Who Competes for Whom?
Monopsony in Ability-Based Segmented Labor Markets

Luca Lorenzini*

UCLA, Anderson School of Management

November 25, 2024

Abstract

This paper investigates how labor market sorting and segmentation influence labor market power, aggregate efficiency, and welfare distribution. Using matched employer-employee data from Germany, I show that low-ability workers are disproportionately employed by smaller, low-productivity firms due to selective hiring by firms offering higher wage premiums. Motivated by this evidence, I develop a model that examines monopsony power driven by heterogeneous firm selection of the workforce. In this framework, more productive firms are larger, apply stricter hiring standards, and selectively employ higher-ability workers, resulting in an ability-based segmentation of the labor market, where not all firms are available as options to every worker. This creates localized competition, with firms primarily competing with others targeting similar workers. Less productive firms exert greater labor market power over lower-ability workers, while more productive firms over higher-ability ones. The model predicts welfare losses of 26% to 53% for workers, especially large for those at the lower end of the ability distribution, who face a restricted choice set as they are excluded by most firms, thereby reducing competition. Entrepreneurs experience welfare gains of 65%. Predicted output loss is 0.1%, significantly lower than the 8% seen in traditional models without labor market segmentation.

*lucalorenzini@ucla.edu.

I am immensely grateful to Hugo Hopenhayn for his guidance, support, and feedback. I am deeply indebted to Michael Rubens, Romain Wacziarg, Nico Voigtländer, Brian Wheaton, and Jonathan Vogel for their invaluable advice and extensive suggestions. I also thank Ariel Burstein, Daniel Haanwinckel, Oleg Itskhoki, Lee E. Ohanian, Bruno Pellegrino, Gianluca Violante, and the workshop participants at the UCLA Anderson GEM Brownbags and the UCLA Macro reading group for their valuable comments. All mistakes are mine. I gratefully acknowledge financial support from the Center for Global Management at UCLA Anderson. This study 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 has previously circulated with the name *Endogenous Oligopsony*. This paper received the Consultaccount Award for Best Paper presented by a PhD student at the 17th PEJ annual meeting.

1 Introduction

The study of labor market power has gained increasing attention in economic research due to its significant implications for wage suppression, job mobility, and overall welfare. The U.S. Treasury's State of Labor Market Competition Report (2022) underscores how employer concentration and anti-competitive practices reduce productivity and exacerbate inequality. This focus has spurred calls for policy measures—such as stronger antitrust enforcement and limiting non-compete agreements—to restore competition, enhance worker bargaining power, and improve economic efficiency[1]. Additionally, wage inequality has emerged as a critical concern, with empirical evidence linking labor market sorting and segmentation[2] as major contributors to rising disparities (Song et al. (2019)). While these features of the labor market are well-documented drivers of wage inequality, the literature has largely neglected to explore how segmentation shapes competition among employers, influencing their labor market power. This paper aims to bridge that gap by investigating the implications of labor market sorting and segmentation for labor market power, aggregate efficiency, and welfare distribution.

Using matched employer-employee data from Germany, I document empirical evidence of what I term *strong sorting* in labor markets: low-ability workers are disproportionately employed by low-productivity firms, despite these firms being relatively smaller in size. I define a local labor market as an occupation-geography group and categorize workers and firms according to their AKM fixed effects rank[3]. I find that low-ability workers are four times more likely to be employed by firms in the bottom decile of the labor market distribution compared to firms in the top decile, despite the larger size of top-decile firms. Conversely, high-ability workers are concentrated in firms within the top decile. This pattern of segmentation extends across both low-skilled occupations (e.g., office clerks) and high-skilled occupations (e.g., chemical engineers). I show that this segmentation arises because high-wage firms, employ more selective hiring practices. Empirical evidence shows that firms offering higher pay premiums within their respective labor markets set significantly higher screening thresholds. These thresholds are defined as the minimum worker ability level—measured by worker fixed effects—required for a job offer. Specifically, a firm in the 10th decile, employing only college-educated workers, has a minimum worker fixed effect that is 1.61 standard deviations in the labor market's worker fixed effect distribution higher compared to firms in the 1st decile employing non-college workers, all else equal. This difference highlights that more productive firms

[1] The Biden administration's executive order on competition in labor markets reflects this policy shift, emphasizing wage growth and economic mobility.

[2] Sorting refers to high-ability workers tending to work in high-wage firms, while segmentation involves high-ability and low-ability workers clustering together in separate firms.

[3] The AKM regression model, named after the initials of its authors Abowd et al. (1999), is widely used in labor economics to estimate fixed effects for both workers and firms. These fixed effects are typically interpreted as representing worker ability and firm productivity, allowing for an analysis of how worker and firm characteristics contribute to wage determination.

adopt stricter hiring criteria in the job market.

The empirical evidence reveals a stark segmentation in labor markets: low-ability workers are overwhelmingly concentrated in low-productivity firms, while high-ability workers cluster within high-productivity firms. Standard models of competition in the labor market often link monopsony power to firm size and productivity. However, the evidence suggests that even small, less productive firms can wield considerable power over low-ability workers, as these firms may be the only options available to them. Motivated by this empirical evidence, I develop a novel, unifying theory of labor market power with ability-based, endogenously segmented labor markets. The model incorporates two-sided heterogeneity, where workers differ in ability and firms in productivity. A key innovation is that the ability-based segmentation arises endogenously in the model: workers' choice sets vary with their abilities, determined by firms' selective hiring decisions. This framework produces a mechanism termed *Endogenous Oligopsony*, where firms selectively hire workers of certain abilities, restricting workers' choice sets and leading firms to specialize in distinct ability segments. This segmentation fosters localized competition, as firms compete more intensely with those targeting similar workers. Intuitively, for low-ability workers, whose choice sets are limited due to firm selectivity, competition is diminished, resulting in greater welfare losses from labor market power. The model reveals that labor market segmentation by worker ability reduces production inefficiencies from oligopsony power, with output losses at only 0.1% compared to the 8% losses observed in a model with homogeneous workers. Intuitively, ability segmentation reduces the correlation between distortions and productivity, thereby mitigating losses from misallocation. Nevertheless, monopsony power leads to significant redistributive effects, disproportionately disadvantaging lower-ability workers who face restricted access to firms and reduced competition, resulting in a welfare loss of 53%.

The model features three agents: workers, firms, and entrepreneurs. There is a continuum of workers, each differing in latent ability[4]. Additionally, there is a continuum of labor markets, each containing a heterogeneous number of firms. Firms within each labor market are granular and differ in their exogenous baseline productivity, which is drawn from a distribution. All firms are owned by a representative entrepreneur, who receives income from firms' profits and capital rents, and makes decisions regarding aggregate capital accumulation.

Worker supply is modeled through *preference heterogeneity*, building on D. Berger et al. (2022a). In this framework, workers choose to work for a firm from their available choice set based on wages and idiosyncratic taste shocks, which capture factors like commuting distance or preferences for a firm's culture. From each worker's perspective, preference heterogeneity implies that jobs are differentiated. Consequently, firms face upward-sloping labor supply curves for each worker type,

[4]This latent ability should be understood as workers within the same labor market (i.e., with similar occupation or education levels) but varying in quality

which they internalize. Firms' employment decisions are modeled through strategic interactions in a Cournot competition framework, where firms compete strategically for workers by setting employment schedules for each ability level, anticipating responses from other employers within a local market. Markdown heterogeneity is modeled following Atkeson and Burstein (2008). Labor market power increases when a firm holds a large market share for workers of a particular type, as it faces a more inelastic labor supply. This occurs because the firm internalizes the lower elasticity of substitution across markets, enabling it to exert greater market power over the workers it disproportionally employs. Optimal wages, therefore, take the form of a firm-worker-specific markdown relative to efficient wages, where the efficient wage equals the worker's marginal product of labor[5].

The key innovation in this paper is the firm's production function, which endogenizes workers' choice sets based on their abilities. A firm's realized productivity[6] becomes endogenous, determined by the *average* output produced by its workforce, where each worker's output depends on their ability and the firm's exogenous productivity. Employing higher-ability workers raises a firm's productivity by increasing the average output of its workforce. The micro-foundation assumes that managers cannot differentiate among employees when allocating shared resources (e.g., office space), which are equally distributed. This endogenous productivity, combined with complementarities, motivates firms to screen workers based on their abilities. Hiring an additional worker of a given quality affects firm production through two channels. First, it increases firm size and thus contributes positively to output. Second, hiring an additional worker of a given quality changes the average worker output, impacting productivity. This second effect may reduce firm production if the worker's output is lower than the current firm average, potentially offsetting the positive impact of the first channel. In such cases, the firm may refuse to hire the worker, even at a one-cent wage. In equilibrium, more productive firms optimize by offering jobs exclusively to higher-ability workers. As a result, workers' choice sets become endogenous, with higher-ability workers receiving more job offers than lower-ability workers.

The model introduces a mechanism denoted *Endogenous Oligopsony*, where firms' selective hiring of workers with varying abilities restricts workers' choice sets, leading firms to specialize in hiring specific segments of the labor force. This segmentation creates localized competition, as firms primarily compete with similarly productive firms targeting the same range of worker abilities. As

[5]Throughout the paper, I refer to $w = MPL$ as the efficient wage, where MPL represents the marginal revenue product of labor. These are efficient wages in the sense that they would generate the Planner's allocation in a decentralized economy. Under labor market power, the wage becomes a markdown over the MPL, expressed as $w = \mu MPL$. Here, μ denotes the firm's endogenous markdown. The closer μ is to one, the smaller the markdown, meaning the wage is closer to the competitive benchmark. Conversely, a lower μ implies a larger markdown, indicating that the firm underpays relative to the competitive level. Generally, μ depends on the firm-level labor supply elasticity ϵ, as given by $w = \frac{\epsilon}{\epsilon+1} MPL$. Thus, a lower labor supply elasticity results in a smaller μ and a greater markdown. It is important to note that larger markdowns do not necessarily imply lower wages, as wages depend on both μ and MPL. A firm could offer a high wage despite applying a substantial markdown if its MPL is sufficiently large.

[6]Or product quality, as these terms are isomorphic in the context of this model.

a result, high-ability workers tend to cluster in high-wage firms (*sorting*), while both high- and low-wage workers are disproportionately concentrated within certain firms (referred to as *segregation* or *segmentation*). Depending on the calibration, this process can lead to an overrepresentation of low-ability workers in low-productivity firms. In such cases, low-productivity firms may appear large from the perspective of low-ability workers and exert greater market power over them, despite being small relative to other firms.

I calibrate the model using a combination of parameters from the literature and internally calibrated values to match the predicted worker-firm market shares to the observed data. I then perform two counterfactual analyses. The first compares the model's outcomes to the efficient allocation, where wages equal the marginal product of labor. The second examines how the results differ relative to a standard model of oligopsonistic competition without labor market ability segmentation (e.g. D. Berger et al. (2022a)), which is nested within this framework.

Relative to the efficient allocation, the model predicts small output losses of 0.1% but substantial redistributive effects. Output losses are mainly stemming from *Size Misallocation*: firms with larger markdowns pay lower wages and, as a result, are under-resourced in terms of total employment relative to the efficient level. Additionally, the segmented model introduces a distortion denoted *Misallocation of Talent*. This occurs because low-productivity firms hold market power over low-ability workers, while high-productivity firms exert power over high-ability workers. Consequently, in the efficient allocation, low-productivity firms employ more low-ability workers, while high-productivity firms hire more high-ability workers. Regarding welfare, workers experience losses in the range of 26% to 53%, while gains of 65% accrue to entrepreneurs. Losses among workers are particularly severe for those at the lower end of the ability distribution, who face restricted choice sets due to exclusion from most firms, thereby reducing competition.

Comparing the model with the standard oligopsonistic model—one without labor market ability segmentation—reveals that neglecting firms' selective hiring leads to an overestimation of production inefficiencies caused by labor market power. In the standard model, where segmentation is absent, inefficiencies are measured at 8%, while with segmentation, they drop to just 0.2%. This reduction occurs because labor market segmentation alters the nature of distortions: low-productivity firms also acquire market power over their workers. Without ability-based segmented labor markets, more productive firms are systematically more distorted and receive less employment, resulting in correlated distortions and big inefficiencies from misallocation. On the other hand, in the model with ability segmentation, both low- and high-productivity firms wield market power. Now, less productive firms are also under-resourced relative to the efficient benchmark. This reduces the correlation between distortions and productivity—the primary driver of output losses from misallocation (Restuccia and Rogerson (2008)).

Building on the model's insights into segmentation and labor market power, this paper also offers significant contributions to the empirical measurement of markdowns. Recent studies use two primary approaches. The first is the *production function* method, which measures the gap between a firm's average marginal product of labor and its average wage to obtain a measure of firm average markdown (Kirov and Traina (2021); Yeh et al. (2022)). The second approach measure market concentration to gauge the strenght of competition in labor markets by measuring payroll HHIs (Herfindahl-Hirschman Index). Notably, while firms' wage-productivity wedges tracked payroll HHI until 2000, they sharply increased thereafter, even as HHI remained stable Yeh et al. (2022). First, this paper extends the production function approach by incorporating unobservable worker heterogeneity. Rather than estimating the average markdowns, these estimates account for the within-firm covariance between markdowns and workers' abilities. I show that this covariance term can be significant, creating a positive correlation between these measurements and firm productivity or size. Nonetheless, the production function method provides a model-consistent measure of the markdowns' implications for firm-level labor share. Second, I demonstrate through numerical examples that the payroll HHI index may not always accurately capture labor market power's effect on labor share. For example, as production function complementarities increase, firms intensify worker selection, leading to a reduction in labor share even as the HHI index remains low.

This paper provides a unified framework for studying labor market power in ability-segmented labor markets. The model introduces the concept of Endogenous Oligopsony, demonstrating how labor market segmentation and firm-level hiring strategies affect both aggregate efficiency and welfare distribution. Through model calibration and counterfactual analyses, I show that while labor market power results in significant redistribution of welfare for workers, its impact on production inefficiencies is relatively modest, especially when compared to models that overlook this type of segmentation. These findings highlight the critical importance of accounting for worker heterogeneity and selective hiring when analyzing labor market power and its broader implications for efficiency and welfare.

Literature- This research contributes to five strands of literature.

Labor Market Power Literature This research aligns with the labor market power literature by introducing worker and preference heterogeneity, complementarities, workforce selection, and oligopsonistic firm behavior into a general equilibrium framework, deriving implications for production efficiency and welfare. The two closest papers are D. Berger et al. (2022a) and Lamadon et al. (2022). D. Berger et al. (2022a) developed a model featuring preference heterogeneity, homogeneous workers, and strategic interaction, while Lamadon et al. (2022) quantified imperfect competition using matched employer-employee data, showing pervasive segmentation in the US

labor market. My contribution is a unified framework that incorporates strategic interactions, accounts for segmentation and sorting patterns, and matches statistics on selection in the labor market. This model rationalizes segmentation and sorting with a parsimonious parameterization, maintaining firm oligopsonistic behavior. Research in imperfect competition in the labor market dates back to Robinson (1933). Traditional benchmark models of monopsony are described in Burdett and Mortensen (1998) and Manning (2003). Recently, labor market power has been model primarily using the preference heterogeneity approach. D. W. Berger, Herkenhoff, and Mongey (2022b) explores the consequences of the minimum wage in that economy, while other works, such as D. W. Berger, Herkenhoff, Kostøl, and Mongey (2023) study preference heterogeneity with search frictions. Sharma (2023), investigate labor market power with heterogeneous preferences and its role in determining the gender wage gap. Additionally, Felix (2021) estimates the effect of trade on local labor market concentration. Azar et al. (2022) provide evidence of substantial job differentiation and argue for a significant role for monopsony power.

Additionally, my paper contributes to the literature on markdown estimation using a production function approach, extending the methodology with workers' unobserved heterogeneity. This approach allows the estimation of firms' average markdown, accounting for the covariance term that can be disentangled using the quantified model. Yeh et al. (2022) and Kirov and Traina (2021) estimate labor market power using a production function approach in the spirit of De Loecker et al. (2020).

Wage Inequality In the empirical part of the paper I provided new empirical regularities about segmentation and selection in the labor market. These findings are crucial for understanding wage inequality and labor market power. High wage premiums in high-wage firms trickle down only to high-ability workers, who already earn higher wages, thus amplifying wage inequality. Regarding labor market power, even small firms may have significant power from a worker's perspective, indicating that labor market power is widespread across both large and small firms. My model can rationalize these empirical regularities along with many other patterns extensively documented in the literature. Regularities include firm wage-setting power (Card (2022)), sorting and segmentation in the labor market(Song et al. (2019), Card, Heining, and Kline (2013)), finite firm-level labor supply elasticities (Dal Bó et al. (2013)), and profitability shock rent-sharing elasticities below unity (Kline et al. (2019)). See Card, Cardoso, et al. (2018) for an extensive review.

Preference Heterogeneity Literature The paper relates to the literature employing the framework of preference heterogeneity. This research contributes by introducing the selectivity of the workforce in general equilibrium, addressing a recognized gap in this literature (Card (2022), Manning (2021)). This research innovates by modeling segmentation in the labor market within a general

equilibrium framework, linking this literature to the literature on assortative matching (Shimer and Smith (2000), Eeckhout and Kircher (2018)). Card, Cardoso, et al. (2018) is the first to introduce this framework to explain empirical results challenging the competitive approach to modeling the labor market. Sorkin (2018) estimates workers' preferences for firms and compensating differentials using U.S. data, while Haanwinckel (2023) builds a task-based general equilibrium model with an imperfectly competitive labor market model using the preference heterogeneity approach.

Misallocation Literature This research contributes to the literature on distortions and misallocation by introducing a novel theory of oligopsonistic competition in the labor market, offering insights into the endogenous formation of distortions in a general equilibrium setting. Quantitative assessments of resource misallocation have been conducted by Restuccia and Rogerson (2008), while Hsieh and Klenow (2009) provides a methodology for estimating the impact of distortions on aggregate productivity. De Loecker et al. (2020) offers a methodology to estimate markups from firm-level data, and David, Hopenhayn, and Venkateswaran (2016) presents a theory of imperfect information and resource misallocation. David and Venkateswaran (2019) provides a unified framework for disentangling sources of capital misallocation. Pellegrino (2019) develops a theory of oligopoly and markups in general equilibrium.

Declining Labor Share Literature This research contributes to the literature on the fall of the labor share (Karabarbounis and Neiman (2014), D. Autor et al. (2020)). The contribution is a new explanation for the puzzle that links several trends in wage inequality, documented across various countries, to the fall in the labor share of GDP. The model successfully rationalizes a set of empirical regularities on wage inequality, sorting, and segregation within a unified general equilibrium framework. This provides a potential link between recent empirical trends in wage inequality and the observed decline in the labor share. Consequently, the model may be used as a valuable tool for elucidating the puzzle of the declining labor share witnessed in many developed economies.

2 Motivating Evidence

2.1 Conceptual Framework

In the class of models of oligopsonistic competition, such as D. Berger et al. (2022a), monopsony power is positively related to a firm's market share for a specific category of worker types. This is because firms weigh two elasticities of substitution—within and between markets—and assign greater weight to the lower between-market elasticity of substitution if they are larger, as described in Atkeson and Burstein (2008). Larger firms are more productive, thereby possessing greater labor market power, which implies that they are characterized by more significant distortions. These distortions, when correlated with productivity, are particularly detrimental to aggregate efficiency, leading to substantial inefficiencies and output losses. Moreover, this has strong implications for policy. Consider the positive effects of a minimum wage policy that reduces monopsony power, raising wages and employment, as discussed in Robinson (1933). In the context of these oligopsonistic models, policies such as the minimum wage have little direct effect in removing monopsony power (D. W. Berger, Herkenhoff, and Mongey (2022b)). The reason is that low-productivity firms, for which the minimum wage is binding, are those characterized by lower labor market power. Hence, the minimum wage is a poor tool for counteracting the inefficiencies created by monopsony power.

It is now well recognized that labor markets are characterized by positive assortative matching. Bender et al. (2018) finds that better-managed firms, when hiring or firing, are more likely to hire high-type workers and fire low-type workers. Lamadon et al. (2022) shows that high-type firms have a within-firm distribution of worker types that is more tilted towards high-type workers. Is assortative matching sufficient to alter the predictions of a classical model of oligopsonistic competition in assessing the impact of labor market power on aggregate outcomes? The answer is no. This evidence is not sufficient to change the prediction of an oligopsonistic competition model: if high-type firms are larger (i.e., hiring significantly more workers), it may still be the case that better firms dominate in comparison to the market option available to a lower-type worker.

To clarify this concept, it is useful to provide two definitions:

Definition 1 (Weak Sorting) *Weak sorting occurs when higher-ability workers are disproportionately more likely to be employed by higher-quality firms, but lower-ability workers are still more likely to be employed by better firms than by worse firms.*

This phenomenon arises because, despite potential complementarities, higher-quality firms typically have larger capacities, allowing them to employ a broader range of worker abilities.

Definition 2 (Strong Sorting) *Strong sorting occurs when low-ability workers are primarily employed by worse firms, while high-ability workers are predominantly found in better firms.*

Anticipating the empirical evidence and the model, this pattern emerges as a result of selection processes within the labor market: better firms have more rigorous selection criteria for the workers to whom they extend job offers.

Determining which of the two definitions better describes labor markets is an empirical question that I address in this section. Anticipating the results, I find that labor markets are characterized by strong sorting, which has significant implications for aggregate outcomes, particularly in terms of resource allocation and welfare distribution.

2.2 Data

The primary dataset used in this study is the Sample of Integrated Employer-Employee Data (SIEED), provided by the German Institute for Employment Research (IAB). The SIEED contains a representative 1.5% sample of all German establishments, covering the complete employment biographies of individuals, including periods when they are not employed by the sampled establishments. Establishments are categorized by ownership type, 2-digit industry, and geographic location (141 geographical labor markets). Worker data include total earnings, days worked annually at each job, and additional information on education, occupation, industry, employment status (part-time or full-time), age, nationality, sex, and job skill requirements. The dataset also includes establishment and worker fixed effects provided by the IAB, estimated on the full administrative sample following Card, Heining, and Kline (2013)[7]. The AKM fixed effects are estimated over five overlapping periods. Accordingly, I divide the sample into the same five periods: 1985–1992, 1993–1999, 1998–2004, 2003–2010, and 2010–2017.

From this point onward, a local labor market is defined as a combination of a geographical region and an occupation group. The preferred unit of analysis is a *division*, defined by establishment ID and occupation group. Results will also be reported using the establishment ID as the unit of analysis. Under the preferred definition, if a firm employs both janitors and engineers, these are treated as two distinct units of observation, each operating in separate labor markets. For clarity, the term *firm* will refer to the unit of observation, whether it is an establishment or a division, depending on the specific part of the analysis.

For the purposes of this research, an annual panel dataset is derived from the original spell-level data following a data cleaning process similar to Card, Heining, and Kline (2013). The

[7] See Lochner et al. (n.d.) for more details.

dataset undergoes several restrictions to refine the sample. Initially, I focus on individuals aged 20-60 who are fully employed in establishments located in West Germany, with real daily wages of at least 10 euros. Because the dataset does not include hours of work, I also exclude vocational training jobs, home workers/freelance, and part-time jobs. Limiting attention to full-time jobs reduces the impact of hours dispersion that could confound the analysis. In cases where overlapping employment spells are present, the spell associated with the highest-paying job is retained as the primary episode. Wages, which are recorded on a daily basis, have been deflated to 2015 euros. Observations subjected to top-coding are adjusted using standard imputation techniques. The final dataset spans the years 1985 to 2017 and includes 3'473'220 unique division IDs and 2'525'434 unique individuals. Detailed summary statistics can be found in Appendix A.

The firm-level descriptive statistics indicate that the mean employment (after the sample reduction) of a division is 5.5, with a standard deviation of 0.95 in the log-employment distribution. When firms are defined by establishment IDs, both the mean and standard deviation increase: the mean employment rises to 13.8, and the standard deviation in the log-employment distribution reaches 1.28. Regarding wages, the average occupation-firm pays a mean log wage of 4.5, which corresponds to a daily wage of approximately €90. The standard deviation in the log wage distribution is about 0.48.

On the worker side, the average age of workers in the sample is 42 years. The average years of schooling for workers is roughly 12 years. Additionally, the average worker tenure (experience) is roughly 23 years.

As a baseline, wage data are residualized with respect to a polynomial in age and job tenure, fully interacted with educational attainment and year, occupation, and nationality fixed effects. This approach anticipates that the structural model will not account for life-cycle effects, on-the-job learning, aggregate productivity growth, or any other worker heterogeneity besides workers' abilities. Furthermore, since the structural model focuses on the unobservable component of worker heterogeneity, log wages are also residualized with respect to educational attainment. This residualization process helps clarify how much variance in log wages can be attributed to raw wages, isolating the portion of the wage structure relevant to the model's predictions. The standard deviation of worker-level log wage is 0.48, while the residualized log wages have a standard deviation of 0.37. Thus, the bulk of the variation in wages survive even when wages are residualized based on observables.

Identifying types:

With this background, I now turn to the econometric framework for disentangling worker-specific and employer-specific heterogeneity. In a given time interval, the dataset contains N person-year observations on N workers and J establishments. The function $J(i,t)$ gives the identity of the unique establishment that employs worker i in year t. I assume that the log daily real wage y_{it} of individual i in year t is the sum of a worker component α_i, an establishment component $\psi_{J(i,t)}$, an index of observable characteristics x'_{it}, and an error component ϵ_{it}:

$$y_{it} = \alpha_i + \psi_{J(i,t)} + x'_{it}\beta + \epsilon_{it} \quad (1)$$

Following Abowd et al. (1999), I interpret and refer interchangeably to the person effect α_i as the worker's ability, encompassing skills and other factors rewarded equally across employers. Similarly, I interpret x'_{it} as a combination of life cycle and aggregate factors that affect worker i's productivity at all jobs. The variable x_{it} includes an unrestricted set of year dummies, quadratic and cubic age terms fully interacted with educational attainment, and occupation fixed effects. I interpret the establishment effect ψ_j as a proportional pay premium (or discount) paid by establishment j to all employees (those with $J(i,t) = j$). This premium, often interpreted in the literature as firm productivity Card, Cardoso, et al. (2018), is used throughout the analysis.

