaled output

$$\widetilde{y}_{ij} := (1 - \alpha(1-\gamma))y_{ij}.$$

Substituting capital back into profits yields the isomorphic problem

$$\pi_{ij} = \max_{\{n_{ij}(a)\},\, h_{ij}} \left\{ \widetilde{y}_{ij} - \sum_{a\in\mathcal{A}} w_{ij}(a) n_{ij}(a) \right\},$$

subject to $n_{ij}(a) \geq 0$ and $h_{ij} = \sum_a n_{ij}(a)$.

Lemma A.1 (Existence)**.** *A profit-maximizing solution exists.*

Proof. Let $x = (n_{ij}(a))_{a\in\mathcal{A}}$. Because $\phi(a, z_{ij})$ is bounded, Φ_{ij} is bounded, and thus $\lim_{\|x\|\to 0} (\widetilde{y}_{ij} - \sum_a w_{ij}(a)n_{ij}(a))$ 0. As $\|x\| \to \infty$, the revenue term grows like $O(\|x\|^{\kappa})$ with $\kappa = \frac{1}{1-\alpha(1-\gamma)} < 1 + \rho$, while the convexity of labor costs implies $\sum_a w_{ij}(a)n_{ij}(a) \geq \|x\|^{1+\rho}$. Hence the objective tends to $-\infty$, and continuity establishes a maximizer. □

Step 3: First-order conditions. Form the Lagrangian

$$\mathcal{L} = Z\Phi_{ij}^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\frac{\alpha\gamma}{1-\alpha(1-\gamma)}} - \sum_a w_{ij}(a)n_{ij}(a) + \lambda\left(h_{ij} - \sum_a n_{ij}(a)\right) + \sum_a \varphi(a)n_{ij}(a),$$

with multipliers $\varphi(a) \geq 0$ enforcing $n_{ij}(a) \geq 0$.

FOC with respect to h_{ij}:

$$\lambda = \frac{\partial \widetilde{y}_{ij}}{\partial h_{ij}} = \frac{\alpha\gamma}{1-\alpha(1-\gamma)} Z\Phi_{ij}^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\frac{\alpha\gamma-1}{1-\alpha(1-\gamma)}}.$$

FOC with respect to $n_{ij}(a)$:

$$Z\frac{1}{1-\alpha(1-\gamma)}\Phi_{ij}^{\frac{\alpha(1-\gamma)}{1-\alpha(1-\gamma)}} h_{ij}^{\frac{\alpha\gamma}{1-\alpha(1-\gamma)}} \phi(a, z_{ij}) - w_{ij}(a) - w'_{ij}(a)n_{ij}(a) - \lambda + \varphi(a) = 0.$$

The multiplier λ is the shadow value of a job slot: because each worker occupies one unit of h_{ij}, λ is the marginal cost of expanding total employment. It therefore enters each worker's marginal product of labor negatively. After substituting λ and simplifying,

$$MPL_{ij}(a) - w_{ij}(a)\left(1 + \tfrac{1}{\epsilon_{ij}(a)}\right) + \varphi(a) = 0.$$

Step 4: Case analysis. Case 1: $MPL_{ij}(a) \leq 0$**.** The FOC cannot hold with $n_{ij}(a) > 0$ unless $\varphi(a) > 0$. Thus complementary slackness implies

$$n_{ij}(a) = 0 \quad \Rightarrow \quad w_{ij}(a) = 0.$$

Case 2: $MPL_{ij}(a) > 0$**.** Then $\varphi(a) = 0$ and the FOC gives

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

Step 5: Wage structure. Combining the cases,

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

This completes the proof of Proposition 1.

Lemma A.2. *Under no capital in production ($\gamma = 1$), the wage structure in* (11) *is sufficient for firm maximization.*

Proof. When $\gamma = 1$, firm output reduces to

$$y_{ij} = \Phi_{ij} \, h_{ij}^{\alpha}, \qquad \Phi_{ij} = \sum_{a \in \mathcal{A}} \phi(a, z_{ij}) \, g_{ij}(a),$$

and profits become

$$\pi_{ij}(n, h) = \Phi_{ij}(n) \, h^{\alpha} - \sum_{a \in \mathcal{A}} w_{ij}(a) \, n(a), \qquad h = \sum_{a} n(a).$$

We have already established two facts: (i) a maximizer exists (coercivity), and (ii) any maximizer must satisfy the first–order condition

$$w_{ij}(a) = \frac{\epsilon_{ij}(a)}{1 + \epsilon_{ij}(a)} MPL_{ij}(a) \quad \text{if } n(a) > 0, \qquad w_{ij}(a) = 0 \text{ if } n(a) = 0. \tag{FOC}$$

To prove sufficiency, it remains to show that the system of equations (FOC) admits at most one solution.

Step 1: Parameterization by (Φ, h)**.** Since $\Phi_{ij}(n)$ is a linear functional of g and $h = \sum_a n(a)$, any allocation can be represented by the pair

$$(\Phi, h) \in \mathbb{R}_+ \times \mathbb{R}_+,$$

with wages determined by $n(a) = g(a)h$.

Step 2: Two regions. For any other candidate maximizer $(\Phi_2, h_2) \neq (\Phi_1, h_1)$ to satisfy the same FOCs, it must fall into one of two regions:

1. Same-direction region: $(\Phi_2 - \Phi_1)(h_2 - h_1) \geq 0$.

2. Opposite-direction region: $(\Phi_2 - \Phi_1)(h_2 - h_1) < 0$.

We rule out both cases.

Step 3: No second solution in the same-direction region (single-crossing). Suppose (Φ_2, h_2) satisfies the FOCs and

$$\Phi_2 > \Phi_1, \qquad h_2 \geq h_1$$

(the other sign pattern is symmetric).

Choose any ability a with $n_2(a) > n_1(a)$ (implied by $\Phi_2 > \Phi_1$). For this ability, the FOC requires

$$w(n_i(a))\Big(1 + \tfrac{1}{\epsilon_i(a)}\Big) = MPL_i(a) = (\phi(a) - \Phi_i)\, h_i^{\alpha-1}, \qquad i = 1, 2.$$

- On the left-hand side, $w(n)$ is increasing in n and $1/\epsilon(n)$ is decreasing in n, so the LHS is strictly larger for (Φ_2, h_2).

- On the right-hand side, since $\Phi_2 > \Phi_1$ and $h_2 \geq h_1$, $(\phi(a) - \Phi_i)h_i^{\alpha-1}$ is strictly smaller for (Φ_2, h_2).

Hence the FOC cannot hold at both points—a contradiction. Therefore no two FOC solutions can lie in the same-direction region.

Step 4: No second solution in the opposite-direction region (strict concavity). Suppose instead that

$$\Phi_2 > \Phi_1, \qquad h_2 < h_1$$

(the reverse case is symmetric). Consider any convex combination

$$(\Phi_\lambda, h_\lambda) = (\lambda\Phi_1 + (1-\lambda)\Phi_2,\ \lambda h_1 + (1-\lambda)h_2), \qquad \lambda \in (0, 1).$$

Since $x \mapsto x^\alpha$ is strictly concave for $\alpha \in (0, 1)$,

$$h_\lambda^\alpha > \lambda h_1^\alpha + (1-\lambda)h_2^\alpha.$$

Thus

$$\Phi_\lambda h_\lambda^\alpha > \lambda\Phi_1 h_1^\alpha + (1-\lambda)\Phi_2 h_2^\alpha.$$

The revenue function is therefore strictly concave in (Φ, h) in this region, and the cost function is strictly convex. Hence profits are strictly concave, so there cannot be two distinct stationary points. Thus no second FOC solution exists in the opposite-direction region.

Step 5: Uniqueness and sufficiency. All feasible (Φ, h) lie in one of the two regions. Since both regions allow at most one stationary point, the system (FOC) has exactly one solution. Because a maximizer exists and must satisfy the FOCs, this solution is the unique global maximizer.

Conclusion. Under $\gamma = 1$, the wage structure in (11) is both necessary and sufficient for firm profit maximization. □

A.4 Proof of Proposition 2

This appendix derives the expressions for the average marginal product, average wage, profits, and the 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. Using the expression for the marginal product of a type-a worker,

$$\overline{MPL}_{ij} \equiv Z\alpha\gamma\big(\mathbb{E}_{g_{ijt}(a)}[\phi(a, z_{ijt})]\big)^{\frac{1}{1-\alpha(1-\gamma)}} h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}}$$

Substituting the first order condition for capital $\left(\mathbb{E}_{g_{ij}(a)}[\phi(a, z_{ij})]\right)^{\frac{1}{1-\alpha(1-\gamma)}} h_{ij}^{\frac{\gamma\alpha}{1-\alpha(1-\gamma)}} = \frac{Rk_{ij}}{(1-\gamma)\alpha}$ gives:

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

where the ratio y_{ij}/h_{ij} follows directly from the capital–substituted production function. This proves the first identity in the proposition.

Step 2: Decomposition of $MPL_{ij}(a)$. The above expression can be rewritten as

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

which expresses each worker's marginal product as a scalar multiple of the firm's average marginal product.

Step 3: Average wage. Since $w_{ij}(a) = \mu_{ij}(a) MPL_{ij}(a)$ for $MPL_{ij}(a) > 0$, the firm's average wage is

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

Define

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

which yields

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

Step 4: decomposition of $\widetilde{\psi}_{ij}$. Start from the definition

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

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} (i.e. summing with weights $g_{ij}(a)$) gives

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

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

Now use the identity $\mathbb{E}[\mu\phi] = \mathbb{E}[\mu]\mathbb{E}[\phi] + \mathrm{cov}_g(\mu,\phi)$, and note $\mathbb{E}_{g_{ij}}[\phi] = \Phi_{ij}$. Thus

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

Substitute back:

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

Upper bound. Since every employed ability satisfies $\psi_{ij}(a) \geq 0$, then each summand obeys

$$g_{ij}(a)\,\mu_{ij}(a)\,\psi_{ij}(a) \leq g_{ij}(a)\,\psi_{ij}(a),$$

since $0 \leq \mu_{ij}(a) \leq 1$. Averaging gives

$$\widetilde{\psi}_{ij} = \sum_a g_{ij}(a)\mu_{ij}(a)\psi_{ij}(a) \leq \sum_a g_{ij}(a)\psi_{ij}(a) = 1,$$

so $\widetilde{\psi}_{ij} \leq 1$ under this sign restriction.

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. Profits are

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

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

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

Absent markdowns ($\widetilde{\psi}_{ij} = 1$), the profit share reduces to the standard Cobb–Douglas expression depending only on α and γ.

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. All four identities in Proposition 2 follow directly from the expressions above. □

A.5 Existence and Efficiency

Lemma A.3 (Existence of Equilibrium)**.** *A steady–state competitive equilibrium exists.*

Proof. The firm side can be expressed as a fixed-point problem in wages. For each firm–ability pair (ij, a), the optimality condition derived in Appendix **??** implies that the equilibrium wage must satisfy

$$w_{ij}(a) = B_{ij,a}(w) := \begin{cases} \dfrac{\epsilon_{ij}(a \mid w)}{1 + \epsilon_{ij}(a \mid w)}\; MPL_{ij}(a \mid w)\,, & \text{if } MPL_{ij} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Here $B(w)$ denotes the full wage–update map collecting all (ij, a) components. Equilibrium corresponds to a fixed point w^* such that $B(w^*) = w^*$.

Step 1 (compact domain). As wages approach zero, labor supply collapses and profits are negative. As wages approach infinity, labor supply becomes prohibitively costly and profits again become negative. Thus profitable configurations lie in a bounded hyperrectangle $\mathbb{T} = [0, \overline{w}]^{|\mathcal{F}||\mathcal{A}|}$ for some finite $\overline{w} > 0$.

Step 2 (self-mapping). For any $w \in \mathbb{T}$, all objects entering $B_{ij,a}(w)$ (labor supply, market shares, ϵ_{ij}, and MPL_{ij}) are continuous in w and remain bounded on $\mathbb{T}$. Hence $B(w) \in \mathbb{T}$ for all $w \in \mathbb{T}$.

Step 3 (continuity). Each $B_{ij,a}(w)$ is continuous because both $\epsilon_{ij}(a|w)$ and $MPL_{ij}(a|w)$ are continuous in w (labor supply, shares, and firm output are continuous functions of wages).

By Brouwer's fixed point theorem, a continuous map from a compact convex set into itself admits a fixed point. Thus there exists $w^* \in \mathbb{T}$ such that $B(w^*) = w^*$. Given these wages, optimal labor supply, firm choices, consumption, and capital satisfy all equilibrium conditions.

Therefore a steady–state equilibrium exists. □

Proposition A.1 (Efficiency Without Markdown)**.** *If all firms pay no markdowns ($\mu_{ij}(a) = 1$ for all i, j, a), so that*

$$w_{ij}(a) = MPL_{ij}(a),$$

then the decentralized equilibrium satisfies the planner's first-order conditions and hence coincides with the planner's allocation.

Proof. Rewriting the planner's FOC with respect to $n_{ij}(a)$ for discrete ability types yields

$$\left(\frac{C(a)}{f_a(a)}\right)^{\sigma} \left(\frac{N(a)}{f_a(a)}\right)^{1/\varphi} \left(\frac{N_j(a)}{N(a)}\right)^{1/\theta} \left(\frac{n_{ij}(a)}{N_j(a)}\right)^{1/\eta} = MPL_{ij}(a). \tag{P}$$

But the left-hand side is exactly the decentralized *inverse labor supply*:

$$w_{ij}(a) = \left(\frac{C(a)}{f_a(a)}\right)^{\sigma} \left(\frac{N(a)}{f_a(a)}\right)^{1/\varphi} \left(\frac{N_j(a)}{N(a)}\right)^{1/\theta} \left(\frac{n_{ij}(a)}{N_j(a)}\right)^{1/\eta}.$$

Thus when $\mu_{ij}(a) = 1$ (so $w_{ij}(a) = MPL_{ij}(a)$), the decentralized allocation satisfies the planner's FOC (P) for all (i, j, a).

Because the planner's objective is strictly concave in (C, N) and feasibility is linear in (C, N), consumption and labor allocations that satisfy the FOCs are unique. Hence the decentralized allocation coincides with the planner's allocation when $\mu = 1$. □

A.6 Proofs on Market Equilibrium

Lemma A.4 (Ability–specific average inverse markdown)**.** *Fix a market j and an ability type a. Let*

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

denote firm i's wage–bill share for workers of type a, and let the inverse markdown be

$$\mu_{ij}(a)^{-1} := \frac{MPL_{ij}(a)}{w_{ij}(a)}.$$

Define the employment–weighted averages

$$\overline{MPL}_j(a) := \frac{\sum_i MPL_{ij}(a)\, n_{ij}(a)}{\sum_i n_{ij}(a)}, \qquad \overline{W}_j(a) := \frac{\sum_i w_{ij}(a)\, n_{ij}(a)}{\sum_i n_{ij}(a)}.$$

Then the wage–bill–weighted average inverse markdown satisfies

$$\sum_i s_{ij}(a)\, \mu_{ij}(a)^{-1} = \frac{\overline{MPL}_j(a)}{\overline{W}_j(a)} = 1 + \frac{1}{\eta} + \Big(\frac{1}{\theta} - \frac{1}{\eta}\Big)\, HHI_j(a), \tag{26}$$

where $HHI_j(a) := \sum_i s_{ij}(a)^2$.

