DISCUSSION PAPER SERIES

IZA

IZA DP No. 7036

Worker Productivity and Wages: Evidence from Linked Employer-Employee Data

Ana Sofia Lopes
Paulino Teixeira

November 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Worker Productivity and Wages: Evidence from Linked Employer-Employee Data

Ana Sofia Lopes
Polytechnic Institute of Leiria and GEMF

Paulino Teixeira
University of Coimbra, GEMF and IZA

Discussion Paper No. 7036
November 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7036
November 2012

ABSTRACT

Worker Productivity and Wages: Evidence from Linked Employer-Employee Data

This study compares the determinants of productivity and wages at both firm and worker level. In the firm-level analysis, we follow Hellerstein, Neumark and Troske (1999) and provide improved estimates based on an extended set of covariates including the intensity of firm-provided training. In the worker-level analysis we take a new turn and generate a proxy for unobserved worker productivity. Our results point to the presence of sizeable spillover effects from schooling and training as their impact is bigger on firm-level productivity equations than on the corresponding worker-level equations. In turn, our fully disaggregated model at worker level shows that, by using all possible combinations of worker attributes, we obtain that the wage differences across different worker groups are mostly productivity based and that the gap can be as high as 33%.

JEL Classification: C23, D24, J31

Keywords: worker productivity, wages, human capital, LEED

Corresponding author:

Paulino Teixeira
Faculdade de Economia
Universidade de Coimbra
Av. Dias da Silva, 165
3004-512 Coimbra
Portugal
Email: pteixeira@fe.uc.pt

1. Introduction

Analysis of wage differentials among different types of workers requires examination of the corresponding differences in productivity, otherwise competitive (or productivity-based) and noncompetitive explanations of the observed differentials cannot be distinguished. In this paper, we follow the pioneering work of Hellerstein, Neumark and Troske (1999) and use a similar firm-level modeling strategy, including a full treatment of firm fixed effects.

Firm-level estimation uses differences across firms to explain differences in productivity, but it ignores differences within heterogeneous firms. The presumption here is that workers with higher levels of schooling, for example, are more likely to generate new ideas that may spill over to their fellow co-workers so that the composition of the workforce does matter for the performance of the entire firm (e.g. Jovanovic and Rob, 1989, and Moretti, 2004). To tackle this issue, we propose in this study to combine worker- and firm-level information in order to evaluate the importance of human capital spillovers within firms, while controlling for both worker and firm fixed effects.

We start by assuming a firm-level Cobb-Douglas production function in which labour is the product of hours worked multiplied by 'quality', where quality is a function of observables. Similarly, and based on a standard Mincerian earnings equation, we derive a wage equation at firm level. Then, as both productivity and wages are a function of the same set of regressors, we jointly estimate the two equations to infer the extent to which the productivity gains from schooling and training, *inter al.*, are shared by workers and firms.

In a second stage, we estimate a *proxy* for individual productivity based on the assumption that the unobserved wage-output per hour ratio at worker level and the observed wage-output ratio at firm level are a function of the same set of regressors. We also make the workable assumption that the logarithm of the latter is equal to the mean of the logarithms of the former (see equation (3.3) below). As a result of this procedure, we obtain that by aggregating the generated individual (worker) productivity level across workers in a given firm the implied aggregate (i.e. the firm productivity level) is highly correlated with the average productivity observed at firm level.

Our main contribution in this study is therefore to derive a unique and distinct analytical framework for the determination of worker level productivity. We also run a number of models testing for the robustness of our results with respect to the selected model assumptions.

Our empirical model uses a 2-year long *LEED* panel comprising 288,129 workers and 1,174 firms from the Portuguese manufacturing sector. In our data, which is extracted from two data sources (i.e. *Balanço Social* and *Quadros de Pessoal*), we are able to follow workers and firms longitudinally. We also observe individual – worker and firm – characteristics, including detailed information on firm-provided training which is extracted from *Balanço Social*. This is a relevant aspect as studies exclusively based on *Quadros de Pessoal* cannot control for the workplace training variable.[1]

This paper is organized as follows. In the next section we present the modeling required to estimate the determinants of both firm-level productivity and wages, as well as the full derivation of our selected *proxy* for worker productivity. Section 3 describes our longitudinal *LEED* dataset and Section 4 presents the findings, including an extensive analysis of robustness of our results to various model assumptions. The main conclusions are drawn in Section 5.

2. Modelling

2.1 Firm productivity

We start by considering a Cobb-Douglas production function given by

$$Y_{jt} = AL_{jt}^{\alpha} K_{jt}^{\beta} e^{(\eta Z_{jt} + \pi \psi_j + \varepsilon_{jt})}, \tag{1.1}$$

where Y_{jt} denotes the value added of firm *j* in period (year) *t*. *A* is an efficiency parameter, K_{jt} is the stock of capital, Z_{jt} is a vector of firm characteristics, ψ_j is the (time-invariant) unobserved heterogeneity of firm *j*, and ε_{jt} denotes the error term (*i.i.d.*).[2] L_{jt} is the labour

[1] Lopes and Teixeira (2012) provide a detailed analysis of workplace training using *Balanço Social*.

[2] ψ_j gives the worker average unobserved ability in firm *j* plus an unknown firm specific effect. See section 2.2 below.

input, given by $L_{jt} = h_{jt} * V_{jt}$, where *h* is hours of work per employee, and *V* is a labour composite as explained next.

Let us first suppose that we observe two workforce characteristics, say, schooling and gender, given, respectively, by the number of workers with a high-school degree and the number of males workers. Then *V* is given by (subscripts *j* and *t* omitted hereafter):

$$V = N^{FsE} + \gamma_E N^{FE} + \gamma_G N^{MsE} + \gamma_E \gamma_G N^{ME}, \tag{1.2}$$

where N^{FsE} (N^{MsE}) denotes the number of female (male) workers with a level of education lower than a high-school degree; N^{FE} (N^{ME}) is the number of females (males) with at least a high-school degree. $\gamma_E = \frac{\partial Y / \partial N^{FE}}{\partial Y / \partial N^{FsE}}$, $\gamma_G = \frac{\partial Y / \partial N^{MsE}}{\partial Y / \partial N^{FsE}}$, and $\gamma_E \gamma_G = \frac{\partial Y / \partial N^{ME}}{\partial Y / \partial N^{FsE}}$, are, respectively, the ratios of the corresponding marginal productivities.[3]

Model (1.2) follows from Hellerstein, Neumark and Troske (1999), and, clearly, it assumes that the selected worker categories are perfect substitutes. Imperfect substitution can, however, be easily tested by using an alternative specification, $V = N^{FsE} \left(N^{FE}\right)^{\gamma_E} \left(N^{MsE}\right)^{\gamma_G} \left(N^{ME}\right)^{\gamma_E \gamma_G}$, for example. Given that our results seem to be robust to alternative specifications of *V*, our model derivation in this section assumes perfect substitution for the sake of simplicity.

We make three additional assumptions: a) the proportion of workers with a high-school degree is the same across gender, that is, $\frac{N^{FE}}{N^F} = \frac{N^{ME}}{N^M}$, where N^F (N^M) is the number of female (male) workers in the firm); b) γ_E is equal for men and women, that is, $\gamma_E = \frac{\partial Y / \partial N^{FE}}{\partial Y / \partial N^{FsE}} = \frac{\partial Y / \partial N^{ME}}{\partial Y / \partial N^{MsE}}$; and c) γ_G is equal for high- and low-education workers, that is, $\gamma_G = \frac{\partial Y / \partial N^{MsE}}{\partial Y / \partial N^{FsE}} = \frac{\partial Y / \partial N^{ME}}{\partial Y / \partial N^{FE}}$. (All these assumptions will be relaxed in section 4.3.)

Under assumptions a), b) and c), model (1.2) yields:

---

[3] For example, if $\gamma_G > 1$, then the marginal contribution of a male worker is higher than the marginal contribution of a female worker, both with less than a high-school degree.

$$V = N^F \frac{N^{sE}}{N} + \gamma_E N^F \frac{N^E}{N} + \gamma_G N^M \frac{N^{sE}}{N} + \gamma_E \gamma_G N^M \frac{N^E}{N} \Leftrightarrow$$

$$\Leftrightarrow V = \left(N^F + \gamma_G N^M\right) * \left(\frac{N^{sE}}{N} + \gamma_E \frac{N^E}{N}\right) \Leftrightarrow$$

$$\Leftrightarrow V = N\left(\frac{N^F}{N} + \gamma_G \frac{N^M}{N}\right) * \left(\frac{N^{sE}}{N} + \gamma_E \frac{N^E}{N}\right), \tag{1.3}$$

where N is the number of workers in the firm, N^E is the number of workers with at least a high-school degree, and $N = N^E + N^{sE}$. From (1.3), we then have:

$$V = N\left(1 - G + \gamma_G G\right) * \left(1 - E + \gamma_E E\right) \Leftrightarrow$$

$$V = N\left[1 + \left(\gamma_G - 1\right) G\right] * \left[1 + \left(\gamma_E - 1\right) E\right], \tag{1.4}$$

with $G = N^M / N$ (i.e. the proportion of male workers in the firm) and $E = N^E / N$ (i.e. the proportion of workers with a high-school degree).