For the main analysis, I use the worker and establishment fixed effects estimated on the full administrative sample provided by the IAB. As a robustness check, I also compute the AKM fixed effect regression using the division ID based on the intersection of establishment and occupation, as described above. The first approach results in a 1.6% loss of occupation-firm groups and 3% of workers, while the second approach results in a 29% loss of establishments and 6.7% of workers. Since results remain qualitatively similar, I opt to use the establishment AKM fixed effect as the division fixed effect to avoid greater data loss from a more restricted connected set[8].

To derive a measure of worker and firm quality relevant to labor market competition, I rank workers and firms within their respective occupation-geography-year fixed effect distributions. This approach ensures that the measure of quality is comparable within each distinct labor market, focusing on the competition among firms and workers in the same category. By segmenting workers and firms within these specific labor markets, I obtain a measure of quality that reflects competition more accurately than a cross-occupation comparison. For instance, comparing the pay premiums of firms hiring janitors to those employing electrical engineers would not reveal meaningful information about competition because these groups operate in separate labor markets. Furthermore, antici-

[8]This is consistent with a productivity that is firm specific and that sills over to different divisions within the firm.

pating the model developed in subsequent sections, this within local labor market ranking offers a model-consistent method to identify the ranking of worker and firm types. Using simulated data from the model in the following sections, I demonstrate that, although the model does not assume a log-additive wage structure, AKM fixed effects closely align with the model-derived rankings of firm and worker types within each local labor market (see Appendix C.4). This alignment suggests that the AKM framework effectively captures the ordering of firm and worker types as specified by the model, even in the presence of a non-log-additive wage structure.

2.3 Size Premium

I begin the empirical analysis by examining the relationship between firm characteristics and firm size. Specifically, I analyze the association between firm pay premiums and firm size premiums. This approach serves two main objectives. First, it provides evidence for interpreting firm pay premiums as indicative of firm productivity, assuming that more productive firms are also larger. Second, it serves as a motivation of a wage-posting model in the spirit of Card, Cardoso, et al. (2018), which predict that more productive firms attract larger workforces by offering higher wages.

To investigate these relationships, I conduct a series of firm-level regressions that examine the relationship between a division's decile within the occupation-geography market and its corresponding size. Given that I can accurately infer total employment only for establishments within the original IAB subsample, I limit the analysis to these establishments. I employ two measures of firm size to enhance robustness and capture different dimensions of a firm's scale. The first measure is the log of employment based on the division ID, which is defined at the occupation-establishment level. The second measure considers total employment at the establishment level, without differentiating by occupation. In this case, to compute the firm decile, I assign each firm to the local labor market based on the predominant occupation within its workforce.

To assess the relationship between firm deciles and size, I estimate several regressions using each size measure as the dependent variable. Each model also includes a range of controls to ensure robustness and account for confounding factors:

$$\text{log_employment}_{it} = \beta_0 + \beta_1 \text{Firm Decile}_{it} + \beta_x X_{it} + \gamma_m + \gamma_t + \epsilon_{it}, \tag{2.1}$$

where $\text{log_employment}_{it}$ represents the firm size measure for firm i at time t, Firm Decile_{it} denotes the firm's position in the decile distribution within the occupation-geography market, and X_{it} includes controls such as workforce composition variables and average firm age. Year fixed effects (γ_t) and market fixed effects (γ_m) are included to account for temporal and market-specific

variations.

Additionally, to capture potential nonlinearities in the relationship between firm decile and firm size, I estimate an alternative specification with fixed effects for each firm decile:

$$\text{log_employment}_{it} = \beta_0 + \text{Firm Decile Fixed Effects} + \beta_x X_{it} + \gamma_m + \gamma_t + \epsilon_{it}. \tag{2.2}$$

In this specification, `Firm Decile Fixed Effects` represent distinct fixed effects for each decile rather than a linear trend. This flexible approach allows each firm decile to influence size independently, capturing a more flexible relationship between size and firm pay premiums.

Results from these regressions are reported in Table 1. Figure 1 illustrates the estimated firm decile fixed effects, showing the decile-specific size premiums across various year groups. By comparing firm deciles' effects on both size measures, these results provide insights into how firm size premiums are associated with firm pay premiums and productivity. The positive relationship found across specifications aligns with wage-posting model predictions, suggesting that higher-decile firms indeed attract more workers, likely due to offering competitive wages in line with their productivity levels.

The empirical results presented in Table 1 indicate a positive relationship between a firm's decile within its occupation-geography market and its employment size. Both establishments and divisions in higher deciles, associated with higher pay premiums, consistently show larger employment sizes. This finding is robust across various specifications, with the size premium remaining stable even after accounting for firm-level characteristics. Moreover, the share of college-educated workers and the proportion of low-skill jobs are both positively associated with firm size. Interpreting the results from the preferred specification in column 4, I find that divisions in the 10th decile of the within-labor-market firm fixed effect distribution are, on average, 53% larger in terms of employment size compared to divisions in the 1st decile, conditional on observables[9]. Furthermore, divisions employing a higher share of college-educated workers and engaging in low-skill tasks are also larger in size. A division entirely composed of college-educated workers performing low-skill tasks is, on average, 17% larger in size. These results suggest that firms offering higher wages attract larger pools of workers, supporting the interpretation of firm pay premiums as proxies for productivity:

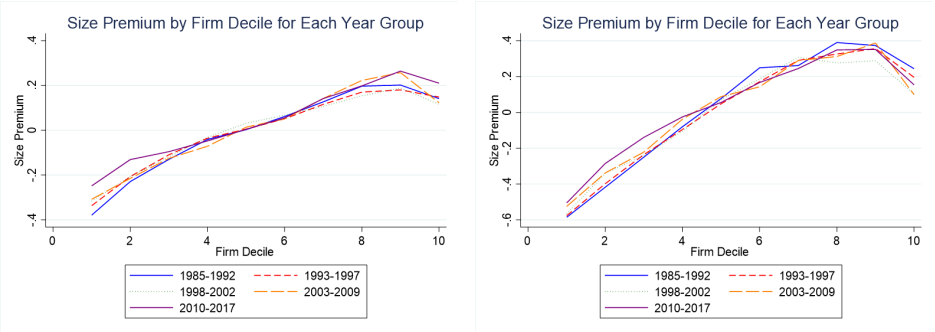
Fact 1. Firm Deciles and Employment Size Premiums. Higher firm deciles within occupation-geography markets are associated with larger employment sizes.

[9]From the point estimates of the 10th decile fixed effects in both Panel A and Panel B, it appears that the pay premium-size relationship is reversing for top deciles. Although the point estimates suggest this trend, the difference between the point estimates of decile 9 and 10 is not statistically significant.

Table 1: Regression Results on Log Employment by Firm Deciles (2010–2017)

	y = **Log Employment, division ID**				y = **Log Employment, establishment ID**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Firm Decile	0.0606***	0.0614***	0.0579***	—	0.1153***	0.1154***	0.0965***	—
	(0.0019)	(0.0019)	(0.0019)	—	(0.0048)	(0.0048)	(0.0046)	—
Share College	—	—	0.0663***	0.0676***	—	—	0.0816	0.0816
	—	—	(0.0220)	(0.0219)	—	—	(0.0544)	(0.0544)
Share Low-Skill Task	—	—	0.1011***	0.1004***	—	—	-0.2733***	-0.2761***
	—	—	(0.0188)	(0.0188)	—	—	(0.0544)	(0.0542)
X: **Additional Controls**	No	No	Yes	Yes	No	No	Yes	Yes
f(age), avg age	No	No	Yes	Yes	No	No	Yes	Yes
Fixed Effects								
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Market	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Firm Decile FE	No	No	No	Yes	No	No	No	Yes
Observations	161,331	161,331	161,331	161,331	31,549	31,549	31,549	31,549
R-squared	0.0325	0.0437	0.1289	0.2445	0.0792	0.1331	0.2445	0.2445

Notes: This table reports regression results examining the relationship between firm deciles and standardized log employment (2010–2017) using two distinct dependent variables: log employment with division ID as firm identifier (Columns 1-4) and log employment with establishment ID as firm identifier (Columns 5-8). Additional controls include share of college-educated workers, share of low-skill tasks, and firm average age. Fixed effects for year, market, and firm deciles are included as indicated. Robust standard errors are in parentheses and are clustered by occupation-firm ID or establishment ID as appropriate. Polynomial terms in firm age are included but not reported.
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.



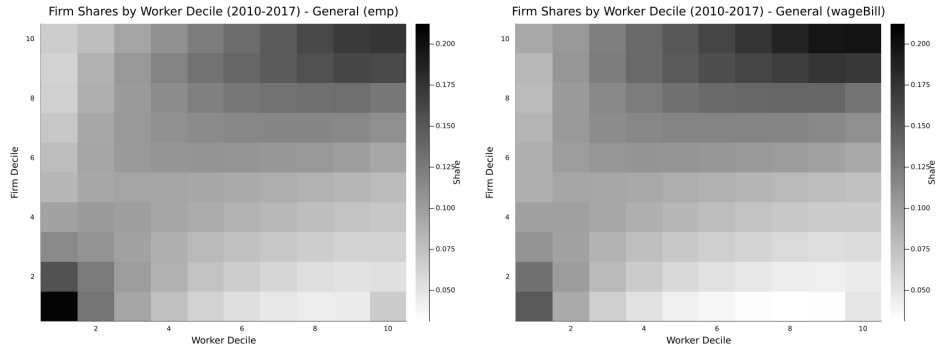
(a) Log employment by firm decile (division ID) (b) Log employment by firm decile (Establishment ID)

Figure 1: Size premium by Firm Decile

Notes: This figure is divided into two panels, labeled A and B. Panel A plots the Size Premium when the dependent variable is the log employment of a division. Panel B plots the Size Premium when the dependent variable is the log employment of an establishment. The x-axis represents firm decile, and the y-axis reports the estimated fixed effect.

2.4 Strong Sorting

For each year, I calculate the employment and wage bill share of workers from a specific decile of their within-year occupation geography fixed effect distribution who are employed in firms corresponding to a particular decile. The categorization is performed on a year-by-year basis. Figure 2 visualizes these shares, with worker deciles on the x-axis and firm deciles on the y-axis. The shading intensity indicates the concentration of workers in firms, with darker shades representing higher employment or wage bill shares. To construct Figure 2, I averaged the market shares over each sample period. Accordingly, the figure displays the market shares for the last period, years 2010-2017. If either worker or firm types are miscategorized, wage bill shares respond more sensitively. For example, if a high-type worker is mistakenly classified as low-type but employed by a high-type firm, their higher wage would disproportionately inflate the share of low-type workers in high-type firms. In contrast, employment shares are less affected by such misclassification, as a single error does not strongly impact results. For this reason, employment shares are preferred for analysis; however, wage bill shares are also reported, and the results are always very similar.



(a) Employment Shares

(b) Wage Bill Shares

Figure 2: Employment and Wage Bill shares

Notes: This figure is divided into two panels, labeled A and B. Panel A plots the employment shares while panel B the wage bill shares of workers categorized by deciles of their within-year occupation geography fixed effect distribution, employed in firms that are also categorized by deciles of their within-year occupation geography AKM fixed effect distribution. The x-axis represents worker deciles, and the y-axis represents firm deciles. The shading intensity indicates the concentration of workers in firms, with darker shades representing higher employment or wage bill shares.

If all workers were employed in equal shares across all firms, the share plot would show uniform intensity at a share of 0.1. If more productive firms are relatively large across all worker types, the darkest regions—indicating high market shares—would concentrate at the top of the graph for all worker deciles. Under *weak sorting*, the plot would show darker shades at the top for both high-

and low-decile workers, though top-decile workers would have the darkest shading. This type of assortative matching would produce macro-implications similar to those of a standard oligopsonistic model.

In contrast, the plot reveals that low-decile workers are disproportionately employed in low-decile firms, despite their smaller size, while high-decile workers are concentrated in high-decile firms. The shift occurs gradually, with higher decile workers increasingly employed in higher decile firms at the expense of low decile firms. This is the central piece of evidence motivating the general equilibrium model in the next section. Figure 2 suggests a type of competition that is localized, where low-productive firms compete with each other for low-ability workers, while high-productive firms compete with each other for high-ability workers. Thus, competition appears to be local in the sense that firms compete primarily with their productivity-level neighbors, as they effectively hire from different segments of the labor market based on worker ability.

Fact 2: The German labor market exhibits strong sorting: low-ability workers are predominantly employed by low-productivity firms, while high-ability workers are mainly employed by higher-productivity firms.

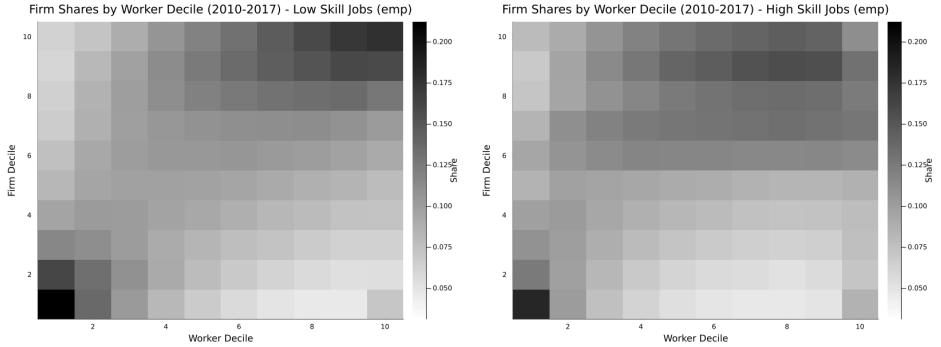
One may argue that this type of assortative matching is not relevant for the macroeconomy if it is confined to high-skill job markets. This consideration is also pertinent for policy, as high-skill labor markets may not be affected by policies such as the minimum wage. Therefore, it is important to examine heterogeneity by skill level to assess the broader applicability of these findings. To do this, I use the variable *niveau* from the dataset, which categorizes jobs into four skill levels: Unskilled/Semiskilled Task, Skilled Task, Complex Task, and Highly Complex Task[10]. I classify low-skill jobs as those in the Unskilled/Semiskilled and Skilled categories, and high-skill jobs as those in the Complex and Highly Complex categories.[11]

I then split the sample into high- and low-skill groups, re-categorizing firms and workers within their occupation-year deciles for each group. Panel A of Figure 3 displays market shares of worker deciles by firm deciles for low-skill jobs, while Panel B shows the same for high-skill jobs, averaged over the last sample period as before. It is evident that strong sorting is not confined to high-skill jobs; if anything, low-skill jobs exhibit even stronger assortative matching. Figure 3 closely resembles Figure 2, indicating similar localized competition.

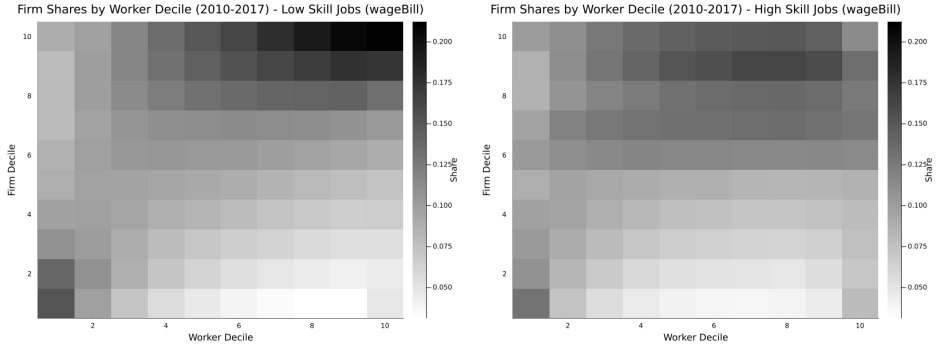
[10]The skill level required for a job is recorded in the fifth digit of the KldB2010 classification code, submitted by employers as part of the "Classification of Occupations 2010" (KldB2010). Employers select the primary job title for each employee based on the main activity performed.

[11]In the dataset, the five most common occupations in the low-skill category are: Drivers in road traffic; Warehousing, logistics, postal, and delivery services; Office clerks and secretaries; Building construction; and Machine operation (non-specialized). For high-skill jobs, the five most frequent occupations are: Business organization and strategy; Technical research and development; Purchasing and sales; Electrical engineering; and Computer science. Note that economics-related occupations are categorized as Highly Complex.

Fact 3: Both German low-skill and high-skill labor markets exhibit strong sorting.



(a) Employment Shares - Low-skill subsample (b) Employment Shares - High-skill subsample



(c) Wage Bill Shares - Low-skill subsample (d) Wage Bill Shares - High-skill subsample

Figure 3: Comparison of Employment and Wage Bill Shares for Low-skill and High-skill Subsamples

Notes: This figure is divided into four panels. Panels (a) and (b) show the employment shares for low-skill and high-skill subsamples, respectively. Panels (c) and (d) present the wage bill shares for the low-skill and high-skill subsamples, respectively. In each panel, workers are categorized by deciles of their within-year occupation geography fixed effect distribution and are employed in firms categorized by deciles of their within-year occupation geography AKM fixed effect distribution. The x-axis represents worker deciles, and the y-axis represents firm deciles. Shading intensity indicates the concentration of workers (or wage bill shares) in firms, with darker shades representing higher values.

Has strong sorting always characterized the labor market? Is there an observable trend? Answering these questions is important because the type of localized competition discussed above may help explain the puzzle of the declining labor share (D. Autor et al. (2020)). Furthermore, this trend is significant because it appears closely connected with the observed rise in wage inequality, shown in the appendix (Table A.5 and Table A.8). If there is increasing dispersion in pay premiums, with high-productivity firms offering larger pay premiums, and if strong sorting intensifies—perhaps due to firms becoming more selective in their workforce over time—then the higher pay premiums

of these firms would primarily benefit high-ability workers, whose wages are already higher. This dynamic would, in turn, amplify wage inequality.

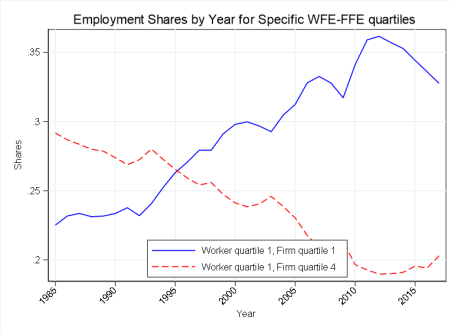
To explore this question, I present the time series of employment shares discussed above, categorizing both workers and firms by quartile rather than by decile. Figure 4 is divided into four panels. Panel A displays the employment share of workers in the first quartile of the within-occupation-geography-year fixed effect distribution, employed by firms in either the first or fourth quartile. First, categorizing by quartile instead of decile does not alter the results of Figure 2; after 1995, labor markets are characterized by strong sorting. Additionally, the figure reveals that strong sorting was not always prevalent: prior to 1995, low-type workers (those in the first quartile of the within-occupation-year fixed effect distribution) were disproportionately employed by high-type firms.

Panel B shows the employment share of workers in the fourth quartile of the within-occupation-year fixed effect distribution for firms in either the first or fourth quartile. Here, high-type workers are disproportionately employed by high-type firms. Panels C and D present the employment shares for the subsamples of low-skill and high-skill jobs, respectively, for firms in either the first or fourth quartile.

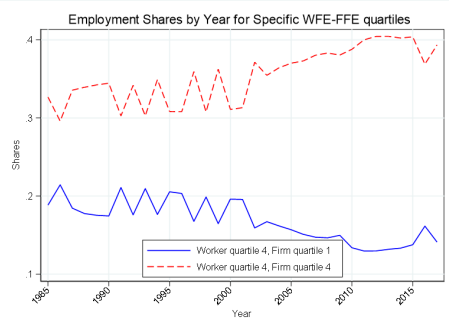
All four panels reveal the same trend: starting in 1992, assortative matching increased over time, peaking around 2011 before a slight reversal. This trend mirrors the patterns observed in wage inequality and firm screening (see subsection 2.5), suggesting that these phenomena share a common underlying cause[12].

Fact 4: Assortative matching in Germany has increased over time, reaching a peak in 2011 and reversing thereafter.

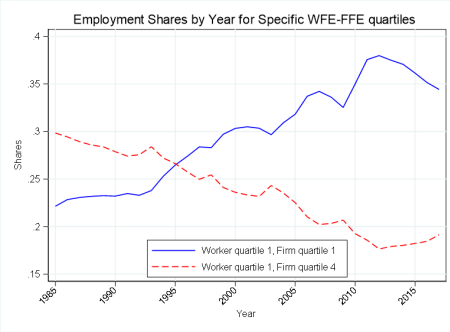
[12]This paper accepts this trend as a given, focusing instead on assessing the implications of segmentation for labor market power. Investigating the determinants of this trend in wage inequality is left as a promising avenue for future research.



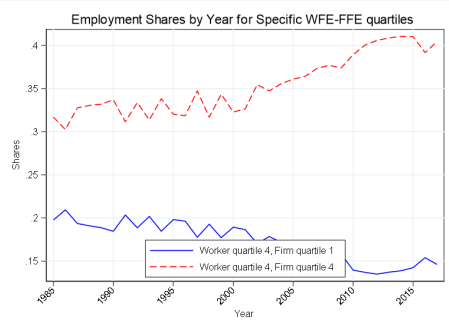
(a) Employment share of low-type workers



(b) Employment share of high-type workers



(c) Subsample of low-skill jobs: low-type workers



(d) Subsample of low-skill jobs: high-type workers:

Figure 4: Trend of employment shares

Notes: This figure is divided into four panels, labeled A, B, C, and D. Panel A illustrates the time trend of the employment shares of workers in the first quartile for firms in either the first or fourth quartile. Panel B displays the employment share of workers in the fourth quartile for firms belonging to either the first or fourth quartile. Panel C illustrates the time trend of the employment shares of workers in the first quartile for firms in either the first or fourth quartile when subsampling only low-skill jobs. Panel D illustrates the time trend of the employment shares of workers in the fourth quartile for firms in either the first or fourth quartile when subsampling only low-skill jobs.

2.5 Screening Thresholds

If certain firms offer higher wage premiums, why don't we observe a disproportionate concentration of workers in those firms? In this subsection, I argue that firms set screening thresholds when hiring in the job market. Specifically, I demonstrate that, within a local labor market, firms with higher fixed effects, a larger share of college-educated workers, and a lower share of low-skill jobs implement higher screening thresholds than firms with lower fixed effects, a smaller share of college-educated workers, and a higher share of low-skill jobs.

To investigate this, I restrict the analysis to firms and workers actively engaged in the job

market, defining these workers as those transitioning between establishments. From 1985 to 2017 there are 1'884'573 unique workers and 2'331'033 unique divisions in the job market. I then define a firm's screening threshold as the minimum worker fixed effect among its new hires from the job market, denoted by $\widetilde{a}_{it} = \min\{\alpha \mid \alpha \in \text{new firm hires}\}$, where α represents the worker's fixed effect. I also use an alternative definition of the screening threshold as the average worker fixed effect among new hires, given by $\hat{a}_{it} = \frac{1}{H_{it}} \sum_{j \in H_{it}} \alpha_j$, where H_{it} is the total number of new hires by firm i. This approach is similar to the measure in Carrillo-Tudela et al. (2023). Results are virtually identical, and the first definition is preferred, as the average is less directly informative about a specific screening level. Finally, to compute the dependent variable, I standardize the screening threshold by subtracting the local labor market mean and dividing by the standard deviation of the worker fixed effect distribution. Standardizing enables comparability across markets and reflects a firm's selectivity relative to the local labor market distribution of worker abilities.

To investigate the relationship between firm deciles and screening thresholds, I estimate two empirical models using the two measures of screening thresholds, standardized within each local labor market.

