Gender Differences in Comparative Advantage Matches: Evidence from Linked Employer-Employee Data[*]

Hugo Sant'Anna [†]

November 1, 2024

Most Recent Version: Link

Abstract

In this paper, I introduce a novel decomposition method based on Gaussian mixtures and k-Means clustering, applied to a large Brazilian administrative dataset, to analyze the gender wage gap through the lens of worker-firm interactions shaped by comparative advantage. These interactions generate wage levels in logs that exceed the simple sum of worker and firm components, making them challenging for traditional linear models to capture effectively. I find that these "complementarity effects" account for approximately 17% of the gender wage gap. Larger firms, high human capital, STEM degrees, and managerial roles are closely related to it. For instance, among managerial occupations, the match effect goes as high as one-third of the total gap. I also find women are less likely to be employed by firms offering higher returns to both human capital and firm-specific premiums, resulting in a significantly larger firm contribution to the gender wage gap than previously estimated. Combined, these factors explain nearly half of the overall gender wage gap, suggesting the importance of understanding firm-worker matches in addressing gender-based pay disparities.

Keywords: gender wage gap, linked employer-employee data, Gaussian mixtures, assortative matching

JEL Codes: J16, J31, J71

[†]University of Georgia, hsantanna@uga.edu

1 Introduction

Human capital has been a crucial component in narrowing wage disparities between men and women. In the United States and similar developed and emerging markets, women are more likely than men to hold college degrees (OECD, 2024). However, there is still substantial gap, and several studies attempted to tackle the issue under different perspectives. Differences in productivity has been argued by Mulligan and Rubinstein (2008). There is also evidence that non-monetary factors, such as preferences for flexible hours, play a major role in generating the gap (Goldin, 2014).

A recent strand of the literature focus primarily on the contribution of firm-specific pay policies that would be important in creating differences across genders. Building on a simple rent-sharing model, these papers break down the wage into two main components: a worker component, solely generated by human capital levels and other worker characteristics, and a firm component, arising from firm heterogeneity such as economic activity, market power, and size. Leveraging from administrative data, Card, Cardoso, and Kline (2016) (CCK) introduced a Kitagawa-Oaxaca-Blinder (KOB) decomposition (Kitagawa, 1955; Oaxaca, 1973; Blinder, 1973) to measure the contribution of firm effects on the gender wage gap, finding that around a fifth of the gap arises from firm premiums. Even though analyses based on rent-sharing models are effective in providing a comprehensive overview of the impact of firm-specific pay premiums on the gender wage gap, these models typically assume that the value of worker characteristics remains constant across firms and vice-versa. Therefore, this "additive separability" assumption is restrictive as it constrains the ability to capture particular worker-firm interactions due to the rank condition. This precludes scenarios where comparative advantages[1] arise from specific worker-firm matches. These models may fail to account for an important source of wage variation when different classes of firms perceive similar workers differently, or when particular matches in the labor market are advantageous

---

[1]In this paper I use "comparative advantage" effects and "complementarity effects" interchangeably.

1

to certain workers, which could significantly contribute to explaining gender wage disparities.

In this paper, I provide the first comprehensive analysis of worker-firm interaction effects on the gender wage gap, explicitly accounting for assortative matching in the labor market. Extending the two-step distributional framework of Bonhomme, Lamadon, and Manresa (2019), I apply k-means clustering and a Gaussian mixture model to the log-hourly wage distribution from massive linked employer-employee data in Brazil. The data provides the universe of formal workers and firms, with a rich set of variables, such as extremely detailed economic activity and occupation codes, gender, race, education level, firm location and more.

My innovative approach groups workers and firms into "types" and "classes" respectively, reducing dimensionality to satisfy the rank condition necessary to explore worker-firm match effects in the wage structure. The model assumes that each group represents categories of workers and firms that are comparable and, when interacted, generate wages by drawing from a Gaussian distribution where parameters are specific to that match. This methodology allows for a wage generation process that deviate from the restrictive additive separable framework, enabling the identification of wage effects that arise solely from specific worker-firm interactions.

For firms, I employ k-means clustering to group them under similar payment distributions. To determine the optimal number of clusters, I utilize a gap statistics analysis (Tibshirani, Walther, and Hastie, 2001), which identifies the point at which within-cluster variance is minimized. For workers, I model the probability density function (PDF) of wages within each firm class as a mixture of log-normal distributions. I demonstrate the robustness of my results across alternative specifications of these combinations. Furthermore, I provide evidence that estimated clusters can be mapped to observable characteristics, validating their economic significance. Since the identification of my model relies on job movers, I test the exogenous mobility assumption by showing job movement is not related to unobservables.

My model is flexible enough to allow for the identification of differential firm valuations of

workers with similar unobserved characteristics. Through Monte Carlo simulations, I identify three distinct channels that contribute to gender wage disparities.

My key contribution is the identification of a "match effect". This component captures the wage effect of specific worker-firm interactions, revealing complementarities that arise when particular worker types are matched with certain firm classes. I simulate a labor market with no complementarity, therefore absent of match effects, and compare it with baseline estimates. While typical separable models struggle to identify this component, I find that women are less likely to benefit from positive complementarity effects in wages. Quantitatively, in a counterfactual world without complementarities, women's average log hourly wages increase by one log point (from 2.10 to 2.11), while men's decrease by three log points (from 2.33 to 2.30). Hence, transitioning from the observed labor market with complementarities to a simulated market without reduces the gender wage gap in log hourly wages from 0.24 to 0.20, a decrease of approximately 17 percent. This result suggests that female workers are more likely to be found in disadvantageous worker-firm interactions that yield negative complementarities in wages and even when they are present in interactions that yield positive complementarities, women tend to benefit less than men.

My results also indicate that the complementarity effect grows with higher levels of human capital and the complexity of occupations. These contributions increase with both education and age. Notably, workers in occupations typically associated with the hospitality industry, such as cleaners and waiters, show no evidence of complementarity effects. However, for individuals in occupations requiring STEM degrees, such as engineers and economists, complementarity contribution becomes positive. For managers, it accounts for as much as one-third of the gender wage gap.

In the spirit of Card, Cardoso, and Kline (2016), I also explore the overall contribution of firms to the gender wage gap by assuming the labor market is under assortative matching[2] and

---

[2]For this paper, I follow (Becker, 1973) to consider assortative matching as the propensity of high quality firms to match with high quality workers.

firms evaluate worker characteristics on top of offering premiums. To perform this analysis, I hold constant the distribution of worker clusters across men and women, measuring the counterfactual wage difference when the distributions of firms and their expected payments vary. This approach reveals two components, in addition to the match component, that mirror the established literature: sorting (women's under-representation at higher-paying firms) and bargaining (equally productive women receiving a smaller share of payments). Additive separable models potentially underestimate these effects since they assume the value of worker characteristics is constant across the labor market, thereby imposing a downward bias to the impact of assortative matching in generating the gender wage gap.

The sorting component, representing the contribution of differences in firm allocations in the labor market, accounts for approximately 37.5 percent of the 24 log point gender wage gap in Brazil. This sorting effect is substantially larger than estimates obtained from additive separable models, about 9 percent. The enhanced magnitude stems from my model's ability to capture heterogeneous firm-specific returns to worker characteristics. Specifically, it reveals that women are disproportionately concentrated in labor market segments where firms offer lower returns to worker and firm characteristics for all workers, regardless of gender. This component is less relevant for young individuals, but increases considerably for older and highly educated individuals, reaching about 40 percent to college graduates.

The final component, representing the contribution of differential payments to similarly productive men and women, accounts for approximately 8.3 percent of the gender wage gap. This "bargaining" effect suggests that even when women overcome sorting barriers, they still face wage disadvantages within firms. Collectively, the "match," "sorting," and "bargaining" components explain more than sixty percent of the observed gap, suggesting that understanding assortative matching in the labor market is essential to mitigate wage disparities.

This paper belongs to the applied literature investigating the channels generating the gender wage gap. While the gap narrowed in recent decades (Blau and Kahn, 2008) due

to increases in female human capital (Goldin, Katz, and Kuziemko, 2006; Black, et al., 2008; Ceci, et al., 2014), Goldin (2014) finds that women's preference for flexible hours over monetary compensation is a relevant factor to narrow the remaining gap. Bertrand, Black, et al. (2019) shows that women are disproportionately underrepresented in jobs with high returns on human capital investment.

The more recent strand in the literature investigates the contribution of firms to the gender wage gap. These papers belong to the applied literature that employs the AKM model (Abowd, Kramarz, and Margolis, 1999; Card, Schmutte, and Vilhuber, 2023), specifically focusing on firm effects on the gap. Generally, these studies use linked employer-employee data in a two-sided separable model with worker and firm identifiers as "plugin" estimators in a log wage linear regression. Card, Heining, and Kline (2013) showed wage dispersion could be largely attributed to these firm components using West Germany data and Card, Cardoso, and Kline (2016) proposed a KOB decomposition using Portuguese data, finding that firms contribute around 20 percent to the gender wage gap [3].

Nevertheless, the AKM-KOB analysis assumes human capital returns for specific workers are constant across the labor market. It also requires special data manipulation to extract a dual connected set of firms through male and female workers changing jobs (Abowd, Creecy, and Kramarz, 2002; Card, Cardoso, and Kline, 2016). I show in my supplementary material that this data restriction may not be innocuous, since the trimming procedure may disproportionately preserve larger firms, which typically exhibit higher wage dispersion. Moreover, wage variance analysis in AKM can be biased by underestimating the role of worker-firm interactions[4]. It was first assumed to be an economic phenomenon; however, Andrews, et al. (2008) showed that this was, in fact, an econometric issue related to small

---

[3]Other papers using the AKM model to explore firm effects KOB decomposition on the gender wage gap: Gallen, Lesner, and Vejlin (2019) in Denmark, Bruns (2019) in West Germany, Jewell, Razzu, and Singleton (2020) in the UK, Masso, Meriküll, and Vahter (2022) in Estonia, and more recently Casarico and Lattanzio (2024) in Italy.

[4]Some earlier examples of biased effects are Barth and Dale-Olsen (2003) using Norse data and Gruetter and Lalive (2009) using French data, where they found negative covariance estimates in joint worker-firm effects.

sample bias. The proposed straightforward correction to this "limited mobility bias" can be computationally prohibitive (Gaure, 2014; Azkarate-Askasua and Zerecero, 2023). Kline, Saggio, and Sølvsten (2020) introduced a Leave-One-Out methodology to fix it.

My paper goes in line with alternative approaches that moves away from AKM's additive separable assumption to avoid biased results and capture match effects. Woodcock (2008) proposed a random effects approach to satisfy the rank condition. More recently, Bonhomme, Lamadon, and Manresa (2019) (BLM) proposed a novel approach that involves clustering both firms and workers into broader categories. This method offers two key advantages. First, reducing the worker firm dimensions is computationally tractable and allows for further exploration of worker-firm match effects. Second, its non-separable nature gives a unique opportunity to observe how firms valuate workers differently under similar circumstances but differing only by gender. Bonhomme, Holzheu, et al. (2023) demonstrated that random effect models such as the BLM are particularly effective in circumventing the AKM limitations, even in short time panels.

I contribute to the literature in many ways. First, to my knowledge, this study is the first to implement the methodology of Bonhomme, Lamadon, and Manresa (2019) in the context of analyzing wage disparities between two groups. Secondly, I empirically show that separable models underestimate the role of firms in generating the gap. More importantly, I find that some interactions in the labor market exhibit comparative advantage effects, generating wage levels that substantially exceed predictions from the traditional models, but with less intensity when these interactions occur with female workers. These matches contribute significantly to wage disparities but are often smoothed out under the separability assumption. Additionally, I find these interactions exist particularly in high-paying, larger firms and they are particularly strong in highly educated workers[5].

My findings have meaningful policy implications. They show that closing the gender wage

---

[5]Studies that explore high paying firms and top earners are, for example, Bertrand, Goldin, and Katz (2010) and Bertrand, Black, et al. (2019), demonstrating the existence of a "glass ceiling" effect.

gap requires improving pay practices in key roles where highly skilled women are employed, particularly in leadership positions. Equally important are efforts to break down barriers in the labor market that push women into lower-paying firms. These firms not only offer smaller wage premiums but also limit the returns on women's skills and education, deepening income inequality over time.

The remainder of the paper is organized as follows: Section 2 provides an explanation of what is a complementarity effect and why additive separable models cannot capture it under typical settings. Section 3 provides an overview of the Brazilian data used in this study. Section 4 explains the BLM clustering method in two-steps. In Section 5, I provide the clustering results. In Section 6, I construct the Monte Carlo simulation counterfactuals. I conclude the paper in Section 7.

2   Additive Separable Models and Complementarity

Researchers are often interested in identifying the returns to unobserved heterogeneity of both workers and firms in labor markets, particularly when administrative data with social identifiers are available. In many empirical studies, these social identifiers are utilized as "plug-in estimators," commonly modeled as fixed effects in a linear equation where the outcome is the wage in logs.

For instance, consider a labor market consisting of N workers and J firms, where workers and firms interact over T periods. Under the assumption of additive separability, the log wage w of worker i at time t, net time varying effects, can be modeled as:

$$\log w_{it} = \alpha_i + \phi_j + \varepsilon_{it} \tag{1}$$

$$\text{s.t. } j = J(i,t) \tag{2}$$

where α_i represents the fixed effect associated with worker i (capturing unobserved worker-specific characteristics such as skills or human capital), and ϕ_j denotes the firm-specific

premium associated with firm j (the wage component determined by firm characteristics, independent of worker-specific attributes). The error term ε_{it} captures idiosyncratic shocks. Firm assignment is indicated by the function $J(i,t)$, which tracks the firm employing worker i at time t.

In this framework, the additive separable model assumes constant returns for both workers and firms. That is, the firm-specific premium ϕ_j is unaffected by the characteristics of the worker employed by the firm, and vice versa. This implies that reshuffling workers across firms does not alter the firm component of wages. Such an assumption is particularly strong and potentially unrealistic in labor markets where comparative advantage is thought to play a role in worker-firm interactions (Shimer and Smith, 2000; Eeckhout and Kircher, 2018).

To allow for complementarity between workers and firms, one could extend the model to include interaction effects. Specifically, the wage equation can be rewritten as:

$$\log w_{it} = \alpha_i + \phi_j + M_{ij} + \varepsilon_{it} \tag{3}$$

$$\text{s.t. } j = J(i,t) \tag{4}$$

where M_{ij} represents the interaction effect between worker i and firm j. This term captures the potential complementarity effect that only arises when worker i is employed at firm j. It reflects the idea that certain worker-firm pairings generate higher (or lower) wages than what would be predicted based solely on the worker's fixed effect α_i and the firm's premium ϕ_j. However, estimating this interaction term in practice is infeasible due to the large dimensionality of the model. The matrix M_{ij} would have $N \times J$ terms, which quickly becomes computationally intractable given that linked employer-employee data typically contain millions of workers and thousands of firms.

Moreover, these models are prone to bias in settings with short panel datasets, where estimating the worker and firm fixed effects becomes difficult. Building on the approach

of Bonhomme, Lamadon, and Manresa, 2019, I employ a methodology that reduces the dimensionality of workers and firms by clustering them into latent groups. This allows for the estimation of complementarity effects while avoiding the rank deficiency problem inherent in models with such high-dimensional interactions. Specifically, workers and firms are grouped based on their interactions in the labor market. I assume each worker-firm interaction draws wages from log-normal distributions, meaning I can relax the linear assumption and the additive separability.

The wage generation function now can be expressed as:

$$\log w_i = f(\alpha_{L(i)} \mid \phi_{K(i,j)}) \tag{5}$$

where $L(i)$ denotes the assignment function of worker i's type, and $K(i,j)$ denotes the assignment function of firm j's class. $f(\alpha_{L(i)} \mid \phi_{K(i,j)})$ denotes a probabilistic function that draws wages from a log-normal distribution that is specific related to the match of worker type l and firm class k. Since wages are assumed to be derived specifically from worker-firm interactions, some matches are allowed to yield comparative (dis)advantage effects on wages, given that returns to firm and worker characteristics, under this model, are not necessarily constant across the labor market.

Therefore, recovering these latent types can be done by exploring the surface of observed wages. The idea behind the model is to recover Gaussian distributions that are "combined" in the full distribution, derived from all the matches occurring from the labor market. Thus the Gaussian mixture model application.

A key assumption for this model is the exogenous mobility assumption, meaning job mobility depend on the type of the worker and the classes of the firms, but not directly on earnings. I discuss the exogenous mobility assumption in my model in Section **??**.

In the context of wage differences due to gender, my approach is to estimate the Gaussian mixture parameters in a pooled dataset, meaning the model does not observe gender at first.

The reasoning is to facilitate the comparability of individuals under the same umbrella of unobserved heterogeneity. I discuss further the model in Section 4.

3 Data

In this section, I provide an overview of the Brazilian administrative data used for the study and the preparatory cleaning for the analysis, followed by a descriptive statistics of the cleaned sample.

3.1 Data Overview and Institutional Background

I use the *Relação Anual de Informações Sociais* (RAIS), an extensive linked employer-employee dataset (LEED) from Brazil spanning from 2010 to 2017. RAIS is mandated and maintained by the Brazilian Ministry of Labor and Employment, serving as a source for the administration of tax and social programs. The dataset offers an universal representation of the formal labor market in Brazil and is characterized by its richness in variables.

A key advantage of using the *RAIS* dataset for this analysis is the relative homogeneity of job-related amenities across firms due to Brazil's robust labor regulations. The Brazilian labor laws, known as the Labor Laws Consolidation (*Consolidação das Leis do Trabalho* (CLT), in Portuguese), mandates a broad range of standardized benefits and protections for all formal workers, regardless of industry or firm size. This regulatory framework significantly reduces variation in non-wage compensation, allowing the analysis to focus more cleanly on wage differentials without the confounding effects of divergent job-related amenities.

For example, Brazilian law requires all formal employees to receive the 13th salary, which is essentially a mandatory annual bonus equivalent to one month's wage, usually paid during Christmas time. Additionally, firms are obligated to provide meal vouchers or food stipends, as well as transportation subsidies for commuting. These benefits are non-negotiable and standardized across the formal labor market. Moreover, formal workers are entitled to

thirty days of paid vacation, overtime pay, and severance protections via the *Fundo de Garantia por Tempo de Serviço*[6] (FGTS), which further ensures that variations in non-wage job characteristics can be minimized.

In Brazil, maternity leave is a legally guaranteed right under the CLT. Female employees are entitled to 120 days of paid maternity leave, funded by the Brazilian Social Security system. In some cases, companies can extend this leave to 180 days through the Empresa Cidadã program, which offers tax incentives to employers. During maternity leave, the employee's job is protected, and she is guaranteed to return to her position or a similar one without loss of salary or benefits. Additionally, Brazilian law prohibits the dismissal of pregnant workers from the moment pregnancy is confirmed until five months after childbirth, with some exceptions under fair cause.

This regulatory uniformity is particularly beneficial for my analysis, as it mitigates concerns that differences in firm payment patterns are due to job amenities that could ultimately explain wage differentials between male and female workers. In contrast, in countries where non-wage compensation varies significantly across firms or sectors, disentangling wage differences from benefit-driven compensation can complicate the analysis of wage gaps.

In my study I focus on São Paulo state, which represents the most economically dynamic region in Brazil, making sure my results are not driven by geographical heterogeneity. For example, a male worker in manufacturing and a female worker in retail, though in distinct sectors, would both receive a standardized package of legal protections and benefits coming not only from federal law, but also from local state law, ensuring that wage comparisons are not distorted by differences in state policies.

Regarding gender dynamics in São Paulo's labor market, it is important to note that, similar to other countries analyzed in the literature, approximately more than 50 percent of the Brazilian women there participate in the labor force, with 71 percent of these women employed full-time. This proportion rises to 90 percent when considering only those employed

---

[6]Roughly translated as Severance Indemnity Fund for Length of Service

in the private sector. Furthermore, the gender wage gap in Brazil mirrors those observed in more developed economies, offering additional comparative insights. In 2016, the median earnings gap between male and female full-time workers was approximately 14.3 percent in Brazil, closely aligned with the average of 13.4 percent observed across OECD countries, and slightly better than the 18.1 percent reported for the United States (OECD, 2024).

3.2   Data Preparation

The RAIS database records each formal employment contract as a separate entry, meaning that for any given year, a worker with multiple contracts, whether with the same employer or different firms, will appear multiple times. To address this, and following the methodologies of Gerard, et al. (2021) and Lavetti and Schmutte (2023), I refine the dataset by retaining only the longest-duration and highest-paid contract for each individual per year. This adjustment shifts the data from a contract-year structure to an individual-year framework, ensuring that the analysis focuses on each worker's primary employment.

To align with a long-run perspective, the sample is further restricted to a *quasi*-full-time workers, defined as those working a minimum of 30 hours per week, and limited exclusively to the private sector. I allow this flexibility to capture a certain degree of non-monetary preference particularly found in female cohorts (Goldin, 2014). This exclusion criteria eliminates part-time employees, public sector workers, and the self-employed from the analysis, thereby focusing on a more homogeneous labor market.

3.2.1   Biennial Grouping and Panel Balancing

The organization of the data for my analysis involves grouping the dataset into jumping biennials. Specifically, the years 2010 and 2012 are paired, 2011 and 2013, and so forth. This method skips intermediate years to avoid transitional anomalies that may occur in short periods, such as firm mergers or changes in identifiers. This "jumping" approach closely mirrors the sample selection method employed by Bonhomme, Lamadon, and Manresa (2019).