Model (1.4) can be easily extended to accommodate characteristics *T*, *S*, and *O*, given, respectively, by the proportion of training participants, workers with at least 10 years of service, workers between 25 and 44 years old, plus several job occupation categories: top managers and professionals (Q_1), other managers and professionals (Q_2), foremen and supervisors (Q_3), highly skilled and skilled personnel (Q_4), and semiskilled personnel (Q_5). In this case, we have:

$$V = N * \left[1 + (\gamma_G - 1)G\right] * \left[1 + \left(\gamma_T - 1\right)T\right] * \left[1 + (\gamma_E - 1)E\right] * \left[1 + (\gamma_S - 1)S\right] * \left[1 + \left(\gamma_O - 1\right)O\right] *$$

$$* \left[1 + (\gamma_{Q_1} - 1)Q_1 + (\gamma_{Q_2} - 1)Q_2 + (\gamma_{Q_3} - 1)Q_3 + (\gamma_{Q_4} - 1)Q_4 + (\gamma_{Q_5} - 1)Q_5\right]. \tag{1.5}$$

By substituting (1.5) into the production function (1.1) and making $L=h*V$ and $H = h * N$, we have:

$$Y = AH^{\alpha} * \left[1 + (\gamma_G - 1)G\right]^{\alpha} * \left[1 + \left(\gamma_T - 1\right)T\right]^{\alpha} * \left[1 + (\gamma_E - 1)E\right]^{\alpha} * \left[1 + \left(\gamma_O - 1\right)O\right]^{\alpha} *$$

$$* \left[1 + (\gamma_{Q_1} - 1)Q_1 + (\gamma_{Q_2} - 1)Q_2 + (\gamma_{Q_3} - 1)Q_3 + (\gamma_{Q_4} - 1)Q_4 + (\gamma_{Q_5} - 1)Q_5\right]^{\alpha} *$$

$$* \left[1 + (\gamma_S - 1)S\right]^{\alpha} * K^{\beta} e^{(\eta Z + \pi\psi + \varepsilon)}. \tag{1.6}$$

Dividing (1.6) by *H*, we obtain *y*, that is, the hourly productivity of labour:

$$y = AH^{(\alpha + \beta - 1)} * \left[1 + (\gamma_G - 1)G\right]^{\alpha} * \left[1 + \left(\gamma_T - 1\right)T\right]^{\alpha} * \left[1 + (\gamma_E - 1)E\right]^{\alpha} * \left[1 + \left(\gamma_O - 1\right)O\right]^{\alpha} *$$

$$* \left[1 + (\gamma_{Q_1} - 1)Q_1 + (\gamma_{Q_2} - 1)Q_2 + (\gamma_{Q_3} - 1)Q_3 + (\gamma_{Q_4} - 1)Q_4 + (\gamma_{Q_5} - 1)Q_5\right]^{\alpha} *$$

$$* \left[1 + (\gamma_S - 1)S\right]^{\alpha} * k^{\beta} e^{(\eta Z + \pi\psi + \varepsilon)}, \tag{1.7}$$

where k denotes capital intensity.

Taking logarithms, we have:

$$Ln\ y = Ln\ A + (\alpha+\beta-1)Ln\ H + \alpha Ln\ \left[1+(\gamma_G-1)G\right] + \alpha Ln\ \left[1+\left(\gamma_T-1\right)T\right] +$$
$$+\alpha Ln\ \left[1+(\gamma_E-1)E\right] + \alpha Ln\ \left[1+\left(\gamma_O-1\right)O\right] + \alpha Ln\ \left[1+(\gamma_S-1)S\right] +$$
$$+\alpha Ln\ \left[1+(\gamma_{Q_1}-1)Q_1+(\gamma_{Q_2}-1)Q_2+(\gamma_{Q_3}-1)Q_3+(\gamma_{Q_4}-1)Q_4+(\gamma_{Q_5}-1)Q_5\right] +$$
$$+\beta Ln\ k + \eta Z + \pi\psi + \varepsilon. \qquad (1.8)$$

Finally, assuming $Ln\ \left[1+\left(\gamma_R-1\right)R\right] \cong \left(\gamma_R-1\right)R,$ for all $R = G,T,E,O,S,Q_1,Q_2,Q_3,Q_4,Q_5$, we obtain, under constant returns to scale (CRS):[4]

$$Ln\ y = Ln\ A + \alpha\left(\gamma_G-1\right)G + \alpha(\gamma_T-1)T + \alpha(\gamma_E-1)E + \alpha\left(\gamma_O-1\right)O +$$
$$+\alpha(\gamma_S-1)S + \alpha\left(\gamma_{Q_1}-1\right)Q_1 + \alpha\left(\gamma_{Q_2}-1\right)Q_2 + \alpha\left(\gamma_{Q_3}-1\right)Q_3 + \alpha\left(\gamma_{Q_4}-1\right)Q_4 +$$
$$+\alpha\left(\gamma_{Q_5}-1\right)Q_5 + \beta Ln\ k + \eta Z + \pi\psi + \varepsilon. \qquad (1.9)$$

2.2 Earnings equation

At the worker level, the augmented Mincerian wage equation can be given by:

$$Ln\ w_{it} = Ln\ A_w + \lambda^i X'_{it} + \eta_w Z_{j(i)t} + \beta_w Ln\ k_{j(i)t} + \alpha_i + \phi_{j(i)} + \xi_{it}, \qquad (2.1)$$

where $Ln\ w_{it}$ is the (log) hourly earnings for worker *i* in period *t*. $Log\ A_w$ is a constant term, α_i denotes the (time-invariant) unobserved ability of worker *i*, while $\phi_{j(i)}$ is the (time-invariant) unobserved effect specific to firm *j*. Again, $Z_{j(i)t}$ denotes the vector of characteristics of firm *j* and $Ln\ k_{j(i)t}$ is the logarithm of capital intensity. X'_{it} comprises all the dummy variables flagging worker-level characteristics.

An equivalent expression for (2.1) is:

$$Ln\ w_{it} = Ln\ A_w + \lambda^i{}_G G'_{it} + \lambda^i{}_T T'_{it} + \lambda^i{}_E E'_{it} + \lambda^i{}_O O'_{it} + \lambda^i{}_S S'_{it} + \lambda^i{}_{Q_1} Q'_{1,it} + \lambda^i{}_{Q_2} Q'_{2,it} + \lambda^i{}_{Q_3} Q'_{3,it} +$$
$$+\lambda^i{}_{Q_4} Q'_{4,it} + \lambda^i{}_{Q_5} Q'_{5,it} + \eta_w Z_{j(i)t} + \beta_w Ln\ k_{j(i)t} + \theta_i + \psi_{j(i)} + \xi_{it}, \qquad (2.2)$$

where E'_{it}, G'_{it}, T'_{it}, O'_{it}, S'_{it}, Q'_{it} are now defined at worker level. E'_{it}, for example, is equal to 1 if worker *i* has at least a high-school degree, 0 otherwise. And similarly for all the

[4] As it will be shown in section 4, the CRS assumption is not rejected by the data. The non-linear version of (1.8) – the one that does not use the approximation $Ln\ \left[1+\left(\gamma_R-1\right)R\right] \cong \left(\gamma_R-1\right)R$ – produces similar results to those based on model (1.9).

other worker-level covariates.[5] We also make $\alpha_i + \phi_{j(i)} = \theta_i + \psi_{j(i)}$, using $\psi_{j(i)} = \phi_{j(i)} + \overline{\alpha}_{j(i)}$, and $\theta_i = \alpha_i - \overline{\alpha}_{j(i)}$, where $\overline{\alpha}_{j(i)}$ is the worker average unobserved ability in firm *j*.

Summing up across workers in firm *j* in period *t*, we have:

$$\sum Ln\ w_{it} = N_{jt} Ln\ A_w + \lambda_G^i N_{jt}^M + \lambda_T^i N_{jt}^T + \lambda_E^i N_{jt}^E + \lambda_O^i N_{jt}^O + \lambda_S^i N_{jt}^S + \lambda_{Q_1}^i N_{jt}^{Q_1} + \lambda_{Q_2}^i N_{jt}^{Q_2} + \\ + \lambda_{Q_3}^i N_{jt}^{Q_3} + \lambda_{Q_4}^i N_{jt}^{Q_4} + \lambda_{Q_5}^i N_{jt}^{Q_5} + N_{jt}\eta_w Z_{jt} + N_{jt}\beta_w Ln\ k_{jt} + N_{jt}\psi_j + N_{jt}\xi_{jt}. \qquad (2.3)$$

(By definition, $\sum \theta_i = 0$.)

Finally, dividing equation (2.3) by N_{jt} (i.e. the number of workers in firm *j* in period *t*), we have:

$$1/N_{jt} \sum Ln\ w_{it} = Ln\ A_w + \lambda_G^i G_{jt} + \lambda_T^i T_{jt} + \lambda_E^i E_{jt} + \lambda_O^i O_{jt} + \lambda_S^i S_{jt} + \lambda_{Q_1}^i Q_{1,jt} + \lambda_{Q_2}^i Q_{2,jt} + \\ + \lambda_{Q_3}^i Q_{3,jt} + \lambda_{Q_4}^i Q_{4,jt} + \lambda_{Q_5}^i Q_{5,jt} + \eta_w Z_{jt} + \beta_w Ln\ k_{jt} + \psi_j + \xi_{jt}. \qquad (2.4)$$

Now, as we want to compare directly the determinants of firm productivity with the determinants of the firm average wage, the dependent variable of model (2.4) is replaced by the log hourly average wage in firm *j* in period *t*, $Ln\ w_{jt}$, with $Ln\ w_{jt} = Ln\ \frac{1}{N_{jt}} \sum_{i=1}^{N_{jt}} w_{it}$, to yield:[6]

$$Ln\ w_{jt} = Ln\ A_w + \lambda_G G_{jt} + \lambda_T T_{jt} + \lambda_E E_{jt} + \lambda_O O_{jt} + \lambda_S S_{jt} + \lambda_{Q_1} Q_{1,jt} + \lambda_{Q_2} Q_{2,jt} + \lambda_{Q_3} Q_{3,jt} + \\ + \lambda_{Q_4} Q_{4,jt} + \lambda_{Q_5} Q_{5,jt} + \eta_w Z_{jt} + \beta_w Ln\ k_{jt} + \psi_j + e_{jt}. \qquad (2.4)'$$

Given that the log function is concave, we have, by Jensen's inequality, $Ln\ w_{jt} > 1/N_{jt} \sum Ln\ w_{it}$. We note, however, that estimation of model (2.4)' produces similar results to those obtained by estimating model (2.4).

