Removing the 'Veil of Ignorance': Nonlinearities in Education Effects on Gender Wage Inequalities

Supporting Information Appendix

Sandeep Mohapatra* Bruno Wichmann* Philippe Marcoul*

*Department of Resource Economics and Environmental Sociology, University of Alberta.

A1 - Control variables for the wage earnings regressions

We differentiate between workers by occupation and industry groups, which is important when estimating human capital earnings functions (Blau and Kahn, 1994). We use the National Classification of Occupations (NCO-1960) at the two-digit level to divide workers into four different skill groups based on their occupation. The occupations are: high skilled (i.e., professional, technical, executive and managerial jobs which employ scientists, managers and computer programmers), medium skilled (i.e., sales and clerical jobs which employ merchants, shopkeepers and custom inspectors), low skilled (i.e., production and transport equipment operators and other mechanical jobs), and unskilled (the base category includes hired farm workers, and traditional service workers such as security guards and garbage collectors). Table 1 shows that 41% of the employed women are engaged in a high skill occupation. In contrast, 38% of employed men are engaged in a low skill occupation. Interestingly perhaps, Table 1 also reveals that more than 39% of women have a Graduate/University education against 28.5% for men.

We use the National Industrial Classification (NIC-2004) at the two-digit level to divide workers into seven categories based on their industry of employment. These are: construction and mining, low-tech manufacturing (e.g. textiles, food and beverages), high tech manufacturing (e.g. automobile and electrical equipment), modern services (e.g. finance, real estate, insurance, utilities, public administration, education, and medical), computers and IT (e.g. software maintenance, web design), and agriculture and traditional services (base category). Table 1 shows that the largest industry is the industry of modern services, which employs 62% of women and 48% of men.

Our analyses control for religious and caste classifications to account for the potential influence of social stratifications on wages. For religions, we use an indicator variable for Muslim. As shown in table 1, 13% of the male population and 8% of the female population are Muslim. In India, castes prevent social mobility because they represent strong social stratifications that can restrict access of specific groups to specific occupations. We control for differences in social stratification of individuals based on caste using three categories: scheduled caste or scheduled tribe (SCST), other backward castes (OBC), and all others (i.e. forward castes, used as the base case). SCSTs are at the bottom of the caste system in India. However, there is also concern that the social and economic standing of OBCs lags behind that of the forward castes.

In addition to the above control variables, we use a binary indicator to distinguish between urban and rural locations. Approximately 73% of our sample are located in urban areas. We also differentiate observations spatially to capture different development levels in Indian various regions. This is important because development may create heterogeneity in the characteristics of jobs and wages and may represent a confounding factor. To capture broad patterns of regional effects, we include a set of region dummies denoting six regions of India: north, south, east, west, north-east and central (base category).

Table 1: Summary Statistics

	Men		**Women**	
	Mean	Std. Dev.	Mean	Std. Dev.
Earnings:				
Ln Wage	4.907	0.788	4.569	0.959
Human Capital:				
Primary Education	0.104	0.306	0.079	0.269
Middle Education	0.198	0.399	0.105	0.306
Secondary Education	0.188	0.390	0.119	0.323
Higher Secondary Education	0.122	0.327	0.103	0.304
Graduate/University Education	0.285	0.451	0.391	0.488
Technical Degree	0.025	0.155	0.041	0.197
Years of Education	9.544	4.688	9.348	5.781
Experience	21.41	11.92	20.80	12.86
Occupation:				
High Skilled	0.193	0.394	0.414	0.493
Medium Skilled	0.291	0.454	0.201	0.401
Low Skilled	0.377	0.485	0.132	0.339
Industry:				
Construction	0.040	0.195	0.014	0.117
Low Tech Mfg.	0.180	0.385	0.119	0.324
High Tech Mfg.	0.079	0.269	0.016	0.124
Modern Services	0.482	0.500	0.624	0.484
Computer & IT	0.008	0.088	0.012	0.108
Social/Religion:				
Scheduled Caste/Tribe	0.219	0.413	0.271	0.444
Other Backward Classes	0.314	0.464	0.303	0.459
Muslim	0.128	0.334	0.082	0.274
Region:				
Urban	0.732	0.443	0.724	0.447
North	0.168	0.374	0.130	0.336
South	0.167	0.373	0.266	0.442
East	0.099	0.299	0.098	0.298
West	0.190	0.392	0.154	0.361
North-East	0.097	0.296	0.131	0.338

Notes: All variables are binary indicators except for Ln Wage (weekly Ruppes), Years of Education, and Experience.