The first model controls for firm decile linearly, capturing a straightforward relationship between the decile position of the firm and its screening threshold:

$$\text{Screening Threshold}_{it} = \beta_0 + \beta_1 \text{Firm Decile}_{it} + \beta_x X_{it} + \gamma_m + \gamma_t + \epsilon_{it}. \tag{2.3}$$

In these equations, $\texttt{Screening Threshold}_{it}$ represents the ability threshold for new hires in firm i at time t, while $\texttt{Firm Decile}_{it}$ denotes the firm's position in the firm decile distribution. The additional firm-level controls, captured by X_{jt}, include the share of college-educated workers, the share of low-skill tasks within the firm, the average age of workers, and the log of new hires. Year fixed effects (γ_t) and market fixed effects (γ_m) are included to control for time and local labor market heterogeneity. These controls are added gradually across the different specifications to assess their impact on the relationship between firm decile and screening thresholds. The share of college-educated workers and low-skill tasks captures variations in workforce composition, as firms with higher shares of college-educated workers may have a higher screening threshold due to differences in job requirements or applicant pools. The average age of workers is included to control for potential productivity effects associated with experience, while the log of new hires controls for firm growth or hiring intensity, which could influence hiring selectivity as found by Carrillo-Tudela et al. (2023). By including these controls, the model aims to isolate the effect of firm decile on screening thresholds from other firm characteristics that could confound the results. The second model allows for a more flexible specification by incorporating a separate fixed effect for each firm

decile, capturing any unobserved heterogeneity across deciles that may not be captured by a linear trend:

$$\text{Screening Threshold}_{it} = \beta_0 + \text{Firm Decile} + \beta_x X_{it} + \gamma_m + \gamma_t + \epsilon_{it}. \tag{2.4}$$

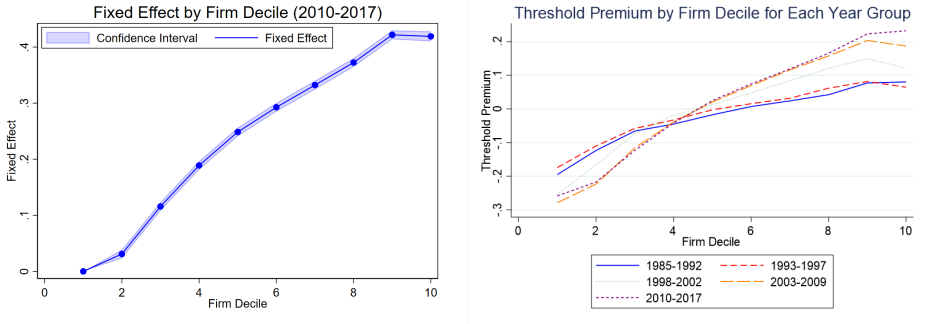
In this specification, `Firm Decile` represents a fixed effect for each decile rather than a linear control. This allows each firm decile to have its own unique effect on the screening threshold, accounting for any nonlinearities in the relationship between firm decile and the screening level.

The results are presented in Table 2. Figure 5 reports the estimated firm decile fixed effect from the specification of Equation 2.4, estimated with the Screening Threshold measure $\widetilde{a}_{it}$. In the first two columns for each threshold type, only year fixed effects are included. The second set of regressions adds market fixed effects, while the last columns introduce further controls for firm composition and age characteristics, including the log of new hires. Notably, the coefficient on `Firm Decile` remains stable when additional controls are added in the linear model, indicating a robust relationship between firm decile and screening thresholds. In the fixed-effect model, each decile is allowed a distinct influence, capturing a more flexible relationship across firm types within the local labor market.

Table 2: Screening Threshold and Firm Deciles: Regression Results (2010–2017)

	$\widetilde{a}_{it}$					$\hat{a}_{it}$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Firm Decile	0.0776***	0.0778***	0.0515***	0.0538***	–	0.0757***	0.0757***	0.0502***	0.0503***	–
	(0.0003)	(0.0003)	(0.0003)	(0.0003)		(0.0003)	(0.0003)	(0.0003)	(0.0003)	
Share College			0.7180***	0.7256***	0.7281***			0.7405***	0.7409***	0.7434***
			(0.0041)	(0.0039)	(.0045)			(0.0038)	(0.0038)	(.0045)
Share Low-Skill Task			-0.4461***	-0.4181***	-0.4183***			-0.4259***	-0.4245***	-0.4247***
			(0.0037)	(0.0035)	(.0039)			(0.0034)	(0.0034)	(0.0039)
Log(New Hires)				-0.5649***	-0.5639***				-0.0281***	-0.0271
				(0.0017)	(0.002)				(0.0016)	(.0015)
X: **Additional Controls**	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
f(age), avg age										
Fixed Effects										
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Market	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Firm Decile	No	No	No	No	Yes	No	No	No	No	Yes
Observations	1,308,987	1,308,987	1,308,987	1,308,987		1,308,987	1,308,987	1,308,987	1,308,987	
R-squared	0.0458	0.0552	0.1154	0.1866		0.0499	0.0583	0.1262	0.1264	

Notes: This table reports regression results examining the relationship between firm deciles and two measures of screening thresholds for the period 2010–2017: $\widetilde{a}_{it} = \min\{\alpha \mid \alpha \in \text{new firm hires}\}$ (Columns 1-4) and $\hat{a}_{it} = \frac{1}{H_{it}} \sum_{j \in H_{it}} \alpha_j$, where H_{it} is the total number of new hires by firm i (Columns 6-9). Each threshold measure is standardized within each local labor market. Controls include the share of college-educated workers, share of low-skill tasks, and log of new hires. Additional controls include a polynomial in age *f(age)* and firm workforce average age *avg age*. Fixed effects for year, market, and firm decile are included as indicated. Robust standard errors are in parentheses and are clustered by firm ID.
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.



(a) 2010–2017 (b) All Time Periods

Figure 5: Estimated Firm Decile Fixed Effects Across Time Periods

Notes: This figure is divided into two panels, labeled A and B. Panel A plots the estimated firm decile fixed effects for the 2010–2017 period, while Panel B shows these effects across all time periods. Both panels use the Screening Threshold measure $\widetilde{a}_{it}$ from Equation 2.4, with controls for year, market effects, and other firm characteristics. Darker shades represent stronger fixed effects, with more detail provided in Table 2.

The regression results indicate a positive and significant relationship between firm decile and

screening thresholds. Across all specifications, higher firm deciles are associated with higher screening thresholds, suggesting that higher-ranked firms set stricter hiring standards. This association holds even as additional controls are introduced, demonstrating the robustness of the relationship. In addition, the share of college-educated workers is positively correlated with screening thresholds, while the share of low-skill tasks is negatively associated. These findings imply that firms with a higher rank in the pay premium distribution, as well as those with a more educated workforce, exhibit greater selectivity in their hiring practices. Moreover, as shown in Panel B of Figure 5, screening intensity has increased over time. Interpreting the results of the preferred specification of Column 5 of Table 2, a firm in the 10th decile with only college-educated workers and high-skill jobs has a minimum worker fixed effect 1.61 standard deviations higher than a firm in the 1st decile with only low-skill workers, all else equal.

Fact 5 Higher-quality firms in Germany implements stricter screening thresholds, hiring workers with higher minimum fixed effects.

3 The Model

Motivated by the empirical evidence, in this section I present a new theory that incorporates preference heterogeneity, firm screening of workers, sorting, and labor market power through strategic wage setting. I begin by discussing the model setup. In the next section, I delve into the analysis of partial and market equilibrium, and finally, I characterize the general equilibrium. Time is dynamic, and the model is evaluated in steady state. Thus, for ease of exposition, I only report the time index when describing the problems of the agents. When analyzing optimality conditions and solutions, I suppress the time index. For all analytical derivations and proofs, see Appendix B.2.

3.1 Environment

Agents - The economy consists of a continuum of representative households (workers) indexed by their heterogeneous ability a. For reasons that will become clearer when analyzing the firms' problem, a has to be interpreted as *latent* heterogeneity. Observable heterogeneity, such as educational attainment (e.g., college graduates), can be easily incorporated, as in D. W. Berger, Herkenhoff, and Mongey (2022b). Workers heterogeneous abilities are distributed as $a \sim G_a(a)$ with density $g_a(a)$ over the support $[\underline{a}, \overline{a}]$. There is a continuum of local labor markets $j \in [0, 1]$, and each market draws a random number of firms $m_j \sim G_m(m)$ — the only *ex-ante* difference across markets. Firms are heterogeneous in their baseline productivity z_{ij}, where ij indexes firm i in local labor market j, drawn from a distribution $z \sim G_z(z)$. Firms are owned by the representative entrepreneur, who receives their profits as a lump sum payment and invests in capital accumulation to rent to firms.

3.2 Workers

Each worker's utility function is characterized by concave preferences over per-capita consumption and per-capita disutility from supplying labor. Labor disutility has a nested-CES functional form, taken directly from D. Berger et al. (2022a) and discussed in detail below. Given that product market power is not the focus of this research, I also assume that consumption goods are perfect substitutes which implies no markups[13]. Thus, I normalize the price of consumption goods to one. Each household a receives an endogenous number of job offers based on their ability. Denote $\mathcal{S}_j(a)$ as the choice set of firms offering a job to workers with ability a in labor market j. Thus, the household chooses the measure of workers $n_{ijt}(a)$ to supply to each firm in their choice set and

[13]Note that it is easy to incorporate monopolistic competition with a CES aggregator of firm-level output. Firms would charge a constant markup and would optimize over a decreasing return revenue function rather than a decreasing return production function. All the results apply to that production function. I consider perfect substitutes goods just for simplicity and ease of exposition.

consumption of each good c_{ijt} to maximize their net present value of utility while taking wages $\{w_{ijt}(a)\}$ and choice sets $\mathcal{S}_j(a)$ as given:

$$\begin{aligned} U_0(a) = \max_{\{n_{ijt}(a)\}_{t=0}^{\infty}, \{c_{ijt}(a)\}_{t=0}^{\infty}} & \sum_{t=0}^{\infty} \beta^t U\left(\frac{C_t(a)}{g(a)}, \frac{N_t(a)}{g(a)}\right) \\ = & \sum_{t=0}^{\infty} \beta^t \left(\frac{C_t(a)}{g(a)(1-\sigma)^{\frac{1}{1-\sigma}}}\right)^{1-\sigma} - \left(\frac{N_t(a)}{g(a)(1+\frac{1}{\varphi})^{\frac{1}{1+\varphi^{-1}}}}\right)^{1+\frac{1}{\varphi}} \end{aligned} \tag{3.1}$$

where the aggregate per capita consumption and labor supply indexes are given by

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

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

and maximization is subject to the household's budget constraint in each period[14]:

$$C_t(a) = \int_0^1 \sum_{i \in \mathcal{S}_j(a)} [w_{ijt}(a) n_{ijt}(a)] dj. \tag{3.2}$$

The labor supply indexes in the model fall within the broader category of *preference heterogeneity* models, as micro-founded by D. Berger et al. (2022a), discussed in Manning (2021), and modified in the appendix B.2 to account for endogenous choice set $\mathcal{S}_j(a)$[15]. In this paradigm, workers not only consider wage offers but also factor in an unobservable idiosyncratic taste shock when deciding which firm to supply their labor to. This random component encompasses various factors,

[14] Throughout the paper, I differentiate between agents earning from wages (workers) and those earning profits (entrepreneurs). This simplification aids tractability and aligns with data (Survey of Consumer Finances (SCF)). The SCF data suggests that, for the majority of workers, income is primarily derived from wages. The capital income over labor income ratio is approximately 0.06 for the vast majority of college workers and 0.03 for high-school workers (D. W. Berger, Herkenhoff, and Mongey (2022b)).

[15] As derived in D. Berger et al. (2022a), the labor supply curves within this framework can be micro-founded through discrete labor supply decisions by individuals: (i) across an employment/non-employment margin, (ii) across markets, and (iii) across firms within markets. Assuming preferences across these three aspects follow a correlated Gumbel distribution, the parameter θ corresponds to the conditional variance across markets, and η to the conditional variance within markets. Notably, η captures the within-market substitutability of firms, while θ the across-market substitutability of firms.

including commuting distance, preferences for firm culture, work environment, and moving costs. Consequently, workers do not strictly adhere to supplying labor to the firm with the highest wage; instead, they opt for the one that maximizes their indirect utility.

I assume that the elasticities of substitution, η and θ, are structured such that jobs within a market are considered closer substitutes than those across markets ($\eta > \theta$). This suggests that labor supply to firms exhibits greater elasticity within markets due to factors such as intra-market mobility costs (e.g., commute costs), with η capturing this intra-market mobility. Conversely, θ accounts for inter-market mobility costs (e.g., moving costs). As $\eta \to \infty$, intra-market mobility costs approach zero, rendering firms within a choice set perfect substitutes. In such scenarios, the representative worker directs workers solely to the firm offering the highest wage in her choice set. Analogously, as $\theta \to \infty$, inter-market mobility costs diminish, leading the household to treat markets as perfect substitutes. In this case, the household allocates its workers exclusively to the market offering the highest wage. It is noteworthy that neoclassical monopsony is nested under the condition $\eta = \theta$.

Optimality Conditions- Optimality conditions evaluated in steady state imply:

$$\begin{aligned}
&\left(\frac{N(a)}{g(a)}\right)^{\frac{1}{\varphi}+\sigma} = W(a)^{1-\sigma} \\
&n_{ij}(a) = \left(\frac{w_{ij}(a)}{w_j(a)}\right)^{\eta} \left(\frac{w_j(a)}{W(a)}\right)^{\theta} N(a) \\
&\Leftrightarrow w_{ij}(a) = \left(\frac{n_{ij}(a)}{n_j(a)}\right)^{\frac{1}{\eta}} \left(\frac{n_j(a)}{N(a)}\right)^{\frac{1}{\theta}} W(a) \quad \text{Inverse labor supply curve.}
\end{aligned} \tag{3.3}$$

Given aggregate labor supply, the firm labor supply curve includes two bookkeeping terms: the market wage index $w_j(a)$ and aggregate wage index $W(a)$. These are defined as the numbers that satisfy

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

Together with optimality conditions 3.3, these definitions imply

$$\begin{aligned}
w_j(a) &= \left(\sum_{i \in \mathcal{S}_j(a)} w_{ij}(a)^{1+\eta}\right)^{\frac{1}{1+\eta}}, \\
W(a) &= \left(\int_0^1 w_j(a)^{1+\theta}\, dj\right)^{\frac{1}{1+\theta}}.
\end{aligned} \tag{3.4}$$

3.3 Representative Entrepreneur

There is a representative entrepreneur indexed by e with monotonic preferences over consumption of the final good, making decisions regarding the investment in the next period capital K_{t+1} and the consumption of each good c_{ijt} to maximize her net present value of utility. The entrepreneur rents capital to firms with demand k_{ijt} and receives their profits π_{ijt} as a lump sum payment. Given an initial capital stock K_0, the entrepreneur solves

$$U_0(e) = \max_{\{K_{t+1}, c_{ijt}(e)\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t U(C_t(e)) = \sum_{t=0}^{\infty} \beta^t \left(\frac{C_t(e)}{1-\sigma} \right)^{1-\sigma} \tag{3.5}$$

Subject to the budget constraint:

$$C_t(e) + K_{t+1} - (1-\delta)K_t = \Pi_t + R_t K_t \tag{3.6}$$

where the aggregate consumption, profits, and capital indexes are given by

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

Optimality conditions: Optimality conditions evaluated in steady state yield the following Euler equation for capital accumulation:

$$1 = \beta(R + 1 - \delta) \tag{3.7}$$

3.4 Firms

Firms exhibit heterogeneity in their baseline productivity z_{ij}, drawn from the distribution $G_z(z)$.

The key novelty of the model is the introduction of the following production function. Let $n_{ijt}(a)$ denote the density of workers with ability a employed in firm ij. Let $\phi(z, a)$ be the output produced by each single worker of ability a in a firm of type z. The production function is specified as:

$$y_{ijt} = \Phi(z_{ijt}, \mathbf{a}_{ijt})(k_{ijt}^{1-\gamma} h_{ijt}^{\gamma})^{\alpha} \tag{3.8}$$

Here, y_{ijt} represents the output produced by the firm, k_{ijt} stands for total capital, and h_{ijt} is the total employment, defined as $h_{ijt} := \int_{\underline{a}}^{\overline{a}} n_{ijt}(a)da$. The term $\mathbf{a}_{ij}$ stands for the within firm distribution of workers' abilities. The term $\Phi(z_{ijt}, \mathbf{a}_{ijt}) := \left[\int_{\underline{a}}^{\overline{a}} \phi(z_{ijt}, a) n_{ijt}(a)da\right] \frac{1}{h_{ijt}}$ is denoted as firm realized endogenous productivity. It is defined as the within-firm average of $\phi(z, a)$, a function dependent on both worker ability and firm baseline productivity. The idea is that firms are characterized by a exogenously assigned productivity z, but their actual productivity is endogenous and depends on the quality of the workers hired. Notice that in the context of this model, productivity and quality are isomorphic concept. Thus, an alternative interpretation is that the firm's product quality depends on the distribution of hired workers' abilities.

Lastly, the function $\phi(z, a)$ is assumed to be a CES aggregator of firm productivity and worker ability:

$$\phi(z, a) = [(1 - \omega_a) z^{\frac{\rho-1}{\rho}} + \omega_a a^{\frac{\rho-1}{\rho}}]^{\frac{\rho}{\rho-1}} \quad \text{For } \rho \leq 1, \quad \omega_a \in [0, 1], \quad \xi \in \mathbb{R}_+ \tag{3.9}$$

The parameters ω and ρ regulates segmentation in the labor market and will be calibrated to match the observed market shares from the data, as discussed in section 5.

Discussion- Appendix B.1 provides a microfoundation for the production function, incorporating elements discussed in Eeckhout and Kircher (2018) and Helpman et al. (2010). The central idea behind the micro-foundation is that workers compete with each other over some resources[16], and the manager cannot assign these resources discriminately based on workers' abilities[17]. The significance of interpreting a as unobservable ability becomes evident: assuming that resources are split equally regardless of workers' observable characteristics may pose challenge in justification.

This production function encompasses various functional forms commonly employed in the literature. In particular, consider the functional form for ϕ in Equation 3.9, and examine variations in the parameters ρ and ω_a. If $\omega_a = 0$, the production function collapses to the classic Cobb-Douglas form $y = zk^{1-\gamma}h^{\gamma}$, also used in D. Berger et al. (2022a). As $\rho \to 1$, the functional form becomes $z^{1-\omega_a} h^{\gamma} k^{1-\gamma} \frac{\int_{\underline{a}}^{\overline{a}} a^{\omega_a} n(a)da}{h}$, resembling the one employed in Helpman et al. (2010).

[16]The production function 3.8 essentially introduces externalities among workers. An alternative interpretation is that there are interdependencies among workers. This idea traces back to the O-Ring technology studied in Kremer (1993). For empirical evidence on externalities, significant studies have been conducted by Moretti (2004), Mas and Moretti (2009), and Gennaioli et al. (2013).

[17]To illustrate, consider the following two familiar examples: a research university typically assigns the same office space to each professor, irrespective of their research ability, and similarly assigns the same office space/desk and training to each PhD student, regardless of their ability.

4 Equilibrium Analysis

4.1 Partial Equilibrium

Given the specified production function, it is instructive to first examine the case in which firms operate within a monopsonistically competitive labor market, as this allows for the clarification of key concepts and the derivation of analytical results. In this setting, firms are assumed to be infinitesimally small, both relative to the macroeconomy and within their respective local labor markets. The analysis then transitions to the primary focus of this paper: the equilibrium under oligopsonistic firm behavior, where firms are granular within local labor markets. In both cases, the equilibrium framework is modeled as Cournot competition.

Monopsonistic Firms

In this setting, firms maximize profits by taking the rental rate of capital R_t and the labor supply function as given. Firms choose their capital k_{ijt} and an employment schedule $n_{ijt}(a) : [\underline{a}, \overline{a}] \to \mathbb{R}_+$ without internalizing their impact on the market employment index $n_j(a)$. The capital and employment schedule are determined as the solutions to the following profit maximization problem:

$$\pi_{ijt} = \max_{\{n_{ijt}(a)\}, k_{ijt}} \Phi_{ijt}(z_{ij}, \mathbf{a}) \left(h_{ijt}^{\gamma} k_{ijt}^{1-\gamma} \right)^{\alpha} - R_t k_{ijt} - \int_{\underline{a}}^{\overline{a}} w_{ijt}(a) n_{ijt}(a)\, da \tag{4.1}$$

subject to the labor market constraint:

$$w_{ijt}(a) = \left(\frac{n_{ijt}(a)}{n_{tj}(a)} \right)^{\frac{1}{\eta}} \left(\frac{n_{jt}(a)}{N_t(a)} \right)^{\frac{1}{\theta}} W_t(a)$$

To further simplify the firm's problem, we derive the first-order condition with respect to capital and substitute the resulting capital demand back into the firm's profit function. This yields the following expression for the firm's profit maximization problem:

$$\pi_{ijt} = \max_{\{n_{ijt}(a)\}} Z_t \Phi_{ijt}(z_{ij}, \mathbf{a})^{\frac{1}{1-(1-\gamma)\alpha}} h_{ijt}^{\frac{\gamma\alpha}{1-(1-\gamma)\alpha}} - \int_{\underline{a}}^{\overline{a}} w_{ijt}(a) n_{ijt}(a)\, da \tag{4.2}$$

where $Z_t := (1 - \alpha(1-\gamma)) \left(\frac{\alpha(1-\gamma)}{R_t} \right)^{\frac{(1-\gamma)\alpha}{1-(1-\gamma)\alpha}}$ is a constant that applies to all firms.

Next, we turn to the characterization of optimal employment $n_{ij}(a)$ by deriving the marginal product of labor (MPL). Substituting the capital demand into the production function, the marginal product of a worker with ability a is given by:

$$MPL(a|\Phi_{ij}(z_{ij},\mathbf{a}),h) = Z\frac{\alpha\gamma}{1-\alpha(1-\gamma)}\Phi_{ij}(z_{ij},\mathbf{a})^{\frac{1}{1-\alpha(1-\gamma)}}h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}}\left[\underbrace{1}_{\text{Size Effect}} - \underbrace{\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_{ij},a)}{\Phi_{ij}(z_{ij},\mathbf{a})}\right)}_{\text{Productivity Effect}}\right] \tag{4.3}$$

For simplicity, I will denote $MPL(a|\Phi_{ij}(z_{ij},\mathbf{a}),h)$ as $MPL_{ij}(a)$. The production function naturally generates firm-specific workforce selection. Hiring an additional worker has two effects. First, it increases total employment h, leading to a positive size effect on production (labelled Size Effect in equation 4.3). Second, it affects the firm's endogenous productivity $\Phi_{ij}(z_{ij},\mathbf{a})$ (labelled Productivity Effect in equation 4.3). This productivity effect can potentially be negative and substantial enough to outweigh the positive size effect on total employment. Specifically, when the term $\left[1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_{ij},a)}{\Phi_{ij}(z_{ij},\mathbf{a})}\right)\right] < 0$, the negative impact on the firm's endogenous productivity dominates the positive size effect, leading the firm to refrain from hiring that worker.

Let $\phi(z,a)$ be a log-supermodular function, increasing in both a and z. Assume that firms operate in monopsonistic labor markets, where markdowns remain constant across both workers and firms. Based on this assumption, I derive the following theoretical results, with proofs provided in Appendix B.3.

Lemma 1 (Log-Supermodularity of the Marginal Product of Labor) *The marginal product of labor, $MPL_{ij}(a)$, satisfies log-supermodularity. Specifically, for any two worker abilities $a_1 > a_0$ and any two firm productivities $z_{1j} > z_{0j}$, the following inequality holds:*

$$\frac{MPL_{1j}(a_1)}{MPL_{0j}(a_1)} > \frac{MPL_{1j}(a_0)}{MPL_{0j}(a_0)},$$

indicating that the marginal product of labor increases more steeply with higher worker ability at more productive firms.

Corollary 1 (Log-Supermodularity of the Employment Schedule) *In a monopsonistic labor market, the employment schedule $n_{ij}(a)$ also satisfies log-supermodularity:*

$$\frac{n_{1j}(a_1)}{n_{0j}(a_1)} > \frac{n_{1j}(a_0)}{n_{0j}(a_0)},$$

This implies that more productive firms are relatively more likely to employ higher-ability workers.

Corollary 2 (Monotonicity of Endogenous Productivity) *In a monopsonistic labor market, firm endogenous productivity Φ_{ij} is monotonically increasing in firm exogenous productivity. Specifically:*

$$z_{1j} > z_{0j} \Rightarrow \Phi_{1j} > \Phi_{0j}$$

Thus, higher exogenous productivity in a firm translates to higher endogenous productivity.

Proposition 1 (Monotonicity of Screening Thresholds) *In a monopsonistic labor market, firm screening thresholds $\widetilde{a}_{ij}$ are weakly increasing with firm exogenous productivity. Specifically:*

$$z_{1j} > z_{0j} \Rightarrow \widetilde{a}_{1j} \geq \widetilde{a}_{0j}$$

Thus, more productive firms set weakly higher minimum ability thresholds for hiring.

Intuitively, due to complementarities in the function $\phi(z, a)$, more productive firms tend to hire workers from a distribution that is skewed toward higher-ability workers, leading to higher firm-level endogenous productivity. From Equation 4.3, what determines the marginal product of labor for a worker of type a in firm ij is the worker's output relative to the firm average worker output. Consequently, a within-firm distribution of worker abilities that is skewed toward high-ability workers causes the output of low-ability workers to fall further below the firm's mean output. This implies that low-ability workers have a larger marginal product of labor in firms where the within-firm distribution is skewed to the left. This will discourage firms from hiring lower-ability workers. Consequently, more productive firms exhibit greater selectivity in their hiring practices. Figure 6 illustrates the screening thresholds $\widetilde{a}$ by firm exogenous productivity $\ln z_{ij}$ for a market with one thousand firms operating under monopsonistic competition. The figure confirms that the minimum ability level of workers hired by the firm increases monotonically with firm productivity.

Oligopsonistic firms

Given the production function, the assumption now is that firms are infinitesimal concerning the macroeconomy but exhibit granularity within each local labor market. Consequently, firms consider aggregate quantities $N_t(a)$ and $W_t(a)$ as exogenous, while internalizing the consequences of their choices on the market indexes $n_{jt}(a)$ and $w_{jt}(a)$. The equilibrium framework is characterized by Cournot competition, where firms factor in the employment decisions of their competitors n^*_{-ijt}.