Proof. Step 1: Ratio of averages. Using the definition of $s_{ij}(a)$,

$$\sum_i s_{ij}(a)\, \mu_{ij}(a)^{-1} = \sum_i \frac{w_{ij}(a) n_{ij}(a)}{\sum_{i'} w_{i'j}(a) n_{i'j}(a)} \cdot \frac{MPL_{ij}(a)}{w_{ij}(a)} = \frac{\sum_i MPL_{ij}(a)\, n_{ij}(a)}{\sum_{i'} w_{i'j}(a)\, n_{i'j}(a)} = \frac{\overline{MPL}_j(a)}{\overline{W}_j(a)}.$$

Step 2: Closed form under nested CES labor supply. Under the nested–CES structure, the inverse markdown for type a at firm ij is

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

Wage–bill averaging 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 $\sum_i s_{ij}(a)[1 - s_{ij}(a)] = 1 - \sum_i s_{ij}(a)^2 = 1 - HHI_j(a)$, the expression 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 (26). □

Lemma A.5 (Finite–sample decomposition of concentration)**.** *Fix a market j and ability a. Let $\mathcal{S}_j(a)$ denote the set of firms that employ type a in market j, let $m_j(a) := |\mathcal{S}_j(a)|$, and define the wage–bill share for ability a at firm $i \in \mathcal{S}_j(a)$ as*

$$s_{ij}(a) = \frac{w_{ij}(a)\, n_{ij}(a)}{\sum_{i' \in \mathcal{S}_j(a)} w_{i'j}(a)\, n_{i'j}(a)}.$$

Under CES assignment within the firm nest, $n_{ij}(a) \propto w_{ij}(a)^{\eta}$, so $s_{ij}(a)$ admits the closed form

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

Define $X_{ij}(a) := w_{ij}(a)^{1+\eta}$, its sample mean

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

and its sample variance (with denominator $m_j(a)$)

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

Let $\mathrm{CV}_j(a)^2 := V_j(a)/\bar{X}_j(a)^2$. Then the ability–specific Herfindahl–Hirschman index satisfies

$$HHI_j(a) = \sum_{i \in \mathcal{S}_j(a)} s_{ij}(a)^2 = \frac{1}{m_j(a)} \Big[1 + \mathrm{CV}_j(a)^2\Big].$$

Proof. Using $s_{ij}(a) = X_{ij}(a)/\sum_{i'} X_{i'j}(a)$,

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

Write $S := \sum_i X_{ij}(a) = m_j(a)\bar{X}_j(a)$. 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

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

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

Lemma A.6 (Equilibrium invariance from $\omega_a = 0$ to $\rho \to 1$)**.** *Let $f_a(a)$ be the exogenous probability density function of workers type within the local labor market. Let $\{q_{ij}(a)\}_{i,j,a}$ denote the equilibrium employment assignment under $\omega_a = 0$ (i.e. the share of labor market supply that works in firm ij) and let*

$$\tilde{z}_{ij} = z_{ij}^{1-\omega_a} \frac{\mathbb{E}_{f_a(a)}[a^{\omega_a}]}{\bar{a}}, \qquad \bar{a} := \sum_{a \in \mathcal{A}} a\, f_a(a),$$

be the renormalized firm productivities under the $\rho \to 1$ limit. Assume $a_\ell > (1-\alpha)\bar{a}$ so that $MPL_{ij}(a) >$

0 *for all employed types.*[59] *Then the same assignment* $\{q_{ij}(a)\}$ *constitutes a market equilibrium under the* $\rho \to 1$ *technology*

$$y_{ij} = \tilde{z}_{ij}\, \bar{a}_{ij}\, \left(k_{ij}^{1-\gamma} h_{ij}^{\gamma}\right)^{\alpha}.$$

Proof. Step 1: Assignment under $\omega_a = 0$. When $\omega_a = 0$ the marginal product does not depend on ability a. Denote the common (ability-independent) marginal product factor by

$$MPL_{ij}^{(0)} := MPL_{ij}(a)\big|_{\omega_a=0} = z_{ij}\, k_{ij}^{\alpha(1-\gamma)} h_{ij}^{\gamma\alpha-1},$$

so that the within-type CES numerator for any ability is proportional to $\left[\mu(s_{ij})\, MPL_{ij}^{(0)}\right]^{\eta}$. Hence the within-type employment share satisfies

$$q_{ij}(a) = \frac{\left[\mu(s_{ij})\, MPL_{ij}^{(0)}\right]^{\eta}}{\sum_{i'} \left[\mu(s_{i'j})\, MPL_{i'j}^{(0)}\right]^{\eta}}.$$

The right-hand side is independent of a; therefore $q_{ij}(a) = q_{ij}$ for all abilities. Consequently each firm hires a representative slice of the ability distribution,

$$n_{ij}(a) = q_{ij}\, f_a(a), \qquad h_{ij} = \sum_a n_{ij}(a) = q_{ij}, \qquad \bar{a}_{ij} = \bar{a}.$$

Step 2: Marginal product under $\rho \to 1$. In the $\rho \to 1$ limit the marginal product of a type-a worker at firm (i, j) can be written

$$MPL_{ij}(a) = \tilde{z}_{ij}\, k_{ij}^{\alpha(1-\gamma)} h_{ij}^{\gamma\alpha-1}\, [\, a - (1-\alpha)\bar{a}_{ij}\,],$$

with $\tilde{z}_{ij}$ the renormalized productivity in the lemma. Under the assignment from Step 1 we have $\bar{a}_{ij} = \bar{a}$ for all firms, so the factor $[a - (1-\alpha)\bar{a}]$ is common across firms for each given a.

Step 3: CES shares under $\rho \to 1$. The within-type employment share for ability a becomes

$$q_{ij}(a) = \frac{\left[\mu(s_{ij}(a))\, \tilde{z}_{ij}\, k_{ij}^{\alpha(1-\gamma)} h_{ij}^{\gamma\alpha-1}\right]^{\eta} \left[a - (1-\alpha)\bar{a}\right]^{\eta}}{\sum_{i'} \left[\mu(s_{i'j}(a))\, \tilde{z}_{i'j}\, k_{i'j}^{\alpha(1-\gamma)} h_{i'j}^{\gamma\alpha-1}\right]^{\eta} \left[a - (1-\alpha)\bar{a}\right]^{\eta}}.$$

The common factor $\left[a - (1-\alpha)\bar{a}\right]^{\eta}$ cancels between numerator and denominator, yielding

$$q_{ij}(a) = \frac{\left[\mu(s_{ij}(a))\, \tilde{z}_{ij}\, k_{ij}^{\alpha(1-\gamma)} h_{ij}^{\gamma\alpha-1}\right]^{\eta}}{\sum_{i'} \left[\mu(s_{i'j}(a))\, \tilde{z}_{i'j}\, k_{i'j}^{\alpha(1-\gamma)} h_{i'j}^{\gamma\alpha-1}\right]^{\eta}}.$$

This is the same fixed-point system (up to the renormalization $\tilde{z}_{ij}$) solved under $\omega_a = 0$. Hence it admits the same solution $q_{ij}(a) = q_{ij}$ for all a, which implies identical h_{ij} and $\bar{a}_{ij}$. Since $MPL_{ij}(a) > 0$ for all employed abilities, the solution is admissible. Therefore the assignment

[59] If this condition does not hold, the only difference is that worker types with negative marginal product are not hired under $\rho \to 1$, whereas they would be hired under $\omega_a = 0$.

$\{q_{ij}(a)\}$ from the $\omega_a = 0$ case remains an equilibrium under $\rho \to 1$. □

Proof of Lemma 6: Monotonicity in firm productivity

Fix an ability type a and consider two firms with $z_{i'j} > z_{ij}$. Under the within-market CES structure,

$$q_{ij}(a) \propto w_{ij}(a)^{\eta},$$

so $q_{i'j}(a) < q_{ij}(a)$ would require $w_{i'j}(a) < w_{ij}(a)$ and therefore $s_{i'j}(a) < s_{ij}(a)$.

Because markdowns are (weakly) decreasing in own wage-bill share,

$$\mu_{i'j}(a) \geq \mu_{ij}(a),$$

and since the marginal product is (weakly) increasing in firm productivity,

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

The wage equation $w_{ij}(a) = \mu_{ij}(a) MPL_{ij}(a)$ therefore implies

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

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

Thus $q_{i'j}(a) \geq q_{ij}(a)$ whenever $z_{i'j} > z_{ij}$, with strict inequality when either $MPL_{ij}(a)$ increases strictly in z_{ij} or $\mu_{ij}(a)$ is strictly decreasing. □

Proof of Lemma 7: Positive assortative matching

Let $z_{i'j} > z_{ij}$ and define the ability-specific share ratio

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

Under the CES structure,

$$\mathcal{R}(a) = \left(\frac{w_{i'j}(a)}{w_{ij}(a)}\right)^{\frac{1+\eta}{\eta}}, \qquad w_{ij}(a) = \mu_{ij}(a) MPL_{ij}(a).$$

Step 1: Implicit equation for $\mathcal{R}(a)$. Combining these expressions gives an implicit equation of the form

$$\log \mathcal{R}(a) - (1+\eta) \log \Psi\big(\mathcal{R}(a)\big) = (1+\eta)\, \Delta(a; z_{ij}, z_{i'j}), \tag{27}$$

where

$$\Delta(a; z_{ij}, z_{i'j}) = \log\left(\frac{MPL_{i'j}(a)}{MPL_{ij}(a)}\right),$$

and

$$\Psi(\mathcal{R}) = \frac{\mu_{i'j}(a, s_{i'j}(a, \mathcal{R}))}{\mu_{ij}(a, s_{ij}(a, \mathcal{R}))}.$$

Here $s_{i'j}(a, \mathcal{R})$ and $s_{ij}(a, \mathcal{R})$ are the wage-bill shares induced by the relative wage vector implied by $\mathcal{R}(a)$.

Step 2: Wage-bill shares as functions of $\mathcal{R}(a)$. Index firms in market j by productivity so that

$$z_{1j} < z_{2j} < \cdots < z_{ij} < z_{i'j} < \cdots < z_{m_j j}.$$

Fix the pair $(ij, i'j)$ of adjacent firms under comparison.

For any firm q, define

$$x_q(a) := \left(\frac{w_{qj}(a)}{w_{ij}(a)}\right)^{1+\eta},$$

so that wage-bill shares satisfy

$$s_{qj}(a, \mathcal{R}) = \frac{x_q(a)}{\sum_{q'} x_{q'}(a)}.$$

Firms below ij ($q < i$). For these firms, wages $w_{qj}(a)$ are unaffected by the comparison between ij and $i'j$, hence $x_q(a)$ does not depend on $\mathcal{R}(a)$. Define

$$C(a) := \sum_{q<i} x_q(a).$$

The reference firm ij. By construction, $x_i(a) = 1$.

Firms above ij ($m > i$). For these firms, relative wages scale multiplicatively with the ratio

$$\frac{w_{i'j}(a)}{w_{ij}(a)} = \mathcal{R}(a)^{\frac{\eta}{1+\eta}},$$

and therefore each $x_m(a)$ can be written as

$$x_m(a) = \left(\frac{w_{mj}(a)}{w_{i'j}(a)} \cdot \frac{w_{i'j}(a)}{w_{ij}(a)}\right)^{1+\eta} = \mathcal{R}(a)\, y_m(a),$$

where

$$y_m(a) := \left(\frac{w_{mj}(a)}{w_{i'j}(a)}\right)^{1+\eta}$$

does not depend on $\mathcal{R}(a)$. Define the aggregate term

$$D(a) := \sum_{m>i} y_m(a).$$

Putting the pieces together,

$$T(a,\mathcal{R}) := \sum_q x_q(a) = C(a) + 1 + \mathcal{R}(a)D(a),$$

and therefore

$$s_{ij}(a,\mathcal{R}) = \frac{1}{T(a,\mathcal{R})}, \qquad s_{i'j}(a,\mathcal{R}) = \frac{\mathcal{R}(a)}{T(a,\mathcal{R})}.$$

Differentiating,

$$\frac{\partial s_{ij}}{\partial \mathcal{R}} = -\frac{D(a)}{T(a,\mathcal{R})^2}, \qquad \frac{\partial s_{i'j}}{\partial \mathcal{R}} = \frac{C(a)+1}{T(a,\mathcal{R})^2} > 0.$$

Step 3: Monotonicity of $\Psi(\mathcal{R})$. Because markdowns are decreasing,

$$\mu'(s) \leq 0.$$

Applying the chain rule,

$$\Psi'(\mathcal{R}) = \frac{\mu'(s_{i'j})}{\mu(s_{i'j})}\frac{\partial s_{i'j}}{\partial \mathcal{R}} - \frac{\mu'(s_{ij})}{\mu(s_{ij})}\frac{\partial s_{ij}}{\partial \mathcal{R}} \;\leq\; 0.$$

Thus the left-hand side of (27),

$$H(\mathcal{R}) = \log \mathcal{R} - (1+\eta)\log \Psi(\mathcal{R}),$$

is strictly increasing in $\mathcal{R}$.

Step 4: Comparative statics across abilities. Take two abilities $a' > a$. Subtracting the implicit equations (27) for a' and a gives

$$H(\mathcal{R}(a')) - H(\mathcal{R}(a)) = (1+\eta)\big[\Delta(a'; z_{ij}, z_{i'j}) - \Delta(a; z_{ij}, z_{i'j})\big].$$

Log-supermodularity and strict monotonicity of $MPL_{ij}(a)$ in (a, z) imply

$$\Delta(a'; z_{ij}, z_{i'j}) > \Delta(a; z_{ij}, z_{i'j}),$$

so the right-hand side is strictly positive. Since H is strictly increasing, this yields

$$\mathcal{R}(a') > \mathcal{R}(a).$$

Step 5: Interpretation. Because $\mathcal{R}(a) = q_{i'j}(a)/q_{ij}(a)$, the inequality $\mathcal{R}(a') > \mathcal{R}(a)$ means higher-ability workers are relatively more employed in the higher-productivity firm. □

Proof of proposition 4

Proposition A.2. *Under the same assumptions, the concentration index $HHI_j(a) = \sum_i s_{ij}(a)^2$ is strictly increasing in ability: for any $a' > a$, $HHI_j(a') > HHI_j(a)$. Higher-ability workers are therefore employed in more concentrated segments of the market and face greater firm-level labor market power.*

We already characterized the ordering of wages, and the behavior of adjacent wage ratios. Before establishing Proposition A.2, we first develop a sequence of intermediate results on the sensitivity of market concentration to these ratios. Then we conclude with the proof.

