# Online Appendix – Not for Publication

## The Great Equalizer: How Competition in the Labor Market Reduces Wage Inequality[*]

Martim Leitão
University of Maryland, College Park

Jaime Montana
Universidade Católica Portuguesa & Université Paris-Dauphine

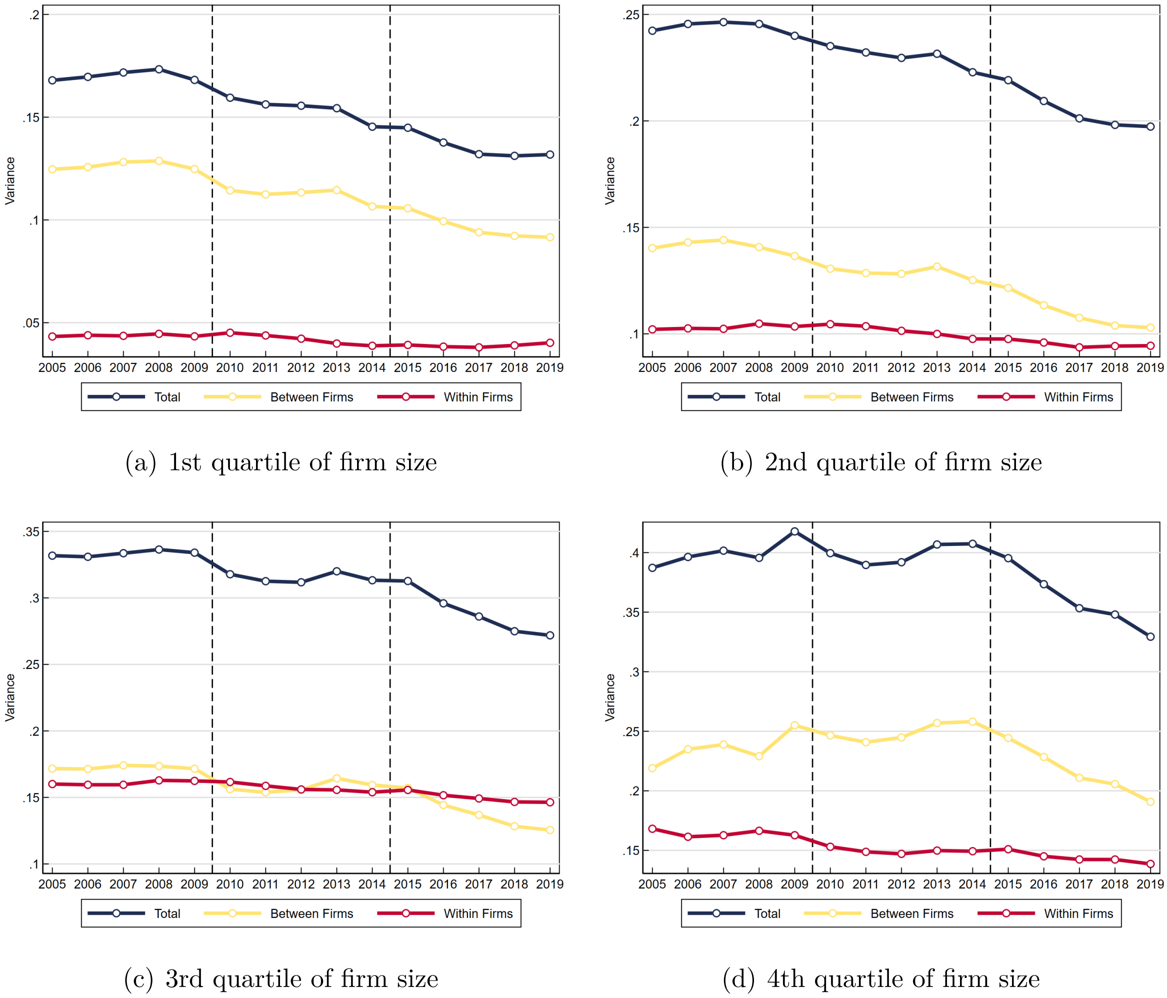
Joana Silva
Universidade Católica Portuguesa, World Bank, & CEPR

[*] mleitao@umd.edu, jdoncel@ucp.pt, joana.silva@ucp.pt. We thank the Portuguese Institute of Statistics (INE) for data access. We remain responsible for any errors. We thank FCT for financial support. We declare that we have no relevant or material financial interest that relate to the research described in this paper.

# Contents

# A Tables and Figures for Online Publication

Figure 1: Within- and Between-Firm Inequality, by Firm Size (2005-19).