In my analysis, it comprises of six sets of balanced panel data spanning from 2010 and 2012 to 2015 and 2017. Each biennial set is balanced and analyzed to estimate worker and firm clusters, with final estimates related to wages presented as a weighted average of these samples. This "rolling" approach has been used to some extent in Card, Cardoso, and Kline (2016) and Lachowska, et al. (2023).

Each biennial panel is balanced, ensuring that the same set of workers and firms are observed consistently within each two-periods. In addition, firms with pronounced gender preferences are excluded from the analysis. Only firms exhibiting a gender ratio of 1 to 4 are included, which helps mitigating any potential bias that could arise from firm gender imbalance.

3.3   Summary Statistics

Table 1 reports descriptive statistics by gender cohorts for the aggregated cleaned data, representing the first year of each biennial sample. Columns (1) and (2) represent the statistics for female and male workers, respectively.

The dataset encompasses a total of 346,617 unique firms. Of these, a substantial portion is relatively large; approximately 204,994 firms employ 10 or more workers, and 58,866 firms have at least 50 workers. The average firm size across the sample is 57 employees, but the median firm size is considerably smaller, at 13 employees, indicating a skewed distribution.

Gender related educational attainment confirms that women are generally more educated than their male counterparts. The data show a higher prevalence of men without high school diplomas, while women are more likely to have completed high school or pursued some college education. As stated previously, this educational dynamic is consistent with recent trends observed in both developed and developing nations, such as the United States and other OECD countries.

Approximately 40 percent of the female sample is under 30 years old, with another 50 percent aged between 31 and 50. In contrast, 37 percent of the male sample is under 30, with

Table 1: Descriptive Statistics by Gender

Features	Female Workers (1)	Male Workers (2)
Firm Characteristics		
Number of Firms	346 617	346 617
Firms with ≥ 10 Workers	204 994	204 994
Firms with ≥ 50 Workers	58 866	58 866
Mean Firm Size	57	57
Median Firm Size	13	13
Worker Characteristics		
Education (%)		
Dropout	22	28
High School Graduates	48	45
Some College	30	27
Age (%)		
< 30	40	37
31–50	50	49
≥ 51	10	14
Sector of Employment (%)		
Primary	2	2
Manufacturing	19	26
Construction	1	2
Trade	24	25
Services	54	45
Occupation (%)		
Scientific and Liberal Arts	11	11
Technicians	11	11
Administrative	34	18
Managers	5	7
Traders	25	22
Rural	1	2
Factory	13	29
Labor Market Outcomes		
Mean Tenure (years)	4.04	4.63
Mean Log-Wage	2.06	2.29
Variance of Log-Wage	0.52	0.65
Worker-Year Observations	9 503 233	10 283 471
Unique Number of Workers	3 497 651	3 725 990
Gender Fraction (%)	48	52

Note: [1]Descriptive statistics calculated from the first year of each biennial sample (2010-2015). [2] Percentages may not sum to 100% due to rounding. [3]The number of firms is the same for both genders since every firm in the cleaned sample employs both male and female workers.

49 percent in the 31 to 50 age bracket. Moreover, men are slightly more represented in the over-50 cohort, constituting 12 percent compared to 8 percent of women. Hence, the average experience in the labor market is 4.6 years for males and 4.0 years for females.

Industry distribution varies significantly between genders. Men dominate in sectors such as manufacturing, agriculture, and trade, whereas women are predominantly engaged in services, an umbrella term that includes sectors such as healthcare, education, hospitality, and financial services.

The occupational distribution also highlights a notable gender sorting: women are almost twice as likely as men to hold administrative positions, representing 34 percent of women compared to 18 almost of men. Men are more frequently employed in manual labor-intensive roles such as in agricultural settings and factories.

Despite these occupational disparities and the educational advantages observed for women, the unweighted gender wage gap remains substantial at approximately 23 log-points. This gap persists even though women are, for instance, equally likely as men to occupy scientific roles, which typically require higher educational qualifications.

3.4   Extended Mincer Equation

As a first step to analyze the gender wage gap, I provide a classical Kitagawa-Oaxaca-Blinder (Kitagawa, 1955; Oaxaca, 1973; Blinder, 1973) decomposition of an extended Mincer equation and an AKM equation, assuming the gap is a mean difference of female and male wages. A "Mincer wage function" can be specified as:

$$w_{it} = \beta_0 + \beta_1 \text{Age}_{it} + \beta_2 \text{Age}_{it}^2 + \beta_3 \text{Education}_{it} + \beta_4 \text{Occupation}_{it} + \beta_5 \text{Activity}_{it} + \varepsilon_{it} \quad (6)$$

where w_{it} is the natural logarithm of hourly wages for individual i in time period t, regressed on the worker's age and their squared age, their education level, the firm's industry, the

worker's occupation, and a idiosyncratic error term. For the Oaxaca decomposition, I run this regression for the male and female observations separately, for each biennial sample.

Assume the matrix of explanatory observables can be expressed as X^g, where g represents the gender sample used in the regression. Also assume β is the vector of estimates. The KOB decomposition can be expressed as:

$$\bar{w}^m - \bar{w}^f = \underbrace{\left(\bar{X}^m - \bar{X}^f\right)\hat{\beta}^f}_{\text{Explained}} + \underbrace{\bar{X}^f\left(\hat{\beta}^m - \hat{\beta}^f\right)}_{\text{Unexplained}} \tag{7}$$

where $\left(\bar{X}_m - \bar{X}_f\right)\hat{\beta}_f$ represents the "explained" component of the decomposition. In simpler terms, this term represents a counterfactual scenario where men and women possess the same returns to covariates, however, they differ in these covariates' distribution. The unexplained component, on the other hand, captures differences in the returns to these characteristics. This is expressed as $\bar{X}_f\left(\hat{\beta}_m - \hat{\beta}_f\right)$, where the difference in coefficients $(\hat{\beta}_m - \hat{\beta}_f)$ measures a scenario where men and women have the same observable characteristics, however, the market values differently each gender. The unexplained portion is often interpreted as the part of the wage gap that cannot be accounted for by observable factors alone, potentially indicating discrimination or other structural labor market imbalances.

Table 2 presents the overall log hourly wage gap in means, the explained, and the unexplained portion of the gender wage gap across the six biennial samples, along with the number of observations for each sample. The overall wage gap remains consistent at 24 log-points for the first three samples. However, the gap slightly decreases in the subsequent samples, with the smallest gap observed in 2015-2017 at 22 log-points.

In this setting, the explained portion of the Oaxaca decomposition accounts for approximately 6.11 log-points, or roughly one-quarter of the total gender wage gap. This indicates that observable factors, such as the allocation of workers across different occupations or sectors, explain about 25 percent of the wage differential in an additively separable labor market.

Table 2: Extended Mincer Equation KOB Decomposition For Each Biennial Sample

Sample	Overall Gap	Explained Gap	Unexplained Gap	N
2010–2012	-0.244	-0.0651	-0.179	5 946 240
2011–2013	-0.244	-0.0637	-0.180	6 145 676
2012–2014	-0.244	-0.0642	-0.180	6 534 444
2013–2015	-0.241	-0.0621	-0.179	6 787 446
2014–2016	-0.230	-0.0571	-0.173	7 086 062
2015–2017	-0.221	-0.0542	-0.167	7 073 540
Weighted Avg[a]	**-0.237**	**-0.0611**	**-0.176**	39 573 408[b]

Note: [a]Weighted average calculated using sample sizes as weights and the gap as $female - male$. [b]Total number of observations across all samples. [1]Extended Mincer equation defined as $\log(y_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Age}_i^2 + \beta_3 \text{Education}_i + \beta_4 \text{Occupation}_i + \beta_5 \text{Activity}_i + \varepsilon_i$. [2]Explained gap represents differences in distribution of characteristics. Unexplained gap represents differences in estimated returns to characteristics.

In Section 6, I extend the analysis by introducing firm identifiers as fixed effects under an AKM framework following Card, Cardoso, and Kline (2016). Under this specification, firm effects explain about 9 percent of the gender wage gap.

The issue with separable models is the assumption that these components should not vary depending on the association happening. Under AKM, these firm effects will occur in any worker reshuffling instance of the labor market.

In the next sections, I propose the distributional framework of Bonhomme, Lamadon, and Manresa (2019) to capture particular interactions in the labor market that does not necessarily follow an additive separable assumption.

4   Empirical Framework: The BLM Model

Estimating the Gaussian mixture requires two main parts. Following Bonhomme, Lamadon, and Manresa (2019), I assume cluster membership of firms is exogenous to the model, allowing their estimation by employing straightforward clustering methods from features observed from the data. Still following BLM, I choose to cluster firms based on their wage cumulative distribution function using k-means clustering (MacQueen, 1967).

In the second part, I take the estimated firm clusters, called "firm classes", to assume that they are Gaussian mixtures of latent worker types in log wages. In the spirit of AKM settings, I leverage individuals changing jobs to identify the Gaussian parameters.

Finally, I use a *maximum a posteriori* estimation to find the most likely worker type for each worker observation. After the classification, I split the data into male and female cohorts.

4.1 Recovering Firm Classes

The first objective is to recover firm clusters, or "firm classes", which are initially unobserved in the data. The approach relies on two key assumptions. First, the mapping of firms to clusters is exogenous to labor market dynamics.

Formally, let $k(j)$ denote cluster assignment of firm j. The exogeneity assumption can be expressed as:

$$P(k(j)|X) = P(k(j)) \tag{8}$$

where X represents labor market conditions and worker characteristics. In plain language, this condition ensures that the probability of a firm belonging to a firm class is unconditional to these labor market features, which allows a direct estimation of firm classes using the clustering method.

Secondly, the wage distribution in the data follows a log-normal shape for workers, conditional on these firm clusters. Consequently, each firm class represents a Gaussian mixture of log-wages. Within these mixtures, each component corresponds to a log-normal distribution arising from the unobserved heterogeneity of worker groups, which is termed "worker types" in this study, following BLM's terminology.

Formally, the assumption states that for a firm j in class k, the log-wage distribution, for a given time period, can be expressed as:

$$f_k(w_i) = \mathbb{1}\{\hat{k}(i) = k\} \sum_{l=1}^{L} q_k(L(i))\mathcal{N}(\theta_{kl}) \tag{9}$$

where, $f_k(\log(w_i))$ is the log hourly wage mixture of firm class k, when observing worker i. With some abuse of notation, L denotes the number of worker types, $q_k(L(i))$ represents the proportion of workers of type $L(i)$ in class k, and $\mathcal{N}(\theta_{kl})$ is the Gaussian probability density function for type l workers in class k, with θ_{kl} representing the parameters of this distribution. The indicator function $\{\hat{k}(i) = k\}$ ensures that we consider only the wage distributions of workers assigned to the specific firm class k.

My approach leverages firm clustering to address the dimensionality challenge inherent in firm heterogeneity analyses. By aggregating individual firms into a more manageable set of "firm classes", I circumvent the need to restrict the dataset to a set of connected firms through workers. However, the identification strategy of this methodology still relies on job movements. It shifts, however, the focus from tracking movements between individual firms to observing transitions across firm classes. Therefore, while this mixture model still fundamentally relies on job mobility, it does so at a more aggregated level. In the supplementary material, I perform a clustered AKM regression to show that on average, the residual change in wages for these movers is close to zero, suggesting the movement pattern is not related to the labor market structure itself.

A crucial assumption of this approach is that each worker type, to be estimated in the second step, exhibits a unique pattern in their "cycling" through firm classes as they navigate job changes. These transitional pathways must be sufficiently distinct to allow for clear identification of worker type parameters (Bonhomme, Lamadon, and Manresa, 2019). The robustness of this assumption in my context is based on the substantial number of observations in the dataset, which provides the statistical power necessary to discern these distinct mobility patterns between worker types and firm classes.

The k-means algorithm aims to group firms with similar payment schedules. Formally:

$$\underset{k(1),\ldots k(J),H_1,\ldots,H_K}{\arg\min} \sum_{j=1}^{J} n_j \int (\hat{F}_j(w) - H_{k(j)}(w))^2 d\mu(w) \qquad (10)$$

where $\hat{F}_j$ represents the empirical CDF of the log-weekly wages w of firm j, μ is a discrete measurement, supported by a finite grid of ventiles from the population. K, the number of firm classes, is known, while the array $k(1),\ldots,k(J)$ represents the partitioning for each firm. H_k represents cluster k's CDF. Finally, n_j is the firm's corresponding workforce size. I perform 1000 repetitions to ensure a global minimum distance estimation.

In simple terms, this procedure minimizes the distance between firms and unobserved classes using as measurements each firm's empirical CDF generated from the ventiles of the observed population log hourly wage distribution. It imposes a weighting parameter to ensure different minimization process for larger firms. For each biennial sample, I assume that the firm class classification is time-invariant.

I choose $K = 10$ as the baseline number of groups since it minimizes the wage variance within each group. I follow Bonhomme and Manresa (2015) and Bonhomme, Lamadon, and Manresa (2019), where the estimation of firm classes does not affect parameter estimation in the Gaussian mixture step. Nevertheless, in the Appendix , I provide a comprehensive cluster choice analysis using gap statistics to find optimal K-Means clustering estimation. I also provide alternative cluster settings as robustness checks in the discussion section.

4.2   Gaussian Mixture Estimation

I assume that observed wages follow a mixture of log-normal distributions, where every "latent" probability distribution represents an interaction of a latent worker "type" with the respective firm class. This approach enables me to reduce the high-dimensional unobserved heterogeneity among individual workers into a manageable set of Gaussian distributions.

I estimate the parameters with the pooled dataset, not observing gender at first. By not accounting for gender at the outset, I ensure that male and female workers assigned to

the same distribution are as similar as possible in terms of unobserved characteristics. The idea is that the algorithm will approximate individuals with sufficiently similar unobserved characteristics that spawn the same distribution of wages, regardless of gender. It allows for a more precise comparison of how these latent worker types interact with firm classes without biasing the results by preemptively imposing gender differences.

This approach allows for a more flexible examination of the wage structure assumption in the labor market. By constructing and comparing expected payment levels for each worker type and firm class interaction, I can empirically assess at which extent the additive separability assumption hold, and capture interactions in the market that deviates from this condition. Finally, I can disaggregate these payment levels by gender to measure the differential complementarity effects on wages, providing insights into how worker-firm interactions contribute to gender wage disparities, especially at matches where the separable form is not observed.

4.2.1 Recovering Worker Types

To identify latent worker types, I posit that the wage distribution for each type depends on their associated firm class and follows a log-normal distribution. This approach incorporates potential complementarities characteristic of specific worker-firm matches. I first, estimate the densities for job movers, and subsequently, I estimate the proportions of stayers using the job mover distributions from the initial period.

I formulate this as a maximum likelihood problem, closely following Bonhomme, Lamadon, and Manresa (2019):

$$\underset{\theta_p,\theta_1,\theta_2}{\arg\max} \sum_{i=1}^{N_m} \sum_{k=1}^{K} \sum_{k'=1}^{K} \mathbb{1}\{\hat{k}_{i1} = k\}\mathbb{1}\{\hat{k}_{i2} = k'\} \log \left(\sum_{l=1}^{L} p_{kk'}(l;\theta_p) f_{kl}^1(w_{i1};\theta_1) f_{k'l}^2(w_{i2};\theta_2) \right) \quad (11)$$

where N_m denotes the number of job movers, K the number of firm classes, and L the number of worker types (set to 10 for interpretability). The indicator functions $\mathbb{1}\{\hat{k}_{i1} = k\}$ and

$\mathbb{1}\{\hat{k}_{i2} = k'\}$ capture the transition of worker i from firm class k to k' between periods 1 and 2. $p_{kk'}(l; \theta_p)$ represents the proportion of type l workers moving from class k to class k', while f_{kl}^1 and $f_{k'l}^2$ are log-normal wage distributions for type l workers in classes k and k' in periods 1 and 2, respectively.

Therefore, Equation 11 captures the parameters of the conditional distributions of the worker types leveraging the job movers.

For job stayers, I estimate:

$$\arg\max_{\theta_q} \sum_{i=1}^{N_s} \sum_{k=1}^{K} \mathbb{1}\{\hat{k}i1 = k\} \log \left(\sum_{l=1}^{L} q_k(l; \theta_q) f_{kl}^1(w_{i1}; \hat{\theta}_1) \right) \tag{12}$$

where N_s is the number of stayers, and $q_k(l; \theta_q)$ is the proportion of type l stayers in class k. I leverage the first year parameters for job movers. I employ the Expectation-Maximization (EM) algorithm with 50 repetitions to estimate these parameters.

To recover the most likely worker type for each observation, I utilize the Maximum A Posteriori (MAP) estimation. Formally, for a worker i in firm class k with wage w_i, the probability of belonging to type l is given by:

$$P(l|w_i, k) = \frac{q_k(l; \hat{\theta}_q) f_{kl}(w_i; \hat{\theta})}{\sum_{l'=1}^{L} q_k(l'; \hat{\theta}_q) f_{kl'}(w_i; \hat{\theta})} \tag{13}$$

The worker type is then assigned as:

$$\hat{l}_i = \arg\max_l P(l|w_i, k) \tag{14}$$

5 Estimated Parameters

In this section, I present the estimated parameters for the mixtures, beginning with firm class estimates, followed by the mixture proportions, and concluding with a detailed analysis of the estimated moments disaggregated by gender.

5.1 Cluster eCDFs

The effectiveness of the algorithm in segregating firms into distinct clusters is evaluated by visualizing the empirical cumulative distribution function of the generated clusters. They are illustrated in Figure 1a.

As depicted, the algorithm managed to delineate mostly clear firm classes, grouping firms with similar pay policy, evidenced by the "clear cuts" of each cluster's CDFs, with the exception being firm class 4.

Figure 1b provides the moments of their log hourly wage distribution, with the means as the bars and the first standard deviation as the error-bars. For each estimated cluster, not only expected payment increase but also their dispersion when going upward in the firm class ranking. For example, the lowest firm class pays, on average, 1.45 in log hourly wages, with a variance of 0.09, while the highest pays 3.67 with 0.47 in variance.

Figure 1c reveals the gender wage gap in means (expressed as $\mathbb{E}[w_{it}^f|k] - \mathbb{E}[w_{it}^m|k]$) as the line plot (the right y axis), and the average size per firm as the bar plot (the left y axis). The expected gender wage gap in means has a tendency to increase when going up in firm class ranking. The lowest paying firms are the most equitable firms in the labor market, with the lowest difference between genders at 11 log-points. The plot also reveals highest-paying firms, which tend to be larger firms[7], exhibiting the largest gender wage disparities, reaching 25 log points. This finding is not entirely unexpected given the substantial variance in wages within firm class 9 or 10. This pattern suggests potential overestimation of the magnitude of firm effects contribution to the gender gap under additive separable models. This overestimation likely stems from the necessary practice of focusing on large firms to ensure sufficient worker mobility within a connected set, while addressing the "double-coincidence" problem of observing both male and female job transitions. However, this approach inadvertently oversamples precisely those firms where gender wage disparities
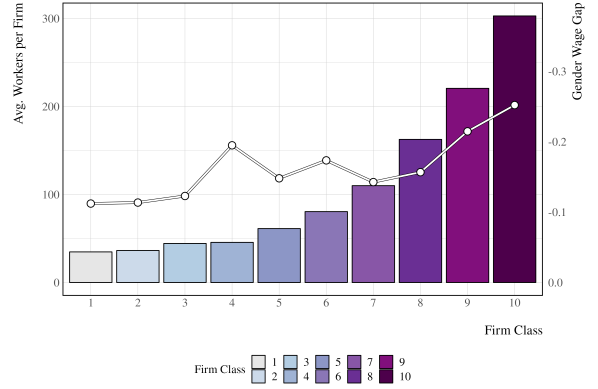
---

[7]For a full descriptive statistics of firm classes, see Table E.2 and E.3
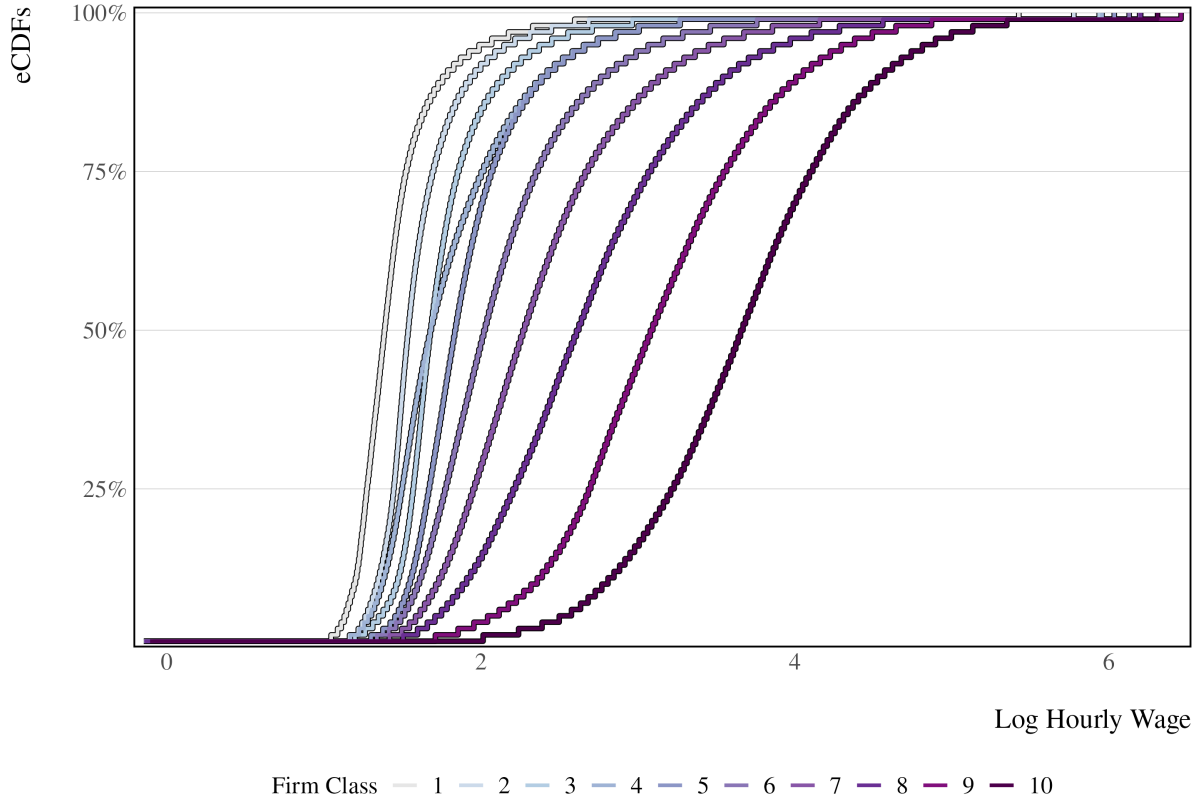
(a) Empirical CDF of Firm Classes
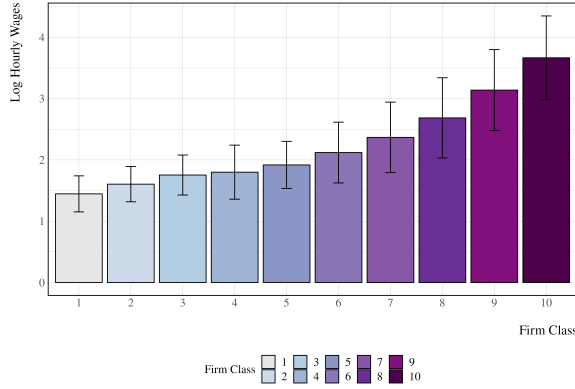
(b) Wage (Mean and Variance) Statistics

(c) Gap and Size Statistics

Figure 1: (a) Firm Class ECDFs, (b) Firm Class Mean and Variance, and (c) Firm Class Size and Gender Wage Gap Statistics

Note: [1]Firm classes estimated by a k-means clustering algorithm using as measurement their empirical cumulative distribution function supported by the ventiles of the population. [2]The Gender wage gap in means (line in Panel C) is calculated as the female minus male: $\mathbb{E}[w_{it}^f|k] - \mathbb{E}[w_{it}^m|k]$.

are most pronounced, potentially skewing overall estimates of firm effects on the wage gap.