Lemma A.7 (Head–tail inequality)**.** *Let $\{r_i\}_{i=1}^{N-1}$ be positive numbers with $r_i \geq 1$, and define the sequence*

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

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

$$x_j(t) = t\, y_j,$$

where $y_j := x_j(t)/t$ is independent of 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

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

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

$$x_1 \leq x_2 \leq \cdots \leq x_k = t \leq x_{k+1} \leq \cdots \leq x_N.$$

Lower bound. For all $i \leq k$, $x_i \leq t$. Hence

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

which gives

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

Upper bound. For $j > k$ we have $x_j = t\, y_j$ with $t \geq 1$ and $x_j \geq x_k = t$. Thus

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

Therefore

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

which yields

$$t \leq \frac{B_0}{T_1}.$$

Combining the two inequalities proves the claim. □

Lemma A.8 (HHI monotonicity in each adjacent ratio)**.** *Fix a market j and ability type a. Let firms be indexed so that $w_{1j}(a) \leq \cdots \leq w_{N_j j}(a)$ and let*

$$r_{i,i+1,j}(a) := \frac{w_{i+1,j}(a)}{w_{ij}(a)} \;\geq\; 1 \qquad (i = 1, \ldots, N_j - 1)$$

denote the adjacent wage ratios.

Construct the synthetic size sequence

$$x_{1j}(a) := 1, \qquad x_{ij}(a) := \prod_{m=1}^{i-1} r_{m,m+1,j}(a) \quad (i \geq 2),$$

so that the wage-bill shares satisfy

$$s_{ij}(a) = \frac{x_{ij}(a)}{\sum_{i'} x_{i'j}(a)}.$$

Define the Herfindahl–Hirschman index of wage-bill concentration as

$$HHI_j(a) = \sum_{i=1}^{N_j} s_{ij}(a)^2 = \frac{\sum_{i=1}^{N_j} x_{ij}(a)^2}{\left(\sum_{i=1}^{N_j} x_{ij}(a)\right)^2} =: H\big(r_{1,2,j}(a), \ldots, r_{N_j-1,N_j,j}(a)\big)\,.$$

Fix a particular adjacent ratio $r_{k,k+1,j}(a) \geq 1$ and treat all other ratios as constants. Then

$$\frac{\partial H}{\partial r_{k,k+1,j}(a)} \;\geq\; 0,$$

with strict inequality whenever the inequalities in Lemma A.7 are strict.

Proof. Fix k and set

$$t := r_{k,k+1,j}(a).$$

By Lemma A.7, the synthetic sequence $\{x_{ij}(a)\}$ can be decomposed as follows:

- For $i \leq k$ (the "head"), $x_{ij}(a)$ is independent of t. - For $i > k$ (the "tail"),

$$x_{ij}(a) = t\, y_{ij}(a),$$

where $y_{ij}(a)$ is independent of t.

Define the corresponding aggregates

$$H_1(a) = \sum_{i \leq k} x_{ij}(a), \qquad T_1(a) = \sum_{i > k} y_{ij}(a),$$

$$A_0(a) = \sum_{i \leq k} x_{ij}(a)^2, \qquad B_0(a) = \sum_{i > k} y_{ij}(a)^2.$$

Using

$$HHI_j(a) = \frac{A_0(a) + t^2 B_0(a)}{\big(H_1(a) + tT_1(a)\big)^2},$$

let

$$H(t) := \frac{A_0 + t^2 B_0}{(H_1 + tT_1)^2}.$$

Differentiation yields

$$H'(t) = \frac{2\left(tH_1 B_0 - T_1 A_0\right)}{(H_1 + tT_1)^3}.$$

Lemma A.7 establishes the bounds

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

Multiplying these inequalities by $H_1 T_1$ gives

$$tH_1 B_0 - T_1 A_0 \;\geq\; A_0 T_1 (t-1) \;\geq\; 0 \qquad \text{for all } t \geq 1.$$

Hence $H'(t) \geq 0$, with strict inequality whenever at least one of the inequalities in Lemma A.7 is strict. Therefore

$$\frac{\partial H}{\partial r_{k,k+1,j}(a)} = H'(t) \;\geq\; 0.$$

□

Lemma A.9 (HHI monotonicity under adjacent-ratio monotonicity, discrete types)**.** *Let $\mathcal{A}$ be a finite or countable subset of $\mathbb{R}$ endowed with the usual order. For each $a \in \mathcal{A}$, index firms in market j so that*

$$w_{1j}(a) \leq \cdots \leq w_{Nj}(a),$$

and define the adjacent wage ratios

$$r_{i,i+1}(a) := \left(\frac{w_{i+1,j}(a)}{w_{ij}(a)}\right)^{1+\eta} \;\geq\; 1, \qquad i = 1, \ldots, N-1.$$

Construct the normalized wage sequence

$$x_{1j}(a) := 1, \qquad x_{ij}(a) := \prod_{m=1}^{i-1} r_{m,m+1}(a) \quad (i \geq 2),$$

so that the wage-bill shares satisfy

$$s_{ij}(a) = \frac{x_{ij}(a)}{\sum_{i'=1}^{N} x_{i'j}(a)}.$$

The Herfindahl–Hirschman index of wage-bill concentration is therefore

$$HHI_j(a) = \sum_{i=1}^{N} s_{ij}(a)^2 = \frac{\sum_{i=1}^{N} x_{ij}(a)^2}{\big(\sum_{i=1}^{N} x_{i'j}(a)\big)^2} =: H\big(r_1(a), \ldots, r_{N-1}(a)\big),$$

where $r_k(a) := r_{k,k+1}(a)$.

Assume:

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

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

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

Moreover, if there exists a subset $\mathcal{A}_0 \subseteq \mathcal{A}$ of positive counting measure such that, for each $a \in \mathcal{A}_0$, there exist $a' > a$ and an index k with

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

then $HHI_j(a)$ is strictly increasing on $\mathcal{A}_0$ (i.e. $HHI_j(a') > HHI_j(a)$ for the corresponding pairs).

Proof. Fix $a, a' \in \mathcal{A}$ with $a' > a$. Assumption (i) gives

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

Let

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

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

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

establishing weak monotonicity.

For strict monotonicity, fix $a \in \mathcal{A}_0$ and the corresponding $a' > a$ and index k. Then

$$r_k(a') > r_k(a), \qquad \frac{\partial H}{\partial r_k}\big(\mathbf{r}(a')\big) > 0.$$

Holding all other ratios at their (weakly larger) values in $\mathbf{r}(a')$, increasing the k-th coordinate from $r_k(a)$ to $r_k(a')$ strictly raises H. Thus

$$HHI_j(a') = H(\mathbf{r}(a')) > H(\mathbf{r}(a)) = HHI_j(a),$$

proving strict monotonicity on $\mathcal{A}_0$. □

Conclusion of Proof. For any $a' > a$, $HHI_j(a') > HHI_j(a)$:

Proof. Lemma 6 establishes that employment shares preserve the productivity ordering of firms. Lemma 7 shows that adjacent ratios $\mathcal{R}(a)$ are strictly increasing in ability under log-supermodularity. Lemma A.8 proves that HHI is (weakly) increasing in each adjacent ratio, and lemma A.9 therefore implies that

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

Thus higher-ability workers are allocated to more concentrated segments of the market. □

Proof of Lemma 8

Lemma A.10. *If the equilibrium marginal product of labor $MPL_{ij}(a)$ is log-supermodular in (a, z_{ij}), then:*

i) realized productivity $\mathbb{E}_{g_{ij}(a)}\big[\phi(a, z_{ij})\big]$ is weakly increasing in firm productivity z_{ij}, and

ii) the screening threshold $\widetilde{a}_{ij}$, defined as the minimum worker ability among employees, is weakly increasing in z_{ij}.

Proof of part (i). Let abilities $a \in \mathcal{A}$ be countable and ordered in the usual way. Consider two firms in the same market j with productivities z_{ij} and $z_{i'j}$ such that

$$z_{i'j} > z_{ij}.$$

Let $g_{ij}(a)$ and $g_{i'j}(a)$ denote their within-firm employment distributions over abilities:

$$g_{ij}(a) := \frac{n_{ij}(a)}{h_{ij}}, \qquad g_{i'j}(a) := \frac{n_{i'j}(a)}{h_{i'j}}.$$

Step 1: Higher-z firms employ stochastically higher-ability workers. Log-supermodularity of $MPL_{ij}(a)$ in (a, z_{ij}) implies positive assortative matching between worker ability and firm productivity by Lemma 7. In particular, for $z_{i'j} > z_{ij}$, the ability distribution at the more productive firm, $g_{i'j}(a)$, first-order stochastically dominates that at the less productive firm, $g_{ij}(a)$:

$$\sum_{a' \geq \bar{a}} g_{i'j}(a') \;\geq\; \sum_{a' \geq \bar{a}} g_{ij}(a') \qquad \text{for all thresholds } \bar{a} \in \mathcal{A}. \tag{28}$$

Intuitively, the high-z firm puts relatively more mass on higher abilities.[60]

Step 2: Realized productivity is increasing in z_{ij}. Realized firm productivity is

$$\mathbb{E}_{g_{ij}(a)}\big[\phi(a, z_{ij})\big] = \sum_{a \in \mathcal{A}} \phi(a, z_{ij})\, g_{ij}(a),$$

and analogously for firm (i', j). Consider the difference in realized productivity:

$$\Delta := \mathbb{E}_{g_{i'j}(a)}\big[\phi(a, z_{i'j})\big] - \mathbb{E}_{g_{ij}(a)}\big[\phi(a, z_{ij})\big].$$

Add and subtract $\mathbb{E}_{g_{i'j}(a)}[\phi(a, z_{ij})]$:

$$\Delta = \underbrace{\Big(\mathbb{E}_{g_{i'j}(a)}[\phi(a, z_{i'j})] - \mathbb{E}_{g_{i'j}(a)}[\phi(a, z_{ij})]\Big)}_{\text{direct } z\text{-effect (same } g)} + \underbrace{\Big(\mathbb{E}_{g_{i'j}(a)}[\phi(a, z_{ij})] - \mathbb{E}_{g_{ij}(a)}[\phi(a, z_{ij})]\Big)}_{\text{composition effect (same } z)}.$$

Direct z-effect. For any fixed ability a, $\phi(a, z)$ is (weakly) increasing in z, so $z_{i'j} > z_{ij}$ implies $\phi(a, z_{i'j}) \geq \phi(a, z_{ij})$ for all a. Thus

$$\mathbb{E}_{g_{i'j}(a)}[\phi(a, z_{i'j})] - \mathbb{E}_{g_{i'j}(a)}[\phi(a, z_{ij})] = \sum_{a} \big(\phi(a, z_{i'j}) - \phi(a, z_{ij})\big)\, g_{i'j}(a) \;\geq\; 0.$$

Composition effect. Fix z_{ij} and note that $\phi(a, z_{ij})$ is (weakly) increasing in ability a. By (28), $g_{i'j}(\cdot)$ first-order stochastically dominates $g_{ij}(\cdot)$, so the expectation of any increasing function of a

[60] Formally, Lemma 7 establishes adjacent-ratio monotonicity of employment shares under log-supermodularity of $MPL_{ij}(a)$. In the discrete case this is equivalent to a monotone likelihood-ratio ordering of $g_{i'j}$ relative to g_{ij}, which in turn implies first-order stochastic dominance.

is higher under $g_{i'j}$ than under g_{ij}. In particular,

$$\mathbb{E}_{g_{i'j}(a)}[\phi(a, z_{ij})] - \mathbb{E}_{g_{ij}(a)}[\phi(a, z_{ij})] = \sum_a \phi(a, z_{ij}) \left[g_{i'j}(a) - g_{ij}(a)\right] \;\geq\; 0.$$

Combining the two terms, we obtain $\Delta \geq 0$, so realized productivity $\mathbb{E}_{g_{ij}(a)}[\phi(a, z_{ij})]$ is weakly increasing in firm productivity z_{ij}.

Proof of part (ii). Fix a market j and consider two firms (i, j) and (i', j) with $z_{i'j} > z_{ij}$. As above, let $g_{ij}(a)$ and $g_{i'j}(a)$ denote their within-firm ability distributions over the ordered support $\mathcal{A}$. Define the screening threshold for each firm as

$$\widetilde{a}_{ij} := \min\{a \in \mathcal{A} : g_{ij}(a) > 0\}, \qquad \widetilde{a}_{i'j} := \min\{a \in \mathcal{A} : g_{i'j}(a) > 0\}.$$

From Lemma 7 and the log-supermodularity of $MPL_{ij}(a)$, the higher-productivity firm has an ability distribution that first-order stochastically dominates that of the lower-productivity firm:

$$g_{i'j}(\cdot) \;\succeq_{\text{FOSD}}\; g_{ij}(\cdot).$$

Writing the cumulative distributions as

$$G_{ij}(a) := \sum_{a' \leq a} g_{ij}(a'), \qquad G_{i'j}(a) := \sum_{a' \leq a} g_{i'j}(a'),$$

FOSD means

$$G_{i'j}(a) \;\leq\; G_{ij}(a) \qquad \text{for all } a \in \mathcal{A}.$$

Suppose, toward a contradiction, that the screening thresholds are ordered in the opposite way:

$$\widetilde{a}_{i'j} < \widetilde{a}_{ij}.$$

By definition of the thresholds,

$$g_{i'j}(\widetilde{a}_{i'j}) > 0, \qquad g_{ij}(a) = 0 \quad \text{for all } a \leq \widetilde{a}_{i'j}.$$

Hence the cumulative distributions satisfy

$$G_{i'j}(\widetilde{a}_{i'j}) = \sum_{a' \leq \widetilde{a}_{i'j}} g_{i'j}(a') \;\geq\; g_{i'j}(\widetilde{a}_{i'j}) \;>\; 0,$$

while

$$G_{ij}(\widetilde{a}_{i'j}) = \sum_{a' \leq \widetilde{a}_{i'j}} g_{ij}(a') = 0.$$

This contradicts the FOSD condition $G_{i'j}(a) \leq G_{ij}(a)$ for all a.