(a) 1st quartile of firm size

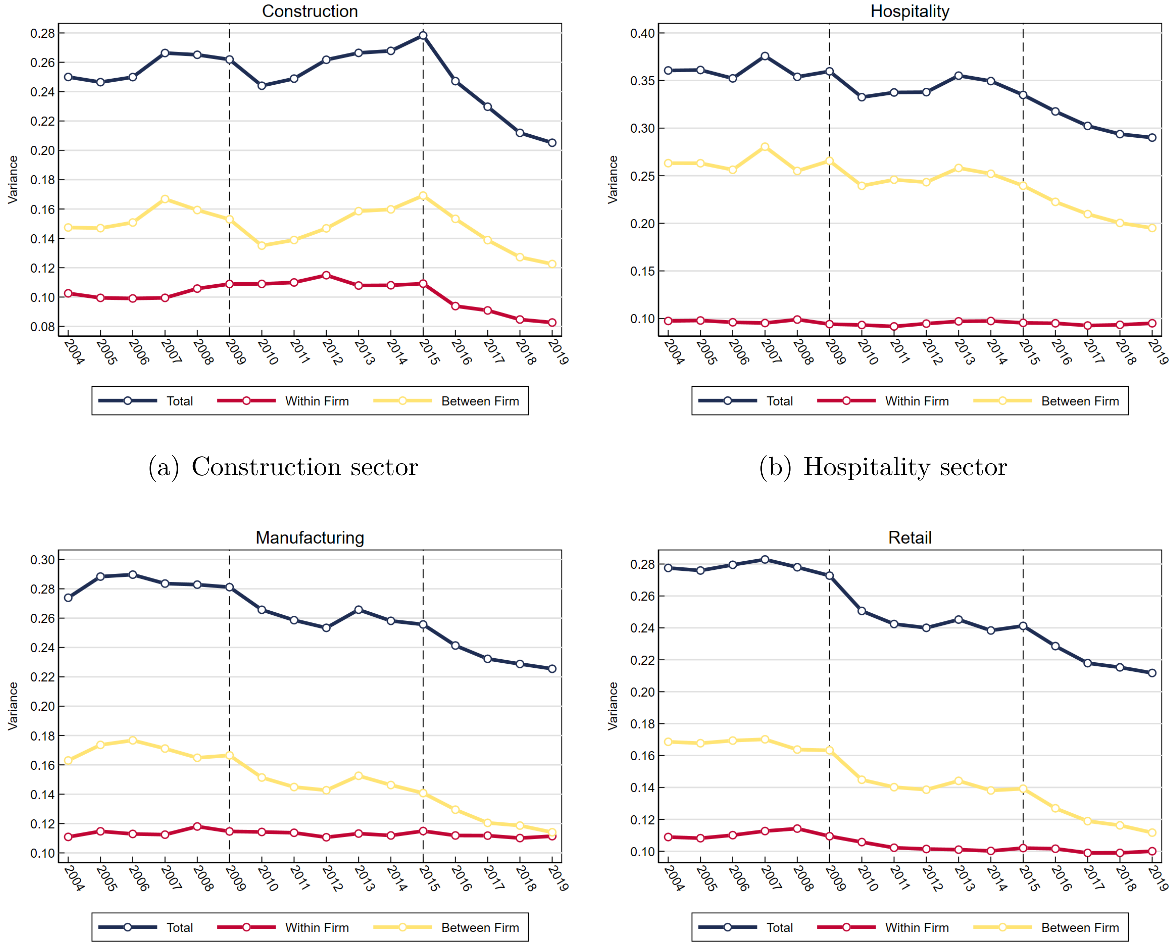
(b) 2nd quartile of firm size

(c) 3rd quartile of firm size

(d) 4th quartile of firm size

**Source:** *Quadros de Pessoal*, $2005 - 19$.
**Note:** Panels (a) to (d) plot the yearly evolution of the variance of hourly wages ("total wage inequality") over 2005-19, decomposed into a within-firm inequality and a between-firm inequality components by quartiles of firm size. Quartiles of firm size are constructed based on the average number of workers in the firm during the entire 2005-19 period. The vertical sum of the within- and between-firm inequality components adds up to overall inequality, for each year. Firm variance is computed based on average log earnings and is weighted by the number of workers in the firm. Within-firm variance is based on the difference between a worker's log hourly earnings and the average wage paid by his or her firm. Additional details on how to implement this estimation are provided in Appendix B.

Figure 2: Within- and Between-Firm Inequality, by Sector (2005-19).



(a) Construction sector

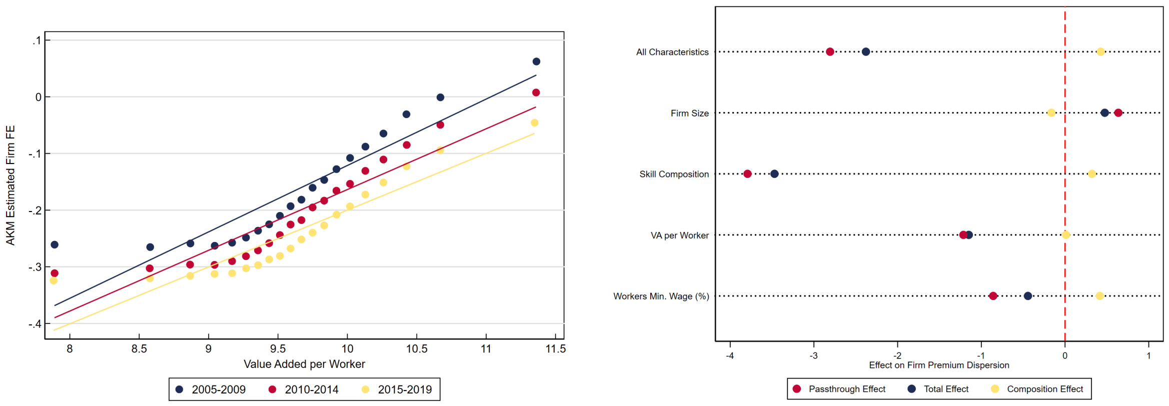
(b) Hospitality sector

(c) Manufacturing sector

(d) Retail sector

**Source:** *Quadros de Pessoal*, 2005 – 19.
**Note:** Panels (a) to (d) plot the yearly evolution of the variance of hourly wages ("total wage inequality") over 2005-19, decomposed into within-firm inequality and between-firm inequality components for selected sectors: construction, hospitality, manufacturing and retail. The vertical sum of the within- and between-firm inequality components adds up to overall inequality, for each year. Firm variance is computed based on average log earnings and is weighted by the number of workers in the firm. Within-firm variance is based on the difference between a worker's log hourly earnings and the average wage paid by his or her firm. Additional details on how to implement this estimation are provided in Appendix B.

Figure 3: Declining Returns to Firm Characteristics and Composition

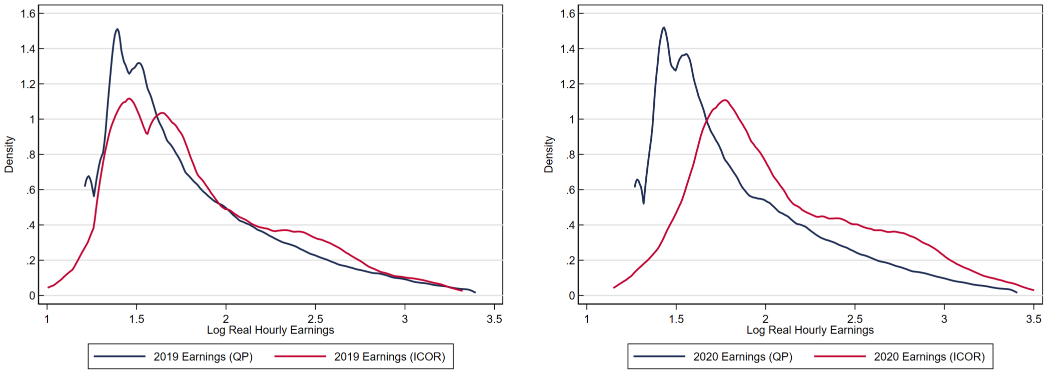


(a) Firm effects *vs.* value added per worker

(b) Passthrough *vs.* composition

**Source:** *Quadros de Pessoal*, $2005-19$.
**Note:** Panel (a) shows the average estimated firm effect in each subperiod against value added per worker (by 20 bins of log mean value added per worker in the subperiod). Value added has been constructed by averaging value added per worker at the firm level over each subperiod, and then taking logs. Overlaid are ordinary least squares best fit lines, whose slope capture returns to value added. Panel (b) presents the key messages from the Oaxaca-Blinder decomposition in a graphical manner. Blue dots represent the contribution of each characteristic to the decline in firm pay premium dispersion. Red dots show the portion of this contribution due to passthrough effects and yellow dots show the portion due to composition effects. The horizontal sum of the yellow and red dots must add up to the blue dots by construction.

Figure 4: Earnings Distributions Comparison between QP and ICOR.