Next we tackle the unobserved firm heterogeneity issue. In matrix notation, equation (2.4)' becomes

$$\overline{Lw} = LA + X\lambda + Z\eta_w + Lk\beta_w + F\psi + e, \qquad (2.5)$$

[5] Worker information is extracted from *Quadros de Pessoal*. By definition, the generated firm-level aggregate matches the corresponding variable taken from *Balanço Social*. The full description of the dataset is provided in Section 3.

[6] We note that, similarly to equation (1.9) above, where, for example, the coefficient γ_T gives the relative productivity of the effect of training, the coefficient λ_T in equation (2.4') gives the relative wage mark-up associated with the same variable. For a quick derivation of the latter effect, see Dearden, Reed and Reenen (2006, p. 402, equation (10)), where it is also assumed for the sake of simplicity the log approximation mentioned at the end of section 2.2 above.

where F is a $JT \times J$ matrix of dummies flagging the J firms. (J is the number of firms in the sample; T is the length of the time series.)

Consider now the matrix of orthogonal projection in F, $P_F = F(F^T F)^{-1} F^T$, and the matrix M_F, given by $M_F = I - P_F$. Multiplying equation (2.5) by M_F, we have:

$$M_F \overline{Lw} = M_F LA + M_F X\lambda + M_F Z\eta_w + M_F Lk\beta_w + M_F F\psi + M_F e, \qquad (2.6)$$

where the first element of the matrix $M_F Z$, for example, is given by $z_{1,1}^1 - \frac{z_{1,1}^1 + z_{1,2}^1}{2}$.[7]

By definition, we have $M_F F\psi = 0$, so we can easily compute $\hat{LA}$; $\hat{\lambda}$; $\hat{\eta}_w$ and $\hat{\beta}_w$. Finally, we estimate the unobserved effects of firms by using fixed effects applied to the difference between the observed average wage of the firm and the expected average wage, given the set of covariates:

$$\hat{\psi} = (F^T F)^{-1} F^T \left(\overline{Lw} - \hat{LA} - X\hat{\lambda} - Z\hat{\eta}_w - Lk\hat{\beta}_w \right). \qquad (2.7)$$

Once obtained the firm fixed effects, $\hat{\psi}_j$, they can be inserted into models (1.9) and (2.4)' to obtain:

$$\begin{aligned} Ln\, y_{jt} &= Ln\, A + \alpha(\gamma_G - 1)G_{jt} + \alpha(\gamma_T - 1)T_{jt} + \alpha(\gamma_E - 1)E_{jt} + \alpha(\gamma_O - 1)O_{jt} + \\ &+ \alpha(\gamma_S - 1)S_{jt} + \alpha(\gamma_{Q_1} - 1)Q_{1,jt} + \alpha(\gamma_{Q_2} - 1)Q_{2,jt} + \alpha(\gamma_{Q_3} - 1)Q_{3,jt} + \alpha(\gamma_{Q_4} - 1)Q_{4,jt} + \\ &+ \alpha(\gamma_{Q_5} - 1)Q_{5,jt} + \beta Ln\, k_{jt} + \eta Z_{jt} + \pi\hat{\psi}_j + \varepsilon_{jt}, \end{aligned} \qquad (2.8)$$

and

$$\begin{aligned} Ln\, w_{jt} - \hat{\psi}_j &= Ln\, A_w + \lambda_G G_{jt} + \lambda_T T_{jt} + \lambda_E E_{jt} + \lambda_O O_{jt} + \lambda_S S_{jt} + \lambda_{Q_1} Q_{1,jt} + \lambda_{Q_2} Q_{2,jt} + \lambda_{Q_3} Q_{3,jt} + \\ &+ \lambda_{Q_4} Q_{4,jt} + \lambda_{Q_5} Q_{5,jt} + \eta_w Z_{jt} + \beta_w Ln\, k_{jt} + e_{jt}, \end{aligned} \qquad (2.9)$$

which means that we have two models with a common left hand side. We are therefore in a position to examine whether a given covariate has a bigger impact on wages than on productivity, or the other way around. By construction, $\hat{\psi}_j$ contains the average (unobserved) worker attributes, $\overline{\alpha}_j$, which means that in our firm-level equations we do control for the possible correlation between unobserved ability of workers and the observed characteristics of workers at the firm level. This is of course a non-trivial aspect of our modelling strategy.

[7] $z_{1,1}^1$ ($z_{1,2}^1$) denotes the first characteristic of firm 1 in period 1 (2).

2.3 An estimate of worker productivity

In this section, we extend the investigation on the determinants of productivity and wages by estimating both worker-level wage and productivity equations. We assume, in particular, that worker productivity, y_{it}, is a function of the same set of covariates as specified by the wage equation (2.2), which means that it depends on both observed and unobserved worker and firm characteristics. Thus, we have (in logs):[8]

$$\begin{aligned} Ln\ y_{it} &= Ln\ A^i + \alpha^i\left(\gamma_G^i - 1\right)G'_{it} + \alpha^i\left(\gamma_T^i - 1\right)T'_{it} + \alpha^i\left(\gamma_E^i - 1\right)E'_{it} + \alpha^i\left(\gamma_O^i - 1\right)O'_{it} + \\ &+ \alpha^i\left(\gamma_S^i - 1\right)S'_{it} + \alpha^i\left(\gamma_{Q_1}^i - 1\right)Q'_{1,it} + \alpha^i\left(\gamma_{Q_2}^i - 1\right)Q'_{2,it} + \alpha^i\left(\gamma_{Q_3}^i - 1\right)Q'_{3,it} + \alpha^i\left(\gamma_{Q_4}^i - 1\right)Q'_{4,it} + \\ &+ \alpha^i\left(\gamma_{Q_5}^i - 1\right)Q'_{5,it} + \beta Ln\ k_{j(i)t} + \eta Z_{j(i)t} + \kappa\hat{\theta}_i + \tau\hat{\psi}_{j(i)} + \mu_{it}. \end{aligned} \qquad (3.1)$$

Unfortunately, y_{it} (or $Ln\ y_{it}$) is unobservable. We next show that under a fairly reasonable set of assumptions it is possible to find an estimate of $Ln\ y_{it}$ (call it $Ln\ y_{it}^*$) so that one can estimate (3.1) and compare with (2.2).

Firstly, note that a) we do observe $Ln\ y_{jt}$, where $y_{jt} = \frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}} y_{it}$ is the productivity level of firm *j* in period *t*; and b) $Ln\left(\frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}} y_{it}\right) > \frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}} Ln\ y_{it}$ (by Jensen's inequality). Now let us assume the following equality**:**

$$Ln\ w_{j(i)t} - \frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}} Ln\ w_{it} = Ln\ y_{j(i)t} - \frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}} Ln\ y_{it}. \qquad (3.2)$$

Clearly, (3.2) implies that the individual wage, w_{it}, and the firm average wage, $w_{j(i)t}$, on the one hand, and the worker productivity, y_{it}, and the firm productivity level, $y_{j(i)t}$, on the other, are interconnected in a similar fashion.[9]

[8] We use equation (A.7) in the appendix to estimate the value of worker i's innate attributes in comparison with the average of innate attributes of workers in firm *j* (i.e. $\hat{\theta}_i$).

[9] We note that by further manipulating (3.2) we have:

From (3.2.), we also have:

$$\frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}} Ln\ w_{it} - \frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}} Ln\ y_{it} = Ln\ w_{j(i)t} - Ln\ y_{j(i)t} \Leftrightarrow$$

$$\Leftrightarrow \frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}}\left(Ln\ w_{it} - Ln\ y_{it}\right) = Ln\ w_{j(i)t} - Ln\ y_{j(i)t} \Leftrightarrow$$

$$\Leftrightarrow \frac{1}{N_{jt}}\sum_{i=1}^{N_{jt}} Ln\ D_{it} = Ln\ D_{j(i)t}, \qquad (3.3)$$

with $D_{j(i)t} = \frac{W_{j(i)t}}{Y_{j(i)t}} = \frac{w_{j(i)t}}{y_{j(i)t}}$ and $D_{it} = \frac{w_{it}}{y_{it}}$.