Therefore the assumption must be false, and we conclude

$$\widetilde{a}_{i'j} \;\geq\; \widetilde{a}_{ij},$$

i.e. the screening threshold is weakly increasing in firm productivity z_{ij}. □

Proof of Lemma 9

Lemma A.11 (Recovering the wage wedge from cost shares)**.** *Let firm ij produce output $y_{ij,t}$ with labor headcount $h_{ij,t}$ 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. As in the model of Section 3, suppose that the marginal product of a worker of type a can be written as*

$$MPL_{ij,t}(a) = \frac{\partial y_{ij,t}}{\partial n_{ij,t}(a)} = \frac{\partial y_{ij,t}}{\partial h_{ij,t}}\,\psi_{ij,t}(a),$$

where $\psi_{ij,t}(a)$ captures heterogeneous marginal products across worker types. If the firm cost-minimizes and the input $x_{ij,t}$ has a marginal price p_x, then the firm-level wedge $\widetilde{\psi}_{ij,t}$ satisfies

$$\frac{1}{\widetilde{\psi}_{ij,t}} = \frac{p_x x_{ij,t}}{\overline{w}_{ij,t}\,h_{ij,t}}\,\frac{\alpha_{l,ij}}{\alpha_{m,ij}}, \tag{29}$$

so that $\widetilde{\psi}_{ij,t}$ is pinned down by the ratio of the material bill to the wage bill and the ratio of output elasticities.

Proof of Lemma 9 (countable ability types). Fix firm ij and period t; for notational compactness I drop the subscripts (i, j, t) where there is no risk of confusion. Let the set of worker types be countable, $a \in \mathcal{A} = \{a_1, a_2, \dots\}$. The firm produces output y using total headcount $h = \sum_{a\in\mathcal{A}} n(a)$ (where $n(a)$ is employment of type a) and a flexible input x. The firm takes as given the marginal price p_x of x and the type-specific wages $w(a)$, and chooses $\{n(a)\}_{a\in\mathcal{A}}, x$ to produce a target output Q at minimum variable cost. The cost minimization problem is

$$\min_{\{n(a)\},\,x} \quad p_x x + \sum_{a\in\mathcal{A}} w(a)\,n(a) \quad \text{s.t.} \quad y\big(h = \sum_{a\in\mathcal{A}} n(a),\, x\big) \geq Q.$$

Form the Lagrangian with multiplier $\lambda > 0$:

$$\mathcal{L} = p_x x + \sum_{a\in\mathcal{A}} w(a)\,n(a) + \lambda\big(Q - y(h, x)\big).$$

The first-order conditions are, for the flexible input x,

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \quad\Longrightarrow\quad p_x \;=\; \lambda\,\frac{\partial y}{\partial x}, \tag{30}$$

and, for each worker type $a \in \mathcal{A}$,

$$\frac{\partial \mathcal{L}}{\partial n(a)} = 0 \quad \Longrightarrow \quad w(a) \;=\; \lambda \, \frac{\partial y}{\partial n(a)}. \tag{31}$$

By assumption the marginal product of a worker of type a can be written

$$\frac{\partial y}{\partial n(a)} = \frac{\partial y}{\partial h} \, \psi(a),$$

where $\psi(a)$ captures the cross-type heterogeneity in marginal products. Multiply (31) by $n(a)$ and sum over $a \in \mathcal{A}$ to obtain an expression for the wage bill:

$$\sum_{a \in \mathcal{A}} w(a) \, n(a) = \lambda \, \frac{\partial y}{\partial h} \, \sum_{a \in \mathcal{A}} \psi(a) \, n(a).$$

Define the average wage $\overline{w}$ and the average wedge $\widetilde{\psi}$ by

$$\overline{w} \;=\; \frac{1}{h} \sum_{a \in \mathcal{A}} w(a) \, n(a), \qquad \widetilde{\psi} \;=\; \frac{1}{h} \sum_{a \in \mathcal{A}} \psi(a) \, n(a).$$

Using these definitions the previous display becomes

$$\overline{w} \, h \;=\; \lambda \, \frac{\partial y}{\partial h} \, h \, \widetilde{\psi} \quad \Longrightarrow \quad \frac{\partial y}{\partial h} \;=\; \frac{\overline{w}}{\lambda \, \widetilde{\psi}}. \tag{32}$$

Combine (30) and (32) to express the two marginal products in terms of λ and $\widetilde{\psi}$:

$$\frac{\partial y}{\partial x} \;=\; \frac{p_x}{\lambda}, \qquad \frac{\partial y}{\partial h} \;=\; \frac{\overline{w}}{\lambda \, \widetilde{\psi}}.$$

Now use the definitions of the output elasticities

$$\alpha_m \;=\; \frac{\partial y}{\partial x} \, \frac{x}{y}, \qquad \alpha_l \;=\; \frac{\partial y}{\partial h} \, \frac{h}{y}.$$

Substitute the expressions for the marginal products above to obtain

$$\alpha_m \;=\; \frac{p_x}{\lambda} \, \frac{x}{y}, \qquad \alpha_l \;=\; \frac{\overline{w}}{\lambda \, \widetilde{\psi}} \, \frac{h}{y}.$$

Take the ratio α_l / α_m:

$$\frac{\alpha_l}{\alpha_m} \;=\; \frac{\dfrac{\overline{w}}{\lambda \, \widetilde{\psi}} \, \dfrac{h}{y}}{\dfrac{p_x}{\lambda} \, \dfrac{x}{y}} \;=\; \frac{\overline{w} \, h}{p_x \, x} \cdot \frac{1}{\widetilde{\psi}}.$$

Rearranging yields the stated identity

$$\frac{1}{\widetilde{\psi}} = \frac{p_x x}{\overline{w}\, h} \frac{\alpha_l}{\alpha_m},$$

which is equation (19) in the text. This completes the proof. □

B Data

B.1 Italian INPS microdata: structure and preparation

B.1.1 Data sources and coverage

The primary 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 under the VisitINPS Scholars program. The VisitINPS panel covers private-sector dependent employment from 1983 through 2024 and contains, for each administrative spell, (i) a harmonized employer identifier, (ii) contract start and end dates, (iii) weeks of contribution, (iv) gross taxable earnings (regular and excess/top-up components), (v) contract type and qualification codes, and (vi) municipality-level location for employer and employee. Worker demographics (year of birth, sex, citizenship) and mortality outcomes are taken from the INPS anagraphic registry and merged onto spells. Commuting-zone identifiers are imputed via the official crosswalk.

B.1.2 Overview of the cleaning pipeline

Data processing follows a multi-stage pipeline: (1) year-by-year ingestion and harmonization of annual spell files, (2) standardization of identifier and string variables, (3) construction of person–spell and firm-year aggregates, (4) definition of a consistent firm unit (location–commodity sector) for analysis, and (5) computation and deflation of pay measures.

B.1.3 Key steps and choices

- **Spell to annual panel.** The raw data arrive in spell format. I transform spells into an annual panel by assigning each worker to the single job (spell) in a calendar year with the highest annual earnings; this constitutes the worker's main episode for the year. This choice reduces ambiguity when multiple spells occur in the same year and is consistent with established practice in the literature.
- **Full-time restriction.** To minimize bias from unobserved hours, I restrict the core sample to full-time employment spells and to workers aged 20–65. Part-time and marginal employment cases are excluded.
- **Wage construction and deflation.** Annual nominal compensation for each spell is constructed as the sum of taxable wages, bonuses, and reported imputed earnings for figurative

events (e.g., short-time work or sickness). All nominal values are deflated to 2022 euros using an internal CPI series.

- **Firm unit.** The analysis defines the empirical firm as the *location–commodity sector* unit (location–commodity sector identifier). When merging enterprise-level financials (CERVED), an enterprise is assigned to the location–commodity sector with the largest cumulative employment, as described in Appendix B.3.
- **Occupation and qualification harmonization.** Qualification codes are grouped into a compact set of labor categories (manual workers, white-collar staff, managers, apprentices, and special categories). Observations falling in excluded categories (e.g., aviation-specific qualifications, managers where not part of the sample) are dropped as described below.

B.1.4 Sample exclusions and final panel

The core empirical sample retains private-sector, full-time, dependent employees aged 20–65 employed by private, for-profit entities (corporations, partnerships, profit-oriented cooperatives, and sole proprietorships). I exclude public-sector employers, non-profit organizations, religious/educational institutions, managers (see note below), apprentices, and specialized categories (e.g., aviation). Observations with implausibly low weekly wages (below €50) are dropped. 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 extract of occupation and education information is available from the INPS archives and coded to the International Standard Classification of Occupations (ISCO). Employers are required to report ISCO codes at contract initiation or when contracts are modified; starting from 2010. Thus anything that uses occupation and education variables is based on the movers subsample. ISCO data are available at the spell level, and thus merge to the INPS microdata accordingly.

B.3 Italian CERVED balance-sheet data

B.3.1 Data description

Firm-level financial information is drawn from the CERVED database, which reports annual balance-sheet variables for the universe of incorporated Italian businesses (1996–2018). Key variables used in the analysis include total assets, revenue, value added, wage bill, legal form, year of foundation, and firm status indicators (active, suspended, closed). CERVED reports at the enterprise level (id_impresa), whereas the labor data use location–commodity sector identifiers (id_azienda).

B.3.2 Matching enterprises to location–commodity sectors

Because CERVED is at the enterprise level, I assign each enterprise to a single location–commodity sector (id_azienda) by selecting the id_azienda with the highest cumulative employment across years for that enterprise. This assignment produces a unique mapping from id_impresa to id_azienda and thereby assigns enterprise financials to the local labor market unit employed in the analysis. Firm financials are therefore matched using year and firm ID to the INPS matched employer employee microdata.

B.3.3 Constructing firm-level variables

After assigning each enterprise to its predominant id_azienda, I construct:

- **Wage bill and employment.** Aggregated from the INPS annual panel at the id_impresa level (summing individual wages and weeks worked), then converted to yearly wage bills and employment headcounts.
- **Value added and capital.** Value added is taken directly from CERVED; capital stock is proxied as the sum of tangible and intangible fixed assets reported on the balance sheet.
- **Firm Intermediates.** Intermediate input expenditures are constructed from the accounting identity: value added equals revenues minus intermediate inputs.

B.4 Sample restrictions (summary)

The baseline sample satisfies the following restrictions:

1. private-sector dependent employment only (excludes self-employed, public employees, agricultural workers, contractors);
2. age 20–65 and full-time employment spells only;
3. exclusion of managers, apprentices, and special categories from the baseline (retained in robustness checks);
4. weekly/annual wages above €50 and below imposed lower bounds; and
5. firms restricted to private, for-profit legal forms for the main analysis.

Limitations

- **ISCO coverage:** occupation and education data are most complete for movers and for the post-2010 period; analyses relying heavily on ISCO variables therefore focus on movers or later subperiods.

- **Enterprise-to-location assignment:** assigning an enterprise to a single location–commodity sector simplifies linkage with CERVED but may obscure multi-site enterprises with heterogeneous local operations.
- **Hours unobserved:** restricting to full-time jobs mitigates, but does not eliminate, the absence of direct hours measures.

B.5 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 observable in administrative data, I rely on a two-way fixed-effects wage decomposition in the spirit of Abowd et al. (1999), Card, Heining, 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, and I show that their within–local-labor-market rankings closely track the corresponding rankings of the structural types a and z at the baseline calibration.

Residualizing wages. I begin by removing systematic wage variation that reflects observable characteristics rather than persistent heterogeneity. Let $\log w_{a,ij,t}$ denote the log real wage of worker a in firm ij and year t. I residualize $\log w_{a,ij,t}$ by projecting it onto year and profession (white- vs. blue-collar) fixed effects, gender, foreign status, two-digit sector fixed effects, and a flexible polynomial in age and experience fully interacted with both profession and sector. This allows for highly flexible life-cycle and sector-specific earnings profiles and ensures that the remaining variation reflects persistent worker and firm components rather than systematic compositional differences. Let $\tilde{w}_{a,ij,t}$ denote the residual from this regression.

Clustering firms into latent pay types. Directly estimating a firm fixed effect ψ_{ij} for every employer is problematic in 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 the distribution of $\tilde{w}_{a,ij,t}$ for all workers employed at ij over the analysis window. These statistics are standardized and subjected to a K-means clustering algorithm (with $K = 50$). Each employer ij is then mapped to a latent firm-type index $g(ij) \in \{1, \ldots, K\}$. The firm-cluster pay premium serves as the empirical indirect proxy of 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

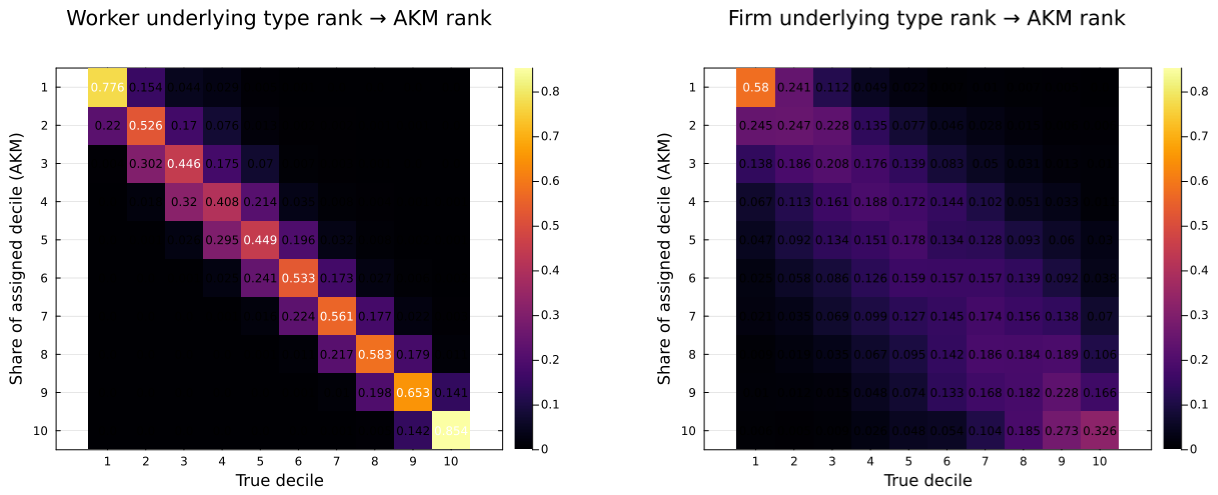
$$\log w_{a,ij,t} = \alpha_a + \psi_{g(J(a,t))} + x'_{at}\beta + \epsilon_{a,ij,t}, \tag{33}$$

where α_a is a worker-specific component and $\psi_{g(J(a,t))}$ is a firm-type premium associated with the cluster $g(J(a,t))$. Here $J(a,t)$ denotes the employer of worker a in year t. The covariate vector x_{at} matches the one used in the residualization step, ensuring consistency between the two stages. Estimation proceeds via high-dimensional fixed effects regression, absorbing both worker identifiers and firm-type identifiers.

Interpretation. The estimated worker effects $\hat{\alpha}_a$ provide a non-parametric 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 and approximate the model's firm heterogeneity z_{ij}. Although the AKM specification in (33) is a linearized representation of the richer wage-setting environment in the model, it offers a tractable way to recover stable worker and firm components that correspond closely to 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 synthetic panel from the model (details on how the simulation is performed are available in later sections). For each structural type decile, I compute the share of individuals (or firms) placed into 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 classified into the same AKM decile. When misclassification occurs, it is almost entirely to adjacent deciles, indicating 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 valid *indirect-inference proxies* for the latent worker and firm types (a, z_{ij}) used throughout the empirical analysis.