5.2 Assortative Matching of Estimated Parameters

Figure 2 displays the unconditional distribution of workers across firm classes (top row) and worker types (bottom row) for each gender. Both male and female workers exhibit a concentration of employment in firm class 6, but the proportion is slightly higher for men, with 15 percent of the male workforce in this class compared to 13 percent for women. Additionally, the distribution for men shows a more noticeable skew towards higher-productivity firms. Specifically, 17 percent of men are employed in the top two firm classes (9 and 10), whereas about 14 percent of women are employed in these high-productivity firms. This suggests that men are more likely to be employed in firms that offer higher wage premiums, which may contribute to the observed gender wage gap through the sorting channel.

The differences in distribution become more pronounced when examining worker types. The female distribution is heavily skewed to the left, with nearly 24 percent of women concentrated in worker type 3 versus 17 percent among male workers. In contrast, the male distribution is more evenly spread across worker types, exhibiting a more balanced, albeit still slightly left-skewed, pattern.

In this paper, worker types represent comparable unobserved heterogeneity. Meaning female and male type 3 are individuals where their wages are likely drawn from the same set of Gaussian distributions. The firm class distribution has a more straightforward interpretation, as the proportion of firms with similar payment policies, mirroring patterns of productivity and industry.

When I discuss the gender wage gap decomposition, I hold the distribution of worker types constant since channels of worker type heterogeneity may arise from a multitude of mechanisms in the labor market, such as non-monetary preferences or human capital levels.

Figure 3 displays firm classes along the horizontal axis against the stacked conditional proportions of corresponding worker types, separately for female and male workers. These

proportions are recovered by grouping types for each male and female sample conditional on each firm class after the *maximum a priori* classification.

Worker types and firm classes are numbered according to expected payment. Therefore, type 10 represents on average the highest paid worker in the data, a proxy for individuals that overall possess high human capital value. The visual representation clearly illustrates an assortative matching pattern, revealing that higher-paying firms predominantly employ higher types of workers for both genders. However, there are notable differences between male and female sorting patterns.

For female workers, there is a strong concentration of lower-type workers in lower-class firms. For instance, in firm class 1, 29 and 23 percent of the workforce comprise of type 1 and type 2 workers, with another 36 percent being type 3 and 4 together. Moving to higher firm classes, this composition shifts dramatically: in class 10, less than 5 percent are type 1 workers, while 15 and 26 percent belongs to type 10 and 9 workers.

On the other hand, male workers shows a slightly different trend. Type 1 and 2 workers comprise together 44 percent of firm class 1 workforce, slightly less concentrated than for females. In the highest firm class, while also presenting negligible proportions of the lowest type, 51 percent of the workforce is comprised of type 10 and 9 workers.

Therefore, while assortative matching is evident for both genders, the patterns reveal some disparity in how men and women with sufficiently similar unobserved heterogeneity are sorted across firm classes. Women appear to face some friction in ascending the firm classification hierarchy, resulting in a more pronounced concentration in lower-tier firms even when their latent productivity (as captured by worker types) is comparable to that of their male counterparts.

5.2.1 Theil Index

To quantitatively assess whether the male distribution in the labor market is slightly more symmetrical compared to the female, I perform a Theil Index calculation to measure the
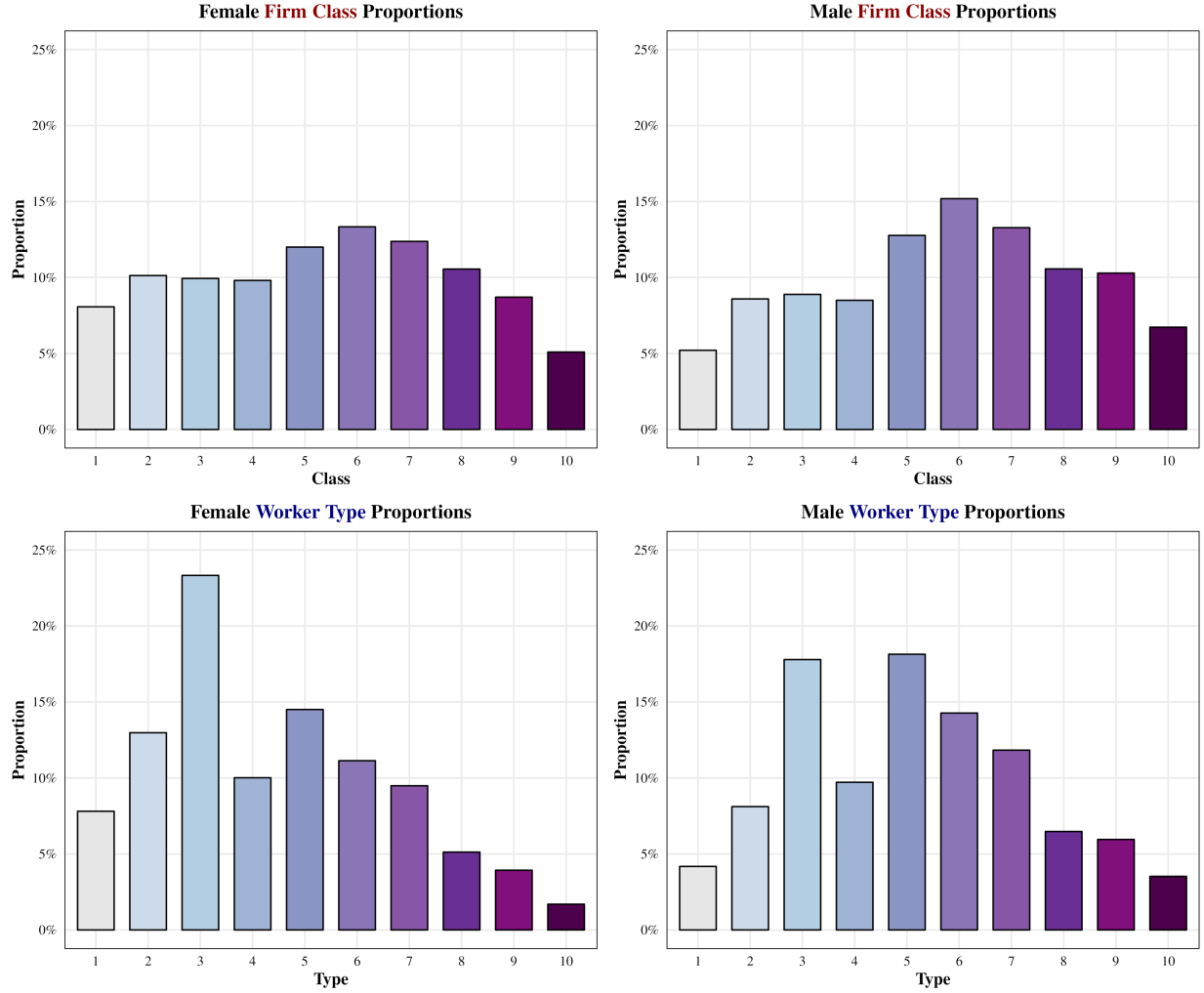
Figure 2: Worker Type and Firm Class Unconditional Probabilities per Gender

Note: [1] Firm class estimated using k-means clustering of the cumulative distribution function of payments. Worker types estimated using a Gaussian mixture model where I assume each latent worker type interact with firm classes by drawing wages from a log-normal distribution.
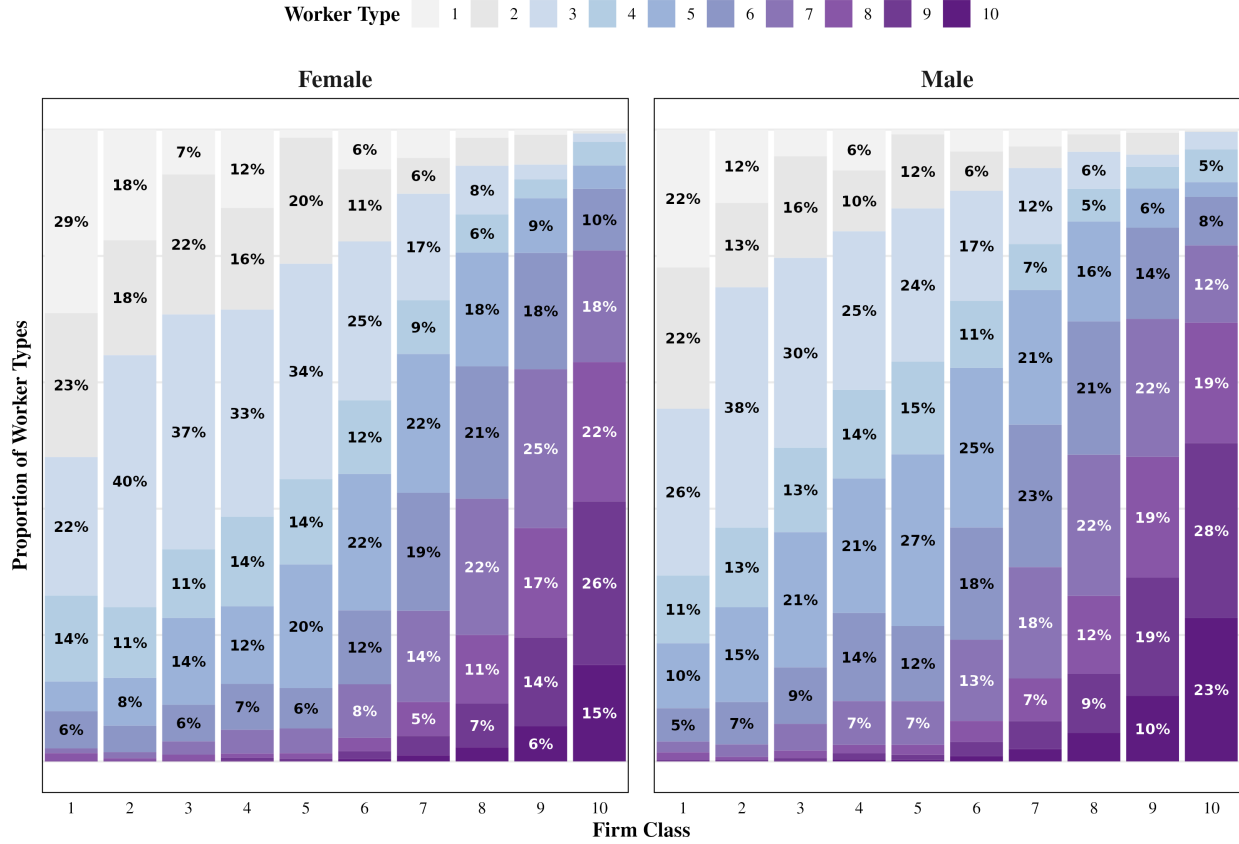
Figure 3: Proportion of Estimated Worker Types and Firm Classes

Note: [1]Proportions of worker types recovered using a finite Gaussian mixture of log hourly wages conditional on observed firm classes. Firm classes recovered using k-means clustering algorithm on firm's log hourly wage's CDFs. [2]Worker type membership assigned using a *maximum a priori* estimation. [3]Types and classes ordering based on expected log hourly hours.

inequality, where the metric is the number of workers per match. I separate the firm classes into low and high classes, where low comprises of firm class 1 to firm class 5, while high comprises of firm class 6 to firm class 10.

The simple Theil Index formula is:

$$T = \frac{1}{M} \sum_{m=1}^{M} \frac{N_m}{\bar{N}_m} \log\left(\frac{N_m}{\bar{N}_m}\right) \tag{15}$$

where m is each match in the labor market, M the total number of matches, in this case, 100. N_m is the total number of workers for each match, while $\bar{N}_m$ is the average number of workers per match in the labor market.

The Theil index for the male distribution is 0.43. For the female distribution is 0.48, slightly larger, suggesting that the female distribution of workers in the labor market is more sorted towards the left, concentrated in overall less paying matches.

5.3   Payment Schedules

The BLM method not only demonstrates the flexibility to capture assortative matching but also enables researchers to discern the underlying wage structure arising from firm-worker interactions by assessing whether certain compensation patterns result in wage levels that surpass predictions from additive separable models.

Figure 4 presents the payment schedules by firm class and worker type. Panel (a) represents estimated average payments directly under the Gaussian mixture model. Panel (b), on the other hand, is a counterfactual scenario where worker and firms do not yield complementary wage effects in their interactions. I construct this counterfactual by performing an "AKM" two-way fixed effect model such as:

$$w_{it} = \alpha_{L(i)}^{g} + \psi_{K(i,t)}^{g} + \varepsilon_{it} \tag{16}$$

where w_{it} is the log hourly wage of worker i in time period t, $\alpha_{L(i)}$ is the fixed effect of worker

i's type l, represented under the assignment function $L(i) = l$, $\psi^g_{K(i,t)}$ is the fixed effect of firm class k, also represented under an assignment function $K(i,t)$. ε_{it} is the idiosyncratic error term. To make sure I preserve gender disparities, I regress twice for each gender sample.

I introduce a weighting parameter to mitigate the influence of extreme values on my estimation. It leverages the fact that common interactions in the labor market tend to possess small complementarity effects[8]. Consequently, the objective function for this minimization problem can be expressed as:

$$\min \sum_{i,t} n_{M(k,l)}(w_{it} - \alpha^g_{L(i)} + \psi^g_{K(i,t)})^2 \tag{17}$$

where $n_{M(k,l)}$ represents firm class k and worker type l match's proportion of the number of workers.

Each panel in Figure 4 shows each line representing an expected payment "path" of each worker type when hired by a particular firm class.

The Gaussian mixture model is able to capture different wage levels that do not necessarily follow a linear trend, as shown in Figure 4a.Top firm classes tend to offer substantially higher wage levels to individuals, with particularly pronounced effects for workers in the lower to middle range of the skill distribution. High-earning individuals exhibit remarkable wage stability across firm classes, maintaining their elevated earnings even when matched with low firm classes, with a slight decrease. There is also severe wage compression at the left tail of the distribution. In particular, "worker type 8" experiences severe wage compression if matched with extreme low firm classes such as 1 or 2.

When worker-firm interactions are assumed to be "additive separable", particular interactions are smoothed out as shown by Figure 4b, the lines become parallel, which is the quintessential feature of the additive separability assumption: workers and firms contribute to the wage generation function by adding their respective "values". That means worker type 10, for example, if transferred from firm class 10 to 1, should not lose the part of their wage

---

[8]Figure 3 reveals that "extreme complementarity" matches are approximately 5 percent of the total.

that belongs purely to their components.

K = 10, L = 10
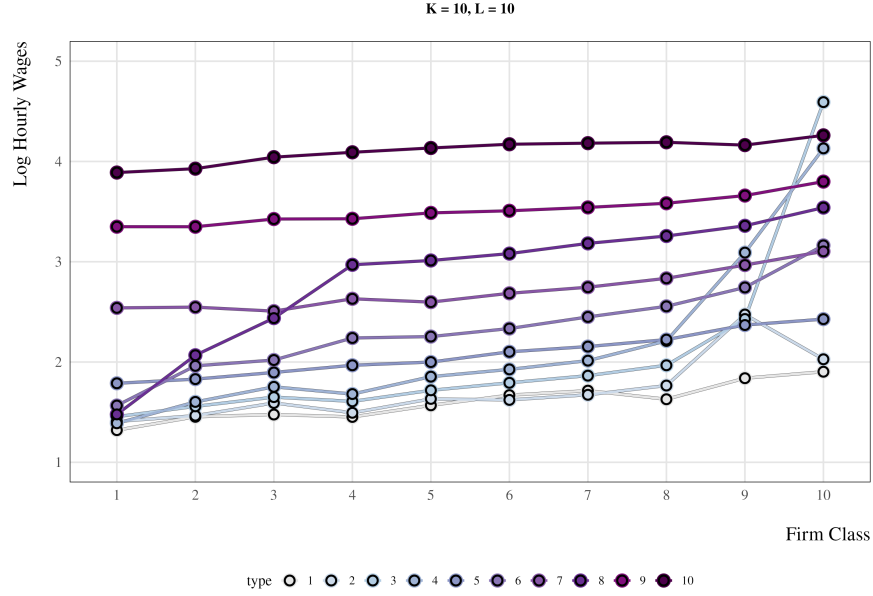
(a) Gaussian Mixture Estimated Wages per Match

K = 10, L = 10
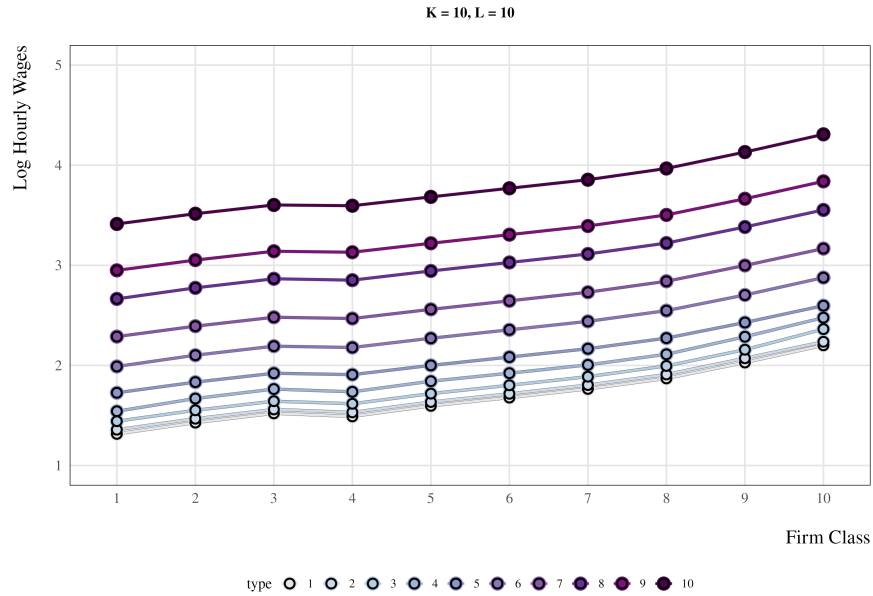
(b) Additive Separable Prediction of Wages per Match

Figure 4: Payment Schedules of worker-firm interactions under Gaussian mixture estimates and predicted linear model.

Note: Panel (a) generated by using estimated means and variances of each Gaussian component of the mixture. Panel (b) generated by running a two-way fixed effect estimation with firm classes and worker types as fixed effects, weighted by the number of workers per each match.