Then, using (2.8) and (2.9), we obtain:[10]

$$\begin{aligned}
Ln\ D_{jt} &= Ln\ w_{jt} - Ln\ y_{jt} = \left(Ln\ A_w - Ln\ A\right) + \left[\lambda_G - \alpha\left(\gamma_G - 1\right)\right]G_{jt} + \left[\lambda_T - \alpha(\gamma_T - 1)\right]T_{jt} + \\
&+ \left[\lambda_E - \alpha(\gamma_E - 1)\right]E_{jt} + \left[\lambda_O - \alpha\left(\gamma_O - 1\right)\right]O_{jt} + \left[\lambda_S - \alpha(\gamma_S - 1)\right]S_{jt} + \left[\lambda_{Q_1} - \alpha\left(\gamma_{Q_1} - 1\right)\right]Q_{1,jt} + \\
&+ \left[\lambda_{Q_2} - \alpha\left(\gamma_{Q_2} - 1\right)\right]Q_{2,jt} + \left[\lambda_{Q_3} - \alpha\left(\gamma_{Q_3} - 1\right)\right]Q_{3,jt} + \left[\lambda_{Q_4} - \alpha\left(\gamma_{Q_4} - 1\right)\right]Q_{4,jt} + \\
&+ \left[\lambda_{Q_5} - \alpha\left(\gamma_{Q_5} - 1\right)\right]Q_{5,jt} + \left(\beta_w - \beta\right)Ln\ k_{jt} + \left(\eta_w - \eta\right)Z_{jt} + \left(1 - \pi\right)\hat{\psi}_j + \upsilon_{jt}, \qquad (3.4)
\end{aligned}$$

while using (2.2) and (3.1) yields:

$$\begin{aligned}
Ln\ D_{it} &= Ln\ w_{it} - Ln\ y_{it} = \left(Ln\ A_w - Ln\ A\right) + \left[\lambda_G^i - \alpha^i\left(\gamma_G^i - 1\right)\right]G'_{it} + \left[\lambda_T^i - \alpha^i\left(\gamma_T^i - 1\right)\right]T'_{it} + \\
&+ \left[\lambda_E^i - \alpha^i\left(\gamma_E^i - 1\right)\right]E'_{it} + \left[\lambda_O^i - \alpha^i\left(\gamma_O^i - 1\right)\right]O'_{it} + \left[\lambda_S^i - \alpha^i\left(\gamma_S^i - 1\right)\right]S'_{it} + \left[\lambda_{Q_1}^i - \alpha^i\left(\gamma_{Q_1}^i - 1\right)\right]Q'_{1,it} + \\
&+ \left[\lambda_{Q_2}^i - \alpha^i\left(\gamma_{Q_2}^i - 1\right)\right]Q'_{2,it} + \left[\lambda_{Q_3}^i - \alpha^i\left(\gamma_{Q_3}^i - 1\right)\right]Q'_{3,it} + \left[\lambda_{Q_4}^i - \alpha^i\left(\gamma_{Q_4}^i - 1\right)\right]Q'_{4,it} + \\
&+ \left[\lambda_{Q_5}^i - \alpha^i\left(\gamma_{Q_5}^i - 1\right)\right]Q'_{5,it} + \left(\beta_w^i - \beta^i\right)Ln\ k_{j(i)t} + \left(\eta_w^i - \eta^i\right)Z_{j(i)t} + \\
&+ \left(1 - \kappa\right)\hat{\theta}_i + \left(1 - \tau\right)\hat{\psi}_j + \left(\xi_{it} - \mu_{it}\right). \qquad (3.5)
\end{aligned}$$

$$Ln\ w_{j(i)t} - \frac{1}{N_{jt}} Ln \prod_{i=1}^{N_{jt}} w_{it} = Ln\ y_{j(i)t} - \frac{1}{N_{jt}} Ln \prod_{i=1}^{N_{jt}} y_{it} \Leftrightarrow Ln \prod_{i=1}^{N_{jt}} w_{it} - N_{jt} Ln\ w_{j(i)t} = Ln \prod_{i=1}^{N_{jt}} y_{it} - N_{jt} Ln\ y_{j(i)t} \Leftrightarrow$$

$$\Leftrightarrow Ln \prod_{i=1}^{N_{jt}}\left(\frac{w_{it}}{w_{j(i)t}}\right) = Ln \prod_{i=1}^{N_{jt}}\left(\frac{y_{it}}{y_{j(i)t}}\right) \Leftrightarrow \prod_{i=1}^{N_{jt}}\left(\frac{w_{it}}{w_{j(i)t}}\right) = \prod_{i=1}^{N_{jt}}\left(\frac{y_{it}}{y_{j(i)t}}\right).$$

[10] By using (3.4), we are therefore able to obtain the impact of each observed characteristic on w_{jt}/y_{jt}. For example, the term $\lambda_s - \alpha\left(\gamma_s - 1\right)$ is expected to be positive, as tenure is assumed to have a bigger impact on wages than on the productivity level. By the same token, if ψ_j flags non-competitive high-wage firms, then it will be expected to be associated with a larger w/y ratio.

Then, adding up to all *i, i=1, 2, …,* N_{jt} , and dividing by N_{jt} yields:[11]

$$
\begin{aligned}
\tfrac{1}{N_{jt}} \sum Ln\, D_{it} &= \left(Ln\, A_w - Ln\, A\right) + \left[\lambda_G^i - \alpha^i\left(\gamma_G^i - 1\right)\right] G_{jt} + \left[\lambda_T^i - \alpha^i\left(\gamma_T^i - 1\right)\right] T_{jt} + \\
&+ \left[\lambda_E^i - \alpha^i\left(\gamma_E^i - 1\right)\right] E_{jt} + \left[\lambda_O^i - \alpha^i\left(\gamma_O^i - 1\right)\right] O_{jt} + \left[\lambda_S^i - \alpha^i\left(\gamma_S^i - 1\right)\right] S_{jt} + \left[\lambda_{Q_1}^i - \alpha^i\left(\gamma_{Q_1}^i - 1\right)\right] Q_{1,jt} + \\
&+ \left[\lambda_{Q_2}^i - \alpha^i\left(\gamma_{Q_2}^i - 1\right)\right] Q_{2,jt} + \left[\lambda_{Q_3}^i - \alpha^i\left(\gamma_{Q_3}^i - 1\right)\right] Q_{3,jt} + \left[\lambda_{Q_4}^i - \alpha^i\left(\gamma_{Q_4}^i - 1\right)\right] Q_{4,jt} + \\
&+ \left[\lambda_{Q_5}^i - \alpha^i\left(\gamma_{Q_5}^i - 1\right)\right] Q_{5,jt} + \left(\beta_w^i - \beta^i\right) Ln\, k_{jt} + \left(\eta_w^i - \eta^i\right) Z_{jt} + (1-\tau)\hat{\psi}_j + \left(\xi_{it} \text{-} \mu_{it}\right). \qquad (3.6)
\end{aligned}
$$

In other words, by aggregating (3.5) we end up with $Ln\, D_{jt}$ since $\frac{1}{N_{jt}} \sum_{i=1}^{N_{jt}} Ln\, D_{it} = Ln\, D_{j(i)t}$ (by (3.3)). This means that we can obtain $Ln\hat{D}_{it}$ by substituting all the estimated coefficients in (3.6) into equation (3.5). Then, given $Ln\, D_{it} = Ln\, w_{it} - Ln\, y_{it}$, we have $Ln\, y_{it} = Ln\, w_{it} - Ln\hat{D}_{it}$, which in turn can be used to obtain an estimate ($Ln\, y_{it}^*$) of $Ln\, y_{it}$, given by:

$$Ln\, y_{it}^* = Ln\, w_{it} - Ln\hat{D}_{it} \,. \qquad (3.7)$$

3. Data

Our linked employer-employee dataset (*LEED*) was obtained by combining *Quadros de Pessoal* (worker-level information) and *Balanço Social* (firm-level information), both from *Gabinete de Estudos e Planeamento* (GEP) of the Portuguese Ministry of Labor. These two datasets are linked using the unique identification number allocated to each firm. *Quadros de Pessoal* covers the entire population of firms with at least one employee excluding public administration, while firms in *Balanço Social* have at least 100 employees. By construction, all firms in *Balanço Social* are necessarily in *Quadros de Pessoal* database.

Balanço Social includes information on value added, annual worker earnings, the number of employees, hours worked, sector of activity, and region. It also contains information on firm average characteristics of workers (for example age, gender, schooling,

[11] By definition, we have $\tfrac{1}{N_{jt}} \sum G'_{it} = \tfrac{N_{jt}^M}{N_{jt}} = G_{jt}$. And similarly for all other variables.

tenure, and skill categories). A key feature of *Balanço Social* is that it contains detailed information on firm-provided training, including the number of training sessions (on- and off-the-job), the number and share of training participants by occupation level and the number of training hours.

The information on individual worker attributes is extracted from *Quadros de Pessoal*. It contains monthly earnings, hours of work, age, gender, schooling level, skill, tenure, occupation, and whether the individual is a full or part-time worker, *inter al*.[12] Based on information provided by *Balanço Social*, we have also imputed training participation at worker level. The imputed training variable is then used in worker-level regressions. (The imputation procedures are available from the authors upon request.)

The estimation sample was obtained by applying several filters to the raw *LEED* dataset. In particular, we dropped all firms located in Madeira and Açores. We also focused on manufacturing. On the whole we have in our sample 288,129 workers and 1,174 firms. All firms have at least 100 employees and were observed in 1998 and 1999.[13]

The summary statistics of the estimation sample are presented in Table 1, both at firm and worker level. Firstly, the wage dispersion is pronounced, both across firms and workers. In column (2), the coefficient of variation is equal to 0.4, which is a lot bigger, for example, than the value observed in Germany, at 0.1 (Addison, Teixeira and Zwick, 2010, Table 1a). Similar heterogeneity is detected in productivity levels across firms. Secondly, the difference between worker- and firm-level (weighted) means, in columns (1) and (2), respectively, is small and it is solely due to the fact that 'atypical' workers were dropped from the corresponding sample (see footnote 13). Finally, the standard deviation of earnings, age, schooling, gender and tenure in column (1) are roughly ½ of the corresponding value in column (2), an indication that there is a sizeable sorting of individuals across firms.

---

[12] By aggregating worker information at firm level we were able to check the corresponding information extracted from *Balanço Social*.

[13] Job switchers, part-time workers, individuals aged less than 16 or more than 65, apprentices, and individuals with earnings less than the statutory minimum wage were eliminated from the worker sample.

4. Results

4.1 Productivity and wages at firm-level

Columns (1) and (2) of Table 2 present the results from models (1.9) and (2.8) – the productivity equations, with and without control for unobserved fixed firm effects – while columns (3) and (4) give the corresponding estimates of the wage equations (models (2.4)' and (2.9), respectively). Columns (1) and (3) – and (2) and (4) – are estimated simultaneously to test the independence of the error terms. As a matter of fact, the corresponding χ^2 statistic of the Breusch-Pagan test rejects the null hypothesis of independence of the two models ($P > \chi^2 = 0.0005$). Based on the $\overline{R}^2$ statistic, the included variables explain approximately 64% of the productivity variability in column (1), increasing slightly to 68% in column (2).