(a) Workers

(b) Firms

Figure B.1: **Alignment between structural deciles and AKM deciles in simulated data.** Each panel reports, for a given structural type decile, the distribution of AKM decile assignments: darker shading indicates a lower share of observations placed in that AKM decile. The mass lies predominantly along the diagonal, indicating strong rank preservation. Misclassification occurs almost exclusively into adjacent deciles, and the mapping is more precise for workers than for firms.

B.6 Production Function Estimation and Revenue Productivity

This section describes how I estimate firm-level revenue productivity using the merged INPS–CERVED panel for Italy. The objective is to recover a measure of (revenue-based) total factor productivity, ω_{ijt}, that is consistent with the model's firm heterogeneity. These estimates are also used to calibrate the model parameters α and γ (the decreasing-returns parameter and output elasticities in production).

B.6.1 Sample and variable definition

The estimation sample is the merged INPS–CERVED firm-year panel for the period 2014–2018 (year group 5), 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,
$$y_{ijt} \equiv \log(VA_{ijt}).$$
- **Capital:** K_{ijt}, measured as the sum of tangible and intangible fixed assets; I use $\log K_{ijt}$ as the capital input.

- **Labor:** L_{ijt}, measured as the firm's headcount derived from the INPS employment panel; the corresponding regressor is $\log L_{ijt}$.
- **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.

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 1.1 million firm-year observations for the 2014–2018 window.

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 are included to absorb sector–year-specific price levels and common shocks.

B.6.2 Empirical specification and identification

Within each three-digit sector s, I estimate a Cobb–Douglas value-added production function of the form

$$y_{ijt} = \beta_{\ell s} \log L_{ijt} + \beta_{ks} \log K_{ijt} + \delta_t^{(s)} + \omega_{ijt} + \varepsilon_{ijt}, \tag{34}$$

where $\delta_t^{(s)}$ are sector-specific year fixed effects, ω_{ijt} is (log) firm revenue productivity, and ε_{ijt} is an idiosyncratic error. The coefficients $\beta_{\ell s}$ and β_{ks} are sector-specific output elasticities with respect to labor and capital.

Input choices $(L_{ijt}, K_{ijt}, M_{ijt})$ are potentially correlated with ω_{ijt}. I address this using the control-function approach of Levinsohn and Petrin (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, this control function is approximated by 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 proxy for productivity. The ACF moment conditions use the Markov structure of ω_{ijt} and lagged inputs to disentangle the contemporaneous correlation between inputs and productivity. Estimation is carried out separately in each sector s, and sectors with very small samples are excluded to avoid poorly identified parameters.

B.6.3 Elasticities and productivity residuals

For each sector s, I recover the estimated elasticities $\hat{\beta}_{\ell s}$ and $\hat{\beta}_{ks}$ and the sector-specific year fixed effects $\hat{\delta}_t^{(s)}$. The fitted value from (34) 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

$$\hat{\omega}_{ijt} = y_{ijt} - \hat{y}_{ijt}. \tag{35}$$

By construction, $\hat{\omega}_{ijt}$ is orthogonal to observed inputs and sector–year effects, and captures firm-level deviations in value added not explained by measured factor inputs.

Across all sectors in the 2014–2018 window, the employment-weighted mean of the estimated labor elasticity $\hat{\beta}_{\ell s}$ is approximately 0.73, while the mean capital elasticity $\hat{\beta}_{ks}$ is around 0.10. The distribution of $\hat{\omega}_{ijt}$ exhibits substantial dispersion (standard deviation ≈ 0.64), indicating sizeable heterogeneity in firm revenue productivity even within narrowly defined industries.

These residual productivity measures $\hat{\omega}_{ijt}$ constitute the main firm-level outcome used in the Italian production and event-study analyses (e.g., Fact 4 in the main text). They provide an empirically grounded proxy for the model's firm productivity component z_{ij}, and are used to discipline the calibration of the production parameters (α, γ).

B.7 Additional descriptive statistics (period 2014–2019)

This subsection reports detailed tabulations for the baseline period 2014–2019. The tables summarize contract types, working-time status, qualification categories, reasons for termination, weekly wage distribution, legal-form composition (and the subset retained for-profit), firm-age distribution, coverage of AKM fixed-effect measures, wage dispersion and variance decompositions, employment-size distributions, and selected AKM fixed-effect percentiles. All tabulations are produced by the project's descriptive-stat routines; underlying logs are available for inspection.

Employment contract, working-time status, and qualification (2014–2019). This table reports the distribution of contract types, working-time status, and broad qualification categories for 2014–2019. The workforce in the main-episode sample is overwhelmingly full-time; permanent contracts account for roughly 76.4% of observations while fixed-term contracts account for the remaining 23.6%. Manual workers comprise the largest qualification group (about 58.5%), followed by white-collar employees (36.2%) and managers (5.4%).

Table B.1: Contract type, working-time status and qualification (2014–2019)

Variable / category	Freq.	Percent
Contract type		
Permanent	53,761,295	76.43%
Fixed-term	16,580,977	23.57%
Working-time status		
Full time	70,342,272	100.00%
Qualification / occupational category		
Manual workers	41,118,037	58.45%
White-collar staff	25,452,797	36.18%
Managers & quadro	3,771,438	5.36%

Notes: Tabulations are on the cleaned annual panel for 2014–2019. Contract-type and working-time categories refer to the main-episode job selected per worker-year. Source: descriptive-stat logs.

Weekly wage distribution (2014–2019). Table B.2 summarizes the distribution of real weekly wages over 2014–2019 (deflated to 2022 euros). The median weekly wage is approximately 438, and the mean is 531, indicating a moderately right-skewed wage distribution.

Table B.2: Weekly wage: mean and median (2014–2019)

Statistic	Value
Observations	64,657,422
Mean (weekly wage,)	531.30
Median (weekly wage,)	437.83

Notes: Summary statistics computed on main-episode worker–year observations. Real wages deflated to 2022 euros.

Legal form (enterprise) and retained for-profit sample. This table reports enterprise legal-form frequencies from the CERVED-INPS linkage and documents the subset retained for the baseline (private, for-profit entities). Corporations are the dominant legal form (73.4% of observations), and applying the for-profit filter removes roughly 4.26 million enterprise observations from the raw enterprise panel.

Table B.3: Legal forms (grouped) and sample selection

Grouped legal form	Freq.	Percent (of full sample)
Corporations	41,950,440	73.35%
Partnerships	3,725,053	6.51%
Cooperatives (for-profit)	3,883,923	6.79%
Sole proprietorships / family	3,391,917	5.93%
Other / public / non-profit	4,345,678	7.42%
Kept (private, for-profit forms)	(after filter)	*See note*

Notes: The table reports the full-sample legal-form frequencies (total obs = 57,196,011). In the baseline sample we keep observations with legal_form in the for-profit categories (corporations, partnerships, for-profit cooperatives and relevant sole proprietorships); the selection removed 4,259,282 observations from the raw enterprise panel. Source: CERVED linking and selection logs.

Firm age (2014–2019). Table B.4 reports the distribution of firm age in the matched INPS–CERVED panel. The median firm is 16 years old, and the mean is approximately 18.9 years, reflecting substantial heterogeneity in firm longevity within the private, for-profit sector.

Table B.4: Firm age: mean and median (years)

Statistic	Value
Observations (firm-years)	52,951,333
Mean firm age	18.86
Median firm age	16

Notes: Statistics computed on firm-year observations after restricting to private, for-profit enterprises. Firms are weighted by headcounts.

Wage dispersion and variance decomposition (2014–2019). Table B.5 reports three complementary variance decompositions for the 2014–2019 period. Using residualized wages that control for age, tenure, education and year effects, the within- and between-firm components are 0.0844 and 0.0612 (total 0.1456). When additionally controlling for occupation fixed effects (Mincer–occupation residual), defining the firm as occupation-firm, the within component falls to 0.0709 while the between component rises to 0.0974. For raw log wages, the between-firm component is much larger (0.137), accounting for most of the total variance (0.231).

Table B.5: Wage dispersion and variance decomposition (2014–2019)

Measure	Within	Between	Total
Residualized (Mincer residual)	0.0843606	0.0612299	0.1455903
Residualized (Mincer–occupation residual)	0.0709282	0.0974218	0.1683483
Raw log wage decomposition	0.0941585	0.1369044	0.2310629

Notes: "Mincer residual" controls for a polynomial in age and tenure interacted with education and year fixed effects. "Mincer–occupation residual" additionally absorbs occupation fixed effects. The decompositions are implemented by computing within- and between-firm squared deviations and averaging them within the 2014–2019 period. For the occupation resisidualized measure, the firm is intended as occupation-firm.

Employment size (firm and occupation–firm) and residualization. Table B.6 reports employment levels at the firm and occupation–firm cell. Firm-level employment is the headcount associated with the main job spell selected for each worker–year, with median firms employing 2 workers and a mean of roughly 9, reflecting a highly skewed size distribution. Occupation–firm cells are mechanically smaller (median = 1, mean 3.2), as they capture employment within occupation-by-firm units used in the occupation–firm classification. Residualized log employment is constructed by regressing log employment on a flexible polynomial in firm age while absorbing year and detailed sector fixed effects (including sector-specific age profiles), and taking the resulting residuals as the size-adjusted measure.

Table B.6: Employment distribution and dispersion (2014–2019)

	Firm-level employment	**Occupation–firm employment**
Median employment	2	1
Mean employment	8.98	3.22
SD(log employment)	1.118	0.799
SD(log employment residual)	1.034	0.776
SD(employment level)	131.31	25.96
Observations	5,566,246	11,814,983

Notes: Firm-level employment is measured as headcounts from the main job spell per worker–year. Occupation–firm employment refers to counts within occupation-by-firm units used in the occupation–firm AKM specification. Residualized log employment is obtained by projecting log employment on a flexible polynomial in firm age, absorbing year and detailed sector fixed effects.

Covariance structure of AKM fixed effects (2014–2019). Table B.7 reports the variances of worker and firm fixed effects and their covariance, for both the firm-level AKM specification (WFE2–FFE2) and the occupation–firm AKM specification (WFE4–FFE4). The variances reflect the dispersion in underlying worker and firm premia, while the covariance captures the degree of assortative matching between workers and firms induced by the AKM estimates.

Table B.7: Variance and covariance of AKM worker and firm fixed effects (2014–2019)

	Var(WFE)	Var(FFE)	Cov(WFE, FFE)
Firm-level AKM (WFE2, FFE2)	0.093231	0.029576	0.013840
Occ–firm AKM (WFE4, FFE4)	0.089517	0.045125	0.012010

Notes: Variances and covariances computed using the final sample weighted by number of workers. WFE2/FFE2 correspond to worker/firm fixed effects estimated at the firm level; WFE4/FFE4 correspond to worker/firm fixed effects estimated at the occupation–firm level.

Occupation structure by qualification (2014–2019). Table B.8 reports the main occupational aggregates in the 2014–2019 INPS panel, split by qualification class (manual/blue-collar vs. white-collar). Several fragmented occupation codes in the raw data—notably "Technical professions," "Qualified professions" and "Elementary occupations"—are merged into unified groups. The resulting categories capture the occupational structure of the Italian private-sector workforce while preserving the blue/white-collar distinction used throughout the analysis.

Table B.8: Largest merged occupations by qualification class (2014–2019)

Merged occupation group	**Blue-collar**	**White-collar**	**Total**	**Share**
Office clerks	266,704	3,823,426	4,090,130	10.7%
Commercial / qualified sales	1,408,154	1,948,638	3,356,792	8.8%
Craft and skilled trades (merged)	6,901,828	221,332	7,123,160	18.7%
Machine and plant operators (merged)	2,777,506	108,598	2,886,104	7.6%
Drivers and transport operators	2,261,252	61,953	2,323,205	6.1%
Technical professions (merged)	808,384	4,195,917	5,004,301	13.1%
Qualified professions (merged)	2,959,788	240,235	3,200,023	8.4%
Elementary occupations (merged)	4,603,052	159,132	4,762,184	12.5%
Financial and administrative clerks	753,810	637,139	1,390,949	3.7%
Social sciences / education professionals	50,605	632,234	682,839	1.8%
Math/physics/engineering specialists	19,937	570,238	590,175	1.6%
Total (all occupations)	23,770,971	14,309,276	38,080,247	100%

Notes: Merged categories combine fragmented INPS/ISCO 2-digit occupation codes belonging to the same broad domain. Blue-collar counts correspond to "manual workers", white-collar counts to "white-collar staff". Shares are computed relative to all worker–firm observations in 2014–2019 (38 million). Full uncollapsed tables are available in the project logs.

B.8 German SIEED

This section elaborates on the Sample of Integrated Employer-Employee Data (SIEED) and the methodology applied to process this data. The data access was provided via remote access use 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. I extensively based my data cleaning procedure using publicly accessible code from Card, Heining, et al., 2013 who uses 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, even during periods when they are employed outside the sample establishments. To ensure robust coverage of the sample, I do not limit the dataset to the panel establishments. The dataset provides variables such as the worker's establishment, average daily wage, and an extensive range of other characteristics, including employment status, age, gender, tenure, occupation, and education. Throughout, I employ the 3-digit occupational classification according to "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 a component of the 'employment details' submitted by the employer. If more than one job title with different classification codes apply 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 sub-group 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 codes in the KldB2010, 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, Heining, 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. Initially, 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. Additionally, I exclude jobs where real daily earnings are less than 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 apply established methods following Card, Heining, et al., 2013. This method involves fitting a series of Tobit models to log daily wages and imputing uncensored values for censored observations using the estimated parameters and random draws from the associated censored distribution. I fit 16 Tobit models (across 4 age and 4 education groups) after applying the sample restrictions mentioned above. Following Card, Heining, et al., 2013, I include controls such as age, firm size, firm size squared, a dummy for firms with more than ten employees, the mean log wage of co-workers, and the proportion of co-workers with censored wages.

B.9 Descriptive Statistics

This section summarizes key features of the processed SIEED dataset used in the empirical analysis. I document firm-level and worker-level characteristics, coverage of fixed-effect measures, and the evolution of wage and employment dispersion across year groups.

Tables B.9 and B.10 report firm-level and worker-level characteristics by year group. Table B.9 presents mean log wages and average years of schooling aggregated at the firm level, while Table B.10 provides average years of schooling, age, and experience for individual workers. The year groups correspond to the non-overlapping periods used in the AKM estimation and reveal long-run trends in educational attainment, worker age, and wage levels.

To assess the completeness of fixed-effect measures, Tables B.11 and B.12 report the share of missing firm and worker fixed effects. The fixed effects provided by the IAB on the full administrative dataset (FFE_o and WFE_o) exhibit very low rates of missingness. In contrast, the occupation–establishment AKM identifiers generate somewhat higher missingness rates for firm effects, reflecting the greater fragmentation of the firm definition. Nevertheless, both measures retain broad coverage of the sample.