31
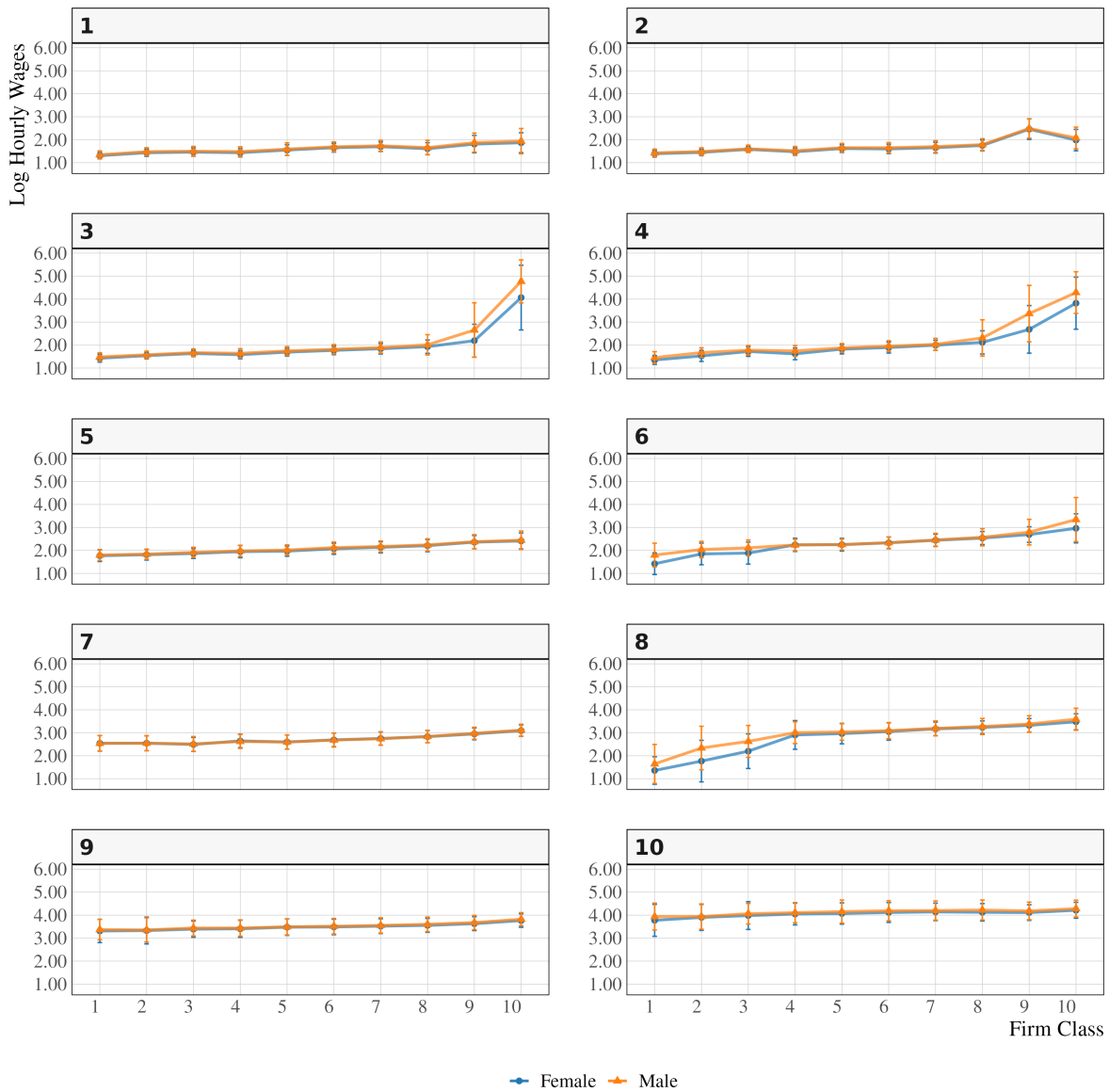
Figure 5: Pay Schedules by Gender, Firm Class, and Worker Type

Note: [1]Each panel represents a worker type payment schedule for each firm class, in log hourly wages, grouped by gender. [2]Points indicate the mean log hourly wage, and error bars represent one standard deviation of the estimated wage distribution.

5.3.1 Gender-wise Payment Schedules

A natural question that arises is to what extent these particularities matter for the creation

of pay differentials between male and female workers. I construct Figure 5 by expanding

Figure 4a, separating each worker type into a male and a female components.

Complementarity effect arising from those matches implies significantly higher wages for men as compared to women, in particular for the worker types 3, 4, and 8, for which the deviation from a "separable setting" is most extreme. For example, type 3 matched with firm class 10 yields almost 1 log hourly wage gap. Type 4 under the same match yields about 50 log-points in gap.

Type 8, and to a lesser extent type 6, exhibit negative deviations as they approach the lower extreme of firm productivity. When these worker types, characterized by moderate-to-high human capital accumulation, are found in low-productivity firms, a compression effect on wages emerges. In these cases, the expected wage falls below the sum of the expected firm and worker effects. Female workers are more susceptible to these unfavorable matches compared to their male counterparts.

5.3.2 Workers Under Comparative Advantage

To understand better the differences being under a comparative advantage match and otherwise in the labor market, I compare individuals with sufficiently close payments but differing in matches. Specifically, I compare the type 3 and type 10 male and female workers when hired by the class of firm 10. Table 3 provide a descriptive statistics of these workers.

The table highlights distinct differences in education, age distribution, and occupations between both types of workers when under firm class 10, male and female. Both groups, regardless of gender, possess a high concentration of college degree individuals, with "type 10 workers" having slightly more. Age is also similar, whereas less than 30 years old female workers are more likely to be found under type 3, while 31-50 are marginally more likely to be found under type 10.

For occupation, while both types display a higher concentration of scientific and liberal

33

Table 3: Workers Under Complementarity and Non-complementarity Matches in Firm Class 10

	Type 3 Workers		Type 10 Workers	
	Female	Male	Female	Male
	(1)	(2)	(3)	(4)
Education And Age				
Dropout	0.04	0.01	0.00	0.01
High School Graduates	0.12	0.06	0.03	0.04
College	0.84	0.92	0.97	0.95
Age (<30)	0.15	0.06	0.10	0.09
Age 31-50	0.69	0.66	0.74	0.70
Age (≥ 51)	0.13	0.24	0.14	0.18
Occupation Statistics				
Scientific and Liberal Arts	0.20	0.25	0.36	0.37
Technicians	0.04	0.03	0.08	0.11
Administrative	0.18	0.07	0.19	0.13
Managers	0.52	0.62	0.35	0.36
Traders	0.05	0.01	0.02	0.02
Rural	0.00	0.00	0.00	0.00
Factory	0.01	0.01	0.01	0.01
Mean experience (years)	6.375	7.302	8.417	8.288
Mean Log-Wage	4.308	4.866	4.204	4.270
Variance of Log-Wage	1.836	0.715	0.094	0.116
Worker-years observations	5718	18 401	73 869	157 282
Number of Workers	4895	15 542	44 309	93 073
Fraction of Women	0.24	0.76	0.32	0.68

Notes: [1]Under complementarity matches are type 3 and type 4. Without complementarity is type 10 match, which in wage levels is comparable to matches under complementarity. [2] Education, age, and occupation statistics are fractions that may not necessarily add to one due to rounding.

arts [9], there is a much higher concentration of managers.

Even with similarities, there is some evidence that individuals in complementarity effect matches might have leading positions and are particularly valuable for firms to employ. Such individuals would be suffering more extreme wage compressions if hired elsewhere, while worker type 10 experiences a more predictable wage path along firm classes.

6 Discussions: Monte Carlo Simulation and Variance Decomposition

In this section, I introduce a novel decomposition of the gender wage gap that accounts for complementarity effects in the labor market. I decompose the gender wage gap into three distinct components. The first component captures the contribution of complementarity effects, which I isolate by constructing a counterfactual labor market without comparative (dis)advantage matches.

The remaining two components are inspired by Card, Cardoso, and Kline (2016). The second component, referred to as the "sorting" component, reflects the impact of firm allocation on the wage gap. I calculate this by simulating male and female labor markets where all factors are held constant except for the distribution of firms.

The final component, the "bargaining" component, represents the wage gap contribution arising when equally productive individuals are employed by firms of the same class, but a gender-based differential persists. I isolate this effect through a simulation in which male and female labor markets share identical distributions of workers and firms, while the means and variances of each gender-wise Gaussian distribution remains as observed in the original data.

The "bargaining" and "complementarity" components share certain similarities in nature.

---

[9]These categories are generated from the Brazilian Code for Occupation, and tend to have similarities with other codes internationally. "Scientific and Liberal Arts' is a generic code that summarizes economists, engineers, lawyers, professors, among others, whose jobs under normal circumstances require at least a degree from a superior institution of learning or education.

The complementarity effect can be viewed as a subset of the bargaining effect in the context of a CCK framework, as both reflect differences in returns for similar individuals within the same firm. However, the Gaussian mixture model allows me to distinguish between these two components, as it identifies labor market matches where wages deviate from the assumption of additive separability. As a result, complementarity effects emerge only in these specific labor market settings, whereas bargaining is more applicable in contexts where the additive separability condition holds.

6.1 Monte Carlo Simulations

To setup the Monte Carlo Simulations, I first calculate the realized moments of every worker type and firm class match in the labor market for the male and the female sample. Then I calculate the unconditional probabilities of worker types and firm classes for male and female[10]

To create a separable market, I match workers following a "diagonal pattern in matches" in Figure 4a. That means type worker 10 is guaranteed to work in firm class 10 as long as there is a spot available. When firm class 10 job slots are filled, firm class 9 starts hiring the best available, until all jobs are filled with workers. Figure 6 shows the resulting conditional probabilities of worker types given firm classes and gender under a separable market.

In a labor market characterized by strictly additive separability, reshuffling matches is expected to have negligible effects on overall wage levels (Graham, Imbens, and Ridder, 2014). Therefore, I leverage this fact, and the fact that "diagonal" matches do not yield large complementarity effects, to construct a labor market that behaves under the additive separability assumption. As a robustness check, I also perform my analysis using the weighted linear regression predicted fixed effects shown in Figure 4b.

---

[10]Means and the standard deviations of each match are shown in Figure 5. The unconditional probabilities are shown in Figure 2.
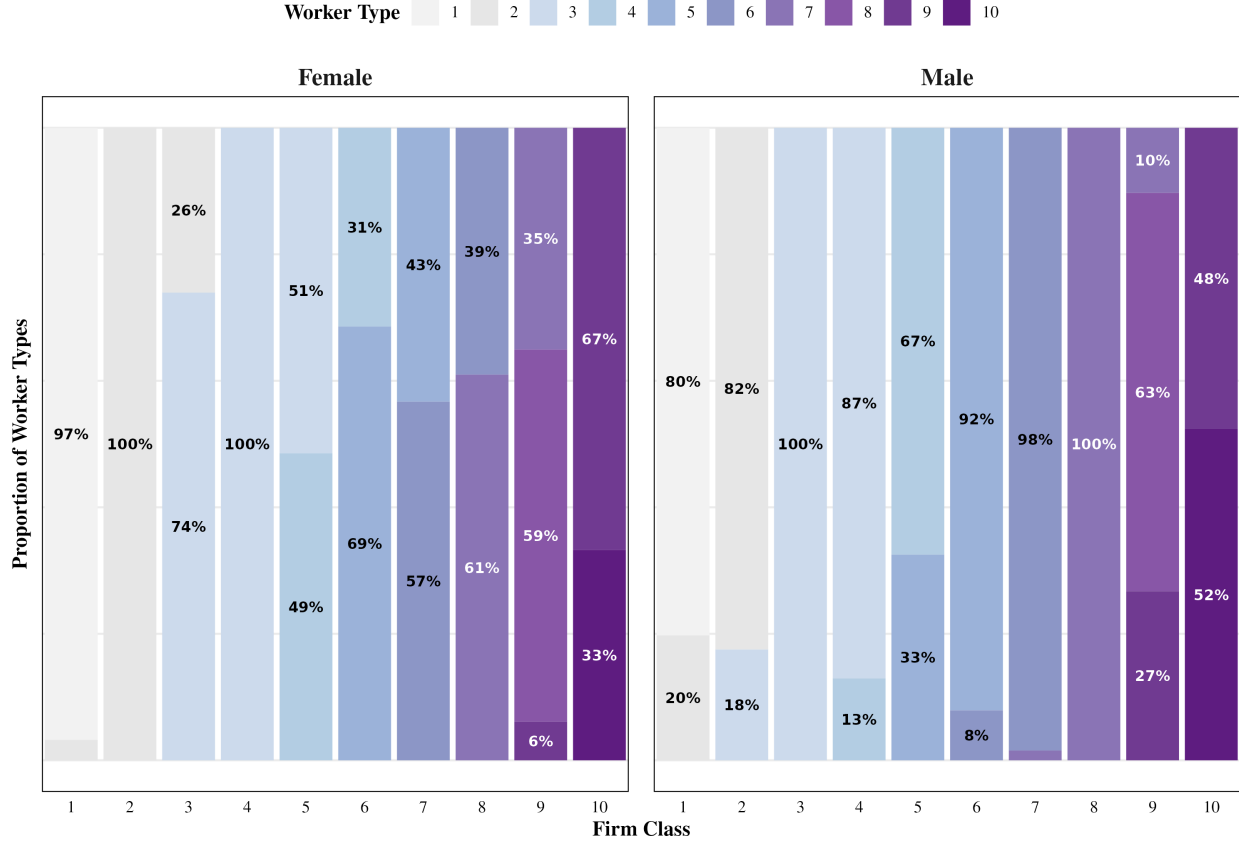
Figure 6: Conditional Probabilities of Worker Types Given Firm Classes and Gender, Under a Separable Market

Note: Probabilities calculated by generating a labor market where "top workers" are guaranteed to match with "top firms", until all positions are filled, following Becker (1973)'s principle of assortative matching.

The difference in the gender wage gap between the separable labor market and the original setting reflects the contribution of complementarity effects that are not captured in the separable model. For CCK's "bargaining" and "sorting" components, I conduct simulations within the separable labor market to ensure that these components remain distinct and do not overlap.

6.1.1 Simulation Results

Table 4 presents the results of the simulations. The first row displays the observed mean log hourly wages for female and male workers, along with the difference in log points, for

the overall dataset, as well as broken down by education cohort and age group. The first column reports the difference in means calculated using the Gaussian distribution, which closely mirrors the observed gender wage gap. The second column reflects the gender wage gap in the counterfactual world where the labor market is additively separable.

The next three columns represent the contributions of the three proposed components. The first column shows the difference between the baseline and the separable market, indicating the contribution of complementarity effects to the gender wage gap. The "sorting" component reflects the contribution of firm allocation to the wage gap, holding all else constant in the separable market except the distribution of firms. Lastly, the "bargaining" component captures the contribution to the wage gap when all else is held constant in the separable market, except for the observed moments (means and variances) of each Gaussian distribution.

Overall, the complementarity effect contributes approximately 16 to 17 percent to the gender wage gap, indicating that disparities arising from comparative advantage matches play a significant role in generating wage differences between male and female workers. Labor market allocation accounts for about 37.5 percent, while differences in bargaining without considering complementarities contribute roughly 8.3 percent. Together, these components explain nearly two-thirds of the gender wage gap. These findings demonstrate that non-separable, two-sided heterogeneity models, such as the Gaussian mixture approach of BLM, more effectively capture the substantial role that firms play in contributing to the gender wage gap, both horizontally as vertically.

The gender wage gap is smaller among individuals with a high school education but reaches its peak among those with college degrees. Furthermore, changes in the gender wage gap are not primarily driven by firm allocations; instead, they are largely explained by complementarity effects, especially for individuals with college degrees. This suggests that women with high levels of human capital, who are in positions of comparative advantage, are particularly susceptible to wage disparities arising from these effects.

Another evidence of the human capital accumulation and complementarity effect positive

Table 4: Gaussian Mixture Decomposition of Gender Wage Gaps

Group	Baseline Market Gap (1)	Separable Market Gap (2)	Contribution to Gender Wage Gap		
			Complementarity Contribution (3)	Sorting Contribution (4)	Bargaining Contribution (5)
All	-0.24	-0.20	-0.04 (16.7%)	-0.09 (37.5%)	-0.02 (8.3%)
Education					
No high-school	-0.30	-0.27	-0.03 (10.0%)	-0.13 (43.3%)	-0.03 (10.0%)
High-school	-0.23	-0.22	-0.01 (4.3%)	-0.08 (34.7%)	-0.02 (8.7%)
College	-0.35	-0.30	-0.05 (14.3%)	-0.16 (45.7%)	-0.04 (11.4%)
Age					
<30	-0.09	-0.08	-0.01 (11.1%)	-0.03 (33.3%)	-0.02 (22.2%)
31-50	-0.30	-0.26	-0.04 (13.3%)	-0.13 (43.3%)	-0.03 (10.0%)
51>	-0.33	-0.28	-0.05 (15.2%)	-0.13 (39.4%)	-0.03 (9.1%)

Notes: [1]All values represent log wage gaps (female - male). Baseline Gap is the observed gap. [2]Separable Market Gap assumes interactions do not yield complementarity effects. [3]Complementarity Contribution is the difference between Baseline and Separable Market gaps. [4]Sorting Contribution is the reduction in gap after equalizing means and variances of worker-firm interactions. [5]Bargaining Contribution is the reduction in the gap after equalizing firm probabilities. [7]Both sorting and bargaining are calculated under a separable market. [6]Numbers in parentheses show the percentage of the Baseline Gap explained by each component.

correlation is the age analysis. While sorting increases its role in wage differentials for individuals older than 30 years, it stabilizes for older than 51, while complementarity effect contribution keeps increasing slightly.

Table 5: Gaussian Mixture Decomposition of Gender Wage Gaps - Firm sizes and occupations

Group	Baseline Market Gap (1)	Separable Market Gap (2)	Complementarity Contribution (3)	Sorting Contribution (4)	Bargaining Contribution (5)
			Contribution to Gender Wage Gap		
All	-0.24	-0.20	-0.04 (16.7%)	-0.09 (37.5%)	-0.02 (8.3%)
Firm Size					
Firms <10	-0.12	-0.12	0.00 (0.0%)	-0.02 (16.7%)	-0.03 (25.0%)
Firms 10-50	-0.14	-0.13	-0.01 (7.1%)	-0.03 (21.4%)	-0.01 (7.1%)
Firms 51>	-0.21	-0.17	-0.04 (19.0%)	-0.08 (38.1%)	-0.02 (9.5%)
Occupations					
Hotel and Restaurants	-0.12	-0.12	0.00 (0.0%)	-0.04 (33.3%)	-0.03 (25.0%)
Economists and Engineers	-0.39	-0.35	-0.04 (10.3%)	-0.16 (41.0%)	-0.05 (12.8%)
Managers	-0.33	-0.22	-0.11 (33.3%)	-0.08 (24.2%)	-0.03 (9.1%)

Notes: [1]All values represent log wage gaps (female - male). Baseline Gap is the observed gap. [2]Separable Market Gap assumes interactions do not yield complementarity effects. [3]Complementarity Contribution is the difference between Baseline and Separable Market gaps. [4]Sorting Contribution is the reduction in gap after equalizing means and variances of worker-firm interactions. [5]Bargaining Contribution is the reduction in the gap after equalizing firm probabilities. [7]Both sorting and bargaining are calculated under a separable market. [6]Numbers in parentheses show the percentage of the Baseline Gap explained by each component.

To further expand my study, I present Table 5 which shows the results of simulations based on samples of different firm sizes and occupations particularly relevant to this study. I focus on three categories of occupations. The first, "Hotels and Restaurants," includes workers directly involved in the hospitality sector, such as waiters, kitchen staff, and cleaners.

The second category, "economists and engineers", is self-explanatory. The rationale for

selecting these professions lies in the fact that, in Brazil, these fields are highly regulated, meaning that only individuals with the appropriate college degree are legally permitted to practice. This choice offers two key advantages: it controls for college diplomas that are more uniform in their practice than, for instance, medical doctors, but also focusing on degrees that typically lead to higher compensation in the labor market.

Finally, the "Managers" category includes workers in leadership positions. These workers are likely to possess high levels of firm-specific human capital, giving them a strong comparative advantage in the labor market.

If comparative advantage drives complementarity effects on the gender wage gap, then strategic positions in the labor market, those involving valuable human capital accumulation and leadership roles, and the size of the firm, associated with bargaining power, should reveal particularly high levels of these effects.

The firm size panel of Table 5 reveals that the gender wage gap increases with firm size, particularly due to the complementarity and the sorting effect, meaning women not only are more likely to be found in low-paying firms when controlling for larger firms, but also in positions where men are receiving much higher complementarity compensations.

The last three rows of Table 5 represent the results of the simulation of different occupations. Hotel and restaurants are typically occupations assumed in the literature to possess high turnover rate and zero firm premium in wages Card, Cardoso, and Kline (2016) and Casarico and Lattanzio (2024). Therefore, these occupations are expected to have negligible comparative advantage effects. Accordingly, my results suggest that the hotel and restaurants labor market is mostly governed by the additive separable assumption, given that the simulated separable market yielded the exact same wage differentials as the baseline market gap[11], confirming that there is no complementarity effect. However, as the expected human capital accumulation is increased, the gap increases. While a considerable portion of the gap is due to firm allocations for economists and engineers, the complementarity contribution

---

[11]For the male and female wage levels for all simulations, refer to Table **??**

represents 10 percent of the gender wage gap.