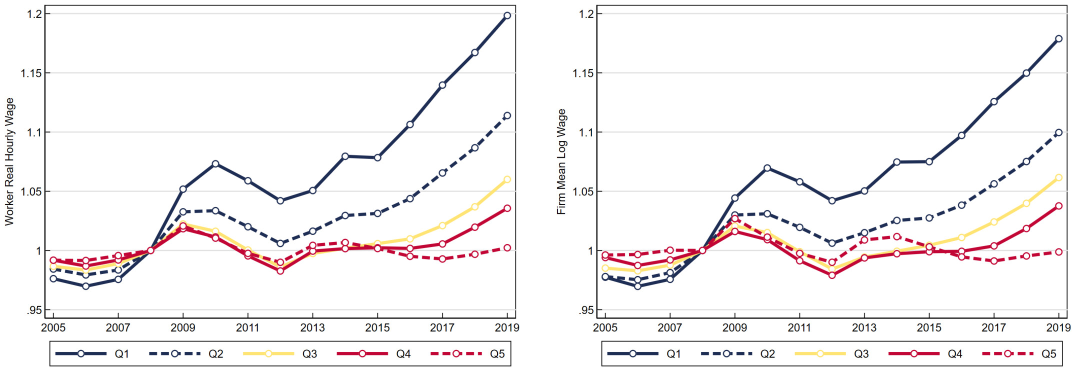


(a) 2019

(b) 2020

**Sources:** *Quadros de Pessoal*, $2019, 2020$; *EU-SILC*, $2019, 2020$.
**Note:** This figure provides evidence supporting the quality of the data. These figures present Kernel density comparisons for the wage distributions of 2019 and 2020, in *Quadros de Pessoal* and *Inquérito às Condições de Vida e Rendimento*. These Figures were built using the log of real hourly wages (in gross terms) of full-time dependent workers between ages 18 and 65. We use the consumer price index to convert both series to real terms. Observations from ICOR are weighted by means of cross-sectional sample weights provided by Statistics Portugal. In both years, and in both data sets, we have trimmed the $1^{st}$ and $99^{th}$ percentiles of real hourly wages.

Figure 5: Change in Percentiles of Annual Earnings Overall and Between Firms



(a) Overall earnings

(b) Between firms

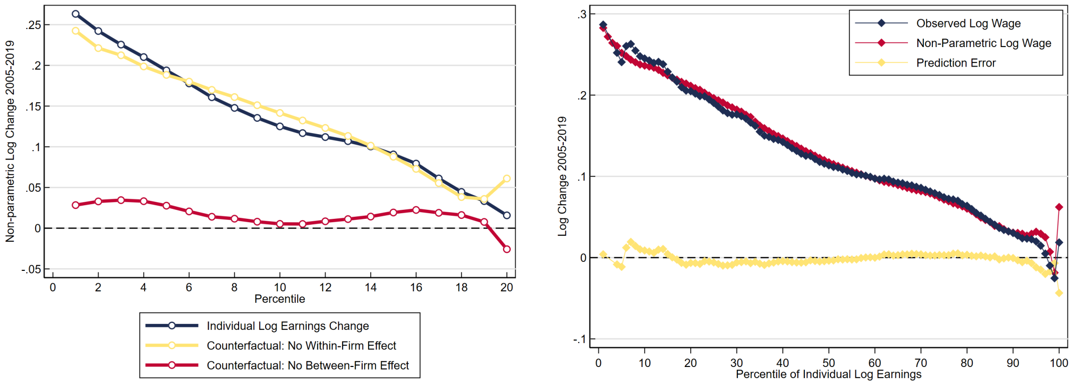
**Source:** *Quadros de Pessoal*, 2005 – 19.
**Note:** Panel (a) plots the dynamics of log hourly earnings for workers in five quintiles. To construct this figure, we average log hourly earnings by wage bin and year, and plot this metric over time. We have normalized this average to 1 in 2008. The widening of the curves – with lower quintiles growing faster – suggests wage inequality is decreasing as years go by. Panel B repeats this procedure but using a worker's firm average log hourly wages. Panel (b) is built by first finding each firm's mean log wage in each year. Then, we proceed to average this value within each year and earnings bin (weighted by employment). We have normalized this average to 1 in 2008. The construction of the metrics behind this figure closely mirrors that of Figure 6. The widening of these curves over time suggests inequality in average firm pay is decreasing over time. Moreover, the fact that the patterns observed in Panel (a) track those observed in Panel (b) suggests that the evolution of average firm pay drove the reduction in inequality.

Table 1: Robustness Checks of the Variance Decomposition

| | 2005 | | 2019 | | Δ | |
|---|---|---|---|---|---|---|
| | Total Var | Between-firm | Total Var | Between-firm | Δ Total Var | Δ Between (%) |
| All | 0.328 | 0.214 | 0.255 | 0.149 | -0.073 | 88.677 |
| Demean: Region | 0.301 | 0.187 | 0.243 | 0.137 | -0.058 | 85.911 |
| Demean: Broad industry | 0.255 | 0.140 | 0.213 | 0.107 | -0.042 | 80.336 |
| Demean: 2-digit industry | 0.240 | 0.125 | 0.190 | 0.084 | -0.050 | 83.567 |
| Demean: Gender | 0.315 | 0.204 | 0.247 | 0.141 | -0.069 | 91.399 |
| Demean: Birth cohort | 0.312 | 0.202 | 0.247 | 0.145 | -0.064 | 88.491 |
| Demean: Nationality | 0.327 | 0.212 | 0.254 | 0.148 | -0.073 | 88.904 |
| Demean: Education | 0.252 | 0.144 | 0.205 | 0.106 | -0.047 | 80.645 |

**Source:** *Quadros de Pessoal*, 2005 – 19.
**Note:** This table provides robustness checks for the within-between firm variance decomposition. Total Var stands for total variance of log hourly real wages in a given year, while Between-firm stands for variance in average firm pay in a given year (weighted by employment). Δ Total Var denotes the absolute value change in total variance, while the last column presents the fraction of this change accounted for by changes in between-firm variance. Except for the first row, all statistics are computed using earnings demeaned within a given group, *before* all variances are calculated. This table shows that even within narrowly defined sectors or demographic groups, most of the decline in earnings inequality occurred between firms.

Figure 6: Evolution of Portuguese Wage Inequality (2005-19).