Table B.13 reports the standard deviation of raw log wages and residual wages by year group. The first residual measure controls for a flexible polynomial in age and tenure interacted with education, together with year fixed effects ("Mincer residual"). The second residual measure additionally includes occupation fixed effects ("Mincer–occupation residual"). The results show a substantial increase in wage dispersion over time, much of which persists even after controlling for worker characteristics and occupations.

Table B.14 decomposes the variance in residual earnings into within-firm and between-firm components. The between-firm component accounts for most of the increase in total variance, consistent with rising heterogeneity across firms in the German labor market.

Table B.15 summarizes the dispersion in firm size by year group. I report the standard deviation of log employment and the dispersion of log-employment residuals after controlling for the firm's worker composition (e.g., shares of college-educated and low-skill workers, age structure). Firm-size dispersion remains relatively stable across periods.

Table B.16 reports the dispersion of firm and worker fixed effects across year groups for both the IAB-provided measures (FFE_o, WFE_o) and the occupation–establishment AKM estimates (FFE, WFE). Firm fixed effects display a rising trend in dispersion, while worker fixed-effect dispersion evolves more modestly.

Finally, Table B.17 lists the five most frequent low-skill and high-skill occupations in the sample. Skill levels are defined by the fifth digit of the KldB 2010 code. Low-skill occupations correspond to tasks requiring limited formal training, whereas high-skill occupations involve more specialized technical or administrative responsibilities.

Overall, the descriptive statistics document strong trends in workforce aging, increasing educational attainment, and rising wage dispersion—driven primarily by between-firm differences. These patterns motivate the empirical focus on firm heterogeneity, worker sorting, and the role of

firm pay policies in shaping wage inequality.

Table B.9: Firm-Level Mean Log Wage and Years of Schooling by Year Group

Year Group	Mean Log Wage	Mean Years of Schooling
1985–1992	4.4620	11.0829
1993–1997	4.5071	11.2403
1998–2002	4.4957	11.4114
2003–2009	4.5003	11.6206
2010–2017	4.5170	12.0601

Table B.10: Worker Characteristics by Year Group

Year Group	Mean Years of Schooling	Mean Age	Mean Experience
1985–1992	11.0700	34.2861	15.8632
1993–1997	11.2211	35.7612	17.2290
1998–2002	11.4017	37.5421	18.8821
2003–2009	11.6275	39.6829	20.8466
2010–2017	12.0396	42.4551	23.2875

Table B.11: Missing Firm Fixed Effects (FFE)

	Total observations	Share missing (FFE_o)	Share missing (FFE)
Full sample	3,473,220	1.68%	28.74%

Table B.12: Missing Worker Fixed Effects (WFE)

	Total observations	Share missing (WFE_o)	Share missing (WFE)
Full sample	2,525,434	3.43%	6.77%

Table B.13: Standard Deviation of Wages and Residuals by Year Group

Year Group	SD (Mincer residual)	SD (Mincer–occupation residual)	SD (Raw log wage)
1985–1992	0.3128	0.2856	0.3673
1993–1997	0.3212	0.2946	0.3687
1998–2002	0.3629	0.3236	0.4205
2003–2009	0.4271	0.3688	0.4930
2010–2017	0.4150	0.3677	0.4837

Notes: "Mincer residual" refers to log-wage residuals net of a polynomial in age and tenure interacted with education and year fixed effects. "Mincer–occupation residual" additionally includes occupation fixed effects.

Table B.14: Variance Decomposition of Residual Earnings by Year Group

Year Group	Within-firm variance	Between-firm variance	Total variance
1985–1992	0.0205	0.0611	0.0816
1993–1997	0.0197	0.0671	0.0868
1998–2002	0.0212	0.0836	0.1049
2003–2009	0.0222	0.1141	0.1363
2010–2017	0.0232	0.1122	0.1354

Table B.15: Firm Size and Employment Variability by Year Group

Year Group	SD (Log employment)	SD (Log employment residual)
1985–1992	1.0175	0.9680
1993–1997	1.0123	0.9595
1998–2002	0.9904	0.9438
2003–2009	0.9903	0.9510
2010–2017	0.9926	0.9550

Table B.16: Firm and Worker Fixed-Effect Dispersion by Year Group

Year Group	SD (FFE_o)	SD (FFE)	SD (WFE_o)	SD (WFE)
1985–1992	0.2221	0.3257	0.2717	0.2067
1993–1997	0.2263	0.3207	0.2892	0.2011
1998–2002	0.2398	0.3291	0.3220	0.1987
2003–2009	0.2794	0.3674	0.3605	0.1967
2010–2017	0.2453	0.3609	0.3870	0.1994

Table B.17: Top 5 Low-Skilled and High-Skilled Occupations by Frequency

Low-skilled occupations		High-skilled 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

Table B.18: Worker–Firm Decile Employment Shares: Italy (Industry, Occupation) and Germany

	Firm Decile									
Worker Decile	1	2	3	4	5	6	7	8	9	10
	.150	.136	.102	.105	.114	.093	.097	.079	.070	.058
1	.213	.138	.108	.096	.099	.088	.082	.074	.059	.044
	.182	.146	.114	.099	.089	.083	.075	.069	.068	.076
	.088	.118	.104	.113	.128	.105	.116	.093	.078	.061
2	.124	.125	.116	.110	.120	.108	.102	.092	.067	.038
	.125	.124	.107	.102	.098	.096	.092	.089	.088	.080
	.066	.104	.094	.110	.130	.108	.126	.105	.091	.071
3	.093	.102	.107	.109	.125	.118	.116	.107	.081	.043
	.094	.102	.096	.099	.099	.103	.104	.104	.105	.094
	.054	.094	.084	.103	.127	.109	.133	.117	.103	.080
4	.077	.087	.096	.103	.124	.123	.126	.120	.095	.049
	.076	.087	.087	.093	.099	.106	.110	.114	.118	.109
	.047	.086	.077	.096	.122	.107	.137	.125	.115	.091
5	.068	.076	.087	.096	.120	.122	.132	.132	.111	.058
	.064	.076	.080	.088	.096	.108	.114	.122	.130	.122
	.042	.080	.070	.089	.117	.103	.138	.132	.127	.105
6	.061	.068	.079	.090	.115	.119	.134	.140	.126	.070
	.056	.068	.074	.084	.094	.106	.116	.129	.141	.133
	.037	.075	.063	.083	.110	.099	.137	.136	.141	.122
7	.056	.060	.071	.083	.111	.114	.133	.145	.140	.086
	.050	.061	.070	.079	.090	.104	.117	.133	.149	.147
	.036	.071	.056	.077	.102	.094	.133	.137	.154	.144
8	.053	.053	.062	.075	.103	.108	.132	.150	.156	.108
	.048	.056	.066	.076	.087	.102	.118	.135	.156	.159
	.037	.068	.050	.069	.092	.085	.123	.135	.169	.175
9	.051	.047	.053	.064	.090	.097	.126	.152	.176	.146
	.050	.053	.061	.071	.083	.098	.115	.136	.162	.172
	.051	.067	.041	.056	.073	.070	.107	.120	.183	.236
10	.071	.045	.040	.046	.066	.073	.101	.137	.193	.229
	.074	.053	.060	.070	.078	.093	.110	.129	.161	.174
	.061	.090	.074	.090	.111	.097	.125	.118	.123	.114
Total	.087	.080	.082	.087	.107	.107	.119	.125	.120	.087
	.103	.094	.086	.088	.092	.100	.109	.120	.135	.136

Each cell reports the share of total employment between workers in a given *worker decile* (rows) and firms in a given *firm decile* (columns). Workers and firms are assigned to deciles based on the empirical distribution of their respective worker and firm fixed effects within their *local labor market*—defined by either industry or occupation for Italy, and occupation for Germany. The first two values in each cell report employment shares for Italy, while the third corresponds to Germany. For Italy, the first value is based on defining local labor markets by industry and the second by occupation. Shares sum to one across all worker–firm pairs. The final row reports the total share of employment accounted for by each firm decile.

C Empirical Evidence

C.1 Market Shares

Robustness: Alternative Firm-Type Classifications. As a robustness check, I recompute the worker–firm employment-share matrices using several alternative measures of firm heterogeneity. First, I classify firms into deciles of revenue total factor productivity (TFPR) computed from revenue-based production-function residuals. Second, I construct firm deciles based on the average co-worker type, where co-worker type is measured by the mean log wage of a worker's colleagues within the same firm. I conduct this exercise separately for industry-defined and occupation-defined local labor markets. Third, I implement an analogous robustness exercise using firm average logwage for classification in German administrative data.

Across all robustness specifications, the qualitative segmentation patterns documented in the main text remain strong. Low-type workers continue to disproportionately match with low-type firms, and high-type workers concentrate in high-type firms, regardless of whether firm type is defined by TFPR, mean co-worker wage, or German firm-side residual wages. Although levels vary across definitions, the bottom–bottom and top–top employment shares remain systematically larger than cross-decile shares, confirming that the observed segmentation is not an artifact of any particular firm classification scheme.

C.2 Ability Thresholds

This appendix provides additional details and robustness checks for the ability-threshold analysis reported in Section 4.2. The empirical design follows the same framework as in the main text but relaxes sample restrictions and considers alternative measures of firm quality.

The sample again includes all firms and worker transitions occurring in local labor markets with at least ten firms and ten workers, ensuring meaningful decile rankings while retaining smaller markets. For each firm i in market j and year t, the hiring threshold is defined as the minimum worker fixed effect among new hires. Thresholds are standardized within market–year cells to ensure comparability across heterogeneous markets.

The regression specification used for robustness checks extends equation (15) by including the full set of firm-level controls:

$$\widetilde{a}_{ij,t} = \beta_0 + f(\text{Firm Decile}_{ij,t}) + \beta \log(\text{New Hires}_{ij,t}) + \beta_x X_{ij,t} + \gamma_m + \gamma_t + \varepsilon_{ij,t}, \tag{36}$$

where $X_{ij,t}$ includes the firm's workforce composition (share college-educated, share blue-collar), average worker age, and a polynomial in firm age. These controls are omitted from the preferred specification in the main text but included here to show that results are not sensitive to observable firm characteristics. As before, standard errors are clustered at the appropriate firm level.

Table B.20 reports the full set of estimates, including specifications with and without controls and the corresponding estimates for Germany. Across all specifications, the relationship between

Table B.19: Worker–Firm Decile Employment Shares: Italy (Industry, Occupation) and Germany (Robustness)

	Firm Decile									
Worker Decile	1	2	3	4	5	6	7	8	9	10
1	.126	.113	.110	.106	.103	.102	.097	.092	.086	.067
	.217	.145	.108	.093	.079	.071	.067	.063	.068	.090
	.255	.162	.126	.098	.080	.067	.058	.052	.047	.056
	.329	.210	.137	.096	.071	.055	.042	.033	.021	.009
2	.105	.106	.110	.110	.106	.107	.106	.095	.089	.069
	.154	.134	.117	.106	.091	.084	.082	.075	.077	.080
	.159	.152	.139	.116	.100	.085	.074	.066	.058	.052
	.132	.170	.156	.135	.113	.096	.078	.062	.041	.018
3	.094	.099	.103	.108	.111	.111	.111	.101	.092	.070
	.121	.120	.115	.109	.101	.096	.092	.084	.085	.078
	.119	.131	.131	.121	.111	.098	.088	.078	.068	.056
	.079	.119	.130	.135	.128	.119	.106	.090	.065	.030
4	.086	.093	.101	.109	.109	.114	.114	.107	.095	.072
	.099	.106	.109	.109	.107	.105	.102	.093	.091	.079
	.095	.112	.118	.118	.116	.107	.100	.091	.079	.063
	.054	.086	.105	.121	.128	.131	.126	.116	.091	.044
5	.082	.087	.098	.103	.109	.113	.116	.111	.104	.079
	.084	.095	.103	.107	.111	.112	.108	.101	.097	.082
	.080	.097	.106	.112	.115	.113	.111	.103	.091	.072
	.039	.064	.083	.105	.120	.133	.137	.137	.120	.063
6	.078	.083	.091	.101	.108	.114	.116	.116	.108	.084
	.074	.087	.098	.104	.113	.115	.111	.108	.103	.088
	.068	.086	.095	.104	.111	.116	.119	.112	.104	.085
	.029	.048	.067	.089	.110	.129	.144	.155	.146	.084
7	.073	.078	.089	.099	.106	.118	.120	.118	.113	.087
	.066	.080	.094	.103	.111	.113	.114	.115	.109	.096
	.059	.076	.085	.098	.106	.115	.123	.119	.117	.101
	.021	.036	.052	.074	.096	.122	.145	.168	.173	.113
8	.069	.071	.085	.095	.105	.115	.123	.125	.117	.095
	.061	.076	.091	.099	.105	.110	.114	.120	.116	.107
	.054	.066	.075	.088	.098	.112	.121	.128	.134	.123
	.017	.028	.041	.059	.082	.110	.139	.175	.198	.152
9	.065	.064	.076	.088	.099	.114	.125	.132	.129	.110
	.059	.076	.089	.091	.096	.104	.110	.123	.125	.126
	.051	.059	.066	.078	.088	.103	.113	.132	.149	.163
	.013	.021	.031	.046	.067	.095	.127	.172	.218	.213
10	.060	.056	.064	.077	.088	.106	.121	.139	.149	.140
	.074	.081	.079	.078	.082	.094	.100	.119	.129	.165
	.063	.058	.059	.068	.075	.086	.095	.118	.152	.228
	.012	.017	.024	.035	.051	.076	.108	.156	.219	.304
Total	.084	.085	.093	.099	.104	.111	.115	.113	.108	.087
	.101	.100	.100	.100	.100	.100	.100	.100	.100	.099
	.100	.100	.100	.100	.100	.100	.100	.100	.100	.100
	.193	.131	.107	.103	.104	.113	.124	.144	.166	.180

Each cell reports employment shares for a worker–firm decile pair. Workers and firms are assigned to deciles based on the distribution of worker and firm-type heterogeneity measure within their local labor market. The four stacked entries (top to bottom) correspond to: (A) Italy—TFPR deciles (industry–CZ LLM), (B) Italy—average co-worker log wage (industry–CZ LLM), (C) Italy—average co-worker log wage (occupation–CZ LLM), and (D) Germany—firm mean log wage. Shares are rounded to three decimals. The final row reports column totals for (A)–(D).

firm quality and hiring thresholds remains strongly positive, and the coefficient on $\log(\text{New Hires})$ remains large and negative, consistent with the mechanical sampling logic discussed in the main text.