For managers, the distance in wage differentials between the baseline market and the separable market is the largest, with the complementarity contribution accounting for about 33 percent of the gap. Moreover, the sorting contribution drastically reduces, from 16 to 8 log-points, falling from 41 percent in contribution to 24 percent.

Under an additive separable model, such as linear regression, the results would suggest that labor market allocations are the primary drivers of the gender wage gap among managers[12]. However, in a non-separable model, I can identify that a substantial portion of the previously unexplained differential is due to specific labor market matches that generate complementarity effects. Because additive separable models assume constant returns to unobserved heterogeneity of workers and firms, these contributions are difficult to capture accurately.

6.2 Robustness Checks

I perform a series of exercises to show my results are not sensitive to particular choice of parameters. While keeping the optimal number of firm classes according to the gap statistics ($K = 10$), I vary the number of worker types, which are the number of Gaussians observed in each firm class. I test with $L = 6$, $L = 10$, and $L = 12$. I also provide an alternative simulation of the separable market where I use a weighted ordinary least squares with firm classes and worker types as fixed effects. Following Equation 17, the weighted parameter is the fraction of workers of each worker type-firm class match. For the weighted OLS, I maintain $L = 10$.

For the alternative number of worker types, results were consistent across all specifications, with the exception of $L = 12$, that seemed to underestimate complementarity effects, putting slightly more contributions to the sorting and the bargaining contribution. The alternative

---

[12]In another example, Card, Cardoso, and Kline (2016) found that firm-related factors contributed approximately 4 percent to the gender wage gap after controlling for managers.

separable market simulation yielded virtually the same estimates as the original, with a more conservative estimation of the complementarity effects contribution to the gender wage gap.

Despite some differences in estimation, overall, the results indicate that my measurements are not driven by errors arising from the Gaussian mixture estimation, local maxima, or a particular setting.

z

6.3   Variance Decomposition

A variance decomposition of log wages in works related to Abowd, Kramarz, and Margolis (1999) (AKM models). It decomposes the variance of log wages into five distinct components: (1) the contribution of worker fixed effects, (2) the contribution of firm fixed effects, (3) twice the covariance between worker and firm effects, (4) the variance of time-varying covariates and their associated covariances, typically captured by period dummies interacted with time-varying human capital indicators, and (5) the residual variance.

The AKM model, however, tends to negatively correlate worker and firm effects (Andrews, et al., 2008), implying a downward bias estimate for assortative matching. Bonhomme, Lamadon, and Manresa (2019) proposed using the dimension reduction technique relying in the Gaussian mixture model to mitigate the bias.

Card, Cardoso, and Kline (2016) found that approximately 10 percent of wage variance can be attributed to assortative matching for both male and female workers. In this section, I apply the framework of Bonhomme, Lamadon, and Manresa (2019) to examine the extent to which assortative matching may be underestimated in the wage variance decomposition used in AKM gender analysis.

I compare three models. The first specification uses clustered firm and individual worker identifiers, related to Bonhomme and Manresa (2015), this clustering approach allows the researcher to maintain the linear assumption but reduces the negative bias in assortative matching. I name this model "clustered AKM", or C-AKM, in which fixed effects for firms are

now the firm classes[13]. In the second setting I employ the full BLM approach by leveraging both worker types and firm classes.

Finally, I test the variance decomposition analysis under a classical CCK approach which uses individual firm and worker identifiers as fixed effects. It requires the largest dual connected set of firms and bias correction. In this study, I use the bootstrapping approximation of Azkarate-Askasua and Zerecero (2023) to correct the assortative matching bias.

Formally, the regression setting is:

$$w_{it} = \Phi^g_{K(i,t)} + \Lambda^g_{L(i)} + x'_{it}\beta + \varepsilon_{it} \tag{18}$$

And the variance decomposition can be formally stated as:

$$\underbrace{\text{Var}(w_{it})}_{\text{Log Hourly Wage Variance}} = \underbrace{\text{Var}(\Phi^g_{K(i,t)})}_{\text{Firm Class Variance}} + \underbrace{\text{Var}(\Lambda^g_{L(i)})}_{\text{Worker Type Variance}} + \underbrace{\text{Var}(\varepsilon_{it})}_{\text{Residual Variance}} \tag{19}$$

$$+ \underbrace{\text{Var}(x'_{it}\beta) + 2 \cdot \text{Cov}(\Lambda^g_{L(i)}, x'_{it}\beta) + 2 \cdot \text{Cov}(\Phi^g_{K(i,t)}, x'_{it}\beta)}_{\text{Time-Varying Covariates Variance and Associated Covariances}}$$

$$+ \underbrace{2 \cdot \text{Cov}(\Phi^g_{K(i,t)}, \Lambda^g_{L(i)})}_{\text{Worker Type and Firm Class Covariance}}$$

where w_{it} represents the logarithmic hourly wage of worker i in period t, decomposed as follows: $\Phi^g_{K(i,t)}$ represents the firm effects, where $K(i,t)$ is the assignment function. Here, K can denote either firm classes or individual firm identifiers. The term $\Lambda^g_{L(i)}$ represents individual worker or worker type effects, where $L(i)$ can refer to either a worker type or an individual identifier. Time-varying covariates are represented by $x'_{it}\beta$, while gender heterogeneity is accounted for by the superscript g. Finally, ε_{it} denotes the idiosyncratic error term.

The covariance between worker and firm effects is of particular interest in understanding

---

[13]See Appendix Section C and D for discussions on the bias (also often dubbed "limited mobility bias" in the literature) and using the clustered AKM method to perform a KOB decomposition on the gender wage gap.

the dynamics of assortative matching and its impact on the gender wage gap. This component can be potentially underestimated due to the limited mobility bias. I test three variance decompositions from three different settings. The C-AKM model, by clustering firms, potentially reduces the noise in firm effect estimates, allowing for a more stable estimation of the worker-firm covariance, however, it still relies on individual fixed effects. The CCK model under the bootstrapping correction provides a "lower bound" of these estimates, given that the bootstrapping correction is an approximation, not a total mitigation. The BLM model, employing both worker and firm clusters, provides a framework that effectively circumvents the limited mobility bias by coarsening job movements in the dataset at the cost of noisier results.

Table 6 presents the variance decomposition results. The first two columns show the results for the BLM decomposition, columns (3) and (4), the clustered firm AKM methodology. Finally, columns (5) and (6) represents the results for the classical AKM approach from CCK.

The first row presents the total variance of log hourly wages by gender. Both the C-AKM and BLM models yield similar magnitudes, as they utilize the full set of worker observations. In contrast, the CCK model relies on the connected set of firms through job movers, which tends to overrepresent larger firms, resulting in higher wage variance estimates.

Male firm effect contribution to the wage variance ranges from 5 percent in the male sample under CCK, to 10 percent under BLM. On the other hand, female firm effect contribution ranges from 3.6 in the clustered AKM model, to 8.8 percent under BLM.

Although AKM models attribute the largest portion of wage variance to worker effects (over 70 percent) for both genders, in the BLM approach worker effects account for a smaller, though still significant, share of wage variance: 41.3 percent for women and 32.1 percent for men.

The reduction in worker effect contribution can be explained by the larger worker-firm covariance contribution, at 26.0 and 18.8 percent of the total variance for women and men. Compared to the "AKM" model, the BLM and the C-AKM model were more effective in

Table 6: Variance Decomposition of log hourly Wages

	BLM		Clustered AKM		CCK	
	Female (1)	Male (2)	Female (3)	Male (4)	Female (5)	Male (6)
Var(log hourly wage)	0.529	0.650	0.529	0.650	0.642	0.797
*Panel A: **Variance Estimates***						
Firm effects	0.046	0.066	0.020	0.023	0.042	0.044
Worker effects	0.367	0.357	0.377	0.467	0.479	0.626
Time-varying covariates	0.008	0.011	0.009	0.011	0.006	0.008
Cov(Worker, Firm)	0.137	0.122	0.114	0.136	0.106	0.108
Residual	0.062	0.094	0.010	0.012	0.009	0.011
*Panel B: **Share of Total Variance (%)***						
Firm effects	8.8	10.2	3.7	3.6	6.6	5.5
Worker effects	55.2	54.9	71.2	71.9	74.6	78.5
Time-varying covariates	1.6	1.6	1.6	1.7	1.0	1.1
Cov(Worker, Firm)	26.0	18.8	21.5	21.0	16.5	13.6
Residual	11.6	14.5	1.9	1.8	1.4	1.3

Notes: [1]AKM, BLM, and CCK stand for Abowd, Kramarz, and Margolis (1999), Bonhomme, Lamadon, and Manresa (2019), and Card, Cardoso, and Kline (2016), respectively. [2]Clustered AKM represents firm clustered using a kmeans algorithm and individual worker identifiers as parameters. [3]Panel A showcases the magnitude of estimated variance components, while Panel B presents these components as percentages of the total log hourly wage variance. [4]Results are a weighted average based on the six biennial samples' number of observations.

capturing assortative matching effects.

Worker-firm covariance is consistently higher for women across all specifications, suggesting that assortative matching is indeed a meaningful contributor to wage dispersion in the labor market, especially for female workers. My results are particularly relevant in the context of recent discussions on the rise of assortativity in labor markets, as highlighted by Song, et al. (2019), who documented the increasing trend of assortative matching in the United States. If men are systematically more likely to find high-paying matches in the labor market, while women are concentrated in lower-paying positions, this could exacerbate the gender wage gap.

7   Conclusion

Additive separable models are unable to capture labor market interactions that generate wages based on comparative advantage, where the match between firms and workers results in compensation that exceeds (or are less than) the simple sum of individual worker and firm contributions.

In this paper, I deviate from linear additive models. I use a linked employer-employee dataset covering all firms and workers from São Paulo, Brazil (2010-2017), to apply the two-sided unobserved heterogeneity framework introduced by Bonhomme, Lamadon, and Manresa (2019). This approach allows me to investigate the contribution of specific worker-firm interactions to the gender wage gap, assuming each interaction generates wages drawn from log-normal distributions. This method allows me to capture the complementarity effects arising from particular worker-firm assortative matching.

Employing Monte Carlo Simulations, I propose a novel decomposition of the gender wage gap into three components. Following Card, Cardoso, and Kline (2016), the sorting component, representing labor market allocations, and bargaining component, representing differences in negotiation of equally productive workers under the same firm. The third

47

component, the complementarity component, is a special case of "bargaining", however, under these matches wage levels do not correspond to the predicted from additive separable model.

I find a positive relationship between human capital and these complementarity effects. They are more pronounced for male workers compared to female workers, accounting for approximately 17 percent of the overall gender wage gap. This contribution go as far as a third of the gender wage gap for individuals in leadership positions. Controlling for occupations that generally require lower levels of human capital, such as occupations related to the hospitality sector, yielded negligible results.

I also find that these interactions are more present at the tails of the wage distribution where larger firms operate and wage dispersion is higher. For firms larger than 50 employees, about a fourth of the gender wage gap is explained by these complementarity effects.

My study demonstrates that firms and the broader labor market structure play a more significant role in shaping the gender wage gap than previously recognized. I demonstrate that firms not only provide varying wage premiums but also evaluate human capital and other worker characteristics in heterogeneous ways. This differential valuation of worker attributes across firms contributes substantially to gender-based wage disparities.

The pronounced complementarity effects observed in managerial positions suggest that policies aimed at increasing transparency in the labor market and promoting key leadership roles among female workers are essential for reducing gender wage disparities.

Future research could leverage on the increased availability of linked employer-employee data and computational power. They could extend the analysis by providing a dynamic framework and exploring how worker-firm interactions evolve over time in response to earnings shocks and their influence on mobility decisions. Incorporating collective bargaining data would further enhance our understanding of the non-monetary factors that shape gender-specific sorting patterns. Expanding this methodological approach to different countries could provide valuable cross-national insights into the extent to which gender wage gaps are driven by universal factors or are shaped by specific institutional and cultural contexts.

References

Abowd, J. M., R. H. Creecy, and F. Kramarz (Mar. 2002). "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data". en. In: *Longitudinal Employer-Household Dynamics Technical Papers*. Number: 2002-06 Publisher: Center for Economic Studies, U.S. Census Bureau. URL: https://ideas.repec.org//p/cen/tpaper/2002-06.html (visited on Dec. 8, 2023).

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). "High Wage Workers and High Wage Firms". en. In: *Econometrica* 67.2, pp. 251–333. ISSN: 1468-0262. DOI: 10.1111/1468-0262.00020. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00020 (visited on Sept. 20, 2023).

Andrews, M. J., L. Gill, T. Schank, and R. Upward (June 2008). "High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?" In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 171.3, pp. 673–697. ISSN: 0964-1998. DOI: 10.1111/j.1467-985X.2007.00533.x. URL: https://doi.org/10.1111/j.1467-985X.2007.00533.x (visited on May 17, 2024).

Azkarate-Askasua, M. and M. Zerecero (Jan. 2023). *Correcting Small Sample Bias in Linear Models with Many Covariates*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4322300. URL: https://papers.ssrn.com/abstract=4322300 (visited on Feb. 5, 2024).

Barth, E. and H. Dale-Olsen (2003). "Assortative matching in the labour market? Stylised facts about workers and plants". en. In: *Institute for Social Research, Oslo, Norway*.

Becker, G. S. (July 1973). "A Theory of Marriage: Part I". In: *Journal of Political Economy* 81.4. Publisher: The University of Chicago Press, pp. 813–846. ISSN: 0022-3808. DOI: 10.1086/260084. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/260084 (visited on Feb. 11, 2024).

Bertrand, M., S. E. Black, S. Jensen, and A. Lleras-Muney (Jan. 2019). "Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway". In: *The Review of Economic Studies* 86.1, pp. 191–239. ISSN: 0034-6527. DOI: 10.1093/restud/rdy032. URL: https://doi.org/10.1093/restud/rdy032 (visited on July 30, 2024).

Bertrand, M., C. Goldin, and L. F. Katz (July 2010). "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors". en. In: *American Economic Journal: Applied Economics* 2.3, pp. 228–255. ISSN: 1945-7782. DOI: 10.1257/app.2.3.228. URL: https://www.aeaweb.org/articles?id=10.1257/app.2.3.228 (visited on July 30, 2024).

Black, D. A., A. M. Haviland, S. G. Sanders, and L. J. Taylor (July 2008). "Gender Wage Disparities among the Highly Educated". en. In: *Journal of Human Resources* 43.3. Publisher: University of Wisconsin Press Section: Articles, pp. 630–659. ISSN: 0022-166X, 1548-8004. DOI: 10.3368/jhr.43.3.630. URL: https://jhr.uwpress.org/content/43/3/630 (visited on Sept. 12, 2024).

Blau, F. D. and L. M. Kahn (2008). "Women's Work and Wages". en. In: *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan UK, pp. 1–14. ISBN: 978-1-349-

95121-5. DOI: 10.1057/978-1-349-95121-5_2207-1. URL: https://doi.org/10.1057/978-1-349-95121-5_2207-1 (visited on Sept. 12, 2024).

Blinder, A. S. (1973). "Wage Discrimination: Reduced Form and Structural Estimates". In: *The Journal of Human Resources* 8.4. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System], pp. 436–455. ISSN: 0022-166X. DOI: 10.2307/144855. URL: https://www.jstor.org/stable/144855 (visited on Aug. 22, 2023).

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler (Apr. 2023). "How Much Should We Trust Estimates of Firm Effects and Worker Sorting?" In: *Journal of Labor Economics* 41.2. Publisher: The University of Chicago Press, pp. 291–322. ISSN: 0734-306X. DOI: 10.1086/720009. URL: https://www.journals.uchicago.edu/doi/full/10.1086/720009 (visited on Mar. 7, 2024).

Bonhomme, S., T. Lamadon, and E. Manresa (2019). "A Distributional Framework for Matched Employer Employee Data". en. In: *Econometrica* 87.3, pp. 699–739. ISSN: 1468-0262. DOI: 10.3982/ECTA15722. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15722 (visited on Aug. 21, 2023).

Bonhomme, S. and E. Manresa (2015). "Grouped Patterns of Heterogeneity in Panel Data". en. In: *Econometrica* 83.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA11319, pp. 1147–1184. ISSN: 1468-0262. DOI: 10.3982/ECTA11319. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11319 (visited on Sept. 25, 2023).

Bruns, B. (Apr. 2019). "Changes in Workplace Heterogeneity and How They Widen the Gender Wage Gap". en. In: *American Economic Journal: Applied Economics* 11.2, pp. 74–113. ISSN: 1945-7782. DOI: 10.1257/app.20160664. URL: https://www.aeaweb.org/articles?id=10.1257/app.20160664 (visited on May 29, 2024).

Card, D., A. R. Cardoso, and P. Kline (May 2016). "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women *". In: *The Quarterly Journal of Economics* 131.2, pp. 633–686. ISSN: 0033-5533. DOI: 10.1093/qje/qjv038. URL: https://doi.org/10.1093/qje/qjv038 (visited on Aug. 21, 2023).

Card, D., J. Heining, and P. Kline (Aug. 2013). "Workplace Heterogeneity and the Rise of West German Wage Inequality*". In: *The Quarterly Journal of Economics* 128.3, pp. 967–1015. ISSN: 0033-5533. DOI: 10.1093/qje/qjt006. URL: https://doi.org/10.1093/qje/qjt006 (visited on Sept. 25, 2023).

Card, D., I. Schmutte, and L. Vilhuber (Apr. 2023). "Introduction to the Special Issue: Models of linked employer–employee data: Twenty years after "High Wage Workers and High Wage Firms"". en. In: *Journal of Econometrics* 233.2, pp. 333–339. ISSN: 03044076. DOI: 10.1016/j.jeconom.2023.01.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407623000337 (visited on Aug. 5, 2024).

Casarico, A. and S. Lattanzio (Apr. 2024). "What Firms Do: Gender Inequality in Linked Employer-Employee Data". In: *Journal of Labor Economics* 42.2. Publisher: The University of Chicago Press, pp. 325–355. ISSN: 0734-306X. DOI: 10.1086/723177. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/723177 (visited on Apr. 8, 2024).

Ceci, S. J., D. K. Ginther, S. Kahn, and W. M. Williams (Dec. 2014). "Women in Academic Science: A Changing Landscape". en. In: *Psychological Science in the Public Interest* 15.3. Publisher: SAGE Publications Inc, pp. 75–141. ISSN: 1529-1006. DOI: 10.1177/

1529100614541236. URL: https://doi.org/10.1177/1529100614541236 (visited on Sept. 12, 2024).

Coudin, E., S. Maillard, and M. Tô (2018). *Family, firms and the gender wage gap in France*. eng. Working Paper W18/01. IFS Working Papers. DOI: 10.1920/wp.ifs.2018.W1801. URL: https://www.econstor.eu/handle/10419/200290 (visited on July 7, 2024).

Cruz, G. and T. Rau (Apr. 2022). "The effects of equal pay laws on firm pay premiums: Evidence from Chile". In: *Labour Economics* 75, p. 102135. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2022.102135. URL: https://www.sciencedirect.com/science/article/pii/S0927537122000288 (visited on July 7, 2024).

Eeckhout, J. and P. Kircher (2018). "Assortative Matching With Large Firms". en. In: *Econometrica* 86.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA14450, pp. 85–132. ISSN: 1468-0262. DOI: 10.3982/ECTA14450. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14450 (visited on Feb. 11, 2024).

Gallen, Y., R. V. Lesner, and R. Vejlin (Jan. 2019). "The labor market gender gap in Denmark: Sorting out the past 30 years". In: *Labour Economics* 56, pp. 58–67. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2018.11.003. URL: https://www.sciencedirect.com/science/article/pii/S0927537118301234 (visited on Mar. 1, 2024).

Gaure, S. (2014). "Correlation bias correction in two-way fixed-effects linear regression". en. In: *Stat* 3.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sta4.68, pp. 379–390. ISSN: 2049-1573. DOI: 10.1002/sta4.68. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.68 (visited on Oct. 23, 2023).

Gerard, F., L. Lagos, E. Severnini, and D. Card (Oct. 2021). "Assortative Matching or Exclusionary Hiring? The Impact of Employment and Pay Policies on Racial Wage Differences in Brazil". en. In: *American Economic Review* 111.10, pp. 3418–3457. ISSN: 0002-8282. DOI: 10.1257/aer.20181596. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20181596 (visited on Mar. 7, 2024).

Goldin, C. (Apr. 2014). "A Grand Gender Convergence: Its Last Chapter". en. In: *American Economic Review* 104.4, pp. 1091–1119. ISSN: 0002-8282. DOI: 10.1257/aer.104.4.1091. URL: https://www.aeaweb.org/articles?id=10.1257/aer.104.4.1091 (visited on Sept. 20, 2023).

Goldin, C., L. F. Katz, and I. Kuziemko (Dec. 2006). "The Homecoming of American College Women: The Reversal of the College Gender Gap". en. In: *Journal of Economic Perspectives* 20.4, pp. 133–156. ISSN: 0895-3309. DOI: 10.1257/jep.20.4.133. URL: https://www.aeaweb.org/articles?id=10.1257%2Fjep.20.4.133&fbclid=IwAR2YgNWGj5pMpZpEpG5a97ahef4Dwi8Wk7rUpLqqRGPMqwOGOnJYmGvcLys (visited on Sept. 12, 2024).

Graham, B. S., G. W. Imbens, and G. Ridder (2014). "Complementarity and aggregate implications of assortative matching: A nonparametric analysis". en. In: *Quantitative Economics* 5.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE45, pp. 29–66. ISSN: 1759-7331. DOI: 10.3982/QE45. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/QE45 (visited on May 14, 2024).