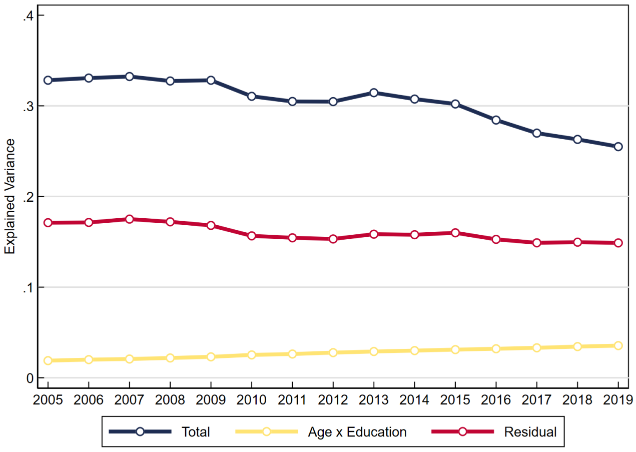


(a) Nonparametric density decomposition

(b) Error from Nonparametric exercise

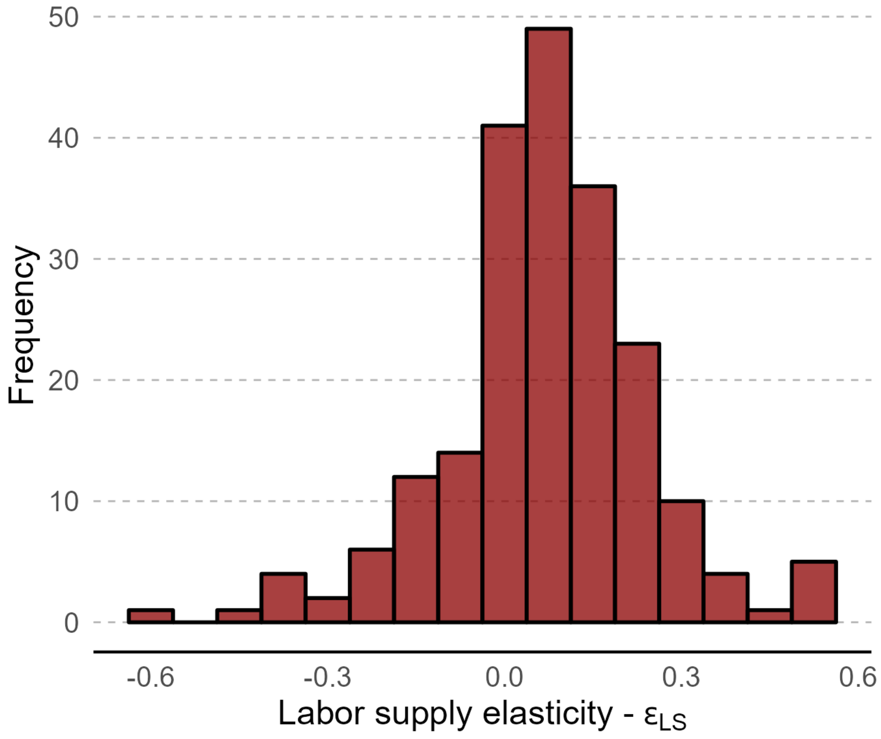
**Source:** *Quadros de Pessoal*, $2005-19$.
**Note:** Panel (a) shows the results of the non-parametric density decomposition described in Machado and Mata (2005), Autor et al. (2005), and Song et al. (2019). To produce this figure, we first compute two sets of statistics each for 2005 and 2019. First, we obtain the percentiles of the distribution of firms' mean log hourly earnings, weighted by employment. Then, within each percentile, we calculate 500 quantiles of the distribution of the difference between log worker hourly earnings and the average earnings in that firm-based percentile. These two sets of bins are subsequently used to produce the counterfactual distributions shown in Panel (a). For additional detail on this procedure, please refer to Song et al. (2019)'s Online Appendix E. Panel (b) shows the prediction error resulting from the non-parametric density decomposition.

Figure 7: Variance Decomposition from Mincer Regression.



**Sources:** *Quadros de Pessoal*, $2005-2019$.
**Note:** This figure presents a variance decomposition built from an underlying Mincer regression decomposition. This figure shows that worker characteristics play a small role in explaining the decline in wage inequality observed in Portugal over the past twenty years. To produce this figure, we start by creating 5 age bins and interact educational attainment with these bins. Then, we regress log wages on this interaction and control for sector and occupation fixed effects. The yellow line plots the variance of the estimated interaction effect, while the red line plots the variance of the residual resulting from this equation.

Figure 8: Distribution of labor supply elasticities across local labor markets

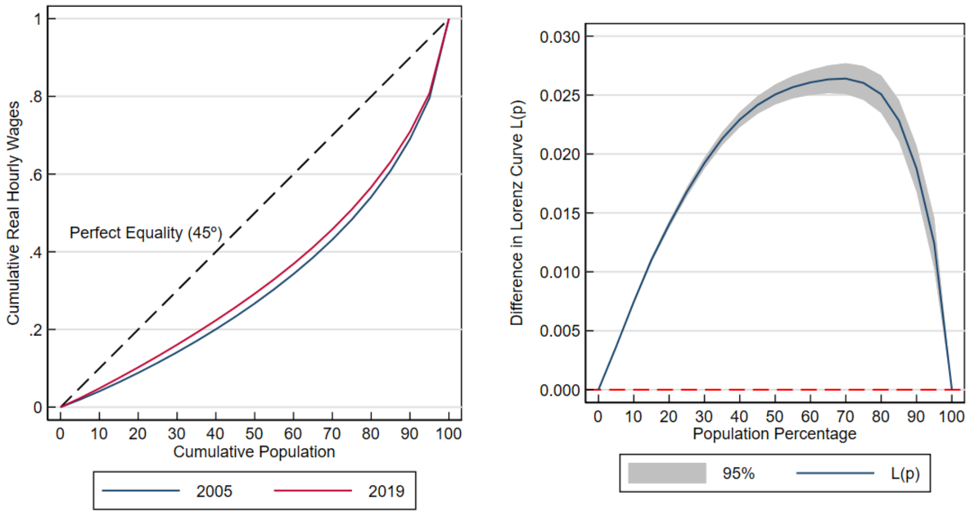


**Source:** *Quadros de Pessoal*, $2005-19$.
**Note:** The figure shows the distribution of the labor supply elasticities,$\epsilon_{LS,j} = -2 \times \beta_1^j$, estimated across local labor markets in Portugal. The coefficient captures how sensitive worker separation is to changes in the wage rate for each local labor market $j$.

# B Stylized Facts on Earnings Inequality in Portugal

In this section, we present the first set of stylized facts on Portugal's rapid decrease in earnings inequality between 2005 and 2019. Wage inequality in Portugal declined continuously over the course of the twenty-first century, by a staggering 20 percent. It would be difficult to determine a priori the direction and effect of firm and institutional characteristics in the evolution of wage inequality in Portugal. Instead, we limit ourselves to reporting some stylized facts that guide our analysis.

Figure 9: Lorenz Curves for Portugal, in 2005 and 2019.