Thus, firms maximize profits taking R_t and the labor supply function as given, choosing

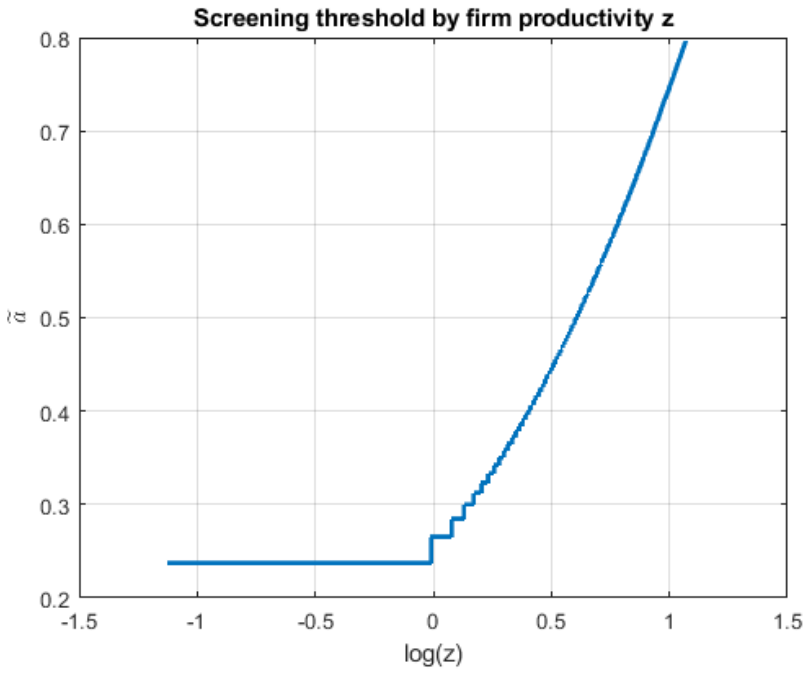


Figure 6: Screening thresholds in a monopsonistic labor market

Notes: This figure illustrates the screening threshold (i.e., the minimum ability hired) by firm log exogenous productivity in a labor market with one thousand firms competing under monopsonistic behavior. The parameter values are chosen to ensure that the function $\phi(z, a)$ exhibits log-supermodularity. This specification, with ϕ increasing in both a and z, is designed to support the theoretical results derived in Proposition 1.

capital k_{ijt} and an employment schedule $n_{ijt}(a) : [\underline{a}, \overline{a}] \to \mathbb{R}_+$[18]. By substituting the firm first order condition for capital back into firms' profits, the firm maximizes:

$$\pi_{ijt} = \max_{\{n_{ijt}(a)\}} Z\Phi_{ijt}(z_{ij}, \mathbf{a})^{\frac{1}{1-(1-\gamma)\alpha}} h_{ijt}^{\frac{\gamma\alpha}{1-(1-\gamma)\alpha}} - \int_{\underline{a}}^{\overline{a}} w_{ijt}(a) n_{ijt}(a) da \tag{4.4}$$

Where $Z_t := (1 - \alpha(1-\gamma))(\frac{\alpha(1-\gamma)}{R_t})^{\frac{(1-\gamma)\alpha}{1-(1-\gamma)\alpha}}$ is a common term to all firms.

If the MPL of the worker is positive, the standard rearrangement of the first-order condition with respect to employment $n_{ij}(a)$ yields a Lerner condition for the wage as an endogenously determined firm-worker specific markdown $\mu_{ij}(a) \leq 1$ on the marginal product of labor. On the other hand, when MPL is negative, the wage is equal 0 as the firm doesn't extend a job offer to

[18]To ease the notation, I have suppressed all indexes and competitors' employment choices in the notation. When I write $w_{ijt}(a)$, it should be understood that the wage is actually a function of employment, competitors' employment, aggregate employment, and aggregate wage (i.e., $w_{ijt}(a, n_{ijt}(a), n^*_{-ijt}, N_t(a), W_t(a))$

the worker.

$$w_{ij}(a) = \begin{cases} \mu_{ij}(a) \cdot MPL_{ij}(a) & \text{if } MPL_{ij}(a) > 0 \\ 0 & \text{if } MPL_{ij}(a) \leq 0 \end{cases} \tag{4.5}$$

Where:

$$\mu_{ij}(a) = \frac{\epsilon_{ij}(a)}{\epsilon_{ij}(a)+1}, \quad \epsilon_{ij} \coloneqq \left[\frac{\partial \log w_{ij}(a)}{\partial \log n_{ij}(a)} \bigg|_{n^*_{-ij}(a)} \right]^{-1} \tag{4.6}$$

$\epsilon_{ij}(a)$ is the firm-worker inverse employment elasticity. Under the assumed structure for labor supply, there is a closed formula for the elasticity:

$$\epsilon_{ij}(a) = \left[\frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta} \right) s_{ij}(a) \right]^{-1}; \quad s_{ij}(a) = \frac{w_{ij}(a) n_{ij}(a)}{\sum_{i \in \mathcal{S}_j(a)} w_{ij}(a) n_{ij}(a)} \tag{4.7}$$

The subscript ij in the firm-level employment elasticity for worker type a, denoted as $\epsilon_{ij}(a)$, highlights that the employment elasticity is contingent upon worker ability, as well as firm and competitor characteristics. Specifically, it is determined on the equilibrium firm wage share of workers with ability a in the labor market in which firm i is located.

The firm markdowns are the model measure of firm market power toward workers of ability a. It is noteworthy that markdowns here depend on a firm's own market share *for the worker category* summarized by a. A firm markdown is larger (i.e., μ is lower) when the firm has a large market share for those workers. Intuitively, this happens because the firm internalizes its impact on the market index characterized by a lower elasticity of substitution θ. When a firm is a small employer for workers of ability a ($s_{ijt} \approx 0$), the markdown is determined entirely by the within-market elasticity of substitution η, as in standard monopsonistic models.

Lastly, it is possible to relate the average wage and average MPL to profits, markdowns, and analyze how the markdown shapes the firm labor share.

Proposition 2 (Firm-Level Quantities) *Define* $\psi_{ij}(a) = \left(1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\phi_{ij}(a)}{\Phi_{ij}}\right)\right)$ *and* $\widetilde{\psi}_{ij} = \overline{\mu}_{ij} + \frac{cov_{ij}(\mu,\phi)}{\alpha\gamma\Phi_{ij}} \leq 1$. *Let* $MPL_{ij}(a)$ *represent the marginal product of labor of type* a *at firm* ij, $\overline{MPL}_{ij}$ *be the average marginal product of labor,* $\overline{w}_{ij}$ *the firm average wage,* $cov_{ij}(\mu,\phi)$ *the within-firm covariance between markdown and workers' productivity, and* ls_{ij} *the firm labor share. Then (proofs*

in Appendix B.3), we have:

$$\overline{MPL}_{ij} = \frac{\alpha\gamma}{1-(1-\gamma)\alpha}\frac{y_{ij}}{h_{ij}}, \tag{4.8}$$

$$MPL_{ij}(a) = \overline{MPL}_{ij} \cdot \psi_{ij}(a), \tag{4.9}$$

$$\pi_{ij} = \left(\frac{1-\alpha(1-\gamma)}{\alpha\gamma} - \widetilde{\psi}_{ij}\right) h_{ij}\overline{MPL}_{ij}, \tag{4.10}$$

$$\overline{w}_{ij} = \overline{MPL}_{ij} \cdot \widetilde{\psi}_{ij}, \tag{4.11}$$

$$ls_{ij} = \alpha\gamma \cdot \widetilde{\psi}_{ij}. \tag{4.12}$$

First, it is worth mentioning that the expression for the average marginal product $\overline{MPL}_{ij}$ is equal to the marginal product of a Cobb-Douglas production function with homogeneous workers. Second, absent markdowns, $\widetilde{\psi}_{ij}$ equals one, and the average wage equals the average marginal product. The firm labor share and profits are equal to those in a Cobb-Douglas production function with homogeneous workers, despite the firm employing many heterogeneous workers. Third, the term $\widetilde{\psi}_{ij}$ determines the wedge between the firm average marginal product of labor and the average wage. It comprises two components. The first component represents the firm's average markdown, thereby causing $\widetilde{\psi}_{ij}$ (and the labor share) to decrease when the firm, on average, exerts higher labor market power. Additionally, there is a second-order effect—captured by the term cov_{ij}—which further magnifies the negative impact of markdowns on the firm labor share for a more negative covariance between workers' output $\phi(z,a)$ and markdowns. This covariance is computed with respect to the within-firm distribution of workers' abilities. Intuitively, when the covariance term is negative, the firm is relatively underpaying workers characterized by the highest $MPL_{ij}(a)$. As shown in appendix B.2, the firm's labor share is a composite of individual worker ability-type labor shares. Consequently, low-ability workers contribute proportionally less to the firm's labor share compared to their high-ability counterparts. To put it simply, in scenarios where the covariance between markdowns μ and worker output ϕ is more negative, the firm tends to underpay workers with higher $MPL_{ij}(a)$, who, in turn, contribute significantly to the within-firm labor share. If absent markdowns the labor share is constant, an increased negative covariance leads to a lower labor share because the earning distribution within the firm places less mass on the right tail of the within firm earning distribution. In an economy characterized by sorting, more productive firms demonstrate a larger covariance between market shares and worker output. This, in turn, contributes to a second-order reduction in their labor share. Finally, equations in Proposition 2 have noteworthy implications for the empirical estimation of markdowns. In their studies, Yeh et al. (2022) and Kirov and Traina (2021) utilize administrative data for U.S. manufacturers and

estimate markdowns using a production function approach by identifying plant-level markdowns through the ratio between a plant's marginal revenue product of labor and its wage. This approach shares similarities with the estimation of markups employed in De Loecker et al. (2020). Their findings reveal substantial wedges, particularly in larger firms, indicating a significant increase in labor market power since the year 2000. Equations in Proposition 2 highlight that, with unobservable workforce heterogeneity, such a measure becomes a model-consistent estimate of $\widetilde{\psi}_{ij}$, which is not the firm average markdown. More productive firms, larger in size in equilibrium, may exhibit a larger $\widetilde{\psi}_{ij}$, not solely due to a larger average markdown but also because of a larger covariance of markdown with worker output. On the other hand, $\widetilde{\psi}_{ij}$ is the relevant model-based measure to examine how markdowns impact the firm-level labor share.

4.2 Market Equilibrium

Once the optimal behavior of firms is established, I now turn to the definition of market equilibrium. Recall that there is a continuum of markets indexed by j, and each market draws a random number of firms $m_j \sim G_m(m)$. The market equilibrium is defined given the labor supply to the market of each worker's ability, which will be derived later in general equilibrium.

A market equilibrium is defined as market shares for each worker ability type $s_{ij}(a) \forall i \in j, \forall a \in [\underline{a}, \overline{a}]$ satisfying the following equations:

$$\forall a \in [\underline{a}, \overline{a}], \quad s_{ij}(a) = \frac{(\mu_{ij}(a) MPL_{ij}(a))^{1+\eta}}{\sum_{i \in \mathcal{S}_j(a)} (\mu_{ij}(a) MPL_{ij}(a))^{1+\eta}} \tag{4.13}$$

$$\forall a \in [\underline{a}, \overline{a}], \quad i \in \mathcal{S}_j(a) \quad \text{if and only if} \quad MPL_{ij}(a) > 0 \tag{4.14}$$

The efficient equilibrium is an equilibrium without distortions. It is defined in the same way as above but with no markdowns (i.e., $\mu_{ij}(a) = 1 \quad \forall a, \forall i \in j$), and it corresponds to the Planner's allocation for a specific set of weights assigned to worker types and entrepreneurs (See Appendix B.4 for details).

One Hundred Firms Example: Consider a labor market with one hundred firms randomly selected firms competing for workers, with calibration as in table 4. Figure 7 illustrates market shares and corresponding markdowns by workers' abilities for a few selected firms based on their productivity percentile in the market.

The figure depicts the key innovations of the model. First, the equilibrium is characterized

by a rationing of the workers' choice sets, as access to the top firm is precluded for workers at the lower end of the ability distribution. Second, the labor market is characterized by strong sorting and segmentation, with low-ability workers disproportionately employed by low-productivity firms. Regarding competition for workers, each firm specializes in a particular segment of the ability distribution and competes primarily with firms that are close in productivity and target the same workers. The wage policies of firms matter only when they are competing for similar segments of the labor force. For example, the employment decisions of low-productivity firms do not affect high-productivity firms, and vice versa, since they are not competing for the same workers. Firms positioned at the lower end of the productivity distribution specialize in employing workers with lower abilities, who may be screened out by more productive firms. Consequently, these firms capture a substantial market share for these workers, coupled with a large markdown, although they are small in terms of total size compared to other firms in the market. More productive firms shift their market share mode to the right, selecting out lower abilities workers. When more productive firms begin extending offers to highly skilled workers, the job offers from less productive firms cannot compete, resulting in the latter becoming the primary employers for workers with higher abilities.

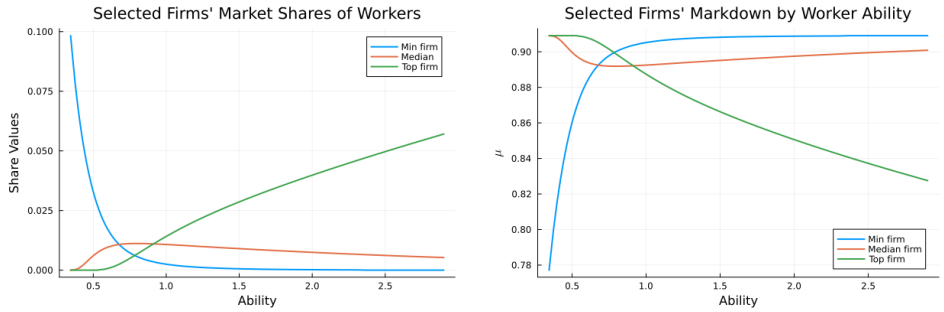
This phenomenon, characterized by firms specializing in different segments of the ability distribution, is termed *Endogenous Oligopsony*. This term emphasizes how selection, sorting, and complementarities make each firm an oligopsonist toward the workers it hires. Specifically, firms compete most intensely for workers with firms that have similar productivities and hire from the same ability segments. This localized competition among similarly productive firms contrasts with labor models where all firms compete equally across the ability distribution. Here, the endogenous segmentation creates narrow oligopsony power, as each firm primarily competes with its closest rivals that hire comparable workers. The degree of competition and resulting labor market power is endogenous and depends on the density of firms at any point along the productivity distribution. A greater number of firms with similar productivity targeting analogous ability segments leads to more intense competition for those workers and lower oligopsony power.

4.3 General Equilibrium

The steady-state general equilibrium[19] is defined as aggregate variables $C(a)$, $C(e)$, $N(a)$, K, labor market shares $s_j(a)$, firm-level shares $s_{ij}(a)$, and prices $W(a)$, $w_j(a)$, R such that the following equilibrium conditions are satisfied:

-Markets are in equilibrium: given the equilibrium $n_j(a)$, $s_{ij}(a)$ solves 4.13 and 4.14 for all j.

[19]The same definition applies for the efficient economy, characterized by $\mu_{ij}(a) = 1 \quad \forall ij$



(a) Panel A: Market shares in 100-firms market (b) Panel B: Markdowns in 100-firms market

Figure 7: Market equilibrium in a one-hundred-firm labor market

Notes: This figure is divided into two panels, labeled A and B, which report market equilibrium outcomes for three selected firms across the simulated productivity distribution in a one-hundred-firm market: the bottom firm, the median firm, and the top firm. Panel A illustrates the wage bill market shares by worker ability, while Panel B shows the corresponding markdowns. The labor supply follows a truncated log-normal distribution, and firm productivities are randomly drawn from a log-normal distribution. Calibration details are available in Table 4. Solid lines represent market shares for different worker abilities, and the vertical dashed lines indicate selected percentiles of the ability distribution.

-Market Labor Supply: The labor supply in each market j is given by $n_j(a) = s_j(a)^{\frac{\theta}{\theta+1}} N(a)$, where $s_j(a) = \frac{w_j(a)^{1+\theta}}{\int_0^1 w_j(a)^{1+\theta}} dj$.

-Aggregate Labor Supply: The aggregate labor supply is determined by $\left(\frac{N(a)}{g(a)}\right)^{\frac{1}{\varphi}+\sigma} = W(a)^{1-\sigma}$.

-Workers' Consumption: The aggregate consumption by workers with ability a is $C(a) = W(a)N(a) = \int_0^1 \sum_{i \in \mathcal{S}_j(a)} w_{ij}(a) n_{ij}(a) dj$.

-Aggregate Capital Supply: The aggregate capital supply is $RK = \alpha(1-\gamma)Y$.

-Entrepreneur's Consumption: The entrepreneur's consumption is given by $C(e) = \Pi + (R - \delta)K$.

5 Preliminary Calibration

This section presents a *preliminary* calibration of model parameters. Ongoing research is directed into estimating labor supply elasticities and other relevant parameters in a more accurate way. The ability distribution is assumed to follow a truncated log-normal distribution with parameters μ_a and σ_a, truncated by removing the top and bottom 0.01% of abilities to eliminate extreme outliers[20]. Firms' productivities are drawn from a log-normal distribution with parameters $\mu_z = 0$ and σ_z. The parameters governing the elasticities of substitution across firms, η and θ, are adopted from D. Berger et al. (2022a). R and δ are set based on common literature practices, and the utility parameters σ and φ are once again drawn from D. Berger et al. (2022a). γ is set so that capital payments are 30% of aggregate production. Relative to the paper by D. Berger et al. (2022a), in this model there are three additional parameters: ρ, ω and σ_a. The internally calibrated parameters are ρ, ω, σ_a, σ_z, and α.

These parameters are jointly set to align with specific moments from the data. α is set to match the Compensation of employees share of income of 53% from the Eurostat data[21]. I simulate a panel dataset based on model-generated data and run an AKM regression on this dataset (see Appendix C.4 for details). From this regression, I obtain model-generated measures for the standard deviations of both firm and worker fixed effects. Additional target moments include the proportion of workers in the first quartile of the ability distribution employed by firms in the first quartile of the productivity distribution, and similarly, the proportion of fourth-quartile workers employed by fourth-quartile firms, all within the final period group 2010–2017.

The moments are jointly targeted, and Table 3 reports the parameter values, the data moments, and the moments estimated from model-generated data, where each moment is most informative about a specific parameter. Intuitively, a larger standard deviation of firm productivities results in a larger standard deviation of the estimated firm fixed effects, while a larger standard deviation of workers' log abilities corresponds to a larger standard deviation of worker fixed effects. Figure 8 provides an illustration of how the chosen moments vary with different parameter values, offering a heuristic argument in support of the identification strategy for the parameters ω_a and ρ. Moreover, the figure clarifies why the Bottom-Bottom and Top-Top shares are selected as target moments. As ρ decreases, complementarities in the function ϕ increase, enhancing sorting and subsequently raising both the top-top and bottom-bottom shares due to greater assortative matching. Consider now the parameter ω_a. When ω_a equals 1, this implies no productivity differences across firms; when ω_a equals 0, it implies worker homogeneity. As ω_a increases, firm heterogeneity becomes less relevant, leading to a decrease in the Top-Top moment, as shown in Panel A of Figure

[20] See appendix C for more details

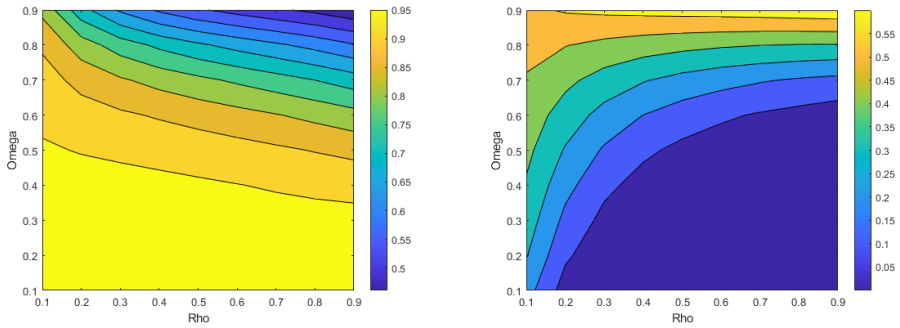
[21] Statistic obtained from Eurostat Annual National Accounts.

8. Conversely, when ω_a increases, worker abilities play a more significant role in firm production, so even minor sorting results in higher firm endogenous productivity, $\Phi_{ij}(z_{ij}, \mathbf{a}_{ij})$, and increased firm selection of lower-ability workers. As a result, the Bottom-Bottom share increases while the Top-Top share decreases. Thus, these two moments capture distinct variations in the data that inform the calibration of ρ and ω_a.

The model performs reasonably well in matching the targeted moments. Figure 9 compares the shares from the model to those from the main motivating evidence shown in Figure 2. Note that these shares are only partially targeted. For the calibration, I specifically targeted the bottom-bottom and top-top quartile shares, while in the figure, I report shares for each decile of the distribution. The similarity between the two figures is remarkable, suggesting that the model effectively captures the strong segmentation documented in the motivating evidence of Section 2.

Parameter	**Value**	**Target Moment**	**Data Moment**	**Model Moment**
σ_a (ability dispersion)	0.37	Worker Fixed Effect Dispersion	0.38	0.44
σ_z (productivity dispersion)	0.33	Firm Fixed Effect Dispersion	0.24	0.24
ω_a (weight in ϕ)	0.9	Market shares by quartile (bottom-bottom)	0.35	0.34
ρ (complementarities)	0.4	Market shares by quartile (top-top)	0.39	0.39

Table 3: Internally calibrated parameters and target moments



(a) Panel A: Top-Top Moment by Parameter Value

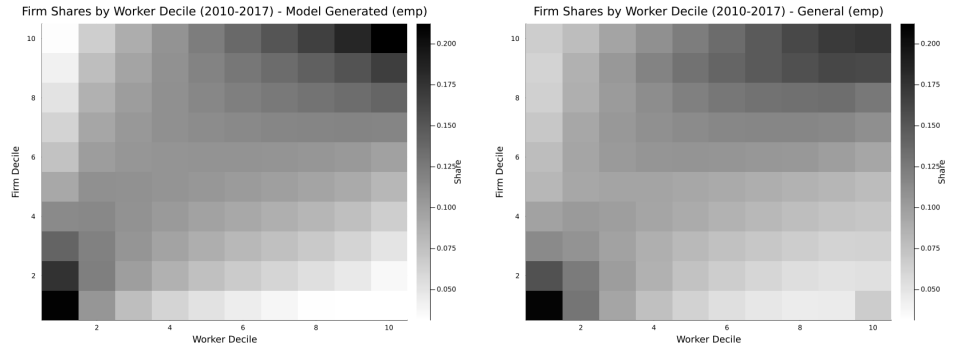
(b) Panel B: Bottom-Bottom Moment by Parameter Value

Figure 8: Moment Curves

Notes: This figure is divided into two panels, labeled A and B, which display moment curves for the shares of top-quartile and bottom-quartile workers employed in top-quartile and bottom-quartile firms, respectively. The x-axis represents the parameter ρ, while the y-axis represents the parameter ω_a. Different colors correspond to different moment values.

Table 4: Baseline Calibration Parameters

Parameter	Value	Parameter	Value
γ	0.7	μ_a, μ_z	0
α	0.99	σ_a	0.37
ρ	0.4	σ_z	0.33
δ	0.08	ω	0.9
R	0.10	η	10
σ	1	θ	0.5
φ	0.65		



(a) Panel A: Employment Shares (Model Generated Data)

(b) Panel B: Employment Shares (Data from 2010-2017)

Figure 9: Comparison of Employment Shares: Model vs Data

Notes: This figure is divided into two panels, labeled A and B, which compare the employment shares across firms in the model-generated data (Panel A) calibrated targeting data from 2010-2017 (parameters in Table 4, and actual data from 2010-2017 (Panel B). The shares are presented by decile, and are not directly targeted in the calibration.

6 Inefficiencies and Welfare

In this section, I explore the implications of markdowns arising in the competitive equilibrium for aggregate production efficiency and welfare. I begin by examining aggregate efficiency, comparing the aggregate output of the economy with labor market power to that of the efficient economy. Following this, I analyze welfare losses segmented by worker ability.

6.1 Aggregate Inefficiencies

I first explore how the market equilibrium gives rise to inefficiencies and examines their impact on aggregate production efficiency in the market equilibrium, under the calibration of Table 4. Firms, when having a wider markdown, pay a lower wage than the efficient wage. Since markdown is

larger when the market share is larger, firms are under-resourced for the workers they employ the most. Thus, there are four variables that are distorted by labor market power: aggregate labor supply, market size, firm endogenous productivity, and firm size. First, markdowns distort the size of firms, with some firms being larger than in the efficient allocation, and others smaller. I refer to this inefficiency as *Size Misallocation*. Second, markdowns distort the within-firm distribution of worker abilities, causing some firms to have excessively large endogenous productivity Φ_{ij}, while others have too small. Since every firm has heterogeneous markdowns for different workers' types, some firms will be under-resourced for workers of low ability, while others will be under-resourced for workers of high ability. I denote this inefficiency as *Misallocation of Talent*. Third, markets characterized by less competition are going to be smaller than in the efficient benchmark. Fourth, potentially, aggregate labor supply decisions may be distorted.

Size Misallocation - The distortions generated by markdowns cause some firms to become larger while others become smaller relative to the efficient benchmark. Size misallocation is a distortion also documented in D. Berger et al. (2022a). In D. Berger et al. (2022a), more productive firms possess greater labor market power and are therefore more distorted relative to the efficient benchmark. This results in correlated distortions, leading to significant inefficiencies due to misallocation.

In the context of this paper, the extent of this effect varies substantially with the parameters of the function (3.9). Intuitively, greater selection of workers implies that firms at the lower end of the productivity distribution increasingly specialize in the lower segment of the worker ability distribution. This, in turn, suggests that such firms impose larger markdowns on these workers. Consequently, bottom firms may also be under-resourced, potentially weakening the monotonic relationship between productivity and size distortion observed in D. Berger et al. (2022a).