Gruetter, M. and R. Lalive (Apr. 2009). "The importance of firms in wage determination". In: *Labour Economics* 16.2, pp. 149–160. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2008.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0927537108001012 (visited on May 17, 2024).

Jewell, S. L., G. Razzu, and C. Singleton (2020). "Who Works for Whom and the UK Gender Pay Gap". en. In: *British Journal of Industrial Relations* 58.1, pp. 50–81. ISSN: 1467-8543. DOI: 10.1111/bjir.12497. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjir.12497 (visited on July 21, 2024).

Kitagawa, E. M. (Dec. 1955). "Components of a Difference Between Two Rates*". In: *Journal of the American Statistical Association* 50.272. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.1955.10501299, pp. 1168–1194. ISSN: 0162-1459. DOI: 10.1080/01621459.1955.10501299. URL: https://doi.org/10.1080/01621459.1955.10501299 (visited on July 30, 2024).

Kline, P., R. Saggio, and M. Sølvsten (2020). "Leave-Out Estimation of Variance Components". en. In: *Econometrica* 88.5, pp. 1859–1898. ISSN: 0012-9682. DOI: 10.3982/ECTA16410. URL: https://www.econometricsociety.org/doi/10.3982/ECTA16410 (visited on July 11, 2024).

Lachowska, M., A. Mas, R. Saggio, and S. A. Woodbury (Apr. 2023). "Do firm effects drift? Evidence from Washington administrative data". en. In: *Journal of Econometrics* 233.2, pp. 375–395. ISSN: 03044076. DOI: 10.1016/j.jeconom.2021.12.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407622000604 (visited on Aug. 5, 2024).

Lavetti, K. and I. Schmutte (Apr. 2023). "Gender differences in sorting on wages and risk". In: *Journal of Econometrics* 233.2, pp. 507–523. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2022.06.012. URL: https://www.sciencedirect.com/science/article/pii/S030440762200183X (visited on Oct. 7, 2023).

MacQueen, J. (1967). "Some Methods for classification and analysis of multivariate observations". en. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Google-Books-ID: IC4Ku_7dBFUC. URL: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

Masso, J., J. Meriküll, and P. Vahter (June 2022). "The role of firms in the gender wage gap". In: *Journal of Comparative Economics* 50.2, pp. 454–473. ISSN: 0147-5967. DOI: 10.1016/j.jce.2021.10.001. URL: https://www.sciencedirect.com/science/article/pii/S0147596721000639 (visited on Mar. 1, 2024).

Mulligan, C. B. and Y. Rubinstein (Aug. 2008). "Selection, Investment, and Women's Relative Wages Over Time*". In: *The Quarterly Journal of Economics* 123.3, pp. 1061–1110. ISSN: 0033-5533. DOI: 10.1162/qjec.2008.123.3.1061. URL: https://doi.org/10.1162/qjec.2008.123.3.1061 (visited on Sept. 30, 2024).

Oaxaca, R. (1973). "Male-Female Wage Differentials in Urban Labor Markets". In: *International Economic Review* 14.3. Publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University], pp. 693–709. ISSN: 0020-6598. DOI: 10.2307/2525981. URL: https://www.jstor.org/stable/2525981 (visited on Aug. 22, 2023).

OECD (2024). *OECD Family Database - OECD*. URL: https://www.oecd.org/social/family/database.htm (visited on Mar. 1, 2024).

Shimer, R. and L. Smith (2000). "Assortative Matching and Search". en. In: *Econometrica* 68.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00112, pp. 343–369. ISSN: 1468-0262. DOI: 10.1111/1468-0262.00112. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00112 (visited on Feb. 16, 2024).

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. von Wachter (Feb. 2019). "Firming Up Inequality*". In: *The Quarterly Journal of Economics* 134.1, pp. 1–50. ISSN: 0033-5533. DOI: 10.1093/qje/qjy025. URL: https://doi.org/10.1093/qje/qjy025 (visited on July 13, 2024).

Tibshirani, R., G. Walther, and T. Hastie (2001). "Estimating the Number of Clusters in a Data Set via the Gap Statistic". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63.2. Publisher: [Royal Statistical Society, Wiley], pp. 411–423. ISSN: 1369-7412. URL: https://www.jstor.org/stable/2680607 (visited on July 5, 2024).

Woodcock, S. D. (Aug. 2008). "Wage differentials in the presence of unobserved worker, firm, and match heterogeneity". In: *Labour Economics*. European Association of Labour Economists 19th annual conference / Firms and Employees 15.4, pp. 771–793. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2007.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0927537107000395 (visited on May 17, 2024).

Appendix A   Cluster Choice Analysis

A common drawback of clustering methods is the optimal number of clusters. To increase the robustness of my analysis, I employ the gap statistic method, a widely used technique in cluster analysis (Tibshirani, Walther, and Hastie, 2001), to explore the within variance of clusters and therefore, choose the right number.

The gap statistic compares the total within-cluster variation for different values of k with their expected values under a null reference distribution of the data. I calculate it through the following steps:

1. For each number of clusters K, I compute the within-cluster sum of squares $W_k = \sum_{k=1}^{K} \sum_{i \in k} (w_i - \bar{w}_k)^2$ where k is given cluster, w_i is the observed worker i log-weekly wage, $\bar{w}_k$ the empirical mean of cluster k (centroid).

2. I generate B reference datasets by sampling uniformly from the range of my observed data. For each reference dataset, I compute W_{kb}, the within-cluster sum of squares when clustering the reference data into k clusters.

3. I then compute the gap statistic as:

$$Gap(k) = \frac{1}{B} \sum_{b=1}^{B} \log(W_{kb}) - \log(W_k)$$

4. I use the bootstrapped standard deviation as the standard error across the B reference datasets. I implement this procedure with $B = 500$ to ensure stable estimates.

5. Finally, I choose the optimal number of clusters as the smallest k such that:

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

where s_{k+1} is the standard deviation for $k+1$ clusters.

Figure A.1 presents the gap statistic values for different numbers of clusters, ranging from 4 to 25. The green dashed line at 10 clusters represents what my analysis suggests as the optimal number of clusters based on the gap statistic. Due to the large number of observations, standard errors were negligible, barely noticeable in the plot. Nevertheless, there is a notable elbow in the curve, indicating diminishing returns to increasing the number of clusters beyond the value of 10, implying that the benefit of additional clusters in explaining data variability becomes less substantial.
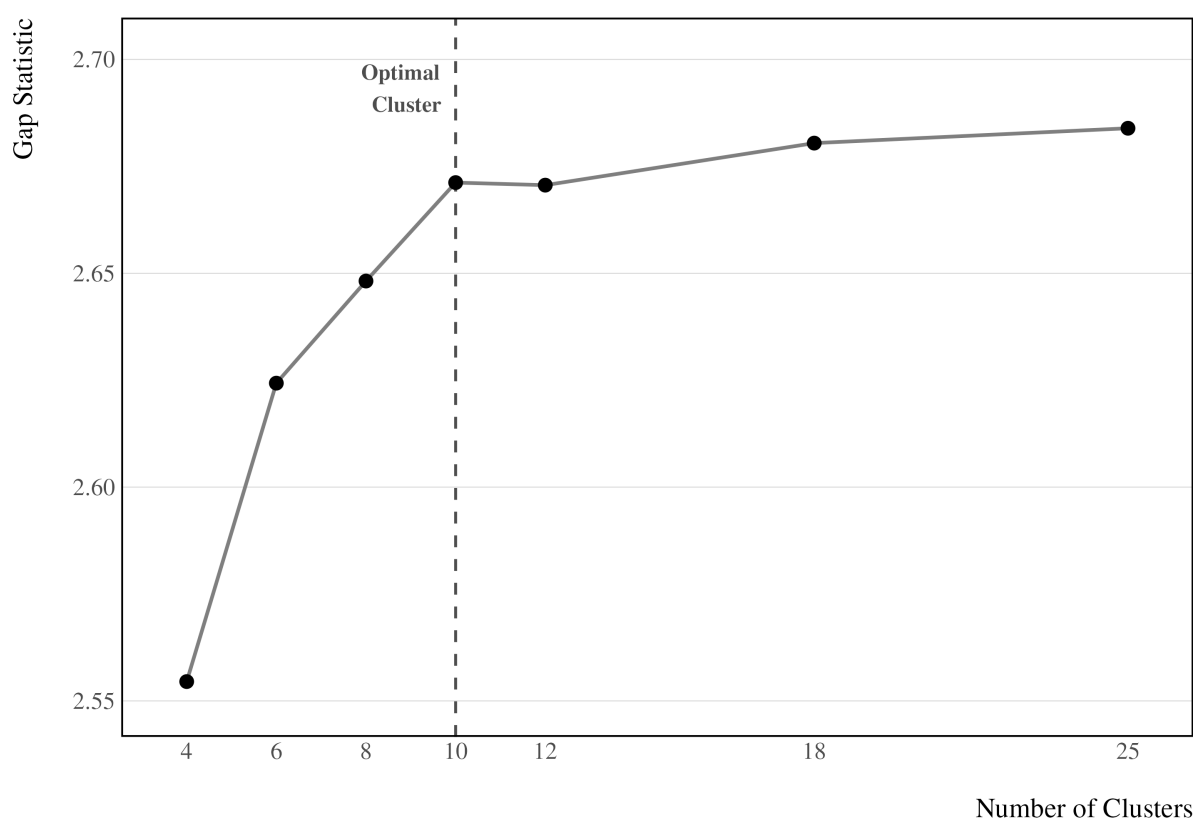
Figure A.1: Point Estimate Gap Statistic by Number of Clusters

Note: [1] Vertical line indicates the number of clusters of choice (10). [2] Firm classes estimated by a k-means clustering algorithm using as measurement their empirical cumulative distribution function supported by the ventiles of the population, over six biennials (from 2010-12 to 2015-17). [3] Bootstrapped standard errors were negligible due to the number of observations, with $p < 0.01$.

Appendix B   Worker Mobility in Firm Clusters

In this subsection, I explore the exogenous mobility assumption of the Gaussian mixture model. This assumption states that the movement of workers should be related to worker types and firm classes, but not directly on earnings. Therefore, the expected wage on unobservables should be zero for job movers.

To test this assumption, first I observe job movers within and across clusters. Movements are considered as long as for each gender workers are changing firms from the first period to the second period. I separate these movements into three categories. Upward movements represent when the worker moves from a lower cluster to an upper cluster. Downward is otherwise. Lateral movements are within cluster job changes.

Figure B.1 is constructed by first running a regression following Equation 31. Than I plot the difference in residuals for every transfer, gender, and data sample cell, discriminated by the movement type. Each dot represents a transfer cell observed in the labor market. The size of the dot indicates how common this particular transfer is.

The figure serves two purposes. First, it addresses obvious trends in job changes, which could indicate that unobservable factors not captured in my model are influencing mobility decisions. A lack of symmetry in the figure would suggest that certain labor market transitions are driven by such unobservables. Second, and equally important, I differentiate these movements by gender to identify any discrepancies that may be endogenous to my model but related to gender differences.

The symmetry plot provides robust evidence supporting the exogenous mobility assumption, which is fundamental to the proper identification of my model. It shows that every movement type possess examples of positive and negative difference in residuals, strongly strongly indicating that job transitions in the labor market are primarily governed by stable firm wage policies and worker characteristics, rather than by time-varying, unobserved factors correlated with wages. Last but not least, the analysis also reveals no apparent gender-specific

patterns that would undermine the exogenous mobility assumption for male or female worker samples separately.
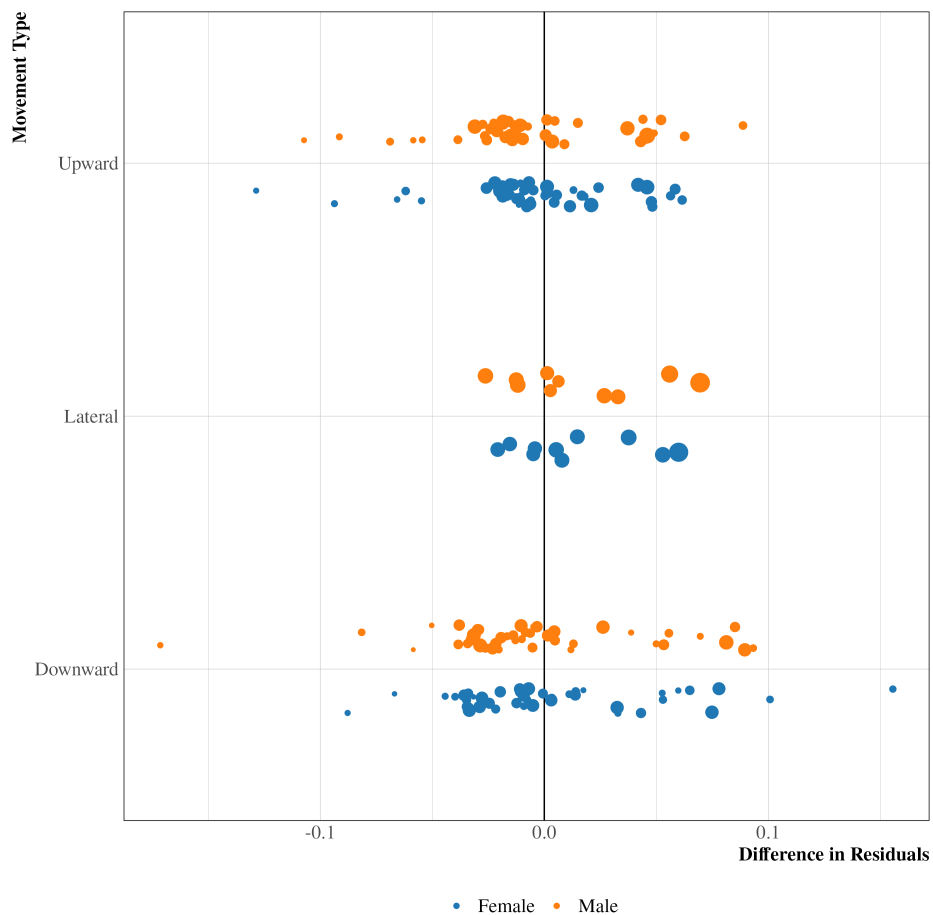
Figure B.1: Symmetry plot of job movers' difference in residuals from first to second period.

Note: [1]Dots represent across and within cluster-gender-sample cell movements. [2]Dot size represents each cell's number of observed movements.

Appendix C   AKM and the Limited Mobility Bias

Here I briefly explain the presence of bias in the AKM estimator of Abowd, Kramarz, and Margolis (1999).

C.1   The AKM Model

The AKM is formally written as:

$$w_{it} = X'_{it}\beta + \alpha_i + \phi_{J(i,t)} + \varepsilon_{it} \tag{20}$$

where w_{it} are the log earnings of worker i in time t, $X'_{it}\beta$ are exogenous covariates such as age or time period, α_i is the unobserved worker heterogeneity, $J(i,t)$ is an assignment function representing the firm where i works at t, meaning $\phi_{J(i,t)}$ represents the unobserved firm heterogeneity, and ε_{it} is the error term.

Following Bonhomme, Holzheu, et al. (2023), assume N is the number of workers, J the number of firms. For convenience, assume $T = 2$ is the number of time periods. The following assumption must hold:

$$\mathbb{E}[\varepsilon_{it}|X_{11}, \ldots, X_{NT}, j(1,1), \ldots, j(N,T), \alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_J] = 0 \tag{21}$$

It is possible, without loss of generality, to rewrite Equation 20 partialing out $X\beta$ and in vector form. Still following Bonhomme, Holzheu, et al. (2023), I have:

$$W = A\gamma + \varepsilon \tag{22}$$

without loss of generality, assume W is subtracted from $X\beta$ and A represents the column-space of worker and firm identifiers.

C.2 Connected Set

In matched employer-employee data, the matrix AA' is typically singular, necessitating an additional data cleaning step to ensure a sample of workers and firms that renders AA' non-singular. This step is crucial for the identification of firm and worker effects in additive wage models. For instance, Card, Cardoso, and Kline (2016) estimates gender-specific firm wage effects by isolating the largest dual connected set from the main sample.
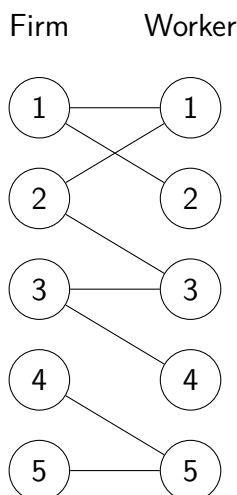
Firm   Worker

Figure C.1: Firm-worker pairs. Firms 1, 2, and 3 are in the largest connected set through workers 1, 2, 3, and 4. Firms 4 and 5 are connected through worker 5 but disjoint from the rest.

This concept of connectivity in the labor market is illustrated in Figure C.1, which provides a simplified representation of worker movements across firms[14]. It depicts a labor market with five firms and five workers over two time periods. Worker 1 moves from Firm 1 to Firm 2, thereby connecting these two firms. Firm 2 is further connected to Firm 3 through the movement of Worker 3. Firms 4 and 5, while isolated from the first three firms, are connected to each other through Worker 5.

In studies employing matched employer-employee data under additive separability models relying on firm and worker identifiers, researchers typically sample the largest connected set of firms. However, when investigating worker heterogeneity between genders, it is necessary

---

[14]For a comprehensive explanation of connected sets and their extraction from data, see Abowd, Creecy, and Kramarz (2002).

to use the dual connected set, defined as the intersection of the largest connected sets for male and female samples. This approach ensures that firm effects are identified and comparable between both gender groups.

C.3 Limited Mobility Bias

The limited mobility bias is a significant concern in the estimation of firm effects, arising from the relative scarcity of job movers in the labor market (Andrews, et al., 2008; Bonhomme, Holzheu, et al., 2023). While this bias does not directly appear in the firm effects estimates from Equation 20, it manifests in the variance analyses that are commonly employed in the literature to decompose wage inequality.

The sample variances or covariances of interest can be expressed in matrix notation as:

$$\sigma^2 = \gamma' Q \gamma \tag{23}$$

where Q is a matrix that depends on the design matrix A.

Andrews, et al. (2008) demonstrated the existence of this bias by decomposing the estimator $\hat{\sigma}^2$:

$$\mathbb{E}[\hat{\sigma}^2 | A] = \gamma' Q \gamma + \operatorname{trace}(A(A'A)^{-1} Q (A'A)^{-1} A' \mathbb{V}[\varepsilon | A]) = \sigma^2 + \xi \tag{24}$$

where ξ represents the bias term.

Directly correcting for this bias is computationally challenging, as it requires inverting a large matrix, often of dimensions in the hundreds of thousands for firms and millions for workers in typical matched employer-employee datasets. Bonhomme, Holzheu, et al. (2023) have shown that common approximations used in the literature may be insufficient, particularly when relying on fixed effects derived from identifiers. This insufficiency stems from the fact that these approximations often fail to fully account for the complex network structure of worker mobility across firms.

To address these challenges, Bonhomme, Lamadon, and Manresa (2019) proposed a dimension reduction framework. This approach groups firms and workers into a smaller number of classes, thereby increasing the relative probability of observed job changes between groups. While this method effectively mitigates the limited mobility bias, it comes at the cost of imposing additional structure on the estimation model.

C.4   Firm Size and Connectivity

Not only does the largest connected set requirement impose a bias due to the rarity of mobility, but it also alters the overall wage distribution of the data. This alteration stems from the fact that the largest connected set tends to include larger firms more frequently than smaller ones. If larger firms differ significantly in their payment schedules and behavior compared to their smaller counterparts, the results derived from such analyses may have limited external validity.

In this section, I provide a formal proof that larger firms are more likely to be included in a connected set of a matched employer-employee dataset. I begin by defining the probability of worker mobility between firms and then demonstrate how this probability scales with firm size.

Without loss of generality, assume $T = 2$. Let $\mathcal{J} = \{1, \ldots, J\}$ be the set of all firms in the economy, and let N_j denote the number of workers in firm j. Define $p_{jj'}$ as the probability that a given worker moves from firm j to another firm j'. For simplicity, assume that this probability is the same across all workers in the labor market.