The coefficient of the capital intensity variable in the productivity equation is similar to the reported values in the literature (e.g. Dearden, Reed and Reenen, 2006, and Hellerstein and Neumark, 1999). We note that in Hellerstein and Neumark (1999) the capital variable is introduced in the wage equation in order to capture firm unobserved effects which are expected to be positively correlated with capital. By comparing columns (3) and (4), there is indeed some evidence in favor of this hypothesis in the sense that the positive and statistically significant effect of capital on firm wages obtained in column (3) virtually vanishes after introduction of $\hat{\psi}_j$ in column (4). However, given the difference in parameter estimates between the two columns – generated by the presence of firm fixed effects – it seems that capital is a poor *proxy* for unobserved firm heterogeneity.

As expected, schooling and training have a positive impact on both productivity and wages, but with the two regressors having a bigger effect on the former outcome than on the latter. The introduction of firm effects reduces the schooling and training coefficients, an indication that firm unobserved heterogeneity is positively correlated with human capital variables.

The hypothesis of constant returns to scale (CRS) is not rejected by the data, with $P > |t| = 0.692$ in the case of model (1.9) and $P > |t| = 0.413$ in case of model (2.8). Under

CRS we have therefore $\alpha\left(\gamma_E - 1\right) = 0.318$, which implies $\gamma_E = {0.318}\big/{0.775} + 1 = 1.41$.[14] In other words, we estimate that workers with at least a high-school degree are 41% more productive than their counterparts with less schooling. Our estimate of the wage gap between these two worker categories is nevertheless much smaller, at 21.7% (column (4), row 1).

Regarding the training variable, and without controlling for firm-specific effects, the semi-elasticity of training with respect to productivity is approximately twice as big as the semi-elasticity of training with respect to wages (0.099 and 0.046 in columns (1) and (3), respectively). The gap is even bigger after controlling for unobserved effects, with the corresponding coefficients being equal to 0.06 and 0.004, respectively. This means that, after controlling for unobserved heterogeneity, training has still a clear impact on productivity – with training participants being 7.7% more productive than non-participants[15] – but not on wages as the corresponding coefficient is not statistically different from zero in column (4).

A quick measure of the percentage of benefits from human capital investment captured by workers can also be easily derived (e.g. Ballot, Fakhfakh and Taymaz, 2006). Thus, using (2.8) and (2.9), we have, for example, $\left(\frac{dy}{dT}\right) * \frac{1}{y} = \alpha\left(\gamma_T - 1\right)$, and $\left(\frac{dw}{dT}\right) * \frac{1}{w} = \lambda_T$. Then, since $\frac{dy}{dT} = \alpha\left(\gamma_T - 1\right) * y$ and $\frac{dw}{dT} = \lambda_T * w$, the worker and firm shares from training are given by $\frac{\lambda_T}{\alpha\left(\gamma_T - 1\right)} * \frac{w}{y}$ and $1 - \frac{\lambda_T \, {w}/{y}}{\alpha\left(\gamma_T - 1\right)}$, respectively.

In our sample, ${w}/{y}$ is, on average, equal to 37%. This means that only 2.5% ($= \frac{\lambda_T}{\alpha\left(\gamma_T - 1\right)} * \frac{w}{y} = \frac{0.004}{0.06} * 0.37$) of the productivity gains from training are captured by workers. In the case of schooling, the worker share is substantially higher, at 25.2% ($= \frac{\lambda_E}{\alpha\left(\gamma_E - 1\right)} * \frac{w}{y} = \frac{0.217}{0.318} * 0.37$). These results confirm our priors as skills acquired through schooling are considerably more general than those obtained via workplace training.

[14] From column (2), we have: $\alpha + \beta - 1 = 0 \Leftrightarrow \alpha = 1 - 0.225 = 0.775$.

[15] Under constant returns to scale, $\gamma_T = 1.077$.

Tenure has also a positive impact on productivity. On average, workers with higher tenure are 14.2% more productive ($= \alpha(\gamma_S - 1) = 0.11 \Leftrightarrow \gamma_S = {}^{0.11}\!/_{0.775} + 1 = 1.142$), in column (2). But, as expected, the impact on wages is larger than on productivity, at 19.9% in column (4). In contrast, the variables *top managers and professionals* and *highly skilled and skilled personnel* seem to have a bigger impact on productivity than on wages. Interestingly, in both the productivity and wage equations, the evidence points to a negative correlation between tenure and unobserved effects, as the coefficient on tenure actually increases after controlling for unobserved effects.

4.2 Productivity and wages at worker level

As described in section 2.3, (log) worker productivity, $Ln\, y_{it}^{*}$, can be obtained by using expression (3.7). Then, by aggregating at firm level, we obtain $\sum_{i=1}^{N_{jt}} \frac{y_{it}^{*}}{N_{jt}}$, which can then be compared with the observed firm-level productivity, y_{jt}. The result is quite striking in the sense that we find a correlation coefficient of approximately 0.90. Clearly, our measure of worker productivity fits the observed firm data.

We therefore use the obtained estimate of worker productivity to run model (3.1), without and with control for firm and worker fixed effects – Table 3, column (1) and column (2), respectively. Again the determinants of both productivity and wages are obtained assuming no independence in the error terms. Summarizing the main results, we note firstly that there is a substantial reduction in all coefficients from column 1 to 2 and from 3 to 4.[16] (The null of absence of unobserved effects is always rejected.) And, interestingly enough, in both equations the impact of firm and worker fixed effects are very similar which means that the corresponding contribution to the productivity and wages is roughly the same.

The test on the equality of coefficients across equations is easily rejected. For example, in case of schooling we have $F_{(1,\ 408.551)} = 547.59$, while the contribution of training to individual productivity is much higher than to individual wages.

[16] Again the exception is tenure, which seems to be negatively correlated with unobserved effects.

By comparing the productivity equations at firm and worker level (i.e. columns (1) and (2) of Tables 2 and 3), we note that the corresponding coefficients tend to be smaller at worker level. The reduction in the schooling coefficient from Table 2 to Table 3, for example, is particularly pronounced. This means that we did find evidence in favor of the existence of spillovers across workers within the same firm. Additionally, we observe that differences between firm- and worker-level estimates are higher when we compare the second column of the two tables. This is an expected result since, in estimations at worker level, we are able to control for differences in innate ability among workers in the same firm (column (2) of Table 3), while in case of estimations at the firm level, it is only possible to control for the firm average of workers' innate ability (column (2) of Table 2).

Finally, and similarly to what we have done in section 4.1, we can derive a measure of the relative benefits captured by workers and firms out of human capital investments. Thus, for each selected covariate, Table 4 shows the workers' share based in the two separate regressions, at firm and worker level, respectively. We found, in particular, that schooling, training, a higher skill job content, and gender imply a higher worker share when models are estimated at worker level. In contrast, in the case of the tenure variable the share is relatively more favourable to workers in firm-level estimation.

4.3 Robustness

In this section we relax two important assumptions: a) that the proportion across groups is constant, namely $\frac{N^{FE}}{N^{F}} = \frac{N^{ME}}{N^{M}}$; and b) that γ_E, for example, is equal for men and women. And similarly for the remaining characteristics.

Let us again consider equation (1.2), where the quality of labour depends on two characteristics, gender and schooling, but now setting $\gamma_E^F \neq \gamma_E^M$, $\gamma_G^{sE} \neq \gamma_G^E$, $\frac{N^{FE}}{N^F} \neq \frac{N^{ME}}{N^M}$ and $\frac{N^{ME}}{N^E} \neq \frac{N^{MsE}}{N^{sE}}$.[17] In this case, equation (1.2) becomes:

$$V = N^{FsE} + \gamma_E^F N^{FE} + \gamma_G^{sE} N^{MsE} + \gamma_E^F \gamma_G^E N^{ME}, \tag{4.1}$$

where $\gamma_E^F \gamma_G^E$ gives the ratio of the marginal productivity of males with at least a high-school degree to the marginal productivity of females with less than a high-school degree (or $\gamma_E^F \gamma_G^E = \frac{\partial Y / \partial N^{ME}}{\partial Y / \partial N^{FsE}}$). Manipulating further, we have:

$$\begin{aligned} &V = N - N^{FE} - N^{MsE} - N^{ME} + \gamma_E^F N^{FE} + \gamma_G^{sE} N^{MsE} + \gamma_E^F \gamma_G^E N^{ME} \Leftrightarrow \\ &\Leftrightarrow V = N\left[1 + \left(\gamma_E^F - 1\right) FE + \left(\gamma_G^{sE} - 1\right) MsE + \left(\gamma_E^F \gamma_G^E - 1\right) ME\right], \end{aligned} \tag{4.2}$$

where $FE = \frac{N^{FE}}{N}$; $MsE = \frac{N^{MsE}}{N}$ and $ME = \frac{N^{ME}}{N}$.