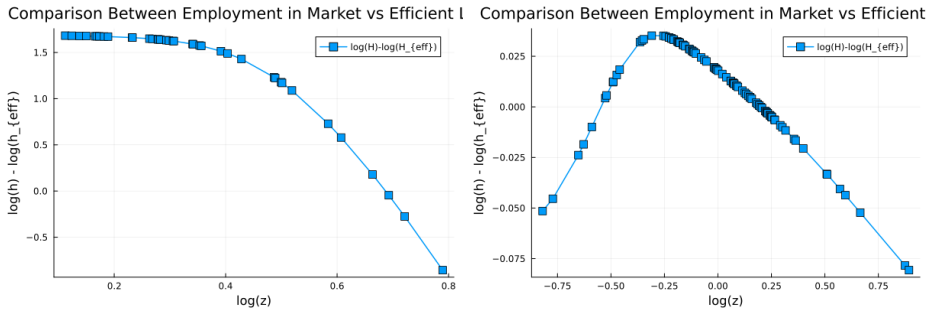
Figure 10 displays the log of the ratio of realized employment in market equilibrium to efficient employment $\ln\left(\frac{h}{h_{\text{eff}}}\right)$, comparing total employment for each firm in the labor market across different calibrations of function (3.9)[22]. The interpretation is that if $\ln\left(\frac{h}{h_{\text{eff}}}\right) = x$, then $h = h_{\text{eff}}e^{x}$, which implies that h is e^{x} times larger or smaller than h_{eff}.

Figure 10 reports Size Misallocation in one of the simulated markets with one-hundred firms. Panel A compares this ratio for a calibration with $\omega_a = 0$, which corresponds to an economy with a Cobb-Douglas production function and homogeneous workers, as in D. Berger et al. (2022a). All firms are over-resourced except the lowest-ranked firm, where $\frac{h}{h_{\text{eff}}} = 0.45$. This indicates that lower-tier firms are 3000% larger than they should be in an efficient economy, while the most productive

[22] In this section, I primarily vary ρ to adjust the degree of complementarities. However, all the results discussed here can be obtained with different combinations of the parameters ω_a, ξ, and ρ. Here, I am varying only ρ to maintain consistency in interpretation.

firm is 55% smaller than in an efficient setting.[23] This calibration results correlated distortions and in substantial aggregate inefficiency due to misallocation, with aggregate GDP being 8% lower than in the efficient allocation.

In contrast, Panel B shows the same ratio using the baseline calibration from Table 4. Compared to Panel A, the relationship changes dramatically. Now, both low- and high-productivity firms are under-resourced. The most productive firm is approximately 8% smaller than it would be in an efficient economy. Rather than being 3000% larger than in an efficient economy, low-productivity firms are now roughly 5% under-resourced. This has substantial implications for market production efficiency. Ignoring firm selection of workers results in significant model misspecification, with important consequences for aggregate inefficiencies. Specifically, labor market segmentation by ability weakens the monotonic relationship between firm size distortions and productivity, thereby mitigating output losses due to labor market power.



(a) Panel A: $\omega_a = 0$ (b) Panel B: Baseline calibration from Table 4

Figure 10: Size Distortion for Different Values of ω

Notes: This figure consists of two panels, labeled A and B, illustrating the relationship between log productivity and size distortion, measured by $\ln\left(\frac{h}{h_{\text{eff}}}\right)$, for different calibrations. Each point represents the log-ratio of total employment in the market equilibrium relative to the efficient allocation. Panel A shows results for $\omega_a = 0$, with productivities truncated at $\log(z) = 0$ to mitigate numerical errors resulting from h_{eff} approaching zero. Panel B displays the baseline calibration from Table 4. In both panels, all other parameters are set at baseline levels, allowing a direct comparison of size distortion under varying values of ω_a.

Misallocation of Talent - The model introduces an additional distortion generated by markdowns. Relative to the efficient market allocation, the *within*-firm ability distribution is now skewed, distorting the firm's endogenous productivity Φ_{ij}. Generally, less productive firms apply larger markdowns to less able workers, while more productive firms apply larger markdowns to more able workers due to their larger market shares. Consequently, less productive firms employ too few less able workers and too many more able workers relative to the efficient allocation, and vice versa for

[23] In Panel A, the firm distribution is truncated at $\log(z) = 0$ because, at lower values of $\log(z)$, the size h_{eff} is very small and approaches zero, introducing numerical error.

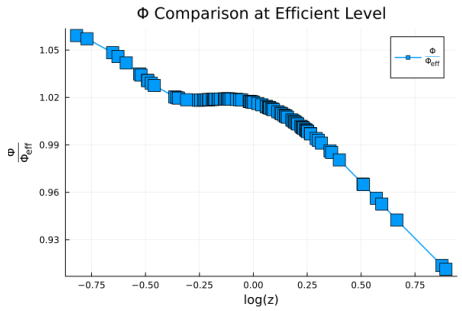
more productive firms. In an efficient allocation, firms with lower z should realize lower productivity Φ_{ij}, while firms with higher z should realize higher productivity by employing more highly skilled workers.

Figure 11 presents this effect across three panels. Panel A displays market shares by firm deciles in the efficient allocation, while Panel B shows market shares under the allocation with markdowns. The efficient allocation in Panel A is characterized by darker shades at the two extremes than the equilibrium in Panel B, indicating that labor market power leads to *undermatching* in the economy, with less assortative matching. Panel C shows the ratio of realized firm endogenous productivity in market equilibrium versus the efficient allocation ($\frac{\Phi_{ij}}{\Phi_{ij,\text{eff}}}$) under the baseline calibration. A monotonic relationship emerges, with lower z firms achieving realized productivity up to 5% higher than efficient levels, while higher z firms exhibit up to an 8% reduction.



(a) Panel A: Market Shares (Efficient Allocation)

(b) Panel B: Market Shares (With Markdowns)



(c) Panel C: $\frac{\Phi_{ij}}{\Phi_{ij,\text{eff}}}$

Figure 11: Misallocation of Talent at Baseline Calibration

Notes: This figure is divided into three panels, labeled A, B, and C. Panel A displays market shares by firm decile in the efficient allocation, while Panel B shows market shares in the equilibrium with markdowns. Panel C shows the ratio of realized firm endogenous productivity to the efficient allocation, $\frac{\Phi_{ij}}{\Phi_{ij,\text{eff}}}$, by firm productivity. All parameters are set at baseline levels from Table 4. The vertical dashed line indicates selected percentiles of the ability distribution. In all simulations, labor supply follows a truncated log-normal distribution, and firm productivities are based on the 100 productivities drawn in Figure 7.

Other Inefficiencies - Two additional inefficiencies emerge when comparing the resulting allocation with the efficient benchmark. First, markets may operate at inefficient sizes. Markets with higher competition are often overcrowded, while those with lower competition are more distorted and underpopulated. Typically, smaller markets experience the greatest under-crowding. Since this inefficiency has a negligible impact on aggregate outcomes, it is documented in the Appendix C.3. Second, there is an inefficiency in aggregate labor supply, as some workers choose to supply fewer hours. With log-utility and worker income derived solely from wages, the substitution and income effects exactly offset each other, rendering labor supply inelastic to wages and, consequently, unaffected by markdowns.

Aggregate Production- The aggregate production in the competitive equilibrium is 0.108% lower then the aggregate production under the efficient allocation. To provide insights into the significance of the four inefficiencies identified in the previous sections, I systematically activate or deactivate each inefficiency, while keeping the rest of the equilibrium unchanged at the efficient level, to compute output $\widetilde{Y}$. This process involves toggling the following inefficiencies one at a time: aggregate labor supply $N(a)$, labor market supplies $s_j(a)$, and the firm-level shares of workers' ability $s_{ij}(a)$. The firm-level shares are first used to compute only the firm endogenous productivity Φ_{ij} and then to compute firm-level total employment h_{ij}. The resulting changes reflect the isolated impact of each inefficiency, computed as $\frac{\widetilde{Y}}{Y_{\text{eff}}}$. Moreover, I also compute the impact of the interaction between the inefficiencies of misallocation and misallocation of talent. The interaction term is calculated as the total impact $\frac{\widetilde{Y}}{Y_{\text{eff}}}$ when using the competitive shares $s_{ij}(a)$ for both Φ_{ij} and h_{ij}, minus the single impacts of Φ_{ij} and h_{ij} alone. The remainder is due to all other combinations of interactions[24]. Table 5 reports gains or losses resulting from the four inefficiencies in aggregate production. Aggregate inefficiencies are small, with the bulk of inefficiency coming from Size Misallocation.

Table 5: Aggregate Efficiency

Inefficiency	**Impact on Aggregate Production**
Aggregate Supply Distortion	0 %
Market Size Distortion	$\approx$ 0 %
Misallocation of Talent	-0.003%
Size Misallocation	-0.109%
Interaction Between size and talent misallocation	+0.004%
Total Impact	**-0.108%**

Notes: This table illustrates the contribution of different inefficiencies to aggregate production. Calibration details are available in Table 4. To compute each single contribution, I impose quantities $N(a)$, $s_j(a)$, and $s_{ij}(a)$ one at a time, obtained in the competitive equilibrium. Subsequently, I calculate the quantity $\frac{\widetilde{Y}}{Y_{\text{eff}}}$, where $\widetilde{Y}$ represents the new aggregate production with the imposed quantity. When imposing the market shares $s_{ij}(a)$, I initially use it solely to compute the firm endogenous productivity Φ_{ij} and then to compute only firm-level total employment h_{ij}. The interaction term is calculated as the total impact $\frac{\widetilde{Y}}{Y_{\text{eff}}}$ when using the competitive shares $s_{ij}(a)$ for both Φ_{ij} and h_{ij}, minus the individual impacts of Φ_{ij} and h_{ij} alone.

6.2 Welfare Distribution

In this section, I conduct a comparative analysis of the impact of labor market power on the distribution of welfare. Labor market power, measured through markdowns (μ_{ij}), deviates the

[24]Notice that the isolated impact is not restricted to be negative. This is because I am tilting one inefficiency alone, but changing that would also affect other inefficiencies. The aggregate impact, when summing all four inefficiencies and their interactions, is negative since the economy is inefficient.

economy from the efficient allocation. This has two key implications for welfare. First, production inefficiencies lead to a reduction in the overall size of production. Additionally, there is a redistribution effect from workers to entrepreneurs. Markdowns decrease the total income for workers, impacting their consumption, while simultaneously increasing profits and the consumption of entrepreneurs. The redistribution effect varies among workers based on their ability types, as different workers are subject to distinct markdowns depending on the level of competition across firms for their labor services. Therefore, I start by describing how markdowns vary with workers' abilities. Then, I compare the welfare distribution in the equilibrium with market power to that in an equilibrium where wages are equal to the marginal product. The distribution of welfare gains or losses, with respect to both workers' ability and entrepreneurs, is measured through the percentage change in per capita consumption, denoted $\frac{C(a)}{g(a)}$ for workers and $\frac{C(e)}{g(e)}$ for entrepreneurs. These adjustments are designed to equalize utilities in steady state, ensuring that for workers, $\lambda(a)$ satisfies $U\left((1+\lambda(a))\frac{C(a)}{g(a)}, \frac{N(a)}{g(a)}\right) = U\left(\frac{C_{\text{eff}}(a)}{g(a)}, \frac{N_{\text{eff}}(a)}{g(a)}\right)$, and for entrepreneurs, $\lambda(e)$ satisfies $U((1+\lambda(e))C(e)) = U(C_{\text{eff}}(e))$.

Markdown Distribution in GE- Denote by $\widetilde{n_j(a)}$ the total employment of workers of ability a in market j, i.e., $\widetilde{n_j(a)} = \sum_{i=1}^{m_j} n_{ij}(a)$, and by $\widetilde{n(a)}$ the total employment, $\widetilde{n(a)} = \int_{j=0}^{1} \widetilde{n_j(a)}$. The weighted average markdown is defined as $\widetilde{\mu(a)} \coloneqq \frac{\int_0^1 \sum_{i=1}^{m_j} \mu_{ij}(a) n_{ij}(a) dj}{\widetilde{n(a)}}$. Figure 12 plots the relationship between the weighted average markdown $\widetilde{\mu(a)}$ by worker log-ability. It's evident how heterogeneous markdowns by worker ability are stronger at the tails of the ability distribution. At the bottom tail of the ability distribution, workers take home 70% of the marginal product of labor. At the top tail, workers receive 76% of the marginal product, while around the median, they take home 79%. Intuitively, markdown is determined by the strength of the competition for workers. At the bottom tail of the worker ability distribution, workers receive a rationing of their choice set $\mathcal{S}_j(a)$ since they are selected out by the majority of the firms. This implies that there are few firms from the bottom tail of the productivity distribution that are willing to hire these workers. Since there are fewer firms clustered at the bottom tail of the productivity distribution, competition for these workers is reduced. A similar argument, albeit with a different mechanism, applies to workers at the top of the log-ability distribution. These workers are disproportionately employed at firms at the top of the productivity distribution. These firms make such high job offers to these workers that the vast majority of firms cannot compete effectively. Consequently, these firms are only in competition with a select few others from the right-tail of the productivity distribution, which are extremely differentiated from each other. As a result, competition for the workforce of workers at the top of the log-ability distribution is reduced.

Welfare Effects - Figure 13 illustrates the relationship between $\lambda(a)$ and log-ability. The black line represents $\lambda(a)$ needed to equalize total utility with efficient utility, accounting for the



Figure 12: Distribution of weighted average markdown $\widetilde{\mu(a)}$

Notes: This figure illustrates the relationship between average worker markdown $\widetilde{\mu(a)}$ and workers log-ability. All parameters are set at the baseline level of table 4.

disutility of supplying additional labor. The red dashed line represents $\lambda(a)$ by log-ability while holding the disutility from labor supply fixed. Since utility is in logarithmic form, the two lines coincide as aggregate labor supply is inelastic to wages.

All workers experience welfare losses relative to the efficient allocation. The median worker would require a 26% increase in consumption to attain steady-state utility in the absence of markdowns. Notably, there is substantial heterogeneity across worker types. Workers at the bottom of the ability distribution would need a roughly 50% increase in consumption to achieve the steady-state utility level absent markdowns. These workers also have lower baseline consumption, even without labor market power distortions. Thus, the 50% number represents a lower bound relative to an allocation by a planner who considers inequality. Workers at the top of the ability distribution would require a roughly 30% increase in consumption to reach the efficient utility level.

Figure 12 demonstrates that endogenous oligopsony leads to stronger markdowns at the tails of the ability distribution, resulting in more pronounced welfare effects. The heterogeneity in welfare losses is driven by differences in markdowns across worker types, stemming from the non-competitive market structure. At the bottom tail, workers face restricted choice sets $\mathcal{S}_j(a)$, as most firms exclude them, resulting in fewer firms hiring them. This lack of competition increases

markdowns, significantly reducing their income. For high-ability workers, a similar competition mechanism applies: high offers from a few highly productive firms render other firms uncompetitive, thereby reducing competition and exacerbating welfare losses.

In contrast, the representative entrepreneur benefits from labor market power. To reach the utility level of the efficient economy, the entrepreneur would need to reduce consumption by 65%.

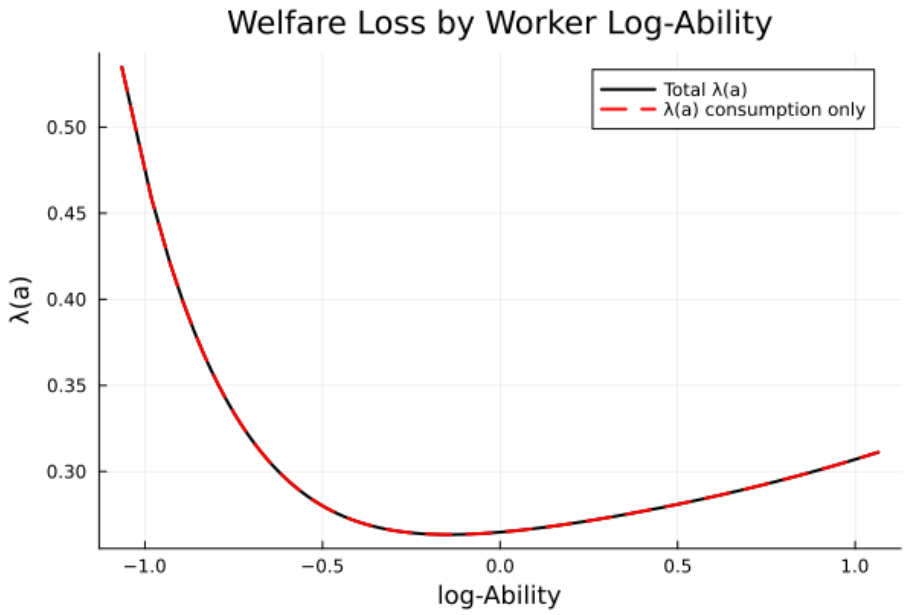


Figure 13: Welfare Loss distribution

Notes: This figure illustrates the relationship between welfare losses measured by $\lambda(a)$ and workers log-ability. All parameters are set at the baseline level of table 4.

6.3 Measurements: Labor Share and Firm Profits

I conclude the analysis with a discussion of the implications of Endogenous Oligopsony for measuring labor market power in the context of the observed decline in labor share across many developed economies[25](Karabarbounis and Neiman (2014)). Recent studies use two primary approaches: the *production function* method, which measures the gap between a firm's marginal product of labor and its average wage (Kirov and Traina (2021); Yeh et al. (2022)), and payroll HHIs (Herfindahl-Hirschman Index) to gauge market concentration (D. Berger et al. (2022a)). Notably, while firms' wage-productivity wedges tracked payroll HHI until 2000, they sharply increased thereafter, even

[25]Here, I am just providing a discussion on the implications of Endogenous Oligopsony for the decline in the labor share. This part of the research represents a primary area of ongoing research for the development of the research.

as HHI remained stable (Yeh et al. (2022)). The labor market wage bill HHI index is defined as $\sum_{i \in j} s_{ij}^2$, where s_{ij} represents the firm wage bill labor market share ($s_{ij} = \frac{w_{ij} n_{ij}}{\sum_{i \in j} w_{ij} n_{ij}}$). The HHI measure is motivated because in a model without worker heterogeneity and selection, there is a monotonic relationship between this HHI index and the market share of labor.

I begin by demonstrating with a numerical contraddicting example that this concentration measure may not always be suitable for estimating the impact of labor market concentration on the labor share. This example considers a labor market using the one-hundred simulated productivities used for the example of Figure 7. I utilize the baseline calibration from Table 4 in one of the simulated one hundred firms market. The parameter governing the degree of complementarities ρ is allowed to decrease (i.e. complementarities increase) from $\rho = 0.9$ to $\rho = 0.1$.

In Figure 14 Panel A, the labor market labor share of GDP is depicted for different values of the parameter ρ. As the degree of complementarity increases, the labor market share of GDP decreases, driven by increased specialization. Panel B plots the unweighted average of firm labor share. Increased complementarities decrease the labor share at each firm, especially at those characterized by a larger covariance between markdowns and workers' ability, as per equations of Proposition 2. Panel C plots the standard deviation of the firm average worker ability, as a measure of labor market segmentation. Panel C shows that as the degree of complementarity increases, workforce selection increases, augmenting segmentation in the labor market, measured by the market standard deviation of firm average workers log ability. Panel D shows the payroll HHI index defined as $\sum_{i \in j} s_{ij}^2$. In this example, as the labor market share declines, the market HHI index also declines, indicating a fall in the concentration of the firm total wage bill share. It is evident how the HHI does not necessarily increase when the labor share decreases, implying that measuring labor market concentration via HHI may not be appropriate. Intuitively, as complementarities increase, both wages and workforce selectivity increase, meaning that the total wage bill firm shares may become less concentrated.

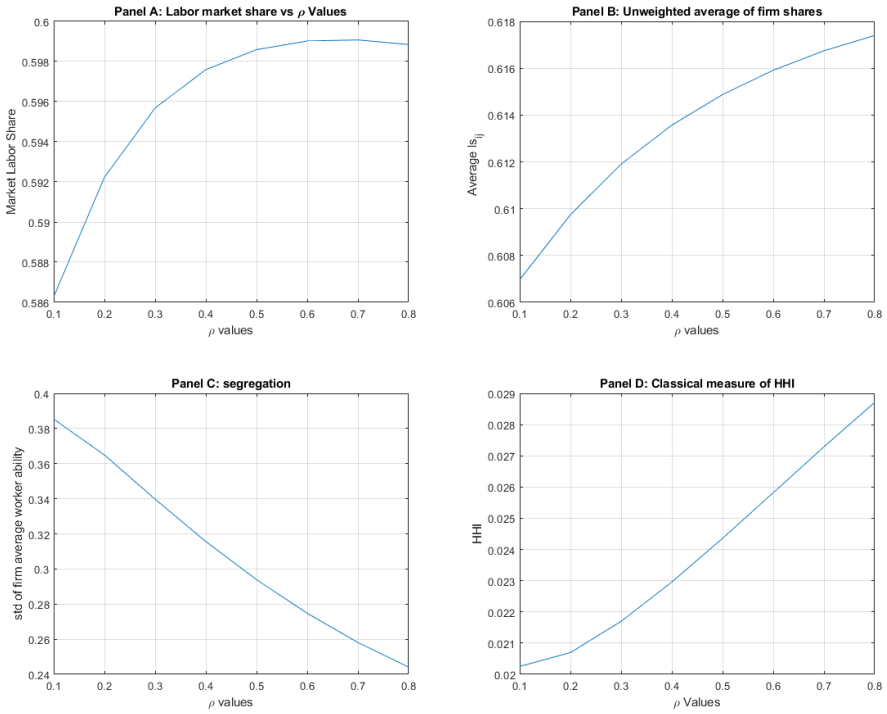


Figure 14: Labor share and complementarities

Notes: This figure is divided into four panels, labeled A, B, C, and D. Panel A illustrates the relationship between the market labor share and increasing complementarities, measured by decreasing ρ. Panel B displays the relationship with the unweighted average firm-level labor share ls_{ij}. Panel C plots the relationship with segregation, measured as the standard deviation of firm-level average worker ability. Panel D shows the relationship with the market HHI index. All parameters are set at the baseline level of Table 4, with the exception of ω, set at 0.7, and ρ, set at each iteration at different values in the range 0.9-0.1.

From Proposition 2, the ratio of the average wage to the average marginal product identifies $\widetilde{\psi}_{ij} := \overline{\mu}_{ij} + \frac{\text{cov}_{ij}(\mu,\phi)}{\alpha\gamma\Phi_{ij}} \leq 1$. This measure distorts the firm's labor share and profits[26]. Moreover, it is effectively what is estimated using the production function approach to assess labor market power, as in Yeh et al. (2022), where the average wage is compared to the average marginal product.

The term $\widetilde{\psi}_{ij}$ incorporates the average markdown as well as a measure of covariance between the markdown and worker output. Consequently, in firms where the distribution skews toward high-ability workers, this covariance will tend to be more negative than in firms where the worker ability distribution is skewed toward lower-ability workers. Figure 15 illustrates the firm's average markdown and the measure $\widetilde{\psi}_{ij}$ plotted against log exogenous productivity in an example market with one hundred firms. The gap between the two measures represents the contribution of this

[26]Profits, conditional on the average marginal product

covariance term. It is evident that this covariance term potentially introduces a negative correlation between $\widetilde{\psi}_{ij}$ and log productivity. The correlation coefficient measured in this example is -0.63. In contrast, the correlation between the average markdown and firm log productivity in this example is -0.01, as reflected in the hump-shaped relationship shown in the figure. Thus, it is clear that this covariance term introduces a wedge between the measurement of markdown and log productivity, confounding correlation measures between the two. Nonetheless, $\widetilde{\psi}_{ij}$ is the relevant measure for assessing the implications of labor market power on labor share and firm profits and thus can be used to interpret the implications of markdowns for these quantities.

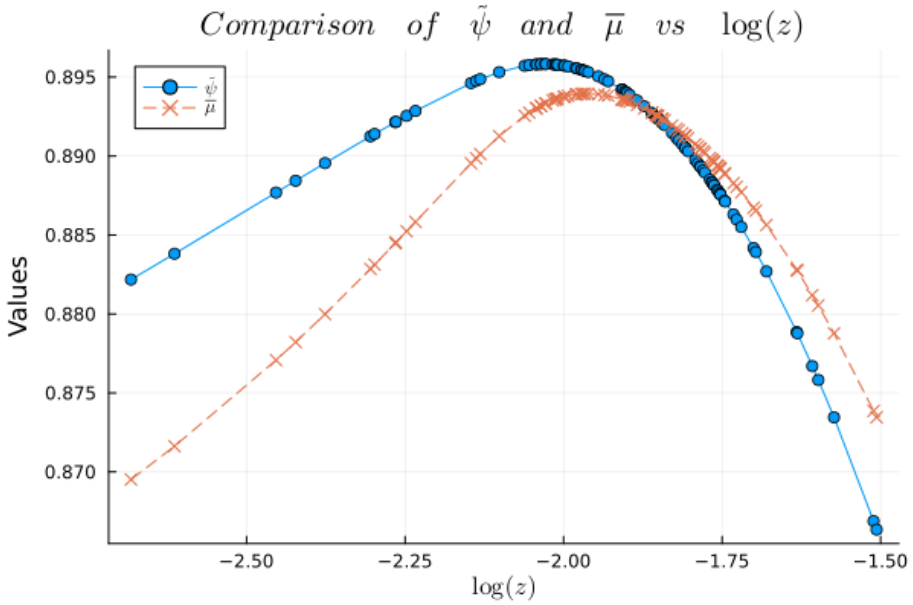


Figure 15: Firm average markdown and $\widetilde{\psi}_{ij}$ against log productivity

Notes: This figure illustrates the relationship between the firm average markdown $\overline{\mu}$, and the measure $\widetilde{\psi}_{ij}$ against firm exogenous log productivity. All parameters are set at the baseline level of table 4.