Definition 1 (Connected Set) *A connected set $\mathcal{C} \subseteq \mathcal{F}$ is a subset of firms such that for any two firms $j, j' \in \mathcal{C}$, there exists a sequence of firms $j_1, \ldots, j_C \in \mathcal{C}$ with $j_1 = j$, $j_C = j'$ and for each connection $c \in \{1, \ldots, C - 1\}$, there is at least one worker who has been employed in both i_c and i_{c+1}.*

Lemma 1 (Probability of Observed Mobility) *The probability of observing at least one worker moving from firm j to firm j' is:*

$$P(j \to j') = 1 - (1 - p_{jj'})^{N_j} \tag{25}$$

Proof. The probability of a single worker not moving from j to j' is $(1 - p_{jj'})$. Assume that for all N_j workers not to move, this must occur independently for each worker. Thus, the probability that no workers are moving is $(1 - p_{jj'})^{N_j}$, and the probability that at least one worker is moving is the complement of this event. ∎

Theorem 1 *The probability of a firm being part of the connected set is increasing in firm size.*

Proof. For firm j to be part of the connected set, it must have at least one worker moving to or from another firm in the set. The probability of firm j being connected is:

$$P(j \in \mathcal{C}) = 1 - \prod_{j' \neq j} (1 - P(j \to j')) \cdot \prod_{j'' \neq j} (1 - P(j'' \to j)) \tag{26}$$

Substituting the result from Lemma 1:

$$P(j \in \mathcal{C}) = 1 - \prod_{j' \neq j} (1 - p_{jj'})^{N_j} \cdot \prod_{j'' \neq j} (1 - p_{j''j})^{N_{j''}} \tag{27}$$

To show that this probability increases with firm size, take the derivative with respect to N_j:

Taking the derivative with respect to N_j:

$$\frac{\partial P(j \in \mathcal{C})}{\partial N_j} = - \left(\prod_{j' \neq j} (1 - p_{jj'})^{N_j} \cdot \prod_{j'' \neq j} (1 - p_{j''j})^{N_{j''}} \right) \cdot \sum_{j' \neq j} \log(1 - p_{jj'}) \tag{28}$$

Since $0 < p_{jj'} < 1$, we have $\log(1 - p_{jj'}) < 0$, and thus $\frac{\partial P(j \in \mathcal{C})}{\partial N_j} > 0$. ∎

Corollary 1 *As firm size approaches infinity, the probability of being in the connected set approaches 1:*

$$\lim_{N_j \to \infty} P(j \in \mathcal{C}) = 1 \tag{29}$$

Proof. As $N_j \to \infty$, $(1 - p_{jj'})^{N_j} \to 0$ since $0 < p_{jj'} < 1$. Therefore, the product $\prod_{j \neq j'}(1 - p_{jj'})^{N_j} \to 0$, and consequently, $P(j \in \mathcal{C}) \to 1$. $\blacksquare$

Other firms approaching infinity Another consequence of this proof is when the size of any other firm N_{j_0} approaches infinity while N_j remains finite, $P(j \in \mathcal{C})$ also approaches 1, but at a slower rate. This is because:

$$\lim_{N_{j_0} \to \infty} P(j \in \mathcal{C}) = 1 - \prod_{j' \neq j}(1 - p_{jj'})^{N_j} \cdot 0 \cdot \prod_{j'' \neq j, j'' \neq j_0}(1 - p_{jj''})^{N_j''} = 1 \tag{30}$$

However, this convergence is slower than when $N_j \to \infty$ because only one term in the product approaches zero, rather than all terms involving N_j.

C.4.1 Empirical Evidence

The theoretical framework is substantiated by empirical evidence presented in Figures C.2 and C.3, which illustrate the differences between the full sample and the largest dual connected set (LDCS) in terms of firm size and wage distributions.

Figure C.2 reveals a stark contrast in the distribution of workforce size between the full sample and the LDCS. The LDCS exhibits a symmetrical distribution shifted significantly to the right, with a mean firm size of approximately 194 workers, compared to the full sample's mean of 27 workers. This rightward shift is accompanied by increased variability, with the standard deviation in the LDCS being almost four times higher than in the original sample. The median firm size in the LDCS is also notably higher, underscoring the overrepresentation of larger firms in the connected set.

The wage distribution, as depicted in Figure C.3, further emphasizes the discrepancies between the full sample and the LDCS. The mean log wage in the full sample is 1.87, with a

standard deviation of 0.446, while the LDCS shows a substantially higher mean log wage of 2.32 and a larger standard deviation of 0.656. This upward shift in both moments indicates a clear upward bias in wage levels within the connected set. The increased standard deviation in the LDCS also points to greater wage dispersion among the firms included in this subset.

These findings have some implications for the estimation and interpretation of gender wage gaps. LDCS exhibits a larger gender wage gap (δ = -0.317) compared to the entire sample (δ = -0.237), suggesting that studies using the connected set may overestimate the overall wage disparity. This overestimation likely stems from the LDCS that captures wage dynamics primarily in larger, more established firms where gender wage differences might be more pronounced. The exclusion of smaller, potentially lower-paying firms that do not meet the connectivity requirements for AKM-style fixed effects estimation contributes to this bias.

Researchers should exercise caution when generalizing results from the connected set to the broader labor market. The LDCS, while providing the necessary conditions for certain econometric techniques, may not fully represent the wage structures and gender dynamics present in smaller or less connected firms. This limitation is particularly important when studying labor markets with a significant proportion of small enterprises or sectors with limited inter-firm mobility.
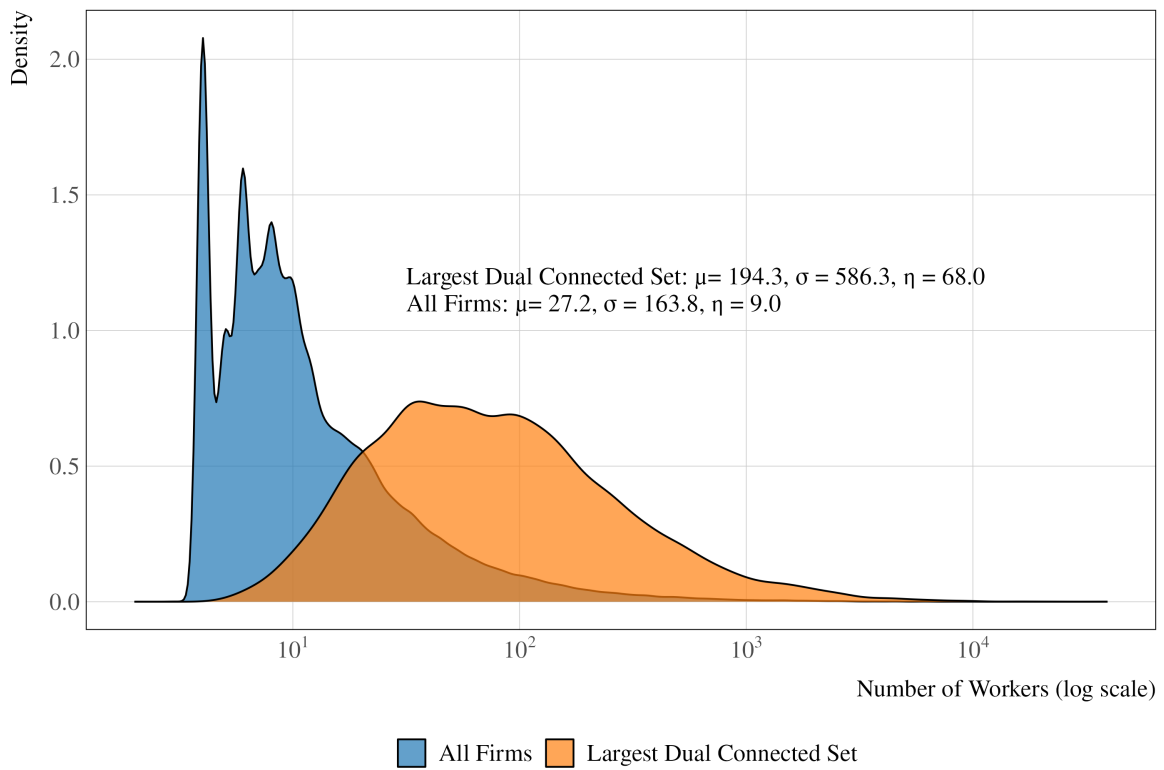
Figure C.2: All Firms and The LDCS Number of Workers Distributions

Note: [1]Distributions generated from the six biennial samples, using the full firm set, and the largest dual connected set of firms. [2]μ is the mean of the distribution, σ represents the standard deviation, η represents the median number of workers per firm.
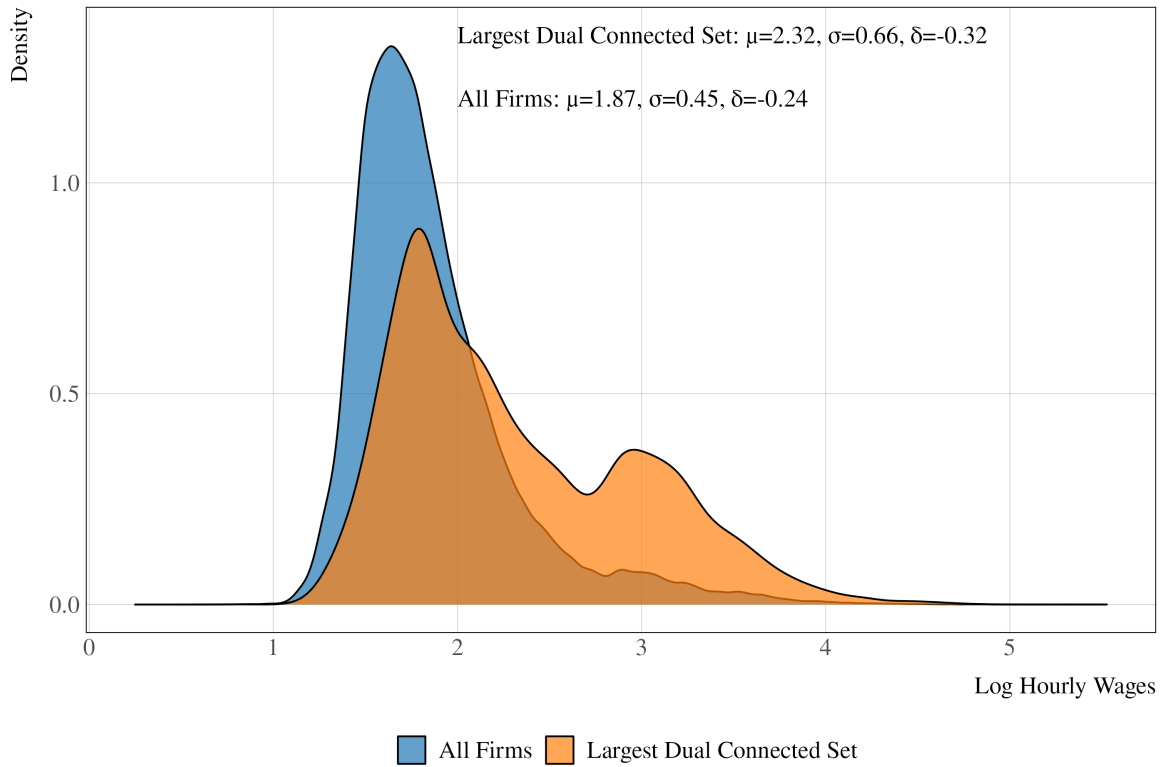
Figure C.3: All Firms and The LDCS Log-Weekly Wage Distributions

Note: [1]Distributions generated from the six biennial samples, using the full firm set, and the largest dual connected set of firms. [2]μ is the mean of the distribution, σ represents the standard deviation, δ represents the average female-male wage gap.

Appendix D   Applying Estimated Clusters in a Linear Framework

Here I provide two exercises to show the mixture model can be used as a plugin estimator of unobserved heterogeneity in a linear regression setting.

First, I employ the estimated firm clusters to estimate firm heterogeneity contribution to the gender wage gap in a classical AKM-KOB from Card, Cardoso, and Kline (2016). The novelty is that I keep social identifiers as the worker plugin estimators, however, I leverage the k-means clusters of firms to avoid data trimming.

Second, I provide a variance decomposition analysis, comparing classical AKM decomposition with the clustered AKM provided previously and the BLM decomposition, where I replace worker social identifiers with their respective mixture membership.

D.1   Estimating Firm Contribution to the Gender Wage Gap

Here, I employ the estimated firm clusters in an AKM-KOB classical linear framework to examine their contribution to the gender wage gap. The advantage of this approach is that by grouping firms with similar wage structures, I am not required to filter the dual connected set from the data.

The model I employ has strong similarities to the typical AKM framework. A key distinction lies in the treatment of firm heterogeneity. Instead of utilizing the standard firm assignment function $J(i,t)$, which maps each worker-year observation to a specific firm, I introduce cluster assignment function $K(i,t)$. This function maps worker-year observations to firm clusters, thereby reducing the dimensionality of the firm entities.

Formally, the wage equation can be expressed as:

$$w_{it} = \alpha_i + \psi^g_{K(i,t)} + X^{g\prime}_{it}\beta + \varepsilon_{it} \tag{31}$$

where w_{it} is the log wage of worker i in year t, α_i is the worker fixed effect, $\psi_{K(i,t)}$ is the effect of the firm cluster to which worker i's employer belongs in year t, X_{it} is a vector of time-varying covariates, and ε_{it} is the error term. Superscript g indicates that I apply the regression to both the male and female samples.

In particular, the key parameter of interest in this analysis is the difference in firm endowment between male and female workers, derived from the KOB decomposition. This parameter represents the differential distribution of workers across firm clusters by gender, serving as a preliminary measure of gender-specific assortative matching patterns in the labor market, although focusing solely on firm heterogeneity.

Specifically, I adapt the KOB decomposition to the context of firm cluster effects, formalizing it as:

$$\underbrace{E[\psi^f_{K(i,t)}|female] - E[\psi^m_{K(i,t)}|male]}_{\substack{\text{Firm Cluster} \\ \text{Contribution to the} \\ \text{Gender Wage Gap}}} = \tag{32}$$

$$\text{Bargaining:} \ \ \frac{1}{2} \underbrace{\sum_{x \in F,M} \left(E[\psi^F_{K(i,t)} - \psi^M_{K(i,t)}|g=x] \right)}_{\substack{\text{Unexplained Portion} \\ \text{(Difference in Returns)}}} \tag{33}$$

$$\text{Sorting:} \ \ + \frac{1}{2} \underbrace{\sum_{x \in F,M} \left(E[\psi^x_{K(i,t)}|g=F] - E[\psi^x_{K(i,t)}|g=M] \right)}_{\substack{\text{Explained Portion} \\ \text{(Difference in Distributions)}}} \tag{34}$$

where the left-hand side of the equation represents the total contribution of firm cluster effects to the gender wage gap. This contribution is decomposed into two components: the bargaining effect and the sorting effect.

The bargaining effect, following the terminology of Card, Cardoso, and Kline (2016), captures the portion of the gap attributable to differences in the estimated firm premia between men and women, holding the distribution of firms constant. This effect is computed

as the average of two counterfactuals: one using the observed female distribution of firms and another using the observed male distribution.

The sorting effect, conversely, measures the portion of the gap that arises from differences in the distribution of men and women across firm clusters, assuming gender-neutral firm effects. This is the difference in endowments of the Oaxaca decomposition. This effect is also computed as the average of two counterfactuals: one using the estimated male returns to firm clusters and another using the estimated female returns.

By averaging these counterfactuals for each component, as suggested by Casarico and Lattanzio (2024), I obtain robust estimates that account for potential sensitivity to the choice of reference group. Unless otherwise specified, the reported bargaining and sorting effects refer to these averaged estimates.

D.1.1 Normalizing Firm Effects

There is established practice in the AKM literature on gender wage gaps in using the hotel and restaurant industry as a reference[15]. This sector is often chosen due to its typical low wage premia and high turnover rates, suggesting minimal rents (Card, Cardoso, and Kline, 2016; Coudin, Maillard, and Tô, 2018).

Although class 2 firms have a slightly higher proportion of hotels and restaurants (1 percent against 0.6 percent in class 1)[16], I argue that class 1 is the most appropriate reference for several reasons. First, conditional average wages on firm fluster are fairly linear, with class 1 exhibiting the lowest average wage premium in my pooled regression, aligning with the theoretical expectation that the reference group should represent firms offering minimal rents. When employing the AKM regression with class 1 as the reference, resulting fixed effects preserve the linear behavior, with no class exhibiting negative estimates.

---

[15]Examples are Cruz and Rau (2022) and Casarico and Lattanzio (2024)

[16]See Figure E.1 for the estimated premia and the proportion of hotels and restaurants per firm class

D.1.2 Oaxaca Decomposition of Firm Cluster Effects

The main findings derived from the estimation of Equation 31 are summarized in Table D.1. These results represent weighted average firm cluster effects estimates obtained from the six separate biennial samples.

Table D.1: Firm Decomposition of the Gender Wage Gap: Overall and by Subgroups

Group	Total Gap	Contribution to Gender Wage Gap		
		Firm Components	Sorting Components	Bargaining Components
	(1)	(2)	(3)	(4)
All	−0.237	−0.033	−0.022	−0.01
		(0.14)	(0.09)	(0.04)
By age group:				
Up to age 30	−0.092	−0.014	−0.009	−0.005
		(0.16)	(0.1)	(0.06)
Ages 31-50	−0.305	−0.041	−0.029	−0.012
		(0.14)	(0.09)	(0.04)
Over age 50	−0.326	−0.076	−0.028	−0.048
		(0.23)	(0.09)	(0.15)
By education group:				
No High school	−0.297	−0.046	−0.035	−0.011
		(0.15)	(0.12)	(0.04)
High school	−0.233	−0.021	−0.025	0.004
		(0.09)	(0.11)	(−0.02)
College	−0.348	−0.075	−0.03	−0.046
		(0.22)	(0.08)	(0.13)

Notes: [1]This table presents the decomposition of the gender wage gap into components attributable to clustered firm-specific factors, using Equation 31. Column (1) shows the total female-male wage gap in means. Column (2) presents the total contribution of firm-specific factors. Columns (3) and (4) further decompose the firm premium contribution into a sorting (explained) and a bargaining (unexplained) components, respectively. [2]Numbers in parenthesis represent the fraction of the overall gender wage gap that is attributed to the source described in column heading. [3]Results are an weighted average of the six biennial samples.

First, considering the overall sample, there is a substantial gender wage gap of 23.7 log points. Firm-specific factors, in Column (2), are estimated around 3.1 log points. As

explained by Card, Cardoso, and Kline (2016), this component can be interpreted as the difference in rent payment relative to firm class 1. The component accounts for 13 percent of this total gap.

This contribution can be further decomposed into the sorting component, the difference in male and female worker distributions considering a gender-neutral relative rent. In this case, the total difference is evaluated at 2.1 log points, corresponding to approximately 9 percent of the total gender wage gap. Likewise, the bargaining channel is the average difference in the estimated premium, assuming both genders possess the same firm share. This channel is measured at 1 log point, about 4 percent of the overall gender wage gap.

Following the literature, the lower rows of Table D.1 show that the gender wage gap increases dramatically with age, with firms playing a role in this increase, since for individuals older than 50, 23 percent of the 32.6 wage gap is due to estimated firm class effects. Notably, while the sorting component remains relatively stable across age groups (ranging from 0.8 to 2.7 log points), the Bargaining component increases substantially, from 0.7 log points for the youngest group to 4.7 log points for the oldest. This pattern suggests that as workers age, differences in how firms compensate men and women in similar positions become increasingly important in explaining the gender wage gap.

The analysis by education level reveals a pattern that aligns with findings from the U.S. labor markets, given the wage gap is more prominent among college-educated workers (34.8 log points), albeit the smallest among those with a high school education (23.3 log points). Workers without a high school diploma fall in between, with a gap of 29.7 log points. Interestingly, the contribution of firm-specific factors to the gap follows a similar pattern, being the highest for college-educated workers (5.6 log points) and the lowest for high school graduates (2.8 log points).

For workers without a high school education and those with a high school diploma, the sorting component dominates the bargaining component. This suggests that for these groups, the allocation of women across firms plays a more significant role in the gender wage gap than

71

within-firm differences in compensation. However, the picture changes for college-educated workers. In this group, the bargaining component (3.0 log points) marginally exceeds the sorting component (2.6 log points), indicating that within-firm differences in compensation between men and women become more pronounced as individuals accumulate human capital. Individuals with higher levels of human capital tend to be allocated to more specialized occupations, often within larger firms. However, women may be concentrated at lower paying occupations compared to male counterparts, resulting in workers with greater human capital accumulation securing positions that ultimately translates into heterogeneous bargaining effects within large firms.

D.1.3 Other Cluster Choices and Classical AKM

Table D.2 presents the robustness analysis of the model, allowing different cluster choices in the clustered AKM (C-AKM), which is the empirical specification of my study, and, for comparison, the classical AKM model. The results span different levels of firm clustering (K = 4, 6, 8, and 10) in my grouped fixed effects approach, as well as the traditional AKM and the baseline clustering under the largest dual connected set.

Under the clustered AKM approach, trimming the data is not required. However, for traditional AKM settings, it is required to extract the largest dual connected set for correct identification. This is reflected by the larger total gap at 31.7 log points, contrasted with the full data 23.7 log points[17].

There is a modest but consistent increase in firm components when moving from K = 4 to K = 6, increasing from 2.4 to 3.1 log points. This trend suggests that finer firm classifications capture additional nuances in firm-specific contributions to the gender wage gap.

---

[17]I reserve the appendix for a comprehensive analysis of the AKM model and the dual connected set requirement. See Section C.

D.1.4   Bargaining Sensitivity to Cluster Normalization

The bargaining channel, however, is the most sensitive to variations in cluster choice, becoming the almost sole driving force of my robustness analysis, with the sorting component practically stable across results. However, it seems that bargaining estimates stabilize between K = 8 and K = 10, indicating that beyond a certain point, further granularity in firm classification yields diminishing returns in terms of explanatory power. Interestingly, $K = 10$ represents the optimal cluster choice in the gap statistics evaluation[18] (Tibshirani, Walther, and Hastie, 2001).

The bargaining channel sensitivity could be attributed to the normalization procedure. Slicing firms in the data based on wage distribution similarities may keep the sorting of male and female workers in the labor market almost intact, given that women are more concentrated in low-paying firms. However, it may potentially underestimate rents coming from firms grouped at the lowest class that are, in fact, different enough to be categorized separately in more granulated settings. This is confirmed by the striking different estimate coming from running a C-AKM on the LDCS sample. The largest dual connect set is overrepresented by larger firms, therefore, it is possible that only larger firms with strong positive rent sharing for men were kept in the sample, severely underestimating the male worker's firm component returns.