Using (4.2) we were able to investigate the hypothesis $\gamma_E^F * \gamma_G^{sE} - \gamma_E^F \gamma_G^E = 0$, in the productivity equation, with and without control for unobserved effects. In both cases the hypothesis seems to be difficult to reject.[18] Using these results we further manipulate (4.2) to yield:

$$\begin{aligned} &V = N\left[1 + \left(\gamma_E^F - 1\right) FE + \left(\gamma_G^{sE} - 1\right) MsE + \left(\gamma_E^F \gamma_G^{sE} - 1\right) ME\right] \Leftrightarrow \\ &V = N\left[1 + \left(\gamma_E^F - 1\right) FE + \left(\gamma_E^F \gamma_G^{sE} - \gamma_G^{sE}\right) ME + \left(\gamma_G^{sE} - 1\right)\left(MsE + ME\right)\right] \Leftrightarrow \\ &V = N\left[1 + \left(\gamma_E^F - 1\right) FE + \left(\gamma_E^F - 1\right) \gamma_G^{sE} ME + \left(\gamma_G^{sE} - 1\right) G\right], \end{aligned} \tag{4.3}$$

since, $MsE + ME = G$. Now, as $FE + ME = E$, equation (4.3) becomes:

---

[17] γ_E^F is the relative marginal productivity of schooling in case of female workers and γ_E^M the relative marginal productivity of schooling for men; γ_G^{sE} corresponds to the marginal rate of technical substitution between men and women for low-schooling workers and γ_G^E is the relative marginal productivity between men and women in case of high-schooling workers (i.e. $\gamma_G^E = \frac{\partial Y / \partial N^{ME}}{\partial Y / \partial N^{FE}}$).

[18] The Wald test is $F_{(1;\ 1,340)} = 0.17$ ($P_F > F_{st} = 0.6831$) and $F_{(1;\ 1,340)} = 0.34$ ($P_F > F_{st} = 0.5584$), if we ignore the unobserved effects.

$$V = N\left[1+\left(\gamma_E^F-1\right)FE+\left(\gamma_E^F-1\right)ME-\left(\gamma_E^F-1\right)ME+\left(\gamma_E^F-1\right)\gamma_G^{sE}ME+\left(\gamma_G^{sE}-1\right)G\right]\Leftrightarrow$$
$$V = N\left[1+\left(\gamma_E^F-1\right)\left(FE+ME\right)+\left(\gamma_E^F-1\right)\left(\gamma_G^{sE}-1\right)ME+\left(\gamma_G^{sE}-1\right)G\right]\Leftrightarrow$$
$$V = N\left[1+\left(\gamma_E^F-1\right)E+\left(\gamma_G^{sE}-1\right)G+\left(\gamma_E^F-1\right)\left(\gamma_G^{sE}-1\right)ME\right]. \quad (4.4)$$

This means that relaxing the initial assumptions is equivalent to add an additional term, $\left(\gamma_E^F-1\right)\left(\gamma_G^{sE}-1\right)ME$, to the productivity equation. In other words, by introducing the $\alpha\left(\gamma_E^F-1\right)\left(\gamma_G^{sE}-1\right)ME$ term in model (2.8) we are in a position to test the robustness of our initial assumption. We use the *Wald* test to check for the statistical significance of $\alpha\left(\gamma_E^F-1\right)\left(\gamma_G^{sE}-1\right)$, and obtained $F_{(1;\ 1,340)}=3.49\,(P_F>F_{st}=0.062)$. We therefore do not find any strong evidence against the null of $\left(\gamma_E^F-1\right)\left(\gamma_G^{sE}-1\right)=0$. Similar results were obtained for the wage equation.

If we extend this investigation for all the observed worker characteristics we end up with a model estimated at firm level that accommodates all possible combinations, or a total of 192 regressors. And the results from the productivity equation show that firms with a higher percentage of male workers, aged 25-44, with at least a high school degree, 10 years of service, and holding a high job occupation (*top managers and professionals*) are indeed a lot more productive than the remaining firms.

We also tested whether the assumption of the same number of hours for all types of workers is satisfactory. Given that workers in our sample are divided into different skill, schooling, tenure, and occupation groups, there is of course the possibility of substantial differences in working hours across these different worker types. A quick inspection of the data reveals indeed that workers with higher working hours have, on average, lower levels of schooling, skill, and tenure, as well as lower wages.

The analysis of the sensitivity of the results to hours amounts to specify a slightly different version of model (2.8). Thus, rather than using the proportion of workers with a specific characteristic, we consider hours worked by each type of worker in total hours worked, to yield:

$$Ln\,y_{jt} = Ln\,A+\alpha\left(\gamma_{G^H}-1\right)G_{jt}^H+\alpha(\gamma_{T^H}-1)T_{jt}^H+\alpha(\gamma_{E^H}-1)E_{jt}^H+\alpha\left(\gamma_{O^H}-1\right)O_{jt}^H+$$
$$+\alpha(\gamma_{S^H}-1)S_{jt}^H+\alpha\left(\gamma_{Q_1^H}-1\right)Q_{1,jt}^H+\alpha\left(\gamma_{Q_2^H}-1\right)Q_{2,jt}^H+\alpha\left(\gamma_{Q_3^H}-1\right)Q_{3,jt}^H+\alpha\left(\gamma_{Q_4^H}-1\right)Q_{4,jt}^H+$$
$$+\alpha\left(\gamma_{Q_5^H}-1\right)Q_{5,jt}^H+\beta Ln\,k_{jt}+\eta Z_{jt}+\pi\hat{\psi}_j+\varepsilon_{jt}, \quad (4.5)$$

where, for example, G^H corresponds to the proportion of hours worked by men in total hours worked. (Note that $G^H = G = \frac{N^M}{N}$ if male employees work exactly the same number of hours as female workers.) γ_{G^H} is the marginal productivity of one additional hour worked by a male worker relatively to one additional hour worked by women. And similarly for the remaining variables in the model. We also note that in order to carry out this exercise it is required to have the number of hours of work by each type of worker. This information is available in *Quadros de Pessoal*.[19]

Table 5 shows that the results do not seem to be too sensitive to hours. Thus, using hours rather than the number of workers, we obtain again a higher impact of schooling and training on productivity than on wages. Tenure also has a larger impact on wages than on productivity. Comparing Table 5 with Table 2, we observe a slightly larger impact of schooling, tenure and unobserved fixed effects on firm productivity. The differences between the two tables are clearly smaller in the case of the wage equation.

Finally, we relax the two assumptions mentioned at the beginning of this section but now using a model estimated at worker level. The results from a model that includes again a total of 192 regressors show that, on average, males with a higher level of schooling and longer tenure, who have participated in training and with a higher job occupation are 33% more productive than the omitted worker category, while the corresponding earnings are 31% higher.[20] Note that the modified model substantially reduces the differences between productivity and wages estimates obtained in Section 4.2 above. This result is very interesting since it indicates that, by using all possible combinations of worker attributes, the individual wage differences tend to mirror the differences in worker productivity.

5. Conclusions

Based on a unique matched employer-employee dataset, this paper firstly examines the determinants of productivity and wages at firm level. The micro foundation for the firm productivity equation is based on a Cobb-Douglas production function in which the labor input is subdivided into several types of observed worker categories. The micro foundation for the corresponding wage equation is based on a standard Mincerian individual earnings

[19] Since firms in *Balanço Social* are necessarily in the *Quadros de Pessoal* database, it is straightforward to impute hours to any single group of workers.
[20] The residual 1/0 dummy flags a female worker, untrained, with less than a high-school degree, with less than 10 years of service, unskilled, and older than 44 years.

regression with worker and firm fixed effects. We then derived firm-level productivity (i.e. output per hour) and wages (i.e. earnings per hour) equations which are assumed to be a function of the same set of regressors. Simultaneous estimation of the two equations reveals that tenure has a greater impact on wages than on productivity, while schooling, training, and skill have a greater impact on productivity. In the case of the productivity gains from training, we show that they are almost totally captured by firms.

Secondly, and in a new departure, we also estimate wage and productivity equations at individual level. We start by developing a model that allows us to obtain an estimate of the (unobserved) individual productivity, and, based on this procedure, we show that the generated firm-level productivity is highly correlated with the observed firm average. Not surprisingly, we confirm that schooling, in comparison with firm-provided training, has a greater impact on wages, with the latter implying that 83% of the productivity gains go to firms, while the former implies a worker share of 32%, a result that largely replicates those obtained from using firm level equations. The introduction of worker and firm unobserved effects into the regression also reduces the schooling and training coefficients, an indication that firm and worker unobserved heterogeneity are, as expected, positively correlated with human capital observed variables. By comparing the productivity equations at worker and firm level, we also note that the corresponding coefficients tend to be smaller in worker level estimations, an indication of the existence of spillovers across workers within firms.

In a separate analysis, we relaxed some model assumptions. In particular, based on a model that accommodates all possible combinations across observed worker attributes, we confirm in a worker-level estimation that prime adult workers – i.e. males with a higher schooling level, longer tenure, with training participation, with a higher job occupation, and with age between 25 and 44 years – are the most productive group, with a favourable productivity gap in the 33% range relatively to the omitted category. The corresponding wage gap is approximately 31% higher. This is an interesting result since it indicates that by using all possible combinations of worker attributes we obtain that the wage differences are mostly productivity based.

Appendix A

Estimation of the earnings equation at worker level – model (2.1) – by *OLS* faces two major obstacles. The first one is related to the possible correlation between observable characteristics, *X'*, and unobserved worker heterogeneity, α_i. (And indeed both the standard Hausman and the F-statistic tests comfortably reject the null of no correlation between the unobservable effects and *X'*.)[21] The second difficulty is related to proper estimation of parameters α_i and $\phi_{j(i)}$, given that both the number of workers and firms is very large.

Let us first rewrite equation (2.4) as follows:

$$\overline{Ln\ w}_{jt} = Ln\ A_w + \lambda X_{jt} + \eta_w Z_{jt} + \beta_w Ln\ k_{jt} + \overline{\alpha}_j + \phi_j + e_{jt}, \qquad (A.1)$$

with $\overline{Ln\ w}_{jt} = \frac{1}{N_{jt}} \sum Ln\ w_{it}$ and $\overline{\alpha}_j + \phi_j = \psi_j$. X_{jt}, we recall, comprises the set of average workforce characteristics in firm *j* and Z_{jt} contains firm-specific attributes (e.g. location and industry).