A2 - Earnings functions with cubic splines

Expected earnings for individual i are determined by

$$E(W_i|X_{ij}, Z_{il}) = \alpha_0 + \sum_{j=1}^{3} f_j(X_{ij}) + \sum_{l=1}^{L} \beta_l Z_{il} \tag{1}$$

where W_i denotes log wages, X_{ij} includes years of education, years of experience, and their interaction, Z_{il} represents the set of L control variables discussed in appendix A1, and α_0 is an intercept term. The functions f_j are smooth functions that respectively link the measures of education and experience to the expected value of wages. They are non-linear counterparts of coefficients in a linear regression.

The functions, f_j, are estimated from the data using cubic regression splines. Our approach selects the regression spline model that best predicts the outcome variable from the explanatory variables. A cubic spline was selected. Splines are piecewise polynomial functions that fit within each intervals of the values of a variable. The intervals are separated by knots at which the splines are joined together to fit a smooth function. We define a cubic spline for each f_j with $n+2$ knots (λ_h, $\lambda_{\min}$, $\lambda_{\max}$) as:

$$\begin{aligned} f_j(X_{ij}) = {} & \varphi_1 X_{ij} + \varphi_2 X_{ij}^2 + \varphi_3 X_{ij}^3 + \varphi_{\lambda_{\min}} \cdot g\left((X_{ij} - \lambda_{\min})^3\right) \\ & + \sum_{h=1}^{n} \varphi_h \cdot g\left((X_{ij} - \lambda_h)^3\right) + \varphi_{\lambda_{\max}} \cdot g\left((X_{ij} - \lambda_{\max})^3\right), \end{aligned} \tag{2}$$

where $g(.)$ is such that

$$g\left((X_{ij} - \lambda_h)^3\right) = \begin{cases} (X_{ij} - \lambda_h)^3 \text{ if } X_{ij} > \lambda_h \\ 0 \text{ otherwise.} \end{cases}$$

Note that the parameters φ and λ are estimated for each f_j function. Parameters in equations (1) and (2) are estimated using an iterative procedure which considers alternative models, each with different functional forms on the non-linear effects and different sets of explanatory variables. The procedure iterates over each variable in X_{ij} and Z_{il} changing the model using statistical significance tests until there is no change in the form of the spline functions and the variables included in the model (Royston and Sauerbrei, 2004, 2008). We include in Z_{il} the full set of control variables including the type of wage employment, the skill level of the employment, social and religious affiliations, and geographical location described in the previous section.

A3 - On the convexity of returns to education

According to the theory of human capital each successive year of schooling yields a smaller wage increase (due to diminishing marginal returns to education accumulation). Thus, a concave relationship between wages or earnings and education is often expected. However, recently, many studies have found that returns to education are highest at the upper-end of the education distribution in developing countries (e.g., Kavuma et al., 2015; Kingdon and Söderbom, 2008; Söderbom et al., 2006; Rankin et al., 2010). Some scholars rationalize the convex empirical returns in developing countries as the result of the supply of individuals with low education increasing more than their demand; and reversely the supply of individuals with high education growing slower than their demand (see Fasih et al., 2012).

Convexity of the returns to schooling is in line with the existence of short term frictions at the supply level. More precisely, assume a slightly different theoretical framework in which demand for qualified workers can fluctuate in a non-anticipated way by workers and where schooling decisions do imply a time lag between the schooling decision and the date at which schooling returns can be realized. When this simple friction is allowed, schooling decision will not respond to a sudden hike in the demand for skilled workers, and wage levels will adjust upward to accommodate the little current supply of these workers leaving this specific labor market in a temporary state of high wage premium. Similarly, when demand for low skilled workers decreases unexpectedly, the latter cannot change quickly enough their schooling choices and supply remains strong; low skill wages have to adjust downward. Note that the small demand for low skills may also be exacerbated when primary schools are of poor quality. Several authors have indeed argued that the quality of primary education is a concern in India (e.g. Agrawal, 2012).