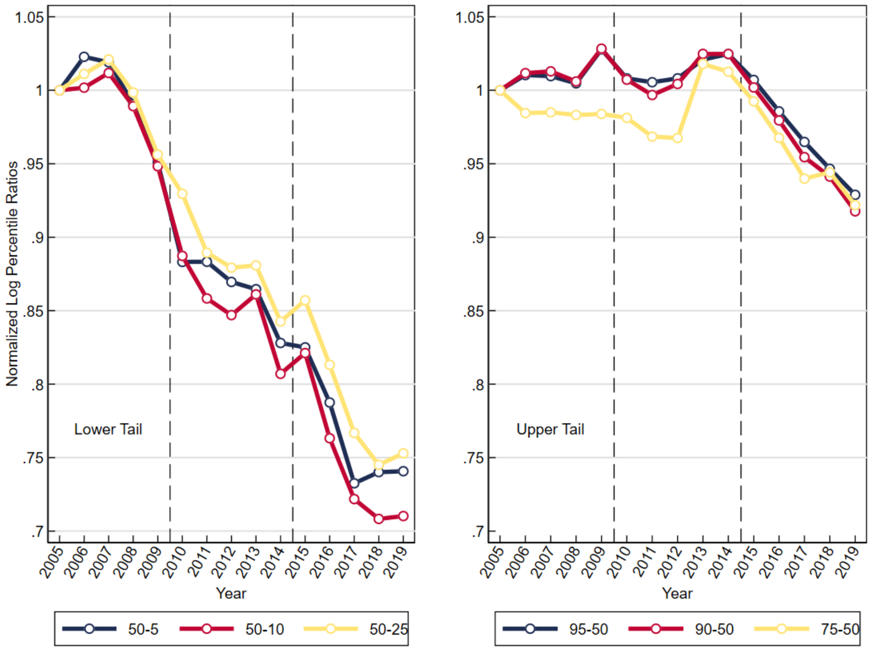
**Sources:** *Quadros de Pessoal*, $2005-2019$.
**Note:** The figure shows the Lorenz curve for labor income in 2005 and 2019 (left), and the difference between them. In the right Panel the shaded area show the confidence interval at the 95% level for the curve. The confidence interval is calculated by bootstrapping.

***i)* Heterogeneity in the Change in Inequality along the Wage Distribution** Although overall inequality has decreased in Portugal over the course of the twenty-first century, various demographic groups along the distribution may have been impacted differently. In what follows, we analyze what happened to (i) the lower tail of the distribution, (ii) the upper tail, and (iii) the distribution as a whole.

Figure 10, presents measures of inequality over 2005-19. The figure shows that the decrease in inequality was driven by the lower tail of the distribution: looking at the normalized log percentile ratios, we see that convergence toward the median of the income distribution occurred at a faster pace for the percentiles below the median, compared to those above the median (corroborating evidence is provided in Figure 5). The fact that we also observe a decrease in inequality in the upper tail of the distribution suggests that the decline in inequality happened along the full support of the income distribution. As a formal assessment of whether inequality *unambiguously* went up or down over the considered period, we evaluate the Lorenz criterion for the log of real hourly wages in Portugal. Specifically, we say that given two distributions, $X^{2005}$ and $X^{2019}$, $X^{2019}$ Lorenz dominates $X^{2005}$ if and only if

$$L_{X^{2019}}(p) \geq L_{X^{2005}}(p)\ \forall p \text{ with } > \text{ for some } p \tag{1}$$

Figure 10: Wage Inequality Dynamics in Portugal: Upper and Lower Tails (2005-19)



**Source:** *Quadros de Pessoal*, $2005-19$.
**Note:** The figure plots different measures of inequality for the lower and upper tails of the distribution for 2005-19. The inequality measures are normalized to 1 for 2005, and present the evolution of the indicators over time. This figure shows that inequality decreased across the entire earnings distribution but was more pronounced at the bottom of the earnings distribution.

If this holds, and if the Lorenz curves do not cross (since this assures the completeness of the criterion), we can state that $X^{2019}$ is *unambiguosly* less unequal than $X^{2005}$. To perform the exercise empirically, we leverage Gastwirth (1971)'s identity to estimate

$$L_{X^{2019}}(p) - L_{X^{2005}}(p) \Leftrightarrow \frac{1}{\mu^{2019}} \int_o^p Q_X^{2019}(t)dt - \frac{1}{\mu^{2005}} \int_o^p Q_X^{2005}(t)dt \qquad (2)$$

The next step is to evaluate whether this differential is positive or negative for $\forall p$. In the expression above, $Q_X(t)$ is the quantile function for the given distribution (”Pen’s Parade”, the inverse of the cumulative distribution function), so that estimating $\int_o^p Q_X(t)dt$ boils down to estimating the generalized Lorenz curve. When scaled down by the mean of the distribution, $\mu$, the Generalized Lorenz curve becomes the Lorenz curve. The application of this criterion to Portuguese data for 2005 and 2019 reveals that the decrease in inequality was unambiguous and took place along the entire wage distribution. We show the application of this criterion in Figure 9. We compare the Lorenz curves of the distributions at the beginning and end of the period considered (Atkinson, 2008; Gastwirth, 1971). This exercise supports the claim that inequality *unambiguosly* decreased in Portugal along the support of the distribution. The Lorenz curve for 2019 stochastically dominates the Lorenz curve for 2005, and there are no intersections.

***ii)* Earnings Dispersion between and within Firms** Next, we decompose wage inequality into the contributions of within- and between-firm inequality. This provides some preliminary understanding on the role of firm heterogeneity. If all firms paid the same wage to all employees, there would be no within-firm inequality, but not necessarily no wage inequality as firms could still differ in the wages that they pay. Likewise, if all firms had the same distribution of wages, there would be no inequality between-firms, but not necessarily no wage inequality as workers within each firm could earn different wages. These are the two extreme cases. With this in mind, we examine which of these factors was more prominent in Portugal between 2005 and 2019, shedding light on whether wage dispersion was mostly driven by systematic differences in pay premiums across firms or differences in pay within each firm. To do so, we decompose the variance of wages into its between and within components. Following Alvarez et al. (2018), Song et al. (2019), and Messina and Silva (2021) wages can be decomposed by construction as:

$$w_t^{i,j,f} \equiv \overline{w_t} + (\overline{w_t^f} - \overline{w_t}) + (w_t^{i,j,f} - \overline{w_t^f}) \tag{3}$$

where $w_t^{i,j,f}$ is the log of real hourly wages of worker $i$ in firm $f$ in year $t$, $\overline{w_t}$ is the average log of the real hourly wage in the economy in year $t$, and $\overline{w_t^f}$ is the average log of the real hourly wage in firm $f$ (where worker $i$ works) in year $t$. The wage of each worker can be seen as the sum of the average remuneration in the economy in that year, the difference paid on average by firms relative to the average wage in the economy, and the difference earned by workers relative to their firm's average wage. To obtain the within- and between-firms components of wage variance in each year, we rearrange and transform this identity into:

$$Var(w_t^{i,j,f} - \overline{w_t}) = Var(\overline{w_t^f} - \overline{w_t}) + Var(w_t^{i,j,f} - \overline{w_t^f}) \tag{4}$$

where $Cov(\overline{w_t^f} - \overline{w_t}; w_t^{i,j,f} - \overline{w_t^f}) = 0$ by construction.[1] Since wage variance is decomposed yearly, equation 4 becomes