7 Conclusion

This paper develops a novel general equilibrium model to study the determinants of labor market power via labor market sorting and segmentation. The general equilibrium model integrates labor market sorting and segmentation, firm workforce selection, preference heterogeneity, and labor market power through strategic wage setting. Central to the model is the mechanism of Endogenous Oligopsony, where firms' selective hiring practices create labor market segmentation and localized competition. This new mechanism has significant implications for both aggregate production efficiency and the distribution of welfare.

Empirical evidence motivates the model, showing that low-type workers are disproportionally employed in low type firms despite these firms being smaller. Moreover empirical evidence show that high-wage firms employ more stringent hiring thresholds, leading to stronger sorting and segmentation. The model captures these dynamics, demonstrating that neglecting firms' selective hiring leads to an overestimation of production inefficiencies due to labor market power. It also reveals substantial heterogeneity in welfare losses across different segments of the workforce, with significant welfare gains accruing to entrepreneurs.

The model contributes to several key areas of economic research. First, it studies the determinant of labor market power via labor market sorting and segmentation, prominent features of labor markets. Second, it provides a unified framework that rationalizes empirical findings on earnings inequality, sorting, and segmentation within labor markets. Third, it extends the literature on preference heterogeneity by incorporating firm selectivity into workforce composition, thereby offering new insights into the transmission of productivity shocks to workers' wages. Fourth, the model has several implications for the empirical measurement of markdowns and for assessing the implications of labor market power for the aggregate labor share of GDP.

Ongoing research builds on this framework, more carefully calibrating the model with data used for the motivating evidence, and calibrating the labor supply elasticities using a trade shock as a demand shifter. Future work will further explore the model's implications for policy interventions, such as minimum wage laws, and their effects on production efficiency and welfare distribution.

Contents

1 Introduction **2**

2 Motivating Evidence **9**

2.1 Conceptual Framework 9

2.2 Data 10

2.3 Size Premium 13

2.4 Strong Sorting 16

2.5 Screening Thresholds 20

3 The Model **25**

3.1 Environment 25

3.2 Workers 25

3.3 Representative Entrepreneur 28

3.4 Firms 28

4 Equilibrium Analysis **30**

4.1 Partial Equilibrium 30

4.2 Market Equilibrium 36

4.3 General Equilibrium 37

5 Preliminary Calibration **39**

6 Inefficiencies and Welfare **41**

6.1 Aggregate Inefficiencies 41

6.2 Welfare Distribution 46

6.3 Measurements: Labor Share and Firm Profits 49

7 Conclusion **53**

A Data **1**

A.1 SIEED 1

A.2 Descriptive Statistics 2

B Theory **5**

B.1 Production Function: micro-foundation 5

B.2 Mathematical Derivations 7

B.2.1 Worker Problem 7

B.2.2 Firm Problem 10

B.3 Proofs 12

B.3.1 General Equilibrium 19

B.4 Planner's Problem 20

C Simulations **21**

C.1 Description 21

C.2 Solution Algorithm 23

C.3 Market Size Inefficiency 24

C.4 Identifying Types in Data 25

References

Abowd, J. M., Kramarz, F., & Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, *67*(2), 251–333.

Atkeson, A., & Burstein, A. (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, *98*(5), 1998–2031.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & Van Reenen, J. (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics*, *135*(2), 645–709.

Autor, D. H., & Scarborough, D. (2008). Does job testing harm minority workers? evidence from retail establishments. *The Quarterly Journal of Economics*, *123*(1), 219–277.

Azar, J. A., Berry, S. T., & Marinescu, I. (2022). *Estimating labor market power* (tech. rep.). National Bureau of Economic Research.

Barron, J. M., Bishop, J., & Dunkelberg, W. C. (1985). Employer search: The interviewing and hiring of new employees. *The Review of Economics and Statistics*, 43–52.

Bender, S., Bloom, N., Card, D., Van Reenen, J., & Wolter, S. (2018). Management practices, workforce selection, and productivity. *Journal of Labor Economics*, *36*(S1), S371–S409.

Berger, D., Herkenhoff, K., & Mongey, S. (2022a). Labor market power. *American Economic Review*, *112*(4), 1147–93.

Berger, D. W., Herkenhoff, K. F., Kostøl, A. R., & Mongey, S. (2023). *An anatomy of monopsony: Search frictions, amenities and bargaining in concentrated markets* (tech. rep.). National Bureau of Economic Research.

Berger, D. W., Herkenhoff, K. F., & Mongey, S. (2022b). *Minimum wages, efficiency and welfare* (tech. rep.). National Bureau of Economic Research.

Burdett, K., & Mortensen, D. T. (1998). Wage differentials, employer size, and unemployment. *International Economic Review*, 257–273.

Card, D. (2022). Who set your wage? *American Economic Review*, *112*(4), 1075–90.

Card, D., Cardoso, A. R., Heining, J., & Kline, P. (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics*, *36*(S1), S13–S70.

Card, D., Heining, J., & Kline, P. (2013). Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly journal of economics*, *128*(3), 967–1015.

Carrillo-Tudela, C., Gartner, H., & Kaas, L. (2023). Recruitment policies, job-filling rates, and matching efficiency. *Journal of the European Economic Association*, *21*(6), 2413–2459.

Dal Bó, E., Finan, F., & Rossi, M. A. (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. *The Quarterly Journal of Economics*, *128*(3), 1169–1218.

David, J. M., Hopenhayn, H. A., & Venkateswaran, V. (2016). Information, misallocation, and aggregate productivity. *The Quarterly Journal of Economics*, *131*(2), 943–1005.

David, J. M., & Venkateswaran, V. (2019). The sources of capital misallocation. *American Economic Review*, *109*(7), 2531–2567.

De Loecker, J., Eeckhout, J., & Unger, G. (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, *135*(2), 561–644.

Eeckhout, J., & Kircher, P. (2018). Assortative matching with large firms. *Econometrica*, *86*(1), 85–132.

Felix, M. (2021). Trade, labor market concentration, and wages. *Job Market Paper*.

Garen, J. E. (1985). Worker heterogeneity, job screening, and firm size. *Journal of political Economy*, *93*(4), 715–739.

Gennaioli, N., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2013). Human capital and regional development. *The Quarterly journal of economics*, *128*(1), 105–164.

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

Helpman, E., Itskhoki, O., & Redding, S. (2010). Inequality and unemployment in a global economy. *Econometrica*, *78*(4), 1239–1283.

Hsieh, C.-T., & Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, *124*(4), 1403–1448.

Karabarbounis, L., & Neiman, B. (2014). The global decline of the labor share. *The Quarterly journal of economics*, *129*(1), 61–103.

Kirov, I., & Traina, J. (2021). *Labor market power and technological change in us manufacturing* (tech. rep.). Working paper, University of Chicago, Chicago, IL, November 11.

Kline, P., Petkova, N., Williams, H., & Zidar, O. (2019). Who profits from patents? rent-sharing at innovative firms. *The quarterly journal of economics*, *134*(3), 1343–1404.

Kremer, M. (1993). The o-ring theory of economic development. *The quarterly journal of economics*, *108*(3), 551–575.

Lamadon, T., Mogstad, M., & Setzler, B. (2022). Imperfect competition, compensating differentials, and rent sharing in the us labor market. *American Economic Review*, *112*(1), 169–212.

Lochner, B., Seth, S., & Wolter, S. (n.d.). Fdz-methodenreport.

Manning, A. (2003). The real thin theory: Monopsony in modern labour markets. *Labour economics*, *10*(2), 105–131.

Manning, A. (2021). Monopsony in labor markets: A review. *ILR Review*, *74*(1), 3–26.

Mas, A., & Moretti, E. (2009). Peers at work. *American Economic Review*, *99*(1), 112–145.

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

Moretti, E. (2004). Workers' education, spillovers, and productivity: Evidence from plant-level production functions. *American Economic Review*, *94*(3), 656–690.

Pellegrino, B. (2019). Product differentiation and oligopoly: A network approach. *WRDS Research Paper*.

Restuccia, D., & Rogerson, R. (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic dynamics*, *11*(4), 707–720.

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

Schmidtlein, L., Seth, S., & Vom Berge, P. (2020). *Sample of integrated employer employee data (sieed) 1975-2018* (tech. rep.). Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for ...

Sharma, G. (2023). Monopsony and gender. *Unpublished Manuscript*.

Shimer, R., & Smith, L. (2000). Assortative matching and search. *Econometrica*, *68*(2), 343–369.

Song, J., Price, D. J., Guvenen, F., Bloom, N., & Von Wachter, T. (2019). Firming up inequality. *The Quarterly journal of economics*, *134*(1), 1–50.

Sorkin, I. (2018). Ranking firms using revealed preference. *The quarterly journal of economics*, *133*(3), 1331–1393.

Teulings, C. N. (1995). The wage distribution in a model of the assignment of skills to jobs. *Journal of political Economy*, *103*(2), 280–315.

Yeh, C., Macaluso, C., & Hershbein, B. (2022). Monopsony in the us labor market. *American Economic Review*, *112*(7), 2099–2138.

A Data

A.1 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). In ensuring methodological rigor, I utilize publicly accessible code from Card, Heining, and Kline (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, and Kline (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, and Kline (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, and Kline (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. .

A.2 Descriptive Statistics

This section presents descriptive statistics for the processed SIEED dataset used in the main analysis, including distributions and sample characteristics for firm and worker fixed effects, employment, wage residuals, and firm size. Each table highlights critical information relevant to the analysis.

Tables A.1 and A.2 provide a summary of firm-level and worker characteristics by year group. Table A.1 presents the mean log wage and years of schooling, averaged at the firm level, for each year group. Table A.2 provides average values of individual worker characteristics, including years of schooling, age, and work experience.

The year groups represent distinct periods, allowing us to observe trends in wages, education levels, and other attributes over time. The 'AKM year group' categories, ranging from 2 to 6, correspond to increasing time periods, which help identify structural changes in the labor market.

Table A.1: 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 A.2: 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

To understand the extent of missing values in firm and worker fixed effects, Tables A.3 and A.4 provide the share of missing firm and worker fixed effects (FFE and WFE, respectively). FFE_o and WFE_o are the firm and worker fixed effect provided by the IAB and estimated on the full administrative data sample. FFE and WFE refer to AKM firm and worker fixed effect estimated assuming that a firm ID is a combination of occupation and establishment ID.

Table A.3: Missing Firm Fixed Effects (FFE)

	Total Observations	Share Missing (FFE_o)	Share Missing (FFE)
Sample	3,473,220	1.68%	28.74%

Table A.4: Missing Worker Fixed Effects (WFE)

	Total Observations	Share Missing (WFE_o)	Share Missing (WFE)
Sample	2,525,434	3.43%	6.77%

Table A.5 presents the standard deviations of log wages for each year group. The variable `log_wage` represents the raw log wage data. `logwage2_resid` refers to wages residualized based on a Mincerian regression, controlling for a polynomial in age and tenure interacted with education, along with year fixed effects. `logwage2_resid2` further adds an occupation fixed effect to this regression. The standard deviations are computed annually and then averaged over each sample period.

Table A.6 decomposes the variance in residual earnings by displaying within- and between-firm components. This decomposition provides insight into the variation attributable to differences within firms over time (within-firm variance) and across firms (between-firm variance), with total variance (`tot_var`) included as a benchmark. The variance components are averaged by year group to summarize changes in earnings dispersion over time.

Both tables indicate an increase in wage variance over time, driven primarily by the between-firm component. Notably, the majority of the standard deviation in log wages persists even after

controlling for observable characteristics.

Table A.5: Standard Deviation of Residuals and Log Wages by Year Group

Year Group	SD (logwage2_resid)	SD (logwage2_resid2)	SD (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

Table A.6: 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 A.7 presents the standard deviation of log employment across firms by year group. The first column displays the standard deviation of log employment, representing firm size variability over time. The second column, labeled "SD (Log Employment Residual)," shows the standard deviation of log employment residuals after controlling for firm observables, including the share of college graduates, share of low-skill jobs, a polynomial in worker age, and average age of workers.

Table A.7: Firm Size and Employment Variability

Year Group	SD (Log Employment)	SD (Log Employment Residual)
2	1.0175	0.9680
3	1.0123	0.9595
4	0.9904	0.9438
5	0.9903	0.9510
6	0.9926	0.9550

Table A.8 presents the dispersion in firm and worker fixed effects by year group. The table reports the standard deviation of firm fixed effects (`FFE` and $\texttt{FFE}_o$) and worker fixed effects (`WFE` and $\texttt{WFE}_o$) over time. The measures $\texttt{FFE}_o$ and $\texttt{WFE}_o$ represent the fixed effects provided by the IAB, while `FFE` and `WFE` represent fixed effects estimated using occupation-establishment as firm ID.

Table A.8: Firm and Worker Fixed Effects Dispersion by Year Group

Year Group	SD (FFE_o)	SD (FFE)	SD (WFE_o)	SD (WFE)
2	0.2221	0.3257	0.2717	0.2067
3	0.2263	0.3207	0.2892	0.2011
4	0.2398	0.3291	0.3220	0.1987
5	0.2794	0.3674	0.3605	0.1967
6	0.2453	0.3609	0.3870	0.1994

Table A.9 displays the five most frequent occupations among low-skilled and high-skilled jobs. The classification of a job as low or high skill is based on the fifth digit of the occupation code, meaning that an occupation may be represented in both skill groups depending on the specific job. Here, I report the five most frequent occupations in both low-skill and high-skill job categories. It is evident that low-skill jobs typically correspond to occupations requiring less formal education or specialized training, while high-skill jobs are associated with occupations that demand advanced skills and knowledge, often acquired through higher education or technical training.

Table A.9: 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

B Theory

B.1 Production Function: micro-foundation

The micro-foundation of this production technology builds upon the frameworks discussed in Eeckhout and Kircher (2018) and Helpman et al. (2010).

Consider each firm producing a single product, denoted as y. Each product is manufactured by a team, conceptualizing each firm as a manager of quality z hiring a team of workers with varying abilities a specialized in a given occupation[27]. Automotive companies have teams of me-

[27]In practice, a firm's products are manufactured by completing and assembling the outputs of various tasks within the company. Task based production function, as in Teulings (1995) and Haanwinckel (2023), introduce significant complexity. Incorporating task-based production in this model is left as an area for future research.

chanical engineers collaborating to design and optimize the mechanical components of automobiles. Biomedical companies have teams of scientists jointly researching and developing medical devices to advance healthcare. Facilities have janitorial teams working together to ensure cleanliness and maintenance.

Assume each worker produces according to a function $f(a, z, \chi)$, where χ represents the resources allocated to the worker and ω_χ the returns on the resources χ:

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

Unlike Eeckhout and Kircher (2018), the assumption here is that firms cannot discriminate based on workers' ability a when allocating resources χ. To fix the idea, firms rent office spaces k at a rental price R and assign an office space equally to each worker. Additionally, assume $\omega_\chi = 1-\gamma$, capturing decreasing returns on the resource allocation. Thus, the worker-level production function follows:

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

The final output is derived by summing the output produced by all employed workers, resulting in the production function:

$$y = \int_{\underline{a}}^{\overline{a}} \phi(z, a)\left(\frac{k}{\int_{\underline{a}}^{\overline{a}} n_{ij}(a)da}\right)^{1-\gamma} n(a)da = \Phi_{ij}(z_{ij}, \mathbf{a})k^{1-\gamma}h^{\gamma} \tag{B.1}$$

Discussion- In Helpman et al. (2010), the considered resource is managerial time, and discrimination is precluded as firms cannot observe workers' ability a. In contrast, in my model, a is observed by the firm, but a rule exists to prevent managers from reallocating this resource to different workers. Examples of such rules may encompass legal or regulatory constraints, social norms[28], or an equal-treatment firm culture[29].

The significance of interpreting the equally split resource as office space and a as unobservable ability becomes evident. Assuming that the equally split resource is supervision time may be unsuitable because time is easily modifiable by the manager. Assuming that resources are split equally regardless of workers' observable ability such as educational attainment might pose chal-

[28] Even in the absence of explicit laws, social norms or corporate policies against discrimination could constrain managers from differentiating resource allocation.

[29] Certain companies or organizations actively cultivate a fair or equal treatment culture as an ethical value, constraining managers from differentiating.

lenges in justification. Nevertheless, given the interpretation of a as unobservable ability, I consider this assumption more realistic compared to the assumption of all resources being freely reallocatable based on workers' abilities, as in Eeckhout and Kircher (2018). To illustrate, consider the following two familiar examples. A research university typically assigns the same office space to each professor, irrespective of their research ability. Similarly, a research university assigns the same office space/desk and training to each PhD student, regardless of their ability. As a glimpse into what follows, this assumption, coupled with complementarities between worker ability a and managerial productivity z[30], creates incentives for firms to selectively choose their workforce, as observed empirically (Bender et al. (2018)). As it will become clear in the next subsection, due to complementarities in ϕ, more productive firms have a larger $\Phi_{ij}(z_{ij}, \mathbf{a})$ and, therefore, will extend job offers only to more able workers. As a consequence, high-ability workers disproportionately sort into high-paying, more productive, and better-managed firms.[31]

Lastly, it is worth mentioning that this production function nests a series of production functions commonly used in the literature. In particular, consider the functional form for ϕ in Equation 3.9, let $\xi = 1$, and examine the following variations in the parameters ρ and ω_a. If $\omega_a = 0$, the production function collapses to the classic Cobb-Douglas $y = zk^{1-\gamma}h^{\gamma}$, also employed in D. Berger et al. (2022a). As $\rho \to 1$, the functional form becomes $z^{1-\omega_a}h^{\gamma}k^{1-\gamma}\frac{\int_{\underline{a}}^{\overline{a}} a^{\omega_a} n(a)da}{h}$, similar to the one used in Helpman et al. (2010). I view this production function as adding an additional layer rather than imposing a constraint on the production function used in previous studies. The parameters will be calibrated differently across different job occupations to match the data, and potentially, the model can revert back to the simple Cobb-Douglas production function, as in D. Berger et al. (2022a).

B.2 Mathematical Derivations

B.2.1 Worker Problem

The micro-foundation of the Nested-CES labor disutility index is adapted from D. Berger et al. (2022a). The difference is that the choice set of workers with ability a is endogenous and denoted $\mathcal{S}_j(a)$.

[30] In this formulation, the complementarity is with respect to the firm managerial productivity z. This can be interpreted literally or as a reduced form to model complementarities with certain types of inputs that are more extensively used in more productive firms. For instance, these inputs may encompass advanced IT technologies, organizational processes, skill-biased capital equipment, etc.

[31] For other studies on firm screening and recruitment policies see D. H. Autor and Scarborough (2008); Barron et al. (1985); Bender et al. (2018); Garen (1985). In particular, Bender et al. (2018) finds that more productive and better-managed firms build a superior stock of employees by selecting higher-ability employees and firing lower-ability employees to a greater extent than other firms.

For each worker type a there is a unit measure of ex-ante identical individuals indexed by $l \in [0,1]$. Each individual has random preferences for working at each firm ij. Their disutility of labor supply is convex in hours worked $h_l(a)$. Worker l of ability a disutility of working $h_{lij}(a)$ hours at firm ij is given by:

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

where the random utility term $\zeta_{lij}(a)$ is distributed independently and identically across individuals according to the following multivariate Nested Gumbel distribution:

$$F(\xi_{i1}, \ldots, \xi_{NJ}) = \exp\left(-\sum_{j=1}^{J} \left(\sum_{i=1}^{M_j} e^{-(1+\eta)\zeta_{ij}^{1+\theta}} \right)^{\frac{1+\theta}{1+\eta}} \right)$$

Each household l of ability a is also earning some income that must satisfy $w_{ijl}(a)h_{ijl}(a) = Y_l(a)$, distributed according to some distribution $F(Y_l(a))$. Then, after drawing their vector $\{\zeta_{lij}(a)\}$, each worker solves:

$$\min_{ij} \{\log h_{lij}(a) - \zeta_{lij}(a)\} \equiv \max_{ij} \{\log w_{ij}(a) - \log y_l(a) - \zeta_{lij}(a)\}$$

Denote by $\mathcal{S}_j(a)$ the choice set of workers of ability a in market j. This is equivalent to the set of firm in the market extending a positive wage $w_{ij}(a)$ to the worker. Then, following McFadden et al. (1973), the household chooses the firm that maximizes her utility. Given the distribution of the taste shock, the probability that worker l of ability a chooses to work for firm ij is:

$$\begin{aligned} Prob_l(w_{ij}(a), w_{-ij}(a)) &= \frac{w_{ij}(a)^{1+\eta}}{\sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)^{1+\eta}} \frac{\left(\sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)^{1+\eta}\right)^{\frac{1+\theta}{1+\eta}}}{\sum_{j=1}^{J} (\sum_{k\in\mathcal{S}_j(a)} w_{kj}(a))^{\frac{1+\theta}{1+\eta}}} \\ &= \frac{w_{ij}(a)^{1+\eta}}{\sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)^{1+\eta}} \frac{\left(\sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)^{1+\eta}\right)^{\frac{1+\theta}{1+\eta}}}{\int_0^1 (\sum_{k\in\mathcal{S}_j(a)} w_{kj}(a))^{\frac{1+\theta}{1+\eta}} dj} \end{aligned}$$

This probability is the same for every worker of the same ability a. Let this probability be denoted $p_{ij}(a)$. To obtain total employment of workers of ability a in firm ij first define the following indexes:

$$w_j(a) = \left[\sum_{i\in\mathcal{S}_j(a)} w_{ij}^{1+\eta} \right]^{\frac{1}{1+\eta}}$$

$$W(a) = \left[\int_0^1 w_j^{1+\theta}\, dj\right]^{\frac{1}{1+\theta}}$$

$$N(a) \coloneqq \left[\int_0^1 n_j(a)^{\frac{\theta+1}{\theta}}\, dj\right]^{\frac{\theta}{\theta+1}}$$

$$n_j(a) \coloneqq \left[\sum_{i\in J_j(a)} n_{ij}(a)^{\frac{\eta+1}{\eta}}\right]^{\frac{\eta}{\eta+1}}, \quad \eta > \theta > 0,$$

Then, find the total fraction of workers of ability a working for firm ij as follows:

$$n_{ij}(a) = \int p_{ij}(a) h_{lij}(a) dF(Y_l(a)) = \frac{w_{ij}(a)^{1+\eta}}{w_j(a)^{1+\eta}} \frac{w_j(a)^{1+\theta}}{W(a)^{1+\theta}} \frac{1}{w_{ij}(a)} \int Y_l(a) dF(Y_l(a))$$

Notice that $\int Y_l(a) dF(Y_l(a))$ is the total income of households of type a. This is also equal to $W(a)N(a)$:

$$W(a)N(a) = \left[\int_0^1 \Big(\sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)^{1+\eta}\Big)^{\frac{1+\theta}{1+\eta}}\right]^{\frac{1}{1+\theta}} \left[\int_0^1 \Big(\sum_{i\in\mathcal{S}_j(a)} n_{ij}(a)\Big)^{\frac{\eta}{\eta+1}\frac{\theta+1}{\theta}}\right]^{\frac{\theta}{\theta+1}}$$
$$= \int Y_l(a) dF(Y_l(a))$$

Where the last equality follows from substituting for $n_{ij}(a)$ and simplifying. Thus, substituting $W(a)N(a)$ in the expression for $n_{ij}(a)$, the firm labor supply function is obtained:

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

Therefore the result is that the supply curves that firms face in this model of individual discrete choice are equivalent to those that the firms face when a representative household solves the following income maximization problem:

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

Subject to the following constraints:

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

B.2.2 Firm Problem

Firms maximize profits taking R and the labor supply function as given, choosing capital k_{ij} and an employment schedule $n_{ij}(a) : [\underline{a}, \overline{a}] \to \mathbb{R}_+$:

$$\pi_{ij} = \max_{\{n_{ij}(a)\}, k_{ij}} \Phi_{ij}(z_{ij}, \mathbf{a})(h_{ij}^{\gamma} k_{ij}^{1-\gamma})^{\alpha} - Rk_{ij} - \int_{\underline{a}}^{\overline{a}} w_{ij}(a) n_{ij}(a) da \tag{B.2}$$

$$\text{s.t.} \quad w_{ij}(a) = \left(\frac{n_{ij}(a)}{n_j(a)}\right)^{\frac{1}{\eta}} \left(\frac{n_j(a)}{N(a)}\right)^{\frac{1}{\theta}} \frac{N(a)^{\frac{1}{\varphi}} g(a)^{1-\sigma}}{C(a)^{-\sigma}}$$

For ease of notation, I will denote the endogenous productivity term just Φ_{ij}. First order with respect to capital:

$$\begin{aligned} \alpha(1-\gamma)\Phi_{ij}(k^{1-\gamma}h^{\gamma})^{\alpha-1}\left(\frac{h_{ij}}{k_{ij}}\right)^{\gamma} &= R \\ \alpha(1-\gamma)y_{ij} &= k_{ij}R \\ k_{ij} &= \frac{\alpha(1-\gamma)y_{ij}}{R} \end{aligned}$$

Substituting back into the production function:

$$y_{ij} = \left(\frac{\alpha(1-\gamma)}{R}\right)^{\frac{(1-\gamma)\alpha}{1-(1-\gamma)\alpha}} \Phi_{ij}^{\frac{1}{1-(1-\gamma)\alpha}} h_{ij}^{\frac{\gamma\alpha}{1-(1-\gamma)\alpha}}$$