In developed countries, Lemieux (2006) empirically argues that estimates of the schooling-wage relationship, consistent with Mincer (1997) and Deschenes (2001), show that log wages are clearly a convex function of years of schooling. For Mincer (1997), convexity comes from an increase in the relative demand for skilled labor in a human capital investment model with heterogeneous workers, like in Becker (1975). The model presents an hedonic equilibrium where marginal returns to schooling can either increase or decrease with years of schooling. An abrupt increase in relative demand of schooling increases the marginal return to schooling for more educated workers relative to less-educated worker. Katz and Murphy (1992) find that the post 1980 period predominantly exhibits excess demand. For Lemieux, this may have resulted in both an increase in the returns to education and in the convexity

of the schooling-wage relationship.

The effects that we describe in the top panel of figure 1 are also consistent with the recent economic evolution of India. The country has witnessed unprecedented changes in its economy notably with a wider opening of its markets to the rest of the world. The emergence of a service economy is associated with a strong increase in the demand for highly qualified workers at the higher secondary and post-secondary levels (Agrawal, 2012). At the same time, the intense mechanization of several sectors of the Indian economy has decreased the demand for low skill jobs, especially in casual employment (Fasih et al., 2012). This phenomenon is not unique to India and several empirical studies in other developing countries have similarly found that returns to education are highest at the upper-end of the education distribution (e.g., Kavuma et al., 2015; Kingdon and Söderbom, 2008).

A4 - Limitations of quantile regressions in informing gender gap policies

Unconditional quantile regressions (UQRs) can be used to estimate the marginal effects of education and experience at different points of the actual distribution of wages. In contrast, quantile regressions allow for the estimation of marginal effects along an wage distribution that is conditional on education and experience levels. Note that these conditional quantile estimates cannot be interpreted the same way marginal effects estimates at the mean are. This is true because estimates based on the conditional distribution are very different from estimates along the actual (or unconditional) wage distribution.

That is, quantile regressions capture heterogeneity in the impact of education along the error terms of wage earnings function and not on expected earnings as in the linear regression model. Consequently, the standard Oaxaca (1973) and Blinder (1973) method for decomposing the gender wage gap which applies to earnings functions estimated from linear regressions, does not apply to earnings functions estimated using quantile regressions.

Nevertheless, Quantile regressions have been increasingly been used to describe distributional effects on wages (e.g. Bollinger et al., 2011; Dahl et al., 2013). This approach, however, cannot be used to decompose the gender wage gap at different quantiles by replacing the OLS coefficients in equation (3) with quantile regression coefficients. This is true because the law of iterated expectations (LIE) does not apply to quantiles as it does to the mean –

that is, conditional quantiles do not equal unconditional quantiles in expectation.

For instance, let W_g denote wages earned by an individual of gender $g = m, f$. The LIE allows the conditional mean functions

$$E(W_g|x_{gk}) = \sum_{k=1}^{K} \beta_k^g x_{gk}$$

to be written as unconditional expectations in the decomposition equation

$$E(W_m) - E(W_f) = \underbrace{(\overline{W}_m - \overline{W}_f)}_{\substack{\text{Mean} \\ \text{Gender Wage Gap}}} = \underbrace{\sum_{k=1}^{K} \beta_k^m (\overline{x}_{mk} - \overline{x}_{fk})}_{\text{Endowment Effect}} + \underbrace{\sum_{k=1}^{K} \overline{x}_{fk} (\beta_k^m - \beta_k^f)}_{\text{Return Effect}},$$

Specifically, taking the expectation of the conditional mean function over the distribution of x yields the unconditional mean: $E_x(E(W_g|x_{gk})) = E(W_g) = E\left(\sum \beta_k^g x_{gk}\right)$. Consequently, the regression coefficient β^g is both the marginal effect of an explanatory variable on the conditional as well as the unconditional means of wages. However, in the case of quantiles, the unconditional distributional differences between men's women's wages cannot be easily broken down into endowment and return effects using standard quantile regressions.

A5 - Multinomial selection in Indian labor markets

Our goal is to estimate human capital earnings functions accounting for selection. Our model specifies log daily wages earned by an individual of gender $g = m, f$ in the regular wage employment sector, R, as