Table B.20: Ability Thresholds and Firm Quality: Full Robustness Results (2014–2019)

	LLM: Industry			**LLM: Occupation**			**Germany**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Firm Decile	0.047***	0.066***	0.066***	0.078***	0.087***	0.092***	0.066***
	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0003)
Log(New Hires)		-0.524***	-0.529***		-0.568***	-0.568***	-0.577***
		(0.001)	(0.001)		(0.001)	(0.001)	(0.002)
Additional Controls	No	No	Yes	No	No	Yes	No
Fixed Effects							
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Market FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,361,957	2,361,957	2,361,957	4,638,030	4,638,030	4,638,030	1,752,033
R-squared	0.076	0.287	0.335	0.075	0.236	0.335	0.119

Notes: This table reports the full robustness specifications for the relationship between firm quality and firms' hiring thresholds. Firm quality is measured by local pay-premium deciles. The dependent variable is the market–year standardized minimum worker fixed effect among new hires. Columns (1)–(3) use local labor markets defined by *industry–commuting zones*, Columns (4)–(6) use *occupation–commuting zones*, and Column (7) reports the corresponding German estimates. "Additional Controls" include workforce composition (share college-educated, share blue-collar), average age, and a polynomial in firm age. Standard errors clustered at the firm level (or firm$\times$occupation level for occupation-based markets).
***Significant at 1%; **Significant at 5%; *Significant at 10%.

C.3 HHI Indices

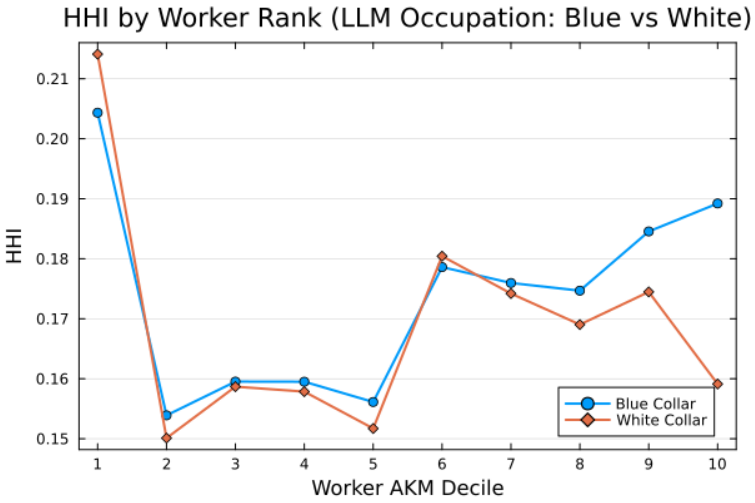


Figure B.2: Wage-Bill HHI by Worker Fixed-Effect Decile (Occupation LLM; Blue vs. White Collar)

Notes: For each local labor market (occupation–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. Panel plots show results separately for blue- and white-collar workers. Local HHI values are aggregated to the national level by weighting each market–decile observation by the total wage bill of that group. Standard errors / confidence bands are omitted from the figure for clarity; full numerical values are reported in the companion tables in Appendix C.1.

The appendix figure reports the wage-bill HHI by worker AKM fixed-effect decile for occupation-defined local labor markets, separately for blue- and white-collar jobs. The two series are nearly identical.

Aggregating across markets, the mean HHI under the occupation-based market definition is approximately 0.18. The index displays a non-monotonic pattern across the ability distribution: it falls from about 0.21 in the lowest decile to roughly 0.5 near the second decile, then rises back to about 0.18 among the highest deciles.

C.4 Event Study: Productivity Effects of Unexpected Worker Deaths

Balance, acceptance rate, and final sample. The matching procedure achieves a very high acceptance rate: 98.9% of treated worker–firm pairs are successfully matched to a placebo event, yielding a final sample of 34,030 matched events (17,015 treated and 17,015 controls). Observations for which no valid match is found are excluded. Table B.21 reports pre-event summary statistics showing that the matched comparison groups are well balanced across all key characteristics.

Among below-average ($L = 1$) worker deaths, treated and control units are nearly identical in average worker fixed effects (-0.140 vs. -0.137), age (45.8 vs. 45.7 years), experience (25.1 vs. 24.4 years), and log wages (-0.134 vs. -0.161). For above-average ($L = 0$) worker deaths, treated and control samples are likewise comparable in worker fixed effects (0.178 vs. 0.174), age (50.1 vs.

50.0 years), experience (28.3 vs. 27.5 years), and log wages (0.282 vs. 0.076). Occupation shares are essentially identical across groups: 20.7% white-collar among low-side events and 27.2% among high-side events. Gender composition is also mechanically the same by construction.[61]

Taken together, these similarities indicate that the matched sampling procedure delivers a well-balanced set of treated and control worker–firm pairs prior to the event.

Table B.21: Pre-Event Balance between Treated and Control Worker–Firm Pairs

	Below-Average Deaths ($L = 1$)		**Above-Average Deaths ($L = 0$)**	
	Control ($T = 0$)	**Treated ($T = 1$)**	**Control ($T = 0$)**	**Treated ($T = 1$)**
Worker fixed effect (mean)	−0.137	−0.140	0.174	0.178
(median)	(−0.127)	(−0.129)	(0.125)	(0.129)
[SD]	[0.173]	[0.179]	[0.260]	[0.267]
Age (years, mean)	45.73	45.80	50.03	50.09
(median)	(47)	(47)	(52)	(52)
[SD]	[9.92]	[9.91]	[9.08]	[9.08]
Experience (years, mean)	24.36	25.11	27.53	28.27
(median)	(25)	(25)	(30)	(30)
[SD]	[11.27]	[10.59]	[11.24]	[10.51]
Log wage (mean)	−0.161	−0.134	0.076	0.282
(median)	(−0.134)	(−0.101)	(0.049)	(0.185)
[SD]	[0.283]	[0.421]	[0.332]	[0.535]
White-collar (%)	20.7	20.7	27.2	27.2
Female (%)	15.4	15.4	11.2	11.2
Observations	17,015		17,015	

Notes: This table reports pre-event summary statistics for treated and matched control worker–firm pairs used in the event-study analysis of worker deaths. Columns (1)–(2) display results for deaths of below-average workers ($L = 1$), and Columns (3)–(4) for deaths of above-average workers ($L = 0$). Treated observations ($T = 1$) correspond to actual worker deaths; control observations ($T = 0$) are matched placebo events drawn from worker–firm pairs that never experienced a valid death. Means, medians (in parentheses), and standard deviations (in brackets) are reported. All variables are measured in the pre-event year ($d - 1$).

Covariates before shock for TFPR event study sample. Table B.22 presents pre-event ($t = -1$) summary statistics—means, medians, and standard deviations—for treated and control firms separately by L group. The matching procedure yields treated and control firms that are highly comparable across all observable dimensions. Firm size, age, and productivity are closely aligned: the average firm employs roughly ten workers, is about 15–16 years old, and has a log revenue productivity of approximately 1.25. Distributions of firm pay premia, employment shares within local labor markets, and worker fixed effects (both average and minimum) are similarly well balanced across treatment status. Overall, the pre-event characteristics provide no indication of systematic

[61]Gender is identical across treated and control observations because exact matching was performed on gender.

differences between treated and matched control firms.

Table B.22: Pre-Event Firm Characteristics at $t = -1$ (Mean (Median) [SD])

	T=0, L=0	T=1, L=0	T=0, L=1	T=1, L=1
Panel A: Productivity and firm pay premium				
Log revenue productivity (ω)	1.254 (1.212) [0.637]	1.254 (1.232) [0.562]	1.261 (1.233) [0.560]	1.246 (1.216) [0.542]
Firm pay premium (FFE2)	-0.023 (0.005) [0.127]	-0.030 (-0.006) [0.133]	-0.022 (0.005) [0.119]	-0.018 (0.005) [0.128]
Panel B: Size, hires, incumbents				
Firm age (years)	15.09 (13) [9.93]	15.15 (13) [10.24]	16.61 (14) [10.52]	15.99 (14) [10.27]
Employment (headcounts)	10.26 (9) [5.46]	10.50 (10) [5.42]	10.28 (9) [5.24]	10.59 (10) [5.24]
Total new hires	1.35 (1) [2.38]	1.35 (1) [2.17]	1.42 (1) [2.10]	1.43 (1) [2.09]
Total incumbents	7.64 (7) [4.35]	7.91 (7) [4.42]	7.66 (7) [4.30]	7.99 (7) [4.38]
Panel C: Market share and worker ability				
Employment share (within market)	0.102 (0.016) [0.209]	0.107 (0.017) [0.217]	0.094 (0.015) [0.200]	0.087 (0.013) [0.194]
Average worker fixed effect	-0.013 (-0.012) [0.128]	-0.017 (-0.020) [0.130]	-0.011 (-0.014) [0.132]	-0.016 (-0.020) [0.129]
Minimum worker fixed effect	-0.270 (-0.246) [0.193]	-0.235 (-0.218) [0.194]	-0.271 (-0.241) [0.205]	-0.230 (-0.222) [0.171]

Notes: Each cell reports Mean (Median) [Standard deviation] for the variable at $t = -1$ by matched group. Groups are defined by treatment status (T) and position indicator L (loss of an above- vs. below-average worker). Sample: firm-productivity event panel (1996–2018). Variables: "omega" = log firm revenue productivity; "FFE2" = firm pay premium; "emp" = firm headcounts; "tot_new_hires" = number of new hires; "tot_incumbents" = number of incumbents; "emp_share" = firm's employment share within its local labor market; "avg_wfe" = mean worker fixed effect; "min_wfe" = minimum worker fixed effect.

D Model Calibration and Quantification

D.1 Numerical algorithm to solve the general equilibrium

This appendix summarizes the numerical algorithm used to compute the decentralized equilibrium (and the efficient benchmark) described in the text. The solver implements three nested steps that mirror the economic structure of the model: (i) solve the within-market firm–worker allocation given an exogenous distribution of employed mass by ability, (ii) aggregate across local markets to recover type–market allocations and market wage indices, and (iii) update aggregate labor supply by ability and iterate until general-equilibrium convergence.

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 vector $S(a)$, which is *not* a probability or a share: it is a labor-supply shifter that scales the mass of type-a workers who participate in the labor market. Thus the total economy-wide supply of type-a workers is

$$N(a) = S(a)\, f_a(a).$$

Workers then choose across local labor markets according to the across-market CES structure with elasticity θ. Let S_{aj} denote the across-market wage-bill shares that characterize their choices

in equilibrium. These are used to determine the mass of type-a workers entering each market:

$$N_j(a) = N(a) \cdot S_{aj}.$$

Inside each market j, there are N firms indexed by i, each with baseline productivity z_{ij}. Worker–firm productivity is given by $\phi(a, z_{ij})$. The object $S_{ij}(a)$ denotes the *within-market* wage-bill share: the share of type-a wage income paid by firm (i, j); the last element in each row represents non-employment.

Model primitives—$\alpha, \gamma, \eta, \theta, \sigma, \varphi, R$—govern production, labor supply, and firm behavior.

Inner step — within-market equilibrium

For each local labor market j, the inner step solves the equilibrium interactions between firms and the workers of all abilities who arrive in the market. Conditional on the effective employed mass $N_j(a)$ described above, the inner step determines firm hiring, wages, and wage-bill shares $S_{ij}(a)$.

Starting from a candidate share matrix, the algorithm performs:

1. **Within-market labor supply (firm choice of workers).** Using the nested-CES structure with elasticity η across firms,
$$n_{ij}(a) \propto S_{ij}(a)^{\,\eta/(1+\eta)} \cdot N_j(a),$$
where the normalization ensures $\sum_i n_{ij}(a) = N_j(a)$.

2. **Total employment and realized productivity.**
$$h_{ij} = \sum_a n_{ij}(a), \qquad \overline{\phi}_{ij} = \sum_a \frac{n_{ij}(a)}{h_{ij}} \phi(a, z_{ij}).$$

3. **Marginal products.** Compute the marginal product schedule $MPL_{ij}(a)$ using equation (10), which captures both scale effects and endogenous productivity effects through $\overline{\phi}_{ij}$.

4. **Wage-setting via markdowns.** Firms pay
$$w_{ij}(a) = \begin{cases} \mu_{ij}(a)\, MPL_{ij}(a), & MPL_{ij}(a) > 0, \\ 0, & MPL_{ij}(a) \le 0, \end{cases}$$
where the markdown is determined by the inverse elasticity of firm-specific labor supply:
$$\epsilon_{ij}(a) = \left[\tfrac{1}{\eta} + \left(\tfrac{1}{\theta} - \tfrac{1}{\eta}\right) s_{ij}(a)\right]^{-1}, \qquad \mu_{ij}(a) = \frac{\epsilon_{ij}(a)}{\epsilon_{ij}(a) + 1}.$$
The efficient benchmark sets $\mu_{ij}(a) \equiv 1$.

5. **Update wage-bill shares.** Workers choose among firms according to:

$$S_{ij}^{\text{new}}(a) \propto \big(\mu_{ij}(a)\, MPL_{ij}(a)\big)^{1+\eta}.$$

   Normalize rows so their entries sum to one (with the final entry as non-employment).

6. **Iterate.** Repeat until $S_{ij}(a)$ and the implied wages and employment converge.

This inner loop computes a fixed point consistent with firm optimization and worker within-market CES choice behavior.

Middle step — aggregation across markets

Given an aggregate labor-supply vector $S(a)$ (outer step) and resulting total supply $N(a) = S(a)f_a(a)$, workers choose across the M local markets according to the CES structure with elasticity θ.

1. **Across-market labor supply.** A candidate set of across-market wage-bill shares S_{aj} implies

$$N_j(a) = N(a) \cdot S_{aj}.$$

2. **Solve each market.** For each j, feed $N_j(a)$ into the inner step to obtain wages $w_j(a)$, employment allocations, and profits.

3. **Across-market share update.** Workers choose markets according to:

$$S_{aj}^{\text{new}} \propto \left(\frac{w_j(a)}{W(a)}\right)^{1+\theta}, \qquad W(a) = \Big(\int_0^1 w_j(a)^{1+\theta}\, dj\Big)^{1/(1+\theta)}.$$

4. **Iterate.** Repeat until the across-market shares S_{aj} converge.

This step maps firm-level outcomes into type–market allocations, producing consistent wage indices.

Outer step — aggregate labor supply (general equilibrium)

The final step ensures aggregate labor supply is consistent with the wage indices generated by markets.

1. Given a trial vector $S(a)$, compute total supply:

$$N(a) = S(a)\, f_a(a).$$

2. Run the middle step to recover updated wage indices $W(a)$.

3. Update labor supply using the inverse labor-supply condition:

$$S(a)^{\text{new}} = W(a)^{\frac{(1-\sigma)\varphi}{1+\sigma\varphi}}.$$

4. Iterate under relaxation until convergence.

At convergence, the allocation satisfies firm optimality, worker choice 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 $n_{ij}(a)$, - wages $w_{ij}(a)$, market wage indices $w_j(a)$, and aggregate $W(a)$, - firm profits π_{ij} and aggregate profits.