Subtracting equation (A.1) from (2.1), we get:

$$\begin{aligned} & Ln\ w_{it} - \overline{Ln\ w}_{j(i)t} = \\ & = \left(Ln\ A_w + \lambda X'_{it} + \eta_w Z_{j(i)t} + \beta_w Ln\ k_{j(i)t} + \alpha_i + \phi_{j(i)} + \xi_{it}\right) - \\ & - \left(Ln\ A_w + \lambda X_{j(i)t} + \eta_w Z_{j(i)t} + \beta_w Ln\ k_{j(i)t} + \overline{\alpha}_{j(i)} + \phi_{j(i)} + e_{j(i)t}\right) \Leftrightarrow \\ & \Leftrightarrow Ln\ w_{it} - \overline{Ln\ w}_{j(i)t} = \lambda (X'_{it} - X_{j(i)t}) + \theta_i + \left(\xi_{it} - e_{j(i)t}\right), \end{aligned} \qquad (A.2)$$

with $\theta_i = \alpha_i - \overline{\alpha}_j$.

It is fair to assume that the difference between the expected individual wage and the expected firm average wage, conditional on *X* and *Z*, or $E\left(Ln\ w_{it} \middle| X'_{it}, Z_{j(i)t}\right) - E\left(\overline{Ln\ w}_{j(i)t} \middle| X_{j(i)t}, Z_{j(i)t}\right)$, depends on the gap between worker's observed attributes and the mean attributes of his/her counterparts in firm *j*, $X'_{it} - X_{j(i)t}$. In this context, it follows that any unexplained wage difference is expected to be due to the difference in unobserved ability from the corresponding firm average. In other words, if the observed

[21] The F-statistic is used to test the significance of γ in the auxiliary regression given by:

$$y_{it} - \hat{\lambda}\bar{y}_i = (1 - \hat{\lambda})\mu + (x_{it} - \hat{\lambda}\bar{x}_i)\beta + (x_{it} - \bar{x}_i)\gamma + e_{it}.$$

ability of a given individual is exactly equal to the observed firm average *and* her/his wage is higher (lower) than the firm average, then her/his unobserved ability must be higher (lower) than that of the fellow co-workers.

In matrix notation, the equation (A.2) is equivalent to

$$(Lw-\overline{Lw})=(X'-X)\lambda+C\theta+(\varepsilon-\overline{\varepsilon}), \tag{A.3}$$

where C denotes a $N^S T\times N^S$ matrix of dummies flagging the worker over the sample period T; N^S is the number of workers in the sample.

Multiplying equation (A.3) by $M_C=I-P_C$, where P_C denotes the matrix that provides an orthogonal projection in C, we obtain

$$M_C(Lw-\overline{Lw})=M_C(X'-X)\lambda+M_C C\theta+M_C(\varepsilon-\overline{\varepsilon}), \tag{A.4}$$

By definition, $M_C C\theta=0$, and therefore

$$M_C(Lw-\overline{Lw})=M_C(X'-X)\lambda+M_C(\varepsilon-\overline{\varepsilon}), \tag{A.5}$$

which, by Frisch-Waugh's theorem, yields the same estimates and residuals as model (A.3). The corresponding estimator of λ can be then written as

$$\hat{\lambda}=\left[(X'-X)^T M_C(X'-X)\right]^{-1}(X'-X)^T M_C\left(Lw-\overline{Lw}\right). \tag{A.6}$$

From equation (A.3) and (A.6), we finally obtain $\hat{\theta}$ as

$$\hat{\theta}=\left(C^T C\right)^{-1}C^T\left(Lw-\overline{Lw}-(X'-X)\hat{\lambda}\right). \tag{A.7}$$

References

Addison, J., Teixeira, P. and Zwick, T., 2010. Works Councils and the Anatomy of Wages. *Industrial and Labor Relations Review* 63(2): 247–270.

Ballot, G., Fakhfakh, F. and Taymaz, E., 2006. Who Benefits from Training and R&D, the Firm or the Workers? *British Journal of Industrial Relations* 44(3): 473-495.

Dearden, L., Reed, H. and Reenen, J., 2006. The Impact of Training on Productivity and Wages: Evidence from British Panel Data. *Oxford Bulletin of Economics and Statistics* 68(4):397-421.

Hellerstein, J., Neumark, D. and Troske, K., 1999. Wages, Productivity, and Worker Characteristics: Evidence from Plant Level Production Functions and Wage Equations. *Journal of Labour Economics* 17(3): 409-446.

Hellerstein, J. and Neumark, D., 1999. Sex, Wages, and Productivity: An Empirical Analysis of Israeli Firm-Level Data. *International Economic Review* 40(1): 95-123.

Jovanovic, B. and Rob, R. 1989. The Growth and Diffusion of Knowledge. *Review of Economic Studies* 56(4): 569–582.

Lopes, A. and Teixeira, P., Productivity, Wages, and the Returns to Firm-Provided Training: Fair Shared Capitalism? *International Journal of Manpower*. (Forthcoming.)

Moretti, E. 2004. Estimating the Social Return to Higher Education: Evidence from Longitudinal and Repeated Cross-sectional data. *Journal of Econometrics* 121(1-2): 175-212.

Table 1: Summary statistics at worker and firm level, weighted data

Variables	Firm Level (1)	Worker Level (2)
(log) Productivity	2.517	2.614
	(0.794)	(0.771)
(log) Earnings	1.325	1.406
	(0.386)	(0.559)
Schooling	0.192	0.202
	(0.158)	(0.401)
Training	0.523	0.413
	(0.692)	(0.492)
Tenure	0.466	0.520
	(0.273)	(0.499)
Age	0.572	0.603
	(0.285)	(0.489)
Gender (male)	0.566	0.624
	(0.285)	(0.484)
Top managers and professionals	0.030	0.050
	(0.036)	(0.218)
Other managers and professionals	0.048	0.033
	(0.062)	(0.178)
Foremen and supervisors	0.055	0.080
	(0.049)	(0.272)
Highly skilled and skilled	0.411	0.544
	(0.256)	(0.498)
Semiskilled	0.302	0.215
	(0.268)	(0.411)
Unskilled	0.099	0.072
	(0.176)	(0.259)
Capital	0.696	
	(1.199)	
Hours per employee	1,770	
	(2,094)	
Productivity bonus	0.239	
	(0.498)	
Proportion of full-time workers	0.895	
	(0.098)	
Proportion of fixed-term	0.077	
	(0.114)	
Foreign ownership	0.323	
	(0.468)	
Medium/large firm	0.644	
	(0.479)	
Norte	0.463	
Centro	0.173	
Lisboa e Vale do Tejo	0.326	
Alentejo	0.024	
Algarve	0.002	
	(0.044)	
Number of employees	800.0	
	(1,132)	
Number of observations	1,716	454,346

Standard deviations in parenthesis.

Notes: The worker-level information is based on *Quadros de Pessoal*, while the firm-level information is extracted from *Balanço Social*. The balanced panel of manufacturing firms covers 1998 and 1999. Firm-level statistics are weighted by the number of workers in each firm so that columns (1) and (2) are comparable. The description of variables is presented in Appendix Table 1.

Table 2: Determinants of productivity and wages, firm-level estimates

Variables	Productivity		Wages	
	Without control for unobserved firm effects (1)	With control for unobserved firm effects (2)	Without control for unobserved firm effects (3)	With control for unobserved firm effects (4)
Schooling	0.640	0.318	0.569	0.217
	(5.49)	(2.79)	(13.39)	(10.84)
Training	0.099	0.060	0.046	0.004
	(4.12)	(2.60)	(5.28)	(0.98)
Tenure	0.092	0.110	0.179	0.199
	(1.60)	(2.01)	(8.59)	(19.68)
Age	0.269	0.056	0.100	-0.133
	(2.53)	(0.54)	(2.58)	(-7.38)
Gender (male)	0.365	0.260	0.265	0.150
	(5.75)	(4.25)	(11.43)	(13.96)
Top managers and professionals	1.583	1.135	0.904	0.414
	(3.58)	(2.68)	(5.60)	(5.55)
Other managers and professionals	0.649	0.234	0.808	0.353
	(2.39)	(0.89)	(8.14)	(7.69)
Foremen and supervisors	0.249	0.086	0.281	0.103
	(1.15)	(0.42)	(3.58)	(2.85)
Highly skilled and skilled	0.146	0.129	0.082	0.063
	(2.01)	(1.85)	(3.08)	(5.13)
Semiskilled	0.035	0.072	0.002	0.043
	(0.46)	(0.99)	(0.06)	(3.31)
Capital	0.262	0.225	0.032	-0.008
	(20.80)	(18.23)	(7.02)	(-3.49)
Productivity bonus	0.006	-0.001	0.004	-0.003
	(0.34)	(-0.02)	(0.63)	(-0.96)
Proportion of full-time workers	0.224	0.349	-0.233	-0.097
	(2.00)	(3.25)	(-5.71)	(-5.12)
Proportion of fixed-term	-0.073	-0.020	-0.037	0.021
	(-0.71)	(-0.20)	(-0.99)	(1.21)
Foreign ownership	0.166	0.112	0.086	0.028
	(5.53)	(3.88)	(7.88)	(5.46)
Medium/large firm	0.014	-0.012	0.033	0.005
	(0.51)	(-0.45)	(3.34)	(1.14)
Norte	-0.012	-0.061	-0.077	-0.131
	(-0.38)	(-1.96)	(-6.52)	(-23.80)
Centro	-0.099	-0.177	-0.061	-0.147
	(-2.71)	(-5.00)	(-4.60)	(-23.57)
Alentejo	-0.218	-0.240	-0.035	-0.058
	(-2.33)	(-2.68)	(-1.02)	(-3.70)
Algarve	-0.518	-0.454	-0.111	-0.042
	(-2.83)	(-2.60)	(-1.67)	(-1.36)
Unobserved firm fixed effect		0.836		1.000
		(12.64)		(78.65)
Number of observations	1,701	1,701	1,701	1,637
F	76.978	86.215	152.03	787.42
$\bar{R}^2$	0.6438	0.6751	0.7812	0.9506

t-statistics in parentheses.