Define $\widetilde{y_{ij}} = (1-\alpha(1-\gamma))y_{ij}$. Define $Z = \left((1-\alpha(1-\gamma))\left(\frac{1-\gamma}{R}\right)^{\frac{1-\gamma}{\gamma}}\right)$ a constant equal to all firms. Thus, substituting capital demand back into firm profits, the maximization problem is isomorphic to a maximization problem subject to the following production function:

$$\widetilde{y}_{ij} = Z\Phi_{ij}^{\frac{1}{1-(1-\gamma)\alpha}} h^{\frac{\gamma\alpha}{1-(1-\gamma)\alpha}}$$

$$\widetilde{y}_{ij} = Z\left[\int_{\underline{a}}^{\overline{a}} \phi(z_{ij},a) n_{ij}(a) da\right]^{\frac{1}{1-(1-\gamma)\alpha}} h_{ij}^{\frac{\gamma\alpha-1}{1-(1-\gamma)\alpha}}$$

Taking the derivative with respect to a marginal increase in the employment of workers of ability a, the expression for the marginal product is obtained:

$$\begin{aligned} MPL_{ij}(a) =& Z\frac{1}{1-(1-\gamma)\alpha}\left[\int_{\underline{a}}^{\overline{a}} \phi(z_{ij},a) n_{ij}(a) da\right]^{\frac{1}{1-(1-\gamma)\alpha}-1} h_{ij}^{\frac{\gamma\alpha-1}{1-(1-\gamma)\alpha}} \phi(z_{ij},a) \\ &+ \frac{\gamma\alpha-1}{1-(1-\gamma)\alpha} h_{ij}^{\frac{\gamma\alpha-1}{1-(1-\gamma)\alpha}-1}\left[\int_{\underline{a}}^{\overline{a}} \phi(z_{ij},a) n_{ij}(a) da\right]^{\frac{1}{1-(1-\gamma)\alpha}} \end{aligned}$$

The first term of the sum can be rewritten as follows:

$$\begin{aligned} & Z\frac{1}{1-(1-\gamma)\alpha}\left[\int_{\underline{a}}^{\overline{a}} \phi(z_{ij},a) n_{ij}(a) da\right]^{\frac{1}{1-(1-\gamma)\alpha}-1} h_{ij}^{\frac{\gamma\alpha-1}{1-(1-\gamma)\alpha}} \phi(z_{ij},a) \\ &= Z\frac{1}{1-(1-\gamma)\alpha}\left[\int_{\underline{a}}^{\overline{a}} \phi(z_{ij},a) n_{ij}(a) da\right]^{\frac{1}{1-(1-\gamma)\alpha}} h_{ij}^{\frac{\gamma\alpha-1}{1-(1-\gamma)\alpha}} \frac{\phi(z_{ij},a)}{\int_{\underline{a}}^{\overline{a}} \phi(z_{ij},a) n_{ij}(a) da} \\ &= Z\frac{1}{1-(1-\gamma)\alpha}\Phi_{ij}^{\frac{1}{1-(1-\gamma)\alpha}} h_{ij}^{\frac{\gamma\alpha}{1-(1-\gamma)\alpha}-1} \frac{\phi(z_{ij},a)}{\Phi_{ij}} \\ &= Z\frac{1}{1-(1-\gamma)\alpha}\Phi_{ij}^{\frac{1}{1-(1-\gamma)\alpha}} h_{ij}^{\frac{\alpha-1}{1-(1-\gamma)\alpha}} \frac{\phi(z_{ij},a)}{\Phi_{ij}} \end{aligned}$$

The second term of the sum:

$$\begin{aligned} \frac{\gamma\alpha-1}{1-(1-\gamma)\alpha} h_{ij}^{\frac{\gamma\alpha-1}{1-(1-\gamma)\alpha}-1}\left[\int_{\underline{a}}^{\overline{a}} \phi(z_{ij},a) n_{ij}(a) da\right]^{\frac{1}{1-(1-\gamma)\alpha}} &= \\ = -\frac{1-\gamma\alpha}{1-(1-\gamma)\alpha} h_{ij}^{\frac{\gamma\alpha}{1-(1-\gamma)\alpha}-1} \Phi_{ij}^{\frac{1}{1-(1-\gamma)\alpha}} &= \\ = -\frac{1-\gamma\alpha}{1-(1-\gamma)\alpha} h_{ij}^{\frac{\alpha-1}{1-(1-\gamma)\alpha}} \Phi_{ij}^{\frac{1}{1-(1-\gamma)\alpha}} & \end{aligned}$$

Combining the two terms, the expression for the MPL of equation 4.3 is obtained:

$$MPL(a|\Phi_{ij}(z_{ij},\mathbf{a}),h) = Z\frac{\alpha\gamma}{1-\alpha(1-\gamma)}\Phi_{ij}(z_{ij},\mathbf{a})^{\frac{1}{1-\alpha(1-\gamma)}}h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}}\left[1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_{ij},a)}{\Phi_{ij}(z_{ij},\mathbf{a})}\right)\right]$$

B.3 Proofs

Recall the following things that will be important for the following proofs:
Log-supermodularity: the function $\phi(z,a)$ is a log supermodular function. Thus for any $a_1 > a_0$ and $z_1 > z_0$ the following holds:

$$\frac{\phi(z_1,a_1)}{\phi(z_1a_0)} > \frac{\phi(z_0a_1)}{\phi(z_0a_0)}$$

Supermodularity: the function $\phi(z,a)$ satisfy supermodularity (i.e. for any $a_1 > a_0$, $z_1 > z_0$ the following holds:

$$\phi(z_1,a_1) - \phi(z_1,a_0) > \phi(z_0,a_1) - \phi(z_0,a_0)$$

Proof: Follows straightforward from ϕ being log-supermodular and increasing in both a and z:

$$\frac{\phi(z_1,a_1)}{\phi(z_1a_0)} > \frac{\phi(z_0,a_1)}{\phi(z_0,a_0)}$$

$$\frac{\phi(z_1,a_1)-\phi(z_1,a_0)}{\phi(z_1a_0)} > \frac{\phi(z_0,a_1)-\phi(z_0,a_0)}{\phi(z_0,a_0)}$$

Which, since $\phi(z_1,a_0) > \phi(z_0,a_0)$, and since the two numerators are positive, implies:

$$\phi(z_1,a_1) - \phi(z_1,a_0) > \phi(z_0a_1) - \phi(z_0,a_0)$$

Positive Employment: With upward sloping labor supply, as long as $MPL_{ij} > 0$ then $n_{ij}(a) > 0$

Notation: For simplicity I will index with i rather than ij and Φ_i the $\Phi_{ij}(z_{ij},\mathbf{a})$ of firm i. Moreover $MPL_i(a)$ will denote the marginal product of labor of a worker of type a in firm i.

Lemma 1: Log-Supermodularity of Marginal Product of Labor
The marginal product of labor $MPL_i(a)$ satisfies log-supermodularity. Specifically, for any two worker abilities $a_1 > a_0$ and any two firm productivities $z_1 > z_0$, the following inequality holds:

$$\frac{MPL_1(a_1)}{MPL_0(a_1)} > \frac{MPL_1(a_0)}{MPL_0(a_0)}.$$

Proof: I am going to prove for the case when $MPL_1(a_0) > 0$ and $MPL_0(a_0) > 0$. The case when one of the two is ≤ 0 is a mirror proof of the one that follows.

Notice that the firm common term of the MPL of the two ability level ($Z\frac{\alpha\gamma}{1-\alpha(1-\gamma)}\Phi_i^{\frac{1}{1-\alpha(1-\gamma)}}h_i^{\frac{\alpha-1}{1-\alpha(1-\gamma)}}$) gets simplified. Then claim is proved if:

$$\frac{1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_1,a_1)}{\Phi_1}\right)}{1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_1,a_0)}{\Phi_1}\right)}>\frac{1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_0,a_1)}{\Phi_0}\right)}{1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_0,a_0)}{\Phi_0}\right)}$$

That is:

$$[1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_0,a_0)}{\Phi_0}\right)][1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_1,a_1)}{\Phi_1}\right)]>[1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_0,a_1)}{\Phi_0}\right)][1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_1,a_0)}{\Phi_1}\right)]$$

Expanding the multiplication and rearranging, the claim is proven if:

$$(1-\alpha\gamma)[\frac{\phi(z_0,a_1)}{\Phi_0}+\frac{\phi(z_1,a_1)}{\Phi_1}-\frac{\phi(z_1,a_1)}{\Phi_1}-\frac{\phi(z_0,a_0)}{\Phi_0}]+(\frac{1}{\alpha\gamma})[\frac{\phi(z_0,a_0)\phi(z_1,a_1)-\phi(z_1,a_0)\phi(z_0,a_1)}{\Phi_1\Phi_0}]>0$$

Rearranging:

$$(1-\alpha\gamma)[\frac{\Phi_1(\phi(z_0,a_1))}{\Phi_1\Phi_0}+\frac{\Phi_0(\phi(z_1,a_1))}{\Phi_0\Phi_1}-\frac{\Phi_0(\phi(z_1,a_1))}{\Phi_0\Phi_1}-\frac{\Phi_1(\phi(z_0,a_0))}{\Phi_1\Phi_0}]+(\frac{1}{\alpha\gamma})[\frac{\phi(z_0,a_0)\phi(z_1,a_1)-\phi(z_1,a_0)\phi(z_0,a_1)}{\Phi_1\Phi_0}]>0$$

So that $\Phi_1\Phi_0$ at the denominator simplifies. Expaning the multiplication:

$$\phi(z_1,a_1)[\frac{\phi(z_0,a_0)}{\alpha\gamma}-(1-\alpha\gamma)\Phi_0]+\phi(z_1,a_0)[(1-\alpha\gamma)\Phi_0-\frac{\phi(z_0,a_1)}{\alpha\gamma}]+(1-\alpha\gamma)\Phi_1[\phi(z_0,a_1)-\phi(z_0,a_0)]>0$$

Now, notice that since $MPL_0(a_0)$ and $MPL_1(a_0)$ are assumed to be both positive we have the following two relations:

$$\phi(z_0,a_0)\geq(1-\alpha\gamma)\Phi_0$$

$$\phi(z_1,a_0)\geq(1-\alpha\gamma)\Phi_1$$

Thus:

$$\begin{aligned}
&\phi(z_1,a_1)\left[\frac{\phi(z_0,a_0)}{\alpha\gamma}-(1-\alpha\gamma)\Phi_0\right] &&+\phi(z_1,a_0)\left[(1-\alpha\gamma)\Phi_0-\frac{\phi(z_0,a_1)}{\alpha\gamma}\right] &&+(1-\alpha\gamma)\Phi_1\left[\phi(z_0,a_1)-\phi(z_0,a_0)\right]>\\
&\phi(z_1,a_1)\left[\frac{\phi(z_0,a_0)}{\alpha\gamma}-\phi(z_0,a_0)\right] &&+\phi(z_1,a_0)\left[(1-\alpha\gamma)\Phi_0-\frac{\phi(z_0,a_1)}{\alpha\gamma}\right] &&+(1-\alpha\gamma)\Phi_1\left[\phi(z_0,a_1)-\phi(z_0,a_0)\right]>\\
&\phi(z_1,a_1)\phi(z_0,a_0)\frac{1-\alpha\gamma}{\alpha\gamma} &&+\phi(z_1,a_0)\left[(1-\alpha\gamma)\Phi_0-\frac{\phi(z_0,a_1)}{\alpha\gamma}\right] &&+(1-\alpha\gamma)\Phi_1\phi(z_0,a_1)-\phi(z_0,a_0)\phi(z_1,a\\
&\phi(z_1,a_1)\phi(z_0,a_0)\frac{1-\alpha\gamma}{\alpha\gamma} &&-\phi(z_1,a_0)\phi(z_0,a_1)(\frac{1-\alpha\gamma}{\alpha\gamma})+\underbrace{A}_{>0}
\end{aligned}$$

Thus the proof is completed if

$$\phi(z_1,a_1)\phi(z_0,a_0)\frac{1-\alpha\gamma}{\alpha\gamma}-\phi(z_1,a_0)\phi(z_0,a_1)(\frac{1-\alpha\gamma}{\alpha\gamma})>0$$

Which holds if:

$$\phi(z_1,a_1)\phi(z_0,a_0)>\phi(z_1,a_0)\phi(z_0,a_1)$$

Which holds by log-supermodularity of ϕ. The proof is completed.

Corollary 1: Log-Supermodularity of Employment Schedule
In a monopsonistic labor market, the employment schedule $n_i(a)$ also satisfies log-supermodularity. This implies that more productive firms are more likely to employ higher-ability workers.

Proof: In a monopsonistic labor market, the proof comes straightforward from log-supermodularity of $MPL_i(a)$. Indeed:

$$\frac{n_0(a_1)}{n_0(a_0)}>\frac{n_1(a_1)}{n_1(a_0)}\longleftrightarrow\frac{MPL_0(a_1)}{MPL_0(a_0)}>\frac{MPL_1(a_1)}{MPL_1(a_0)}$$

Corollary 2: Monotonicity of Endogenous Productivity
In a monopsonistic labor market, firm endogenous productivity Φ_i is monotonically increasing in firm exogenous productivity. That is, if $z_1>z_0$, then $\Phi_1>\Phi_0$.

Proof: The proof is by contradiction. Suppose $\Phi_0\geq\Phi_1$. Denote by $\widetilde{a}_i$ the a such that $MPL_i(\widetilde{a}=0)$ (i.e. the threshold of ability below which the firm doesn't make an offer).

First, notice that $\Phi_0\geq\Phi_1$ implies $\widetilde{a}_0>\widetilde{a}_1$. Indeed:

$$0=\left[1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_0,\widetilde{a}_0)}{\Phi_0}\right)\right]<\left[1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_1,\widetilde{a}_0)}{\Phi_0}\right)\right]<\left[1-\frac{1}{\alpha\gamma}\left(1-\frac{\phi(z_1,\widetilde{a}_0)}{\Phi_1}\right)\right]$$

Where the first inequality comes from $z_1>z_0$ and the second from the assumption $\Phi_1\leq\Phi_0$. Thus, $n_1(\widetilde{a_0})>0$ and $n_0(\widetilde{a_0})=0$.

Denote with $g_i(a)=\frac{n_i(a)}{h_i}$ the within firm density of workers of ability a. Notice that at $\widetilde{a_0}$, $g_1(\widetilde{a_0})>g_0(\widetilde{a_0})=0$. Also notice that by intermediate value theorem there exists an a_0 in a neighborhood of $\widetilde{a_0}$ such that $g_1(a_0)>g_0(a_0)>0$.
Moreover:

$$\Phi_1=\int_{\widetilde{a_1}}^{\overline{a}}\phi(z_1a)g_1(a)da=\int_{\widetilde{a_1}}^{\widetilde{a_0}}\phi(z_1a)g_1(a)da+\int_{\widetilde{a_0}}^{\overline{a}}\phi(z_1a)g_1(a)da$$

$$\Phi_0=\int_{\widetilde{a_0}}^{\overline{a}}\phi(z_0a)g_1(a)da$$

Thus, $\Phi_0 \geq \Phi_1$ implies

$$\int_{\widetilde{a_0}}^{\overline{a}} \phi(z_0 a) g_1(a) da > \int_{\widetilde{a_0}}^{\overline{a}} \phi(z_1 a) g_1(a) da \tag{B.3}$$

Where the strict inequality follows from the fact that $\int_{\widetilde{a_1}}^{\widetilde{a_0}} \phi(z_1 a) g_1(a) da > 0$. Since $z_0 < z_1$ the inequality (B.3) is possible only if $\exists a_1 > \widetilde{a_0} \quad : \quad g_0(a_1) > g_1(a_1)$ Combining with above:

$$\frac{g_0(a_1)}{g_0(a_0)} > \frac{g_1(a_1)}{g_1(a_1)} \longleftrightarrow \frac{n_0(a_1)}{n_0(a_0)} > \frac{n_1(a_1)}{n_1(a_0)} \longleftrightarrow \frac{MPL_0(a_1)}{MPL_0(a_0)} > \frac{MPL_1(a_1)}{MPL_1(a_0)} \tag{B.4}$$

Which contradicts log-supermodularity of MPL, proven above. Thus, we reached a contradiction and the claim is proven.

Proposition 1: Monotonicity of Screening Thresholds
In a monopsonistic labor market, firm screening thresholds $\widetilde{a}_i$ are increasing in firm exogenous productivity. Specifically, if $z_1 > z_0$, then $\widetilde{a}_1 > \widetilde{a}_0$, meaning that more productive firms set higher minimum ability thresholds for hiring.

Proof: The proof comes straightforward from the log-supermodularity of $n_i(a)$. Suppose the $\widetilde{a}_0 \geq \widetilde{a}_1$. This implies $n_0(\widetilde{a}_0) = 0$ and $n_1(\widetilde{a}_0) > 0$. Take a such that $n_0(a) > 0$ (which always exists). Then, we would get the following inequality that contradicts log-supermodularity of MPL:

$$\frac{n_0(a_1)}{n_0(\widetilde{a_0})} > \frac{n_1(a_1)}{n_1(\widetilde{a_0})}$$

A contradiction is reached and the claim is proven.

Firm-level elasticity: The firm-worker measure of markdown depends on the firm level labor supply elasticity $\epsilon_{ij}(a)$ as in equation 4.7. To obtain the markdown expression as a function of $\epsilon_{ij}(a)$, take the first order condition with respect to employment $n_{ij}(a)$, for a worker with positive $MPL_{ij}(a)$:

$$\begin{aligned} MPL_{ij}(a) &= w_{ij}(a) + n_{ij}(a) \frac{\partial w_{ij}(a)}{\partial n_{ij}(a)} \\ \frac{MPL_{ij}(a)}{w_{ij}(a)} &= 1 + n_{ij}(a) \frac{\partial w_{ij}(a)}{\partial n_{ij}(a)} \frac{1}{w_{ij}(a)} \\ \frac{MPL_{ij}(a)}{w_{ij}(a)} &= 1 + \frac{1}{\epsilon_{ij}(a)} \end{aligned}$$

Where $\epsilon_{ij}(a) := (\frac{\partial w_{ij}(a)}{\partial n_{ij}(a)} \frac{n_{ij}(a)}{w_{ij}(a)})^{-1}$. There is a closed formula for the elasticity $\epsilon_{ij}(a)$ in terms of firm market share for workers of ability a. First, it is possible to obtain the firm market share for workers of ability a as a function of employment:

$$
\begin{aligned}
s_{ij}(a) &:= \frac{w_{ij}(a)n_{ij}(a)}{\sum_{i\in\mathcal{S}_j(a)}} \\
&= \frac{\left(\frac{n_{ijt}(a)}{n_{jt}(a)}\right)^{\frac{1}{\eta}}\left(\frac{n_{jt}(a)}{N_t(a)}\right)^{\frac{1}{\theta}}\frac{N_t(a)^{\frac{1}{\varphi}}g(a)^{1-\sigma}}{C_t(a)^{-\sigma}}n_{ij}(a)}{\sum_{i\in\mathcal{S}_j(a)}\left(\frac{n_{ijt}(a)}{n_{tj}(a)}\right)^{\frac{1}{\eta}}\left(\frac{n_{jt}(a)}{N_t(a)}\right)^{\frac{1}{\theta}}\frac{N_t(a)^{\frac{1}{\varphi}}g(a)^{1-\sigma}}{C_t(a)^{-\sigma}}n_{ij}(a)} \\
&= \frac{n_{ij}(a)^{\frac{\eta+1}{\eta}}}{\sum_{i\in\mathcal{S}_j(a)}n_{ij}(a)^{\frac{\eta+1}{\eta}}}
\end{aligned}
$$

Then, we can relate $\epsilon_{ij}(a)$ to $s_{ij}(a)$ as follows:

$$
\begin{aligned}
\epsilon_{ij}(a) &:= \left[\frac{\partial \log w_{ij}(a)}{\partial \log n_{ij}(a)}\right]^{-1} \\
&= \left[\frac{1}{\eta} + \left(\frac{1}{\theta}-\frac{1}{\eta}\right)\frac{\partial \log n_j(a)}{\partial \log n_{ij}(a)}\right]^{-1} \\
&= \left[\frac{1}{\eta} + \left(\frac{1}{\theta}-\frac{1}{\eta}\right)s_{ij}(a)\right]^{-1}
\end{aligned}
$$

Where I used the fact that

$$
\frac{\partial \log n_j(a)}{\partial \log n_{ij}(a)} = \left(\sum_{i\in\mathcal{S}_j(a)} n_{ij}^{\frac{\eta+1}{\eta}}\right)^{\frac{\eta}{\eta+1}-1} n_{ij}^{\frac{1}{\eta}}\frac{n_{ij}(a)}{\left(\sum_{i\in\mathcal{S}_j(a)} n_{ij}^{\frac{\eta+1}{\eta}}\right)^{\frac{\eta}{\eta+1}}} = \frac{n_{ij}(a)^{\frac{\eta+1}{\eta}}}{\sum_{i\in\mathcal{S}_j(a)} n_{ij}^{\frac{\eta+1}{\eta}}}
$$

Since $\frac{1}{\theta}-\frac{1}{\eta}>0$, the larger the share the less elastic the firm-level labor supply and the larger the markdown. Without strategic interact (i.e. when the labor share of each firm approaches zero), the elasticity is completely determined by $\frac{1}{\eta}$.

Other firm-level quantities:
Here, I derive the expressions of equations of Proposition 2.

- $\overline{MPL}_{ij} = \frac{\alpha\gamma}{1-(1-\gamma)\alpha}\frac{y_{ij}}{h_{ij}}$:
  Start with the expression for the $MPL_{ij}(a)$ of a single worker of ability a. Then take the average:

$$\overline{MPL}_{ij} := \frac{\int_{\underline{a}}^{\overline{a}} Z \frac{\alpha\gamma}{1-\alpha(1-\gamma)} \Phi_{ij}(z_{ij},\mathbf{a})^{\frac{1}{1-\alpha(1-\gamma)}} h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}} \left[1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\phi(z_{ij},a)}{\Phi_{ij}(z_{ij},\mathbf{a})}\right)\right] n(a)da}{h_{ij}} = \\ = \frac{\alpha\gamma}{1-(1-\gamma)\alpha} \frac{y_{ij}}{h_{ij}}$$

- $MPL_{ij}(a) = \overline{MPL_{ij}}\psi_{ij}(a)$ where $\psi_{ij}(a) = \left(1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\phi_{ij}(a)}{\Phi_{ij}}\right)\right)$.
  This comes straightforward from the expression of the $MPL_{ij}(a)$

- $\overline{w} = \overline{MPL}_{ij}\widetilde{\psi}_{ij}$:

$$\overline{w}_{ij} = \int_{\underline{a}}^{\overline{a}} \frac{w_{ij}(a)n_{ij}(a)}{h_{ij}} = \\ \int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)MPL_{ij}(a)n_{ij}(a)}{h_{ij}} = \int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)\overline{MPL_{ij}}\psi_{ij}(a)n_{ij}(a)}{h_{ij}} = \\ \overline{MPL_{ij}} \int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)\psi_{ij}(a)n_{ij}(a)}{h_{ij}}$$

  The term in this expression defines $\widetilde{\psi}_{ij}$

$$\widetilde{\psi}_{ij} = \int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)\psi_{ij}(a)n_{ij}(a)}{h_{ij}}$$

- $\widetilde{\psi}_{ij} = \overline{\mu}_{ij} + \frac{\text{cov}_{ij}(\mu,\phi)}{\alpha\gamma\Phi_{ij}} \leq 1$:
  From the expression above take $\widetilde{\psi}_{ij}$:

$$\widetilde{\psi}_{ij} = \int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)\psi_{ij}(a)n_{ij}(a)}{h_{ij}} = \int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)(1 - \frac{1}{\alpha\gamma}(1 - \frac{\phi(z,a)}{\Phi_{ij}}))n_{ij}(a)}{h_{ij}}$$

  First notice that absent markdowns (i.e. $\mu_{ij}(a) = 1 \forall a$), $\widetilde{\psi}_{ij} = 1$ since

$$\int_{\underline{a}}^{\overline{a}} \frac{(1 - \frac{1}{\alpha\gamma}(1 - \frac{\phi(z,a)}{\Phi_{ij}}))n_{ij}(a)}{h_{ij}} = 1 - \frac{1}{\alpha\gamma}(1 - \frac{\Phi_{ij}}{\Phi_{ij}}) = 1$$

Since $\mu_{ij}(a) \leq 1$, $\widetilde{\psi}_{ij} \leq 1$ in the presence of markdowns. Expanding on $\widetilde{\psi}_{ij}$:

$$\widetilde{\psi}_{ij} = \int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)(1 - \frac{1}{\alpha\gamma}(1 - \frac{\phi(z,a)}{\Phi_{ij}}))n_{ij}(a)}{h_{ij}} = \int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)n_{ij}(a)}{h_{ij}}(1 - \frac{1}{\alpha\gamma}) + \int_{\underline{a}}^{\overline{a}} \mu_{ij}(a)\frac{\phi(z,a)}{\alpha\gamma h_{ij}\Phi_{ij}}n_{ij}(a) = \overline{\mu}_{ij} + \frac{\text{cov}_{ij}(\mu,\phi)}{\alpha\gamma\Phi_{ij}}$$

Where the last line used the fact that the term $\int_{\underline{a}}^{\overline{a}} \frac{\mu_{ij}(a)\phi(z_{ij},a)n_{ij}(a)}{h_{ij}}$ is the expected value of $\mathbb{E}_{ij}[\mu\phi]$ taken with respect to the within firm workers ability distribution.