Safeguards enforce non-negativity of MPL, remove indeterminate values, and stabilize convergence via under-relaxation. The efficient benchmark is obtained by imposing $\mu_{ij}(a) = 1$ everywhere.

The entire structure is modular: alternative specifications of $\phi(a, z)$, markdowns, or labor-supply elasticities can be inserted without changing the fixed-point architecture.

D.2 Calibration of the Distribution of Firms Across Local Labor Markets

To discipline the distribution of firms across local labor markets, I simulate the economy over $M = 2000$ markets, each hosting up to 200 potential firms. A local labor market is defined as the intersection between a three-digit industry and a commuting zone, consistent with the structure of the balance-sheet data used in the calibration.[62]

The number of active firms per local labor market is assumed to follow a Pareto distribution with an upper bound of 200 firms. 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 observed data, as illustrated in Figure B.3. In addition, a mass point at one-firm markets is included to capture the substantial share of highly concentrated markets observed in the data.

D.3 Simulation of the Model-Implied Panel Dataset

This subsection describes how I construct the synthetic worker–firm panel used in the quantitative analysis. The guiding idea is straightforward: each simulated firm receives a workforce drawn from the model's general-equilibrium employment structure, and workers move across firms over

[62] This definition mirrors the empirical aggregation used in the CERVED data and ensures that model markets align with observed employer groupings.

Table B.23: Calibration of the Firm-Size Distribution Across Local Labor Markets

Parameter	Value
Location parameter of Pareto distribution	1.000
Tail parameter (shape)	0.966
Scale parameter	3.519
Mass of markets with one firm	0.212
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 observed empirical distribution of firms per market. A mass point at one-firm markets is included to capture the substantial share of highly concentrated markets in the data.

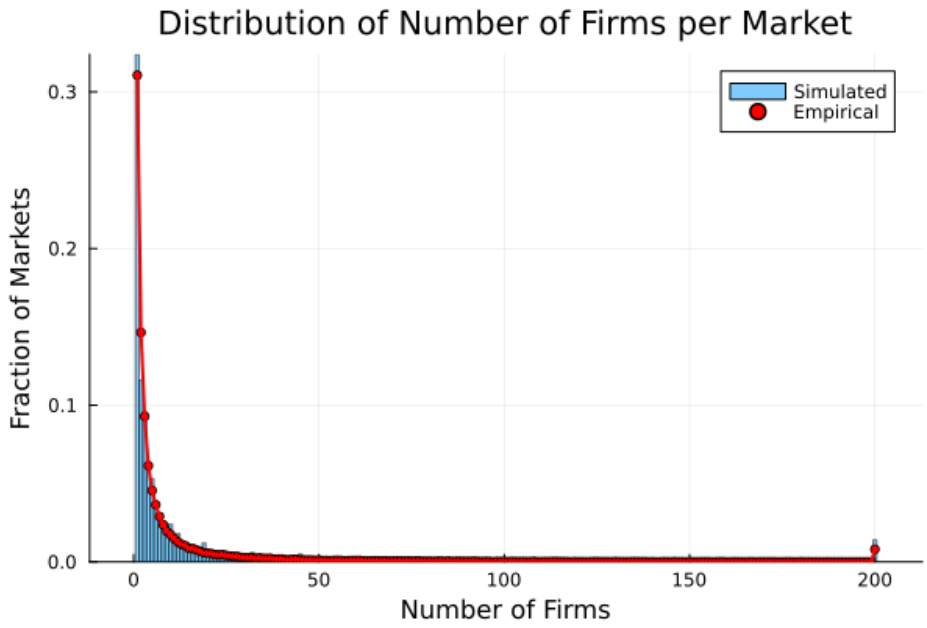


Figure B.3: Distribution of Firms Across Local Labor Markets: Data vs. Model

time according to an empirically disciplined switching probability. The objective is to generate a panel that mirrors the empirical distribution of firms, the number of workers per firm, the time horizon of the data, and the observed mobility patterns of workers, while retaining computational tractability.

The empirical dataset contains 1,555,249 firms and 10,431,213 unique workers over the 2014–2019 period, implying roughly nine workers per firm. The simulation preserves these ratios at a computationally convenient scale by setting the total number of simulated firms and workers proportionally. A second discipline comes from worker mobility: in the data, the share of workers who change employer between years averages 0.2445 over 2014–2019, with year-specific rates ranging between 0.21 and 0.27. I use this switching probability as the reassignment parameter governing worker mobility in the simulated panel.

The simulation takes as inputs the full set of model-implied objects describing every labor

market, every firm within that market, the firm's employment by worker type, the wage it pays to each type, and its basic production characteristics such as output, size, and productivity. No subsampling of firms is performed: all firms generated by the model's general-equilibrium solution are included in the panel.

Workers are assigned to permanent ability types at the start of the simulation by sampling from the model-implied distribution of abilities, accounting for the aggregate labor supply decision and the underlying ability probability density function. The number of simulated workers is chosen so that the average number of workers per firm matches the empirical ratio. In the first period of the panel, each worker is assigned to a firm using the model's employment structure: a worker is first allocated to a labor market in proportion to the number of workers of her ability type that the model assigns to each market; conditional on this market, she is assigned to a specific firm in proportion to the number of such workers the firm employs in the model. In this way, the simulated cross-sectional distribution of workers across firms reproduces the model's equilibrium allocations.

Starting in the second period, workers are allowed to change firms over time. Each worker independently switches to a new employer with probability equal to the empirical switching rate. If a worker switches, the assignment rule is identical to that in the first period: the worker draws a labor market and then a firm according to the same model-implied allocation probabilities. If the worker does not switch, she remains with her current employer. Repeating this procedure for all periods generates a dynamic matched panel whose worker flows closely resemble those observed in the administrative data.

For each period, the simulation records a row for every worker containing her worker identifier, period, ability type, assigned firm, assigned market, wage, log wage, and the matched firm's characteristics (productivity, size, output, wage wedge, and concentration). Stacking these rows over all periods produces a balanced worker–firm panel that is then used to compute model-implied AKM decompositions and to compare the variance and covariance structure of wages to that in the data.

The resulting synthetic panel successfully reproduces the empirical structure that matters for quantitative evaluation: the number of firms, the number of workers per firm, the time horizon, and the observed mobility of workers between employers.

Table B.24: Empirical Targets Used in the Panel Simulation

	Value	Source
Number of firms	1,555,249	Administrative data
Number of workers	10,431,213	Administrative data
Average workers per firm	9.0	Derived from data
Panel years	2014–2019 (6 years)	Administrative data
Mean annual switching rate	0.2445	Worker mobility statistics
Minimum annual switching rate	0.2138 (2019)	Worker mobility statistics
Maximum annual switching rate	0.2685 (2015)	Worker mobility statistics

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.

D.4 Construction of Empirical and Simulated Moments

All constructions are carried out by local labor market j and year t. I maintain the notation used in the main text: worker log wages are $\log w_{a,ij,t}$, worker fixed effects are α_a, and firm pay premia (or firm-type pay premia when discretized) are $\psi_{k(ij)}$, where $k(ij)$ denotes the cluster of firm i located in local labor market j.

AKM estimates and firm clustering. I estimate the two-way fixed-effect decomposition in (14) to recover $\hat{\alpha}_a$ and $\hat{\psi}_{k(ij)}$. For each firm I compute summary statistics of the within-firm log-wage distribution (e.g. selected quantiles) over the entire sample period. I standardize these firm-level moments and apply a K-means clustering algorithm to obtain K mutually exclusive firm groups $g(j) \in \{1, \ldots, K\}$. Each firm is then permanently assigned to one firm type g.

Given these firm types, I estimate an AKM-style two-way fixed effects regression of the form

$$\log w_{a,ij,t} \;=\; \alpha_a \;+\; \psi_{g(ij)} \;+\; \varepsilon_{a,ij,t},$$

Where α_a are worker fixed effect, and $\psi_{g(ij)}$ firm cluster pay premium.

Wage dispersion and covariance moments. The variance and covariance moments used in the calibration are computed directly from the estimated fixed effects and observed wages. Let $\hat{\alpha}_a$ denote the estimated worker effect and $\hat{\psi}_{g(ij)}$ the estimated firm-type effect. I compute, using headcounts as weights (as in the data): (i) the standard deviation of worker fixed effects, $\mathrm{sd}(\hat{\alpha}_a)$; (ii) the standard deviation of firm-type fixed effects, $\mathrm{sd}(\hat{\psi}_{g(ij)})$; (iii) the covariance $\mathrm{cov}(\hat{\alpha}_a, \hat{\psi}_{g(ij)})$ evaluated over worker–firm matches; and (iv) the standard deviation of log wages, $\mathrm{sd}(\log w_{a,ij,t})$. All moments are calculated using the full set of worker–firm–year observations, weighting by worker headcounts. In addition, I decompose the variance of log wages into within- and between-firm components.

Wage-bill HHI by worker fixed-effect decile. A key moment disciplining firm heterogeneity and the strength of worker–firm complementarities is the concentration of the wage bill across firms, conditional on worker fixed effect. Within each local labor market–year pair (j, t), I rank workers by their estimated AKM fixed effect $\hat{\alpha}_a$ and partition them into ten equal-sized within-market deciles.

For each decile, I construct two model-based measures of wage-bill concentration. *First*, I use the HHI implied directly by the full equilibrium allocation of the model. This measure is preferred because it reflects the true underlying model economy and is not affected by the fact that the simulated panel may not draw all firms active in a market. *Second*, I compute an HHI from the simulated worker–firm panel using the identical procedure applied to the data; this measure is slightly higher due to sampling-induced firm undercoverage, but very similar in levels and patterns.

For calibration, I target only the *aggregate* employment-weighted HHI (the "Total" row in Table B.25). The decile profile is left as a partially untargeted validation moment. The maximum absolute deviation across worker deciles between the model-implied HHI and the empirical HHI is modest (approximately 0.02), indicating that the model reproduces well the heterogeneity in concentration across the worker ability distribution. Figure B.4 visualizes this comparison.

Table B.25: Wage-Bill HHI by Worker Fixed-Effect Decile: Data vs. Model

	Data (Empirical HHI)	**Model: Strategy 1** (Model-Implied HHI)	**Model: Strategy 2** (Panel-Constructed HHI)
Decile 1	0.2807	0.2884	0.3149
Decile 2	0.2660	0.2854	0.3123
Decile 3	0.2746	0.2887	0.3138
Decile 4	0.2795	0.2905	0.3128
Decile 5	0.2824	0.2927	0.3118
Decile 6	0.2981	0.2980	0.3169
Decile 7	0.3008	0.3008	0.3184
Decile 8	0.3041	0.3045	0.3206
Decile 9	0.3149	0.3114	0.3253
Decile 10	0.3158	0.3204	0.3376
Total	**0.2917**	**0.2980**	**0.3184**

Notes: This table reports wage-bill Herfindahl–Hirschman Index (HHI) values by worker fixed-effect decile. The first column presents the empirical HHI computed from administrative data. The second column ("Model: Strategy 1") reports the HHI implied by the model's full equilibrium allocation. The third column ("Model: Strategy 2") reports the HHI computed from the simulated worker–firm panel using the same procedure applied to the data. The "Total" row provides the employment-weighted average HHI across worker deciles.

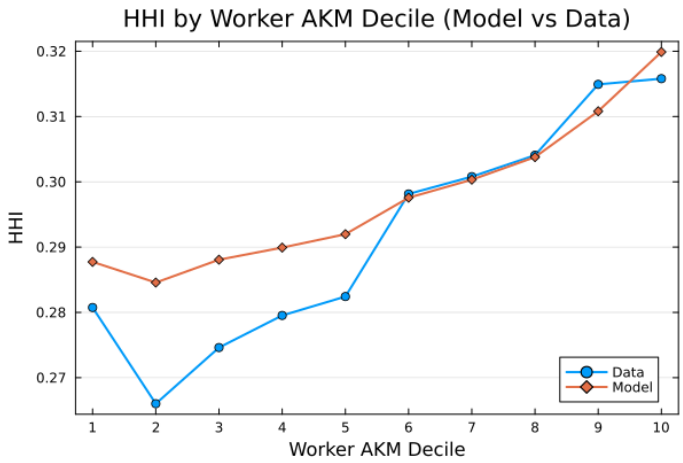


Figure B.4: Wage-Bill HHI by Worker Fixed-Effect Decile: Data vs. Model (Strategy 1)

Notes: This figure plots wage-bill Herfindahl–Hirschman Index (HHI) values by worker fixed-effect decile. The series labelled "Data" depicts the HHI computed from administrative data. The series labelled "Model" depicts the HHI implied by the model's full equilibrium allocation (Strategy 1). HHI values are calculated within each local labor market and year, aggregated to deciles, and averaged using employment weights.

Hiring-threshold moment (untargeted). To discipline the extent of firm screening on worker ability, I construct a hiring-threshold moment based on the distribution of estimated worker fixed effects among new hires.

I first identify incumbents and new hires at the annual frequency. A worker a is classified as an incumbent at firm ij in year t if she was also employed at ij in year $t-1$; otherwise she is classified as a new hire in that firm–year. For each firm ij, and year t, I compute the total number of new hires and retain only firm–market–year cells with strictly positive hiring activity and with at least ten firms and ten workers in the corresponding market–year, so that decile ranking is meaningful.

Within each firm–market–year cell, I define the hiring threshold as the minimum worker fixed effect among the newly hired workers, and I normalize this at the market level.

I then relate these thresholds to measures of firm rank, using the same three measures of firm rank as in the data. First using the estimated AKM firm-type effects, second and third the average ability and wages of incumbent workers. Then, I perform the same regression as I did in the data in 15, by controlling for market year fixed effect and log of new hires.

Employment-share moments by worker and firm rank (partially untargeted). The remaining moments capture the extent of sorting and segmentation between workers and firms. I focus on relatively large labor markets by restricting attention to market–year cells with at least 100 firms and 100 workers.

Within each such market–year (j, t), I construct two sets of worker ranks and two sets of firm ranks. On the worker side, I form deciles based on: (i) the estimated worker fixed effects $\hat{\alpha}_i$; and (ii) the model-based measure of worker ability. On the firm side, I form deciles based on: (i) the estimated firm-type effects $\hat{\psi}_{g(j)}$; and (ii) a model-based measure of firm productivity. To construct

employment shares, I proceed as follows. For each year t and worker-decile–firm-decile pair (d, f) (using the AKM-based ranks), I count the number of distinct workers in decile d employed by firms in decile f, and denote this count by $N_{df,t}$. I also compute the total number of distinct workers in decile d in year t, $N_{d,t}$. The employment share of decile-d workers employed by decile-f firms in year t is

$$s_{df,t} = \frac{N_{df,t}}{N_{d,t}}.$$

I then average these shares over years to compute employment share matrices, using the AKM rankes as model validation.