Therefore, if the researcher desires to commit to the linearity assumption of AKM leveraging from the benefits of clustering firms, the best practice is to employ several number of cluster choices to determine the most appropriate configuration. Ideally, it should minimize within-cluster variance, ensuring that firms within each group are sufficiently homogeneous in their payment behavior, minimizing the normalization cost. Moreover, it should maintain enough heterogeneity between clusters to capture meaningful differences between clusters and provide better economic intuition.

---

[18]See Section A for a more rigorous discussion on cluster choice.

Table D.2: Firm Decomposition: Different Model Specifications

| | | Contribution to Gender Wage Gap | | |
| Group | Total Gap | Firm Components | Sorting Components | Bargaining Components |
	(1)	(2)	(3)	(4)
K = 4	-0.237	-0.024	-0.019	-0.005
		(0.10)	(0.08)	(0.02)
K = 6	-0.237	-0.031	-0.021	-0.010
		(0.13)	(0.09)	(0.04)
K = 8	-0.237	-0.034	-0.022	-0.013
		(0.15)	(0.09)	(0.06)
K = 10 (Baseline)	-0.237	-0.033	-0.022	-0.010
		(0.14)	(0.09)	(0.04)
AKM	-0.317	-0.046	-0.026	-0.020
		(0.14)	(0.08)	(0.06)
K = 10 on LDCS	-0.317	-0.025	-0.027	0.002
		(0.07)	(0.08)	(-0.01)

Notes: [1]K represents the number of firm clusters in the grouped fixed effects models. [2]AKM refers to the traditional Abowd, Kramarz, and Margolis (1999) specification under the largest dual connected set sample. [3]Numbers in parenthesis represent the fraction of the overall gender wage gap that is attributed to the source described in column heading. [4]Column (1) shows the total female-male wage gap in means. Column (2) presents the total contribution of firm-specific factors. Columns (3) and (4) further decompose the firm premium contribution into a sorting (explained) and a bargaining (unexplained) components, respectively.

Appendix E   Additional Tables and Figures
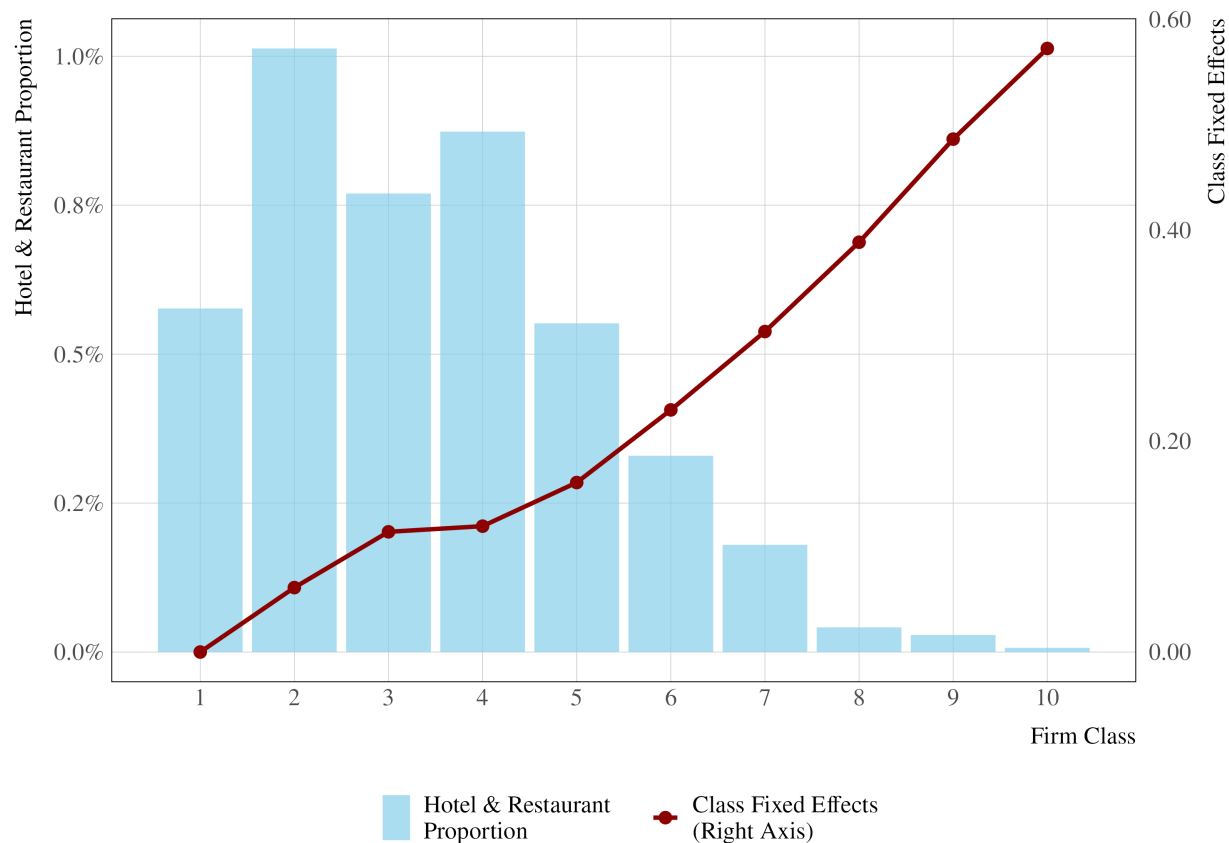
Figure E.1: Estimated Effects and Hospitality Industry Proportions Per Firm Class

Note: [1]Firm classes estimated through a kmeans clustering algorithm based on the distribution of logarithmic wages. [2]Hotels and restaurants extracted from code 55 and 56 of the Brazilian *CNAE* code of economic activities.

Table E.1: Descriptive Statistics by Gender: Largest Dual Connected Set

Features	Female Workers	Male Workers
Firm Characteristics		
Number of Firms	24 500	24 500
Firms with $\geq$ 10 Workers	21 873	21 873
Firms with $\geq$ 50 Workers	12 835	12 835
Mean Firm Size	365	365
Median Firm Size	55	55
Worker Characteristics		
Education (%)		
Dropout	19	22
High School Graduates	46	41
Some College	35	37
Age (%)		
< 30	41	38
31–50	51	50
≥ 51	7	10
Sector of Employment (%)		
Primary	1	2
Manufacturing	19	28
Construction	1	2
Trade	15	15
Services	65	54
Occupation (%)		
Scientific and Liberal Arts	15	16
Technicians	14	14
Administrative	32	19
Managers	5	8
Traders	22	19
Rural	1	2
Factory	18	23
Labor Market Outcomes		
Mean Experience (years)	4.56	5.24
Mean log hourly Wage	2.199	2.516
Variance of log hourly Wage	0.639	0.802
Worker-Year Observations	4 464 653	4 469 690
Number of Workers	1 831 797	1 812 494
Gender Fraction (%)	50	50

Note: [1] Descriptive statistics calculated from the first year of each biennial sample's largest dual connected set (2010-2015). [2] Percentages may not sum to 100% due to rounding. [3] The number of firms is the same for each gender since every firm in the cleaned sample employs both male and female workers.

Table E.2: Descriptive Statistics of Lower Firm Classes

	class 1		class 2		class 3		class 4		class 5	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Firms	74 865	74 865	101 454	101 454	83 715	83 715	79 165	79 165	80 060	80 060
Firms with ≥10 Workers	21 976	21 976	35 851	35 851	32 126	32 126	28 479	28 479	37 413	37 413
Firms with ≥50 Workers	2382	2382	5259	5259	6176	6176	5284	5284	8437	8437
Mean Firm Size	18	18	18	18	22	22	23	23	31	31
Median Firm Size	5	5	6	6	7	7	6	6	9	9
Dropout	0.54	0.52	0.36	0.44	0.28	0.40	0.31	0.40	0.27	0.39
High School Graduates	0.41	0.43	0.57	0.51	0.64	0.54	0.55	0.49	0.60	0.52
Some College	0.05	0.05	0.07	0.05	0.08	0.06	0.14	0.11	0.13	0.09
Age (<30)	0.33	0.40	0.43	0.42	0.45	0.44	0.41	0.39	0.43	0.40
Age 31-50	0.52	0.42	0.46	0.42	0.47	0.42	0.48	0.44	0.48	0.46
Age (≥51)	0.13	0.17	0.09	0.14	0.07	0.12	0.09	0.15	0.08	0.13
Primary Sector	0.05	0.09	0.03	0.05	0.02	0.04	0.03	0.05	0.02	0.03
Manufacturing	0.11	0.11	0.16	0.15	0.17	0.18	0.17	0.20	0.27	0.30
Construction	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Trade	0.15	0.22	0.32	0.35	0.51	0.50	0.25	0.29	0.36	0.36
Services	0.68	0.57	0.49	0.44	0.29	0.27	0.54	0.44	0.35	0.31
Scientific and Liberal Arts	0.02	0.02	0.02	0.02	0.02	0.02	0.05	0.04	0.04	0.03
Technicians	0.03	0.05	0.04	0.05	0.04	0.05	0.06	0.08	0.08	0.07
Administrative	0.20	0.14	0.34	0.15	0.39	0.17	0.36	0.19	0.36	0.16
Managers	0.02	0.03	0.03	0.05	0.04	0.06	0.04	0.05	0.03	0.05
Traders	0.56	0.44	0.39	0.40	0.33	0.36	0.32	0.29	0.27	0.29
Rural	0.04	0.10	0.03	0.05	0.02	0.03	0.02	0.05	0.01	0.02
Factory	0.13	0.22	0.15	0.28	0.16	0.32	0.15	0.30	0.21	0.38
Mean experience (years)	2.789	3.031	2.818	3.136	3.238	3.521	3.405	3.959	3.635	4.033
Mean Log-Wage	1.358	1.466	1.514	1.624	1.660	1.775	1.667	1.860	1.806	1.949
Variance of Log-Wage	0.061	0.101	0.058	0.087	0.074	0.117	0.156	0.206	0.113	0.155
Worker-years observations	775 686	540 010	973 659	888 531	953 073	921 694	942 303	881 295	1 143 801	1 317 477
Number of Workers	437 700	311 953	604 485	544 422	607 113	593 804	626 618	583 517	667 466	757 214
Fraction of Women	0.59	0.41	0.52	0.48	0.51	0.49	0.52	0.48	0.46	0.54

Table E.3: Descriptive Statistics of Upper Firm Classes

	class 6		class 7		class 8		class 9		class 10	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Firms	70 196	70 196	46 157	46 157	25 680	25 680	17 077	17 077	7762	7762
Firms with ≥10 Workers	35 405	35 405	24 479	24 479	14 465	14 465	11 627	11 627	5347	5347
Firms with ≥50 Workers	8915	8915	7124	7124	4946	4946	4855	4855	2448	2448
Mean Firm Size	40	40	55	55	81	81	110	110	148	148
Median Firm Size	10	10	10	10	12	12	20	20	20	20
Dropout	0.19	0.31	0.12	0.23	0.07	0.12	0.02	0.05	0.01	0.02
High School Graduates	0.57	0.53	0.52	0.51	0.40	0.43	0.15	0.23	0.08	0.11
Some College	0.24	0.16	0.36	0.26	0.54	0.44	0.83	0.73	0.91	0.87
Age (<30)	0.41	0.37	0.39	0.36	0.39	0.35	0.40	0.33	0.31	0.25
Age 31-50	0.50	0.49	0.52	0.51	0.53	0.53	0.52	0.54	0.58	0.60
Age (≥51)	0.08	0.12	0.08	0.11	0.07	0.10	0.06	0.11	0.09	0.13
Primary Sector	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Manufacturing	0.29	0.36	0.21	0.34	0.14	0.29	0.12	0.19	0.22	0.27
Construction	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.02
Trade	0.23	0.25	0.16	0.21	0.12	0.15	0.08	0.08	0.13	0.13
Services	0.47	0.36	0.61	0.43	0.72	0.53	0.78	0.70	0.63	0.58
Scientific and Liberal Arts	0.09	0.06	0.14	0.10	0.23	0.17	0.32	0.31	0.36	0.36
Technicians	0.13	0.12	0.19	0.15	0.23	0.20	0.13	0.17	0.14	0.17
Administrative	0.34	0.18	0.34	0.19	0.33	0.21	0.36	0.23	0.29	0.19
Managers	0.04	0.05	0.04	0.05	0.06	0.08	0.11	0.14	0.16	0.21
Traders	0.21	0.20	0.16	0.16	0.10	0.11	0.06	0.06	0.04	0.04
Rural	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Factory	0.19	0.39	0.12	0.34	0.06	0.23	0.02	0.09	0.01	0.03
Mean experience (years)	4.264	4.687	4.664	5.089	5.049	5.567	5.367	6.218	5.889	6.418
Mean Log-Wage	1.990	2.162	2.258	2.399	2.575	2.729	2.986	3.201	3.497	3.749
Variance of Log-Wage	0.210	0.253	0.302	0.340	0.377	0.463	0.374	0.470	0.416	0.479
Worker-years observations	1 263 982	1 561 044	1 168 436	1 361 439	990 776	1 080 347	819 258	1 051 482	472 259	680 152
Number of Workers	675 731	823 446	587 173	686 629	449 740	499 661	334 985	427 467	179 633	251 992
Fraction of Women	0.45	0.55	0.46	0.54	0.48	0.52	0.44	0.56	0.41	0.59

Table E.4: Descriptive Statistics of Lower Worker Types

	type 1		type 2		type 3		type 4		type 5	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Firms	142 765	115 262	195 800	167 929	244 655	237 647	160 499	173 160	170 689	213 850
Firms with ≥10 Workers	18 551	18 551	37 529	37 529	73 454	73 454	33 732	33 732	54 100	54 100
Firms with ≥50 Workers	2262	2262	5428	5428	12 445	12 445	5354	5354	9744	9744
Mean Firm Size	7	7	9	9	14	14	9	9	13	13
Median Firm Size	2	2	3	3	4	4	3	3	3	3
Dropout	0.44	0.46	0.35	0.41	0.30	0.38	0.28	0.35	0.16	0.33
High School Graduates	0.51	0.49	0.57	0.52	0.61	0.54	0.57	0.51	0.58	0.53
Some College	0.05	0.05	0.08	0.07	0.09	0.07	0.15	0.14	0.26	0.13
Age (<30)	0.40	0.52	0.46	0.55	0.45	0.49	0.41	0.41	0.44	0.39
Age 31-50	0.47	0.33	0.44	0.33	0.46	0.38	0.48	0.44	0.48	0.48
Age (≥51)	0.11	0.13	0.08	0.11	0.08	0.12	0.09	0.13	0.07	0.12
Primary Sector	0.04	0.06	0.02	0.04	0.02	0.03	0.02	0.04	0.01	0.03
Manufacturing	0.17	0.17	0.20	0.20	0.23	0.24	0.20	0.24	0.21	0.27
Construction	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
Trade	0.21	0.29	0.29	0.34	0.33	0.34	0.26	0.29	0.26	0.28
Services	0.57	0.47	0.47	0.41	0.42	0.38	0.51	0.42	0.52	0.41
Scientific and Liberal Arts	0.01	0.01	0.02	0.02	0.02	0.02	0.04	0.04	0.07	0.04
Technicians	0.04	0.05	0.04	0.06	0.05	0.06	0.08	0.07	0.14	0.11
Administrative	0.27	0.21	0.34	0.22	0.36	0.19	0.37	0.18	0.42	0.19
Managers	0.01	0.01	0.01	0.02	0.02	0.03	0.03	0.05	0.04	0.03
Traders	0.47	0.38	0.38	0.36	0.33	0.34	0.31	0.28	0.18	0.23
Rural	0.04	0.07	0.02	0.04	0.02	0.04	0.02	0.04	0.00	0.02
Factory	0.16	0.26	0.18	0.29	0.20	0.32	0.16	0.34	0.14	0.37
Mean experience (years)	2.552	2.510	2.645	2.530	3.034	2.993	3.517	3.765	4.190	4.230
Mean Log-Wage	1.395	1.433	1.519	1.550	1.624	1.692	1.770	1.955	2.016	2.028
Variance of Log-Wage	0.032	0.037	0.052	0.062	0.061	0.174	0.248	0.463	0.057	0.054
Worker-years observations	742 040	429 828	1 233 122	834 399	2 216 912	1 829 454	951 776	999 212	1 378 256	1 865 438
Number of Workers	634 290	383 994	1 006 451	710 878	1 517 927	1 330 372	808 545	854 727	990 969	1 316 173
Fraction of Women	0.63	0.37	0.60	0.40	0.55	0.45	0.49	0.51	0.42	0.58

Table E.5: Descriptive Statistics of Upper Worker Types

	type 6		type 7		type 8		type 9		type 10	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Firms	132 151	167 588	87 574	115 373	51 576	62 286	29 157	39 610	14 021	21 175
Firms with ≥10 Workers	40 967	40 967	32 009	32 009	15 816	15 816	11 785	11 785	6470	6470
Firms with ≥50 Workers	7489	7489	6674	6674	3481	3481	2998	2998	1676	1676
Mean Firm Size	13	13	15	15	15	15	21	21	22	22
Median Firm Size	3	3	3	3	2	2	3	3	3	3
Dropout	0.12	0.26	0.05	0.18	0.03	0.08	0.01	0.04	0.00	0.01
High School Graduates	0.43	0.48	0.28	0.40	0.14	0.25	0.07	0.15	0.04	0.07
Some College	0.45	0.26	0.67	0.42	0.83	0.67	0.92	0.82	0.96	0.92
Age (<30)	0.38	0.32	0.35	0.29	0.28	0.24	0.19	0.17	0.08	0.08
Age 31-50	0.53	0.54	0.56	0.57	0.62	0.61	0.68	0.66	0.73	0.69
Age (≥51)	0.07	0.13	0.07	0.12	0.09	0.13	0.11	0.15	0.16	0.21
Primary Sector	0.01	0.02	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01
Manufacturing	0.16	0.28	0.15	0.29	0.16	0.28	0.18	0.29	0.18	0.26
Construction	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.02
Trade	0.19	0.23	0.14	0.18	0.12	0.14	0.11	0.13	0.12	0.13
Services	0.64	0.45	0.69	0.50	0.71	0.55	0.69	0.56	0.67	0.58
Scientific and Liberal Arts	0.17	0.09	0.31	0.18	0.42	0.31	0.44	0.38	0.38	0.36
Technicians	0.22	0.15	0.22	0.20	0.15	0.19	0.11	0.17	0.07	0.09
Administrative	0.35	0.19	0.29	0.19	0.22	0.15	0.19	0.13	0.15	0.10
Managers	0.06	0.06	0.08	0.08	0.12	0.13	0.20	0.20	0.38	0.41
Traders	0.13	0.16	0.07	0.11	0.08	0.08	0.05	0.05	0.02	0.03
Rural	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Factory	0.07	0.33	0.03	0.24	0.01	0.12	0.01	0.06	0.00	0.02
Mean experience (years)	4.943	5.413	5.684	6.145	6.201	6.661	7.124	7.294	8.331	8.247
Mean Log-Wage	2.347	2.400	2.789	2.765	3.196	3.272	3.624	3.662	4.157	4.226
Variance of Log-Wage	0.200	0.202	0.074	0.080	0.232	0.213	0.080	0.085	0.103	0.131
Worker-years observations	1 058 198	1 467 569	901 644	1 216 454	486 706	666 147	373 527	612 301	161 052	362 669
Number of Workers	724 998	996 335	529 627	728 982	315 684	437 215	216 089	351 364	91 719	200 886
Fraction of Women	0.42	0.58	0.43	0.57	0.42	0.58	0.38	0.62	0.31	0.69

Table E.6: Wage Levels for Males and Females under Different Scenarios

Group	Baseline		Separable Market		Constant Returns		Constant Firm Allocation	
	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All	2.10	2.33	2.11	2.30	2.16	2.25	2.19	2.22
Education								
No HS	1.62	1.92	1.65	1.92	1.72	1.84	1.77	1.80
HS	1.90	2.14	1.92	2.13	1.97	2.07	2.01	2.04
College	2.92	3.27	2.87	3.17	2.94	3.10	3.00	3.03
Age								
<30	1.97	2.06	1.98	2.06	2.01	2.03	2.01	2.03
31-50	2.18	2.49	2.18	2.44	2.25	2.38	2.30	2.33
50>	2.09	2.42	2.08	2.36	2.15	2.29	2.21	2.24
Firm Size								
Firms <10	1.66	1.78	1.65	1.77	1.70	1.72	1.70	1.73
Firms 10-50	1.77	1.90	1.77	1.90	1.82	1.85	1.83	1.84
Firms 51>	1.78	1.99	1.79	1.96	1.83	1.91	1.86	1.88
Occupations								
Hotel and Restaurants	1.62	1.74	1.62	1.74	1.71	1.67	1.67	1.70
Engineers & Economists	2.91	3.30	2.88	3.23	2.98	3.13	3.03	3.08
Managers	3.03	3.35	2.90	3.12	2.96	3.05	2.99	3.02

Notes: [1]All values represent base wages in log scale. [2]Baseline is observed wages. [3]Separable Market assumes no complementarity in worker-firm interactions. [4]Constant Returns equalizes means and variances of realized worker-firm interactions. [5]Constant Firm Allocation equalizes firm-specific probabilities.