- $\pi_{ij} = (\frac{1-\alpha(1-\gamma)}{\alpha\gamma} - \widetilde{\psi}_{ij})h\overline{MPL}$:
  To prove this statement recall that profits are

$$\begin{aligned}\pi_{ij} &= [1 - \alpha(1-\gamma)]y_{ij} - h_{ij}\overline{w}_{ij} = \\ &= [1 - \alpha(1-\gamma)]h_{ij}\overline{MPL}_{ij}\frac{1}{\alpha\gamma} - h\overline{MPL}_{ij}\widetilde{\psi}_{ij} = \\ &= (\frac{1-\alpha(1-\gamma)}{\alpha\gamma} - \widetilde{\psi}_{ij})h\overline{MPL}\end{aligned}$$

  Notice that absent markdowns profits are determined solely by the degree of decreasing returns to scale:

$$\pi_{ij} = (\frac{1-\alpha}{\alpha\gamma})h\overline{MPL}$$

  Notice that this is the same expression of profits under Cobb-Douglas production function with homogeneous workers.

- $ls_{ij} = \alpha\gamma\widetilde{\psi}_{ij}$:
  To prove this statement, recall that the firm labor share is defined as total wage payment over total revenue:

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

  Where the first line used the derivation of average marginal product as a function of output and average wage as a function of average marginal product. Again, notice that absent markdowns $\widetilde{\psi}_{ij} = 1$, and the labor share at the firm is constant and equal to $\alpha\gamma$, exactly the same as the labor share with a Cobb-Douglas production function with homogeneous workers.

B.3.1 General Equilibrium

The general equilibrium of this economy can be thought as a three layers structure. In General Equilibrium, each layer is in equilibrium. The first layer is the market equilibrium. The market equilibrium is defined *given* the market labor supply by worker ability. The market equilibrium is defined as market shares for each ability level $s_{ij}(a) \forall i \in j, \forall a \in [\underline{a}, \overline{a}]$ satisfying the following equations:

$$\forall a \in [\underline{a}, \overline{a}], \quad s_{ij}(a) = \frac{(\mu_{ij}(a, s_{ij}(a))MPL_{ij}(a))^{1+\eta}}{\sum_{i \in \mathcal{S}_j(a)} (\mu_{ij}(a, s_{ij}(a)MPL_{ij}(a))^{1+\eta}}$$

$$\forall a \in [\underline{a}, \overline{a}], \quad i \in J_j(a) \quad \text{if and only if} \quad MPL_{ij}(a) > 0$$

To obtain the formula I used the following relationship. First, notice that $w_{ij}(a) = \left(\frac{n_{ij}(a)}{n_j(a)}\right)^{\frac{1}{\eta}} \left(\frac{n_j(a)}{N(a)}\right)^{\frac{1}{\theta}} W(a)$ implies that $w_{ij}(a)^\eta = n_{ij}(a) x_j(a)$ where $x_j(a)$ is a common term to all firms in the market j. Thus:

$$s_{ij}(a) = \frac{n_{ij}(a)}{\sum_{i \in \mathcal{S}_j(a)} n_{ij}^{\frac{1+\eta}{\eta}}} = \frac{(w_{ij}^\eta x_j(a))^{\frac{1+\eta}{\eta}}}{\sum_{i \in \mathcal{S}_j(a)} (w_{ij}^\eta x_j(a))^{\frac{1+\eta}{\eta}}} = \frac{w_{ij}^{1+\eta}}{\sum_{i \in \mathcal{S}_j(a)} w_{ij}^{1+\eta}} = \frac{(\mu_{ij}(a) MPL_{ij}(a))^{1+\eta}}{\sum_{i \in \mathcal{S}_j(a)} (\mu_{ij}(a) MPL_{ij}(a))^{1+\eta}}$$

If the choice set $\mathcal{S}_j(a)$ is empty the share of unemployment is one for worker of ability a because their marginal product is negative at all firms.
As a point of remark notice that the market equilibrium is not block recursive meaning that it depends on the distribution of labor supply to the market.

The second layer is the distribution of labor supply to each market. Given aggregate employment $N(a)$, this consists of a set of market employment indexes $n_j(a) \quad \forall j \in [0, 1]$. A sufficient variable to obtain $n_j(a)$ is the wage share of the market:

$$s_j(a) = \frac{w_j(a) n_j(a)}{\int_0^1 w_j(a) n_j(a) dj} = \frac{w_j(a) (\frac{w_j(a)}{W(a)})^\theta N(a)}{\int_0^1 w_j(a) (\frac{w_j(a)}{W(a)})^\theta N(a) dj} = \frac{w_j(a)^{1+\theta}}{\int_0^1 w_j(a)^{1+\theta} dj}$$

The last layer is the aggregate labor supply index $N(a)$. From the first order condition of the worker:

$$\left(\frac{N(a)}{g(a)}\right)^{\frac{1}{\varphi}} = \left(\frac{C(a)}{g(a)}\right)^{-\sigma} W(a)$$

Rearranging:

$$\left(\frac{N(a)}{g(a)}\right)^{\frac{1}{\varphi}+\sigma} = W(a)^{1-\sigma}$$

Notice that in the special case of log-utility, labor supply is constant and equal to $g(a)$.

In general equilibrium, all these conditions must be satisfied at the same time where the wages are obtained from the wage. These are obtained above as:

$$w_j(a) = \left[\sum_{i\in\mathcal{S}_j(a)} w_{ij}^{1+\eta}\right]^{\frac{1}{1+\eta}}$$

$$W(a) = \left[\int_0^1 w_j^{1+\theta}\, dj\right]^{\frac{1}{1+\theta}}$$

Other G.E quantities: the other G.E. quantities can be obtained as follows:

- Worker a consumption can be obtained from the budget constraint: $C(a) = \int_0^1 \sum_{i\in\mathcal{S}_j(a)} w_{ij}(a)n_{ij}(a)dj = W(a)N(a)$ Where the last equality follows by definition
- Aggregate capital supply: $K = \int_0^1 \sum_{i=1}^{m_j} k_{ij}dj$. Recall that $k_{ij} = \frac{\alpha(1-\gamma)y_{ij}}{R}$. Hence:

$$K = \int_0^1 \sum_{i=1}^{m_j} \frac{\alpha(1-\gamma)y_{ij}}{R} dj = \frac{\alpha(1-\gamma)Y}{R}$$

  Which implies $RK = \alpha(1-\gamma)Y$.

B.4 Planner's Problem

In this section, I outline the social planner's problem, incorporating weights $\{\psi(a)\}$ and $\psi(e)$ to reflect the welfare contributions of workers and the representative entrepreneur, respectively. The planner seeks to allocate resources optimally to maximize aggregate welfare, taking into account both consumption and labor disutility. The objective function is given by:

$$\max \int_a \psi(a)\left[\frac{C(a)^{1-\sigma}}{(1-\sigma)g(a)} - \frac{N(a)^{1+\frac{1}{\varphi}}}{g(a)(1+\frac{1}{\varphi})}\right] + \psi(e)\frac{C(e)^{1-\sigma}}{1-\sigma}. \tag{B.5}$$

The planner's problem is subject to the following resource constraint:

$$\int_a C(a) + C(e) + K = \int_0^1 \sum_{i=1}^{M_j} \Phi_{ij}(h_{ij}^{\gamma} k_{ij}^{1-\gamma})^{\alpha} + (1-\delta)K, \tag{B.6}$$

where $C(a)$ and $C(e)$ denote the consumption of workers and the entrepreneur, respectively, K represents the capital stock, and Φ_{ij} captures firm-specific production parameters.

The first-order conditions (FOCs) for this problem, derived with respect to $n_{ij}(a)$, yield the following equilibrium condition:

$$\left(\frac{C(a)}{g(a)}\right)^{\sigma} \left(\frac{N(a)}{g(a)}\right)^{\frac{1}{\varphi}} \left(\frac{N_j(a)}{N(a)}\right)^{\frac{1}{\theta}} \left(\frac{n_{ij}(a)}{N_j(a)}\right)^{\frac{1}{\eta}} = MPL_{ij}(a), \tag{B.7}$$

where $MPL_{ij}(a)$ is the marginal product of labor for worker type a employed at firm ij. This condition establishes the relationship between workers' consumption, labor allocation, and the marginal productivity of labor.

Equation B.7 highlights that allocation of workers to firms is efficient in a decentralized equilibrium when workers receive a wage that is equal to their marginal product. In particular, notice that the left-hand-side of equation B.7 corresponds to the equation characterizing the inverse labor supply. Thus, equation B.7 can be rewritten as:

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

These results highlight that the efficient allocation of workers to firms is achieved in a decentralized equilibrium where workers are paid their marginal product of labor.

Regarding the distribution of welfare, I select weights such that the planner's allocation of resources corresponds to a decentralized equilibrium, where workers derive their income from wages and entrepreneurs from profits.

C Simulations

C.1 Description

Distribution of number of firms- I assume that there are 1000 markets and each market draws a random number of firms $m_j \sim G_m(m)$. I follow D. Berger et al. (2022a) in parameterizing the distribution $G_m(m)$. The distribution is a mixture of a discrete mass point at $m_j = 1$ and a Pareto

distribution over the support $m_j \in [2, \infty]$. I cap the number of firms per market to be one-hundred. Results are not sensitive to the number of markets or the cap on firms per market. The Pareto's mass point, shape, scale, and location parameters are in table A.10.

Table A.10: Pareto distribution Parameters

Mass at M_j	**Pareto Tail**	**Pareto Scale**	**Pareto Location**
0.16	0.71	38.36	2

Figure A.1 plots the resulting sample distribution of the drawn number of firms per markets.

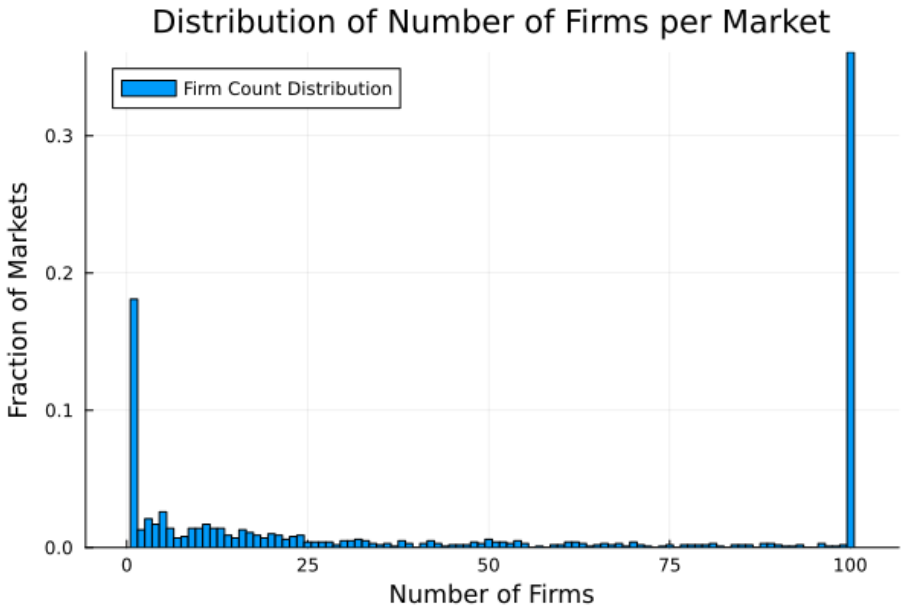


Figure A.1: Sampled number of firms per market

Notes: This figure illustrates the sampling distribution of number of firms per market. Calibration details are available in table A.10.

Distribution of workers abilities- Worker abilities were modeled using a log-normal distribution, which provides a continuous representation of ability levels across the population. To incorporate this distribution into our economic analysis, I discretized the distribution into 500 equal probability levels as follows. First, I truncated the log-normal distribution by removing the top and bottom 0.01% of abilities to eliminate extreme outliers. This allowed us to focus our analysis on the core 99.98% of the distribution. Next, I divided the truncated distribution into 500 ability levels with equal spacing between levels, such that each level represents 0.2% of the probability mass.

With 500 ability levels, I am able to capture the essential shape and variation of the log-normal distribution while enabling practical computation and analysis. By construction, each of the 500 ability levels has a width representing 0.2% probability mass or 0.002 probability. This implies that the probability is evenly spread across the 500 levels, with each level being assigned a probability of 0.002.

C.2 Solution Algorithm

I implement the following solution algorithm to obtain the general equilibrium allocation described in Section 4.3. The algorithm is initialized by guessing i) aggregate labor supply $N(a)$, ii) market shares $s_j(a)$, and iii) firm-level wage bill shares $s_{ij}(a)$ for each discretized worker ability. Note that the number of firm-level shares $s_{ij}(a)$ and market shares $s_j(a)$ has to be increased by one relative to the number of firms per market and the number of markets to account for (potential) unemployment. In equilibrium, the shares of unemployment will be zero because the representative household will not direct workers to markets where they are not employed. The quantities $N(a)$, $s_j(a)$, and $s_{ij}(a)$ will all be updated in the algorithm. The algorithm is characterized by three loops, with one inner loop nested within the other. The general equilibrium is a fixed point for each of these three loops combined.

1. First inner loop: solve for market shares given $s_j(a)$ and $N(a)$
   (a) Start with $s_{ij}^i(a)$.
   (b) Use the shares to compute $n_{ij}(a)$, Φ_{ij}, h_{ij}, and the implied $MPL_{ij}(a)$.
   (c) Use $s_{ij}(a)$ to compute the markdowns $\mu_{ij}(a)$.
   (d) Use $\mu_{ij}(a)$ and $MPL_{ij}(a)$ to compute the updated shares $s_{ij}^{i+1}(a) = \frac{(\mu_{ij}(a)MPL_{ij}(a))^{1+\eta}}{\sum_{i\in J_j(a)}(\mu_{ij}MPL_{ij}(a))^{1+\eta}}$. If $MPL_{ij}(a) < 0$, the updated share is $s_{ij}^{i+1}(a) = 0$.
   (e) Iterate from 1a to 1d until convergence $s_{ij}(a)^{i+1} \approx s_{ij}(a)^i \quad \forall a$.
2. Second inner loop: solve for $s_j(a)$
   (a) Start with $s_j^i(a)$.
   (b) Use $s_j^i(a)$ to compute the market equilibrium as per inner loop 1 described above.
   (c) Given the market equilibrium, compute wages $w_{ij}(a)$ for each firm in the market to compute $w_j(a) = (\sum_{i\in j} w_{ij}(a)^{1+\eta})^{\frac{1}{1+\eta}}$.
   (d) Update the new market share $s_j^{i+1}(a) = \frac{w_j(a)^{1+\theta}}{\int_0^1 w_j^{1+\theta}(a)dj}$.
   (e) Repeat from 2a to 2d until convergence.

3. Outer loop: solve for aggregate employment $N(a)$

   (a) Start with a guess of the total employment $N^i(a)$.

   (b) Use $N^i(a)$ to compute market shares $s_j(a)$ as per inner loop 2 described above.

   (c) Given the market shares $s_j(a)$, compute the aggregate wage index $W(a) = (\int_0^1 w_j(a)^{1+\theta})^{\frac{1}{\theta}}$. Compute the implied consumption $C(a) = W(a)N(a)$.

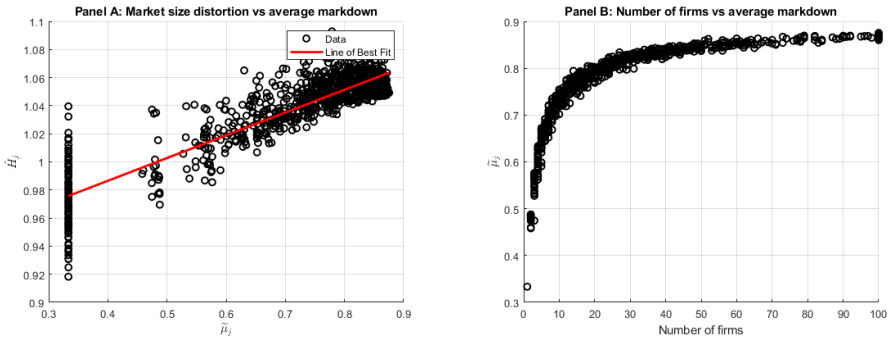
   (d) Update the new total employment $N^{i+1}(a)^{\frac{1}{\varphi}} = \frac{W(a)g(a)^{\sigma-1}}{C(a)^{\sigma}}$.

   (e) Repeat from 3a to 3d until convergence.

C.3 Market Size Inefficiency

This inefficiency is associated with a wedge in the labor supply to market j between the two allocations. Let $b_j = \frac{\sum_{i=1}^{M_j} h_{ij}}{\sum_{i=1}^{M_j} h_{ij,\text{eff}}}$ represent the distortion in total employment in market j. Figure A.2, Panel A, illustrates the relationship between this measure of distortion and the weighted average market markdown $\tilde{\mu}_j$, defined as $\tilde{\mu}_j = \sum_{i=1}^{M_j} \mu_{ij}\omega_{ij}$, where $\omega_{ij} = \frac{h_{ij}}{\sum_{i=1}^{M_j} h_{ij}}$. There is a clear positive relationship between the two quantities. Markets with the lowest $\tilde{\mu}_j$ are between 8% under-resourced and 4% over-resourced. Markets with the highest $\tilde{\mu}_j$ are over-resourced by between 4% and 8%.

Furthermore, Panel B of Figure A.2 depicts the relationship between $\tilde{\mu}_j$ and the total number of firms in the market. Intuitively, markets with a higher number of firms exhibit greater competition for workers, resulting in higher wages—relative to the efficient wage—compared to markets with fewer firms. Consequently, smaller markets with limited competition for workers are under-crowded, while larger labor markets, characterized by greater competition, are over-crowded in terms of the working population.



(a) Market Size vs Average Markdown (b) Average Markdown vs Number of firms

Figure A.2: Market Size Distortion

Notes: This figure is divided into two panels, labeled A, and B. Panel A illustrates the relationship between market size distortion and market weighted average markdown. Panel B illustrates the relationship between total number of firms in the market and the weighted average markdown. All parameters are set at the baseline level of table 4.

C.4 Identifying Types in Data

In the model described in Section 3, wages are not log-additive due to complementarities in the marginal product of labor. Specifically, wages in the model are expressed as:

$$w_{ij}(a) = \mu_{ij}(a) MPL_{ij}(a) = \mu_{ij}(a) Z \frac{\alpha\gamma}{1-\alpha(1-\gamma)} \Phi_{ij}(z_{ij}, \mathbf{a})^{\frac{1}{1-\alpha(1-\gamma)}} h^{\frac{\alpha-1}{1-\alpha(1-\gamma)}} \left[1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\phi(z_{ij}, a)}{\Phi_{ij}(z_{ij}, \mathbf{a})}\right)\right]. \tag{C.1}$$

This expression is not log-additive because both the markdown $\mu_{ij}(a)$ and the term

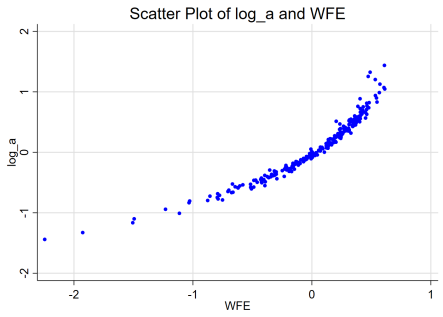
$$\left[1 - \frac{1}{\alpha\gamma}\left(1 - \frac{\phi(z_{ij}, a)}{\Phi_{ij}(z_{ij}, \mathbf{a})}\right)\right]$$

vary by the specific firm-worker match. These non-log-additive elements arise from firm-worker interactions that affect productivity, complicating the wage structure beyond simple additive components.

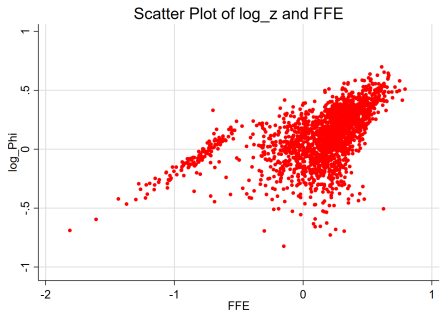
To validate that the AKM decomposition classifies worker and firm types consistently with the model, I simulate a panel dataset of workers and firms. In this dataset, workers re-draw their preference shocks for firms each period, causing them to reallocate to new firms. An AKM regression is then estimated on the simulated data.

Panel A of Figure A.3 shows a scatterplot of estimated worker fixed effects against the logarithm of worker ability. The plot reveals a clear monotonic relationship with little variation, indicating that the AKM decomposition accurately orders workers by their ability. Panel B of Figure A.3 displays a scatterplot of firm fixed effects against the logarithm of firm endogenous productivity Φ_{ij}. While the AKM decomposition generally performs well in ordering firm types, there remains some residual variation, indicating that the firm fixed effects do not perfectly capture firm productivities. However, when I focus on specific labor markets, such as in Panel C (which looks at markets with only one firm) and Panel D (which looks at markets with 100 firms), the AKM decomposition performs significantly better, with a more precise and clear relationship between the firm fixed effects and productivity. Figure A.4 compares the underlying distributions of worker abilities and firm productivities with the estimated distributions of worker and firm fixed effects obtained using the AKM framework. Panel A examines the relationship between worker ability deciles and the deciles of the worker fixed effect distribution, showing the extent to which workers are correctly classified based on their underlying ability. Panel B analyzes the relationship between firm productivity deciles and firm fixed effect deciles, illustrating how firms are categorized relative to their productivity. The figure uses heatmaps to indicate the concentration of correct classifications, with darker shades representing higher shares of correct categorization. The x-axis represents deciles of the AKM distribution, and the y-axis represents deciles of the underlying type distribution. The results reveal systematic patterns in the accuracy of classifications. Workers are correctly categorized 90% of the time, while firms show accuracy ranging from 70% to 40%. Whenever misclassification occurs, it is typically within adjacent categories, indicating that extreme misclassifications are unlikely.

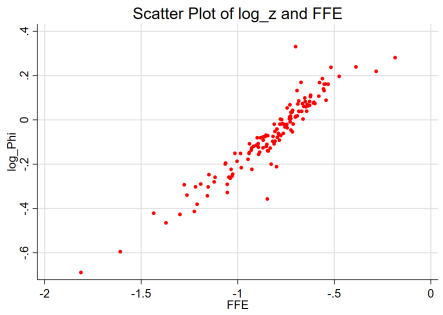
Overall, Figure A.3 and A.4 demonstrate that the AKM framework performs well in classifying both workers and firms within the model, especially if the classification is performed within labor markets. By ordering workers and firms according to their AKM fixed effects within each year and occupation, the model can be effectively mapped onto the data, allowing for a more accurate representation of the underlying wage structure and productivity distribution.



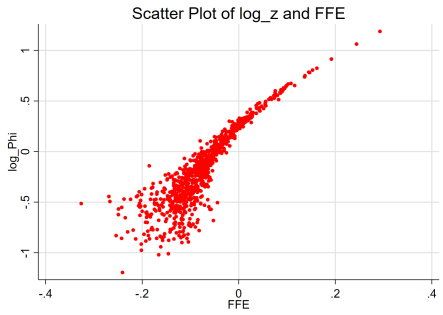
(a) Worker fixed effect vs. log ability



(b) Firm fixed effect vs. log productivity



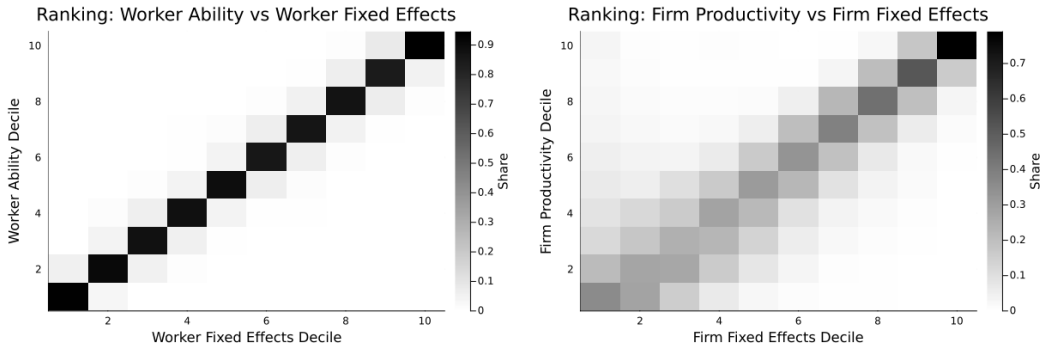
(c) Firm fixed effect in single-firm markets



(d) Firm fixed effect in markets with 100 firms

Figure A.3: Scatterplots of Worker and Firm Fixed Effects

Notes: This figure is divided into four panels. Panel A shows the relationship between the estimated worker fixed effects and worker log ability in the simulated dataset. Panel B illustrates the relationship between the estimated firm fixed effects and the logarithm of firm endogenous productivity Φ_{ij}. Panel C examines the last relationship only ini the subsamples of labor markets with a single firm, showing a more accurate fit of firm productivity. Panel D focuses on labor markets with 100 firms.



(a) Worker Fixed Effect vs Ability

(b) Firm Fixed Effect vs Productivity

Figure A.4: AKM Estimated Type vs Underlying Type

Notes: This figure is divided into two panels, labeled A and B. Both panels compare the underlying distribution of firm and worker types with the distribution of estimated worker fixed effects and firm fixed effects. The x-axis represents deciles of the AKM distribution, while the y-axis represents deciles of the underlying type distribution. The shading intensity indicates the concentration of categorization in the two deciles, with darker shades representing higher shares of correct categorizations. Panel A illustrates the shares of ability deciles that correspond to the same decile in the worker fixed effect distribution. Panel B displays the shares of firms from a specific decile of the firm fixed effect distribution that are categorized within a particular decile of the productivity distribution.