Notes: Column (1) presents the estimates from model (1.9), column (2) from model (2.8), column (3) from model (2.4)' and column (4) presents the estimates from model (2.9). The description of variables is presented in column (1) of Appendix Table 1. The model also includes a constant, 27 industry dummies, and 2 dummies flagging the legal status of the firm.

Table 3: Determinants of productivity and wages, worker-level estimates

Variables	(Estimated) Productivity		Wages	
	Without control for unobserved worker and firm effects (1)	With control for unobserved worker and firm effects (2)	Without control for unobserved worker and firm effects (3)	With control for unobserved worker and firm effects (4)
Schooling	0.196	0.036	0.192	0.031
	(123.20)	(54.57)	(122.19)	(50.94)
Training	0.119	0.015	0.111	0.007
	(91.81)	(28.33)	(86.60)	(14.30)
Tenure	0.102	0.132	0.117	0.146
	(91.81)	(285.74)	(106.26)	(345.07)
Age	-0.016	-0.055	-0.021	-0.060
	(-15.24)	(-123.53)	(-20.21)	(-147.35)
Gender (male)	0.231	0.095	0.219	0.083
	(187.88)	(184.78)	(180.47)	(174.22)
Top managers and professionals	1.039	0.225	1.029	0.203
	(329.36)	(154.40)	(330.38)	(151.74)
Other managers and professionals	0.766	0.199	0.760	0.186
	(221.96)	(133.23)	(223.13)	(135.50)
Foremen and supervisors	0.502	0.149	0.498	0.139
	(191.66)	(133.10)	(192.42)	(135.38)
Highly skilled and skilled	0.213	0.080	0.211	0.077
	(106.39)	(95.91)	(106.88)	(100.90)
Semiskilled	0.051	0.025	0.050	0.023
	(23.57)	(27.87)	(23.08)	(27.63)
Capital	0.313	0.254	0.068	0.013
	(523.71)	(961.05)	(115.38)	(51.59)
Productivity bonus	-0.001	0.001	-0.001	0.001
	(-0.73)	(12.17)	(-0.86)	(13.64)
Proportion of full-time workers	0.233	0.262	0.152	0.185
	(42.10)	(114.92)	(27.88)	(87.99)
Proportion of fixed-term	-0.001	-0.013	0.007	-0.008
	(-0.28)	(-8.25)	(1.86)	(-5.52)
Foreign ownership	0.221	0.110	0.127	0.019
	(169.76)	(197.26)	(98.50)	(37.60)
Medium/large firm	-0.025	-0.043	-0.009	-0.028
	(-20.28)	(-85.80)	(-7.40)	(-59.24)
Norte	-0.004	-0.096	-0.074	-0.165
	(-2.83)	(-176.58)	(-57.43)	(-330.96)
Centro	-0.103	-0.201	-0.099	-0.197
	(-59.69)	(-281.72)	(-58.59)	(-299.99)
Alentejo	-0.264	-0.264	-0.074	-0.072
	(-59.90)	(-144.99)	(-16.53)	(-42.99)
Algarve	-0.536	-0.465	-0.083	-0.015
	(-35.27)	(-74.11)	(-5.54)	(-2.59)
Unobserved worker fixed effect		0.945		0.961
		(1,188.07)		(1,314.87)
Unobserved firm fixed effect		1.048		1.015
		(909.22)		(957.57)
Number of observations	408,723	408,723	408,723	408,723
F	51,453.41	336,757.60	23,270.83	208,624.20
$\bar{R}^2$	0.8271	0.9706	0.6839	0.9533

t-statistics in parentheses.

Notes: Column (1) and column (2) present the estimates from model (3.1) (without and with unobserved effects, respectively), and columns (3) and (4) reproduces model (2.2), but in column (3) we do not control for the unobserved worker and firm fixed effects. The description of variables is presented in column (2) of the Appendix Table 1. See notes to Table 2.

Table 4: Worker share from human capital investment

Variables	Firm-level estimates (1)	Worker-level estimates (2)
Schooling	25.2%	31.7%
Training	2.5%	17.2%
Tenure	66.9%	41.1%
Top managers and professionals	13.5%	33.4%
Highly skilled or skilled personnel	18.1%	35.8%

Note: The corresponding shares are given by $\frac{\lambda_R}{\alpha\left(\gamma_R - 1\right)} * \frac{w}{y}$, $\forall R = E, T, S, Q_1, Q_4$.

Table 5: Determinants of productivity and wages at firm level and with control for hours

Variables	Productivity (1)	Wages (2)
Schooling	0.355	0.218
	(3.73)	(12.98)
Training	0.047	0.001
	(2.12)	(0.26)
Tenure	0.167	0.192
	(3.48)	(22.71)
Age	-0.058	-0.211
	(-0.68)	(-13.92)
Gender (male)	0.268	0.159
	(4.76)	(16.03)
Top managers and professionals	0.331	0.288
	(1.21)	(5.96)
Other managers and professionals	-0.417	0.253
	(-1.27)	(4.35)
Foremen and supervisors	0.376	0.001
	(2.02)	(0.00)
Highly skilled and skilled	0.118	0.132
	(1.08)	(6.83)
Semiskilled	0.062	0.064
	(0.54)	(3.12)
Capital	0.226	-0.009
	(18.70)	(-4.07)
Productivity bonus	-0.011	-0.003
	(-0.67)	(-1.23)
Proportion of full-time workers	0.482	0.031
	(4.74)	(1.75)
Proportion of fixed-term	0.025	-0.006
	(0.26)	(-0.33)
Foreign ownership	0.097	0.020
	(3.63)	(4.23)
Medium/large firm	-0.020	0.001
	(-0.84)	(0.22)
Norte	-0.111	-0.136
	(-3.69)	(-25.74)
Centro	-0.228	-0.163
	(-7.02)	(-28.48)
Alentejo	-0.241	-0.082
	(-3.14)	(-6.09)
Algarve	0.174	-0.064
	(0.86)	(-1.78)
Unobserved firm fixed effect	0.943	
	(14.37)	
Number of observations	1,554	1,554
F	101.62	1,024.04
$\bar{R}^2$	0.7182	0.9625

t-statistics in parentheses.

Notes: Column (1) presents the estimates from model (4.5) and column (2) uses the same specification to estimate wages. See notes to Table 2.

Appendix Table 1: Description of the selected variables

Variable	Firm level	Worker level
(log) Productivity $(Ln\ y)$	Log ratio of annual gross value added to hours worked.	Estimated
(log) Earnings $(Ln\ w)$	Log of monthly earnings divided by hours of work.	Log of total monthly earnings divided by hours of work.
Schooling (E)	Proportion of workers with at least a high-school degree.	Dummy: 1 if the worker has at least a high-school degree; 0 otherwise.
Training (T)	Proportion of workers who have participated in firm provided training.	Dummy: 1 if the worker has participated in firm provided training; 0 otherwise. This variable has been imputed using a procedure available on request.
Tenure (S)	Proportion of workers with 10 or more years of service.	Dummy: 1 if the worker has 10 or more years of service; 0 otherwise.
Age (O)	Proportion of workers between 25 and 44 years old.	Dummy: 1 if the worker has more than 25 years old and less than 44 years old; 0 otherwise.
Gender (male) (G)	Proportion of male workers.	Dummy: 1 if the worker is male; 0 otherwise.
Top managers and professionals (Q_1)	Proportion of top managers and professionals.	Dummy: 1 if the worker is top manager or professional; 0 otherwise.
Other managers and professionals (Q_2)	Proportion of other managers and professionals.	Dummy: 1 if the worker is other manager or professional; 0 otherwise.
Foremen and supervisors (Q_3)	Proportion of foremen and supervisors.	Dummy: 1 if the worker is foreman or supervisor; 0 otherwise.
Highly skilled and skilled (Q_4)	Proportion of highly skilled and skilled personnel.	Dummy: 1 if the worker is highly skilled or skilled; 0 otherwise.
Semiskilled (Q_5)	Proportion of semiskilled personnel.	Dummy: 1 if the worker is semiskilled; 0 otherwise.
Unskilled (Q_6)	Proportion of unskilled personnel.	Dummy: 1 if the worker is unskilled; 0 otherwise.
Capital $(Log\ k)$	(Log) Capital stock per hour of work. The stock of capital is proxied by the annual volume of capital depreciation.	
Productivity bonus	Ratio of non-standard compensation to *basic* earnings.	
Proportion of full-time workers	Proportion of full-time workers.	
Proportion of fixed-term	Proportion of fixed-term contract workers.	
Foreign ownership	Dummy: 1 if the firm is owned partial or totally by foreigners; 0 otherwise.	
Medium/large firm	Dummy: 1 if the number of employees is more than 250; 0 otherwise.	
Norte/Centro/Lisboa e Vale do Tejo/Alentejo/Algarve	Dummy: 1 if the firm is located in Norte/Centro/Lisboa e Vale do Tejo/Alentejo/Algarve; 0 otherwise.	
Unobserved firm fixed effects	It is given by $\hat{\psi}_j$ in model (2.7).	
Unobserved worker fixed effects	Corresponds to $\hat{\theta}_i$ in model (A.7)	

Note: The variables at firm (worker) level are extracted from *Balanço Social* (*Quadros de Pessoal*).