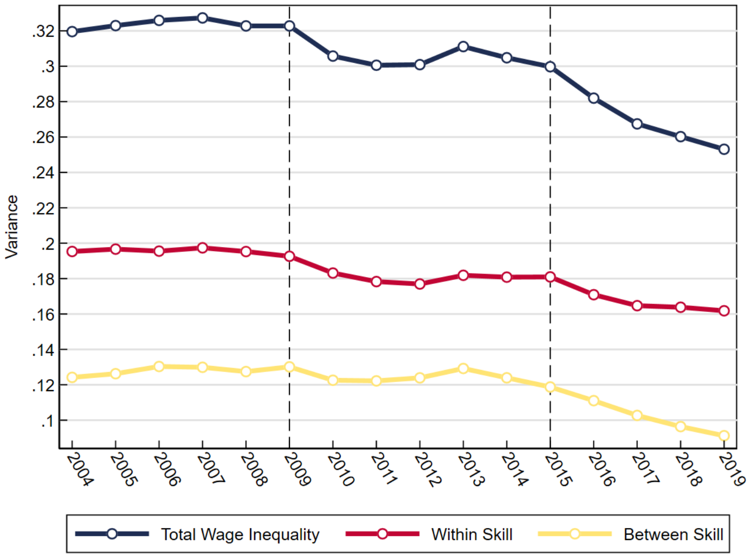
$$Var(w_t^{i,j,f}) = Var(\overline{w_t^f}) + \sum_{f=1}^{N} \omega_f Var(w_t^{i,j,f} | i \in f) \Leftrightarrow \tag{5}$$

$$\underbrace{Var(w_t^{i,j,f})}_{Overall\ Inequality} = \underbrace{Var(\overline{w_t^f})}_{Between\ Firm\ Inequality} + \underbrace{\overline{Var(w_t^{i,j,f} | i \in f)}}_{Within\ Firm\ Inequality} \tag{6}$$

This equation decomposes the yearly overall variance of log real hourly wages into the between-firm component (given by the variance across firm average wages), and a within-firm component (given by the weighted average of within-firm wage variance, with weight $\omega_f$ being the share of employment in firm $f$). Throughout the period, between-firm inequality accounted for over 60 percent of total wage inequality, and within-firm inequality accounted for slightly less than 40 percent (see Figure 1 in the paper). In the subperiods considered (2005-09, 2010-14, and 2015-19) within- and between-firm inequality moved broadly in the same direction driving the overall change in inequality. However, the stronger reduction of inequality in 2010-14 and 2015-19 was mostly driven by the reduction in inequality between-firms. To verify that the observed patterns of between- and within-firm inequality are not driven by specific sectors but are representative of the economy as a whole, we further run this equation for four selected sectors: manufacturing, construction, retail, and hospitality (see Figure 2 in this appendix). Our key insight holds regardless of the broad sector being considered. The same holds if we repeat the decomposition by firm size (see Figure 1 in this appendix).

[1] $Cov(\overline{w_t^f} - \overline{w_t}; w_t^{i,j,f} - \overline{w_t^f}) = E([\overline{w_t^f} - \overline{w_t} - E(\overline{w_t^f} - \overline{w_t})][w_t^{i,j,f} - \overline{w_t^f} - E(w_t^{i,j,f} - \overline{w_t^f})]) = 0$

Figure 11: Between- and Within-Skill Group Inequality (2005-19)



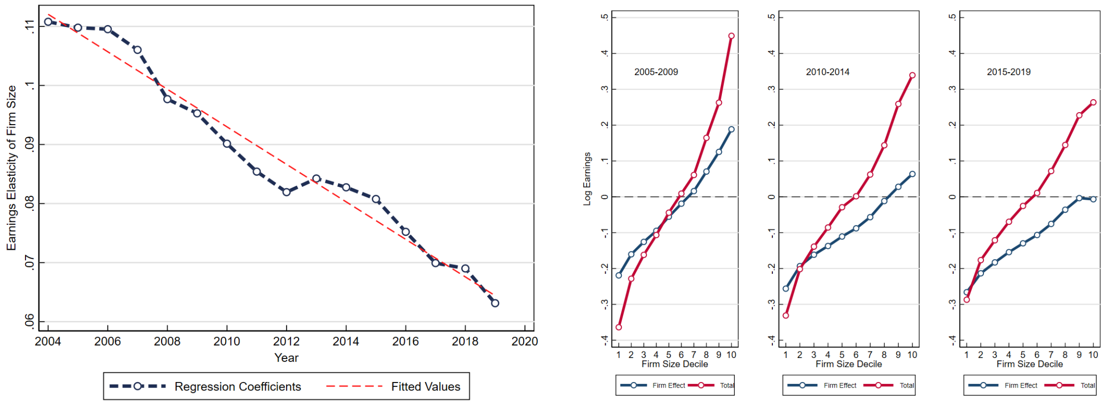
**Source:** *Quadros de Pessoal*, 2005 – 19.
**Note:** The figure shows the inequality decomposition of labor income inequality between and within skill groups. To build this figure, we started by running a Mincer-type regression of log hourly wages on education, tenure, gender, and all possible interactions between these variables. Taking the variance of each side of this estimated model yields the between- and within-skill components of inequality (the variance of the predicted component being between skill inequality, while the variance of the predicted residual can be seen as within skill inequality).

***iii)* Earnings Dispersion between and within Skills** The richness of our data also allows us to calculate inequality between different skill groups and assess how this measure has changed over time. This exercise reveals the prominence of systematic differences in returns to skills across different skill types in determining wage dispersion. To disentangle overall wage inequality, we follow Messina and Silva (2021). We start by running a standard Mincerian equation of the form $w_{it} = \rho_t \mathbf{X}_i + \mu_{it}$, where $w_{it}$ stands for the log hourly wage of worker $i$ in period $t$. $\mathbf{X}_i$ is a vector of covariates including a categorical educational level variable, tenure by five-year bins, and gender (as well as all possible interactions between these). $\rho_t$ is a vector of returns to these covariates, and $\mu_{it}$ is an orthogonal error term, referred to as within-skill group wage inequality. Once estimated, we can apply variances to this relation to obtain

$$\underbrace{Var(w_{it})}_{Overall\ Inequality} = \underbrace{Var(\hat{\rho}_t \mathbf{X}_i)}_{Between-Group\ Skill\ Inequality} + \underbrace{Var(\hat{\mu}_{it})}_{Within-Group\ Skill\ Inequality} \tag{7}$$

where we have used the orthogonality of the error term to impose zero covariance between the residual and the regressors. The variance of wages can thus be decomposed into a between-skill component and a within-skill component. Figure 11 shows the results of implementing this decomposition. In levels, within-skill inequality accounts for the largest share of overall inequality (around 60 percent). In differences, however, between-skill inequality reduction seems to play a role that is roughly as important as within-skill inequality. Over 2005-19, around 50 percent of the reduction in wage inequality is attributed to the reduction in between-skill inequality, against 50 percent explained by within-firm inequality. If we zero in on the reduction in inequality witnessed over 2015-19, the reduction in between-skill inequality accounts for almost 60 percent of the overall reduction in inequality, despite its initially lower level. These findings highlight the importance of considering job title heterogeneity for wage dispersion.

Figure 12: Wages and Firm Size in Portugal (2005-19)



(a) Declining wage-size elasticity

(b) Declining firm premium to large firm

**Source:** *Quadros de Pessoal*, $2005-19$.
**Note:** Panel (a) plots the coefficient that results from projecting labor earnings onto the size of the firm ( by year) using $\log(w_{ij}) = \alpha_o + \alpha_1 \log(N_j) + \epsilon_{ij}$. Panel (b) plots the average firm effects and average log earnings per firm size decile. Firms are assigned to 10 size classes. Following Bloom et al. (2018), we plot the average log earnings in each firm size class relative to total average log earnings over the interval and firm fixed effects components estimated using the AKM equation. We omit worker fixed effects and the residual component for the sake of readability. Each panel displays these results for a different five-year interval. The fact that the blue schedule is flattening over time (say, going from the first panel to the second) suggests that moving from a large to a very large firm is being less rewarded over time. However, at the bottom of the firm size distribution, moving from a small to a medium-sized firm still yields a substantial premium.

***iv)* Decline in the Large Firm Pay Premium** The role of large firms as providers of better working conditions has been acknowledged in the past: in general, large firms offer better monetary and non monetary compensation. It is typical that in larger firms, jobs are more stable, there is greater worker satisfaction, and workers earn higher wages. However, there is evidence for the United States that the large-firm wage premium has been shrinking (Bloom et al., 2018). To assess whether this is the case in Portugal, we perform two exercises on the role of large firm size in the wage premium. First, we calculate the yearly elasticity of firm size with respect to wages.[2] Second, relying on the estimated firm effects from the AKM equation, we plot the (de-meaned) average log earnings and average fixed effects for each firm size decile, as in Bloom et al. (2018). This allows us to assess the wage differential between different types of firms over time.

Panel (a) in Figure 12, shows a declining relationship between firm size and wages. The wage-size elasticity plummeted from around 11 percent in 2004 to under 7 percent in 2019, which corresponds to a reduction of approximately 60 percent. Thus, the pay premium that large firms offer appears to have shrunk in absolute terms. This finding is backed by the results presented in panel (b) in Figure 12, where we explore the relationship between the firm pay premium and firm size along different sub-periods. The fact that

[2] For each year between 2005 and 2019, we run the following specification: $\log(w_{ij}) = \alpha_o + \alpha_1 \log(N_j) + \epsilon_{ij}$.

the blue schedule flattens over time indicates that the returns to working in a large firm have declined over time. As large firms have historically paid significantly higher wages, it is important to understand the implications of a fall in the large firm wage premium for changes in inequality.

# C Data Procedures and Variable Construction

## C.1 Quadros de Pessoal

Quadros de Pessoal is an administrative linked employer-employee job title dataset, for Portugal. The entity responsible for this statistical operation is the *Gabinete de Estratégia e Planeamento* (GEP) from the Ministry of Employment, Solidarity and Social Security (MTSSS), making the data available for Statistic Portugal (*Instituto Nacional de Estatística*). The panel is obtained through an annual administrative census, where employers with at least one dependent worker are required to deliver (electronically or manually) to the responsible entity the information on their employees and their earnings (for example gender of worker, highest education level completed, job titles, collective bargaining agreement, date of birth, occupation, date of hiring, and so forth), as well as information on the firms (for example, sector of activity, and so forth) and establishments. This requirement is meant as a way to verify if firms are complying with labor law. Since the employer is the one actually reporting the data, variables such as worker qualifications are less prone to measurement errors.

In terms of treating the data, each year, we first merge firm and worker data. Worker's observations having a worker ID with less than 6 digits or more than 10 digits are invalid and were therefore discarded. Whenever a worker appears twice within the same year in the Panel with several jobs, his or her highest paying one was selected (since mostly likely, this is his or her primary job). Moreover, we keep, each year, observations for workers having: a job situation corresponding to dependent worker and at least 120 normal monthly hours of paid work (full-time workers). Each year, we also eliminate observations for workers without a complete basic remuneration, belonging to residual categories on job titles and that belong to a collective bargaining agreement corresponding to white zone, employers or relatives, active members of cooperatives and apprentices without link to the employer. We also eliminate observations for workers working at firms in the agriculture, animal production, hunting, forestry or fishing sector (eliminate observations of workers in sector A according to *Classificação Portuguesa das Actividades Económicas Rev.3* (CAE Rev.3) or sector A and B according to CAE Rev.2.1) due to low coverage. Gross monthly earnings from dependent work are obtained by summing the earned remuneration of the worker and some irregular instalments too. We use the consumer price index (CPI) deflator to convert nominal wages in real wages. After treating the datasets each year, the data is then appended and a panel is formed where each worker ID is tracked over time.

**Regional Indicator Variable** To identify the firms' (and workers') broad geographical regions, our setup relies on the Nomenclature of Territorial Units for Statistics at the regional level (NUTS 2). According to this classification, firms can be located in Lisbon, in the North, in Alentejo, in the Center region, in Algarve, in Madeira, or in the Azores. A broader nomenclature exists, NUTS 1, but its level of detail is coarser. Tables for NUTSII and NUTSI can be found on INE's website.

**Education** To determine the level of educational attainment of individuals, the paper focuses on a one-digit classification of highest educational attainment. The education labels were adjusted slightly for 2004 and 2005 to ensure a full harmonization of categories across time.

**Sector Indicator Variable** To determine the firms' sector of activity throughout the years, a crosswalk was used to adjust the classification in place before and after 2007. This was necessary since prior to 2007 activities were classified according to the *Classificação das Atividades Económicas Rev 2.1 (CAE Rev 2.1)*, but from 2007 onward, Portuguese activities have been revised to track international classifications and the new classification in place since then is the *Classificação das Atividades Económicas Rev 3 (CAE Rev 3).* This harmonization crosswalk was built from the underlying two-digit CAE sectors and yielded 31 large categories, later reduced to 29 categories, once agriculture and fishing were discarded. For sake of reference, the coarser level of the classification used for economic sectors since 2007 is given by the sections on CAE Rev.3: A) agriculture, animal production, hunting, forestry and fishing (sector eliminated in our paper), B) extractive industries, C) manufacturing industries, D) electricity, gas, steam, hot and cold water and cold air, E) water collection, treatment and distribution; sanitation , waste management and depollution, F) construction, G) wholesale and retail trade; repair of motor vehicles and motorcycles, H) transport and storage, I) accommodation, catering and similar, J) information and communication activities, K) financial and insurance activities, L) real estate activities, M) consulting, scientific, technical and similar activities, N) administrative and support service activities, O) public administration and defence; compulsory social security, P) Education, Q) human health and social support activities, R) artistic, entertainment, sports and recreational activities, S) other service activities and U) activities of international organizations and other extra-territorial institutions (section T does not appear in our data because *Quadros de Pessoal* excludes employers of domestic service workers and people producing for own consumption).

**Skill Composition Index** To build our skill composition variable, we follow closely Lise and Postel-Vinay (2020). We start by creating a clean crosswalk between ISCO 2008 classification and SOC (Standard Occupational Classification). We then clean O*NET data so as to have a crosswalk between each one of the 35 skill dimensions and SOC codes. Next, we reduce the dimension of this matrix and make it a single vector. That is, we compute the first principal component using Principal Component Analysis (PCA). Call the principal component of each observation $p_i$. Equipped with this object, we normalize the principal component such that it is bounded between zero and one. Formally, let us denote $S$ the set including each non-normalized principal component. We normalize each principal component according to

$$n_i = max\left\{\frac{p_i - min\{S\}}{max\{S\} - min\{S\}}; 0\right\}$$

Still using O*NET data, we convert 8 digit SOC codes into 6 digit SOC codes and adjust our skill measure so as to be the average of each 8 digit measure within each 6 digit code. For example, if profession 11111112 had a skill measure of 0.70 and profession

Table 2: Skill Measure by Occupational Group

| **Highest Skill levels** | | |
|---|---|---|
| 111011 | 1,000 | Chief Executives |
| 192012 | 0,869 | Physicists |
| 119151 | 0,844 | Social and Community Service Managers |
| 119121 | 0,838 | Natural Sciences Managers |
| 212011 | 0,826 | Clergy |
| 193032 | 0,824 | Industrial-Organizational Psychologists |
| 291067 | 0,810 | Surgeons |
| 113131 | 0,807 | Training and Development Managers |
| 113121 | 0,805 | Human Resources Managers |
| 172051 | 0,798 | Civil Engineers |
| **Lowest Skill levels** | | |
| 513023 | 0,079 | Slaughterers and Meat Packers |
| 372012 | 0,076 | Maids and Housekeeping Cleaners |
| 537111 | 0,074 | Mine Shuttle Car Operators |
| 372011 | 0,071 | Janitors and Cleaners, Except Maids and Housekeeping Cleaners |
| 473015 | 0,071 | Helpers–Pipelayers, Plumbers, Pipefitters, and Steamfitters |
| 359021 | 0,061 | Dishwashers |
| 516021 | 0,020 | Pressers, Textile, Garment, and Related Materials |
| 452041 | 0,016 | Graders and Sorters, Agricultural Products |
| 537061 | 0,007 | Cleaners of Vehicles and Equipment |
| 537064 | 0,000 | Packers and Packagers, Hand |

**Sources:** O*NET Dataset, ISCO classification and National Classification of Portuguese Occupations.
**Note:** This table reports the skill composition measure built in this paper associated with selected occupations. We select the ten highest ranked and the ten lowest ranked occupations and display the associated skill score index.

11111120 has a skill measure of 0.76, then profession 111111 will have a skill measure of 0.73. This leaves us with 747 different occupations. Table 2 shows the first ten and last ten entries of this crosswalk, as a sanity check for whether highly skilled professions indeed have a high skill measure associated to them. Once we have this, we merge this information of skills at the SOC occupation level with corresponding ISCO08 codes. We then, trim ISCO08 classification at the 3 digit level and take the mean of the skill measure within each of these 3 digits ISCO08 categories. This step is thus simply generating a correspondance between ISCO08 at the 3 digit level and an associated skill measure for each 3 digit occupational group. It is then possible to bring together ISCO08 data and the portuguese classification of Professions. Once we merge this information with Quadros de Pessoal, we are endowed with a measure of skill intensity for each worker in labor data. Averaging this measure within the firm, we get a measure of firm skill composition.

## C.2 Sistema de Contas Integradas das Empresas

We use *Sistema de Contas Integradas das Empresas* (SCIE), which is longitudinal, firm-level data set collected by Statistics Portugal (INE). This dataset links with QP through the unique firm identification code, designated $NPC\_FIC$ for the most recent years. SCIE covers all firms (companies, individual entrepreneurs, and self-employed) that produce goods or services during the year, excluding firms in the insurance and financial sector, those that produce agricultural products or entities that are not market oriented. From 2005 to 2019, each year has more than 1 million firm observations detailing their economic activity (for example, CAE industry code, geographical location (according to the *Nomenclatura das Unidades Territoriais para Fins Estatísticos*, NUTS, II), birth/death, and number of workers) and accounting statements. Generically, the dataset includes information on financing and accounting variables. Employment and labor productivity (since we can recover value added for each firm in each year, and since we have employment from QP data) variables can also be extracted from SCIE.

# References

**Alvarez, Jorge, Felipe Benguria, Niklas Engbom, and Christian Moser**, "Firms and the Decline in Earnings Inequality in Brazil," *American Economic Journal: Macroeconomics*, January 2018, *10* (1), 149–189.

**Atkinson, Anthony B.**, "More on the Measurement of Inequality," *Journal of Economic Inequality*, 2008, *6* (3), 277. Publisher: Springer Nature BV.

**Autor, David H., Lawrence F. Katz, and Melissa S. Kearney**, "Rising Wage Inequality: The Role of Composition and Prices," NBER Working Paper 11628, National Bureau of Economic Research, MA September 2005.

**Bloom, Nicholas, Fatih Guvenen, Benjamin S. Smith, Jae Song, and Till von Wachter**, "The Disappearing Large-Firm Wage Premium," *AEA Papers and Proceedings*, May 2018, *108*, 317–322.

**Gastwirth, Joseph L.**, "A General Definition of the Lorenz Curve," *Econometrica: Journal of the Econometric Society*, 1971, pp. 1037–1039. Publisher: JSTOR.

**Lise, Jeremy and Fabien Postel-Vinay**, "Multidimensional Skills, Sorting, and Human Capital Accumulation," *American Economic Review*, August 2020, *110* (8), 2328–2376.

**Machado, José A. F. and José Mata**, "Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression," *Journal of Applied Econometrics*, 2005, *20* (4), 445–465. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.788.

**Messina, Julian and Joana Silva**, "Twenty Years of Wage Inequality in Latin America," *World Bank Economic Review*, February 2021, *35* (1), 117–147.

**Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter**, "Firming up Inequality," *Quarterly Journal of Economics*, February 2019, *134* (1), 1–50. Publisher: Oxford University Press, eprint = https://academic.oup.com/qje/article-pdf/134/1/1/28921941/qjy025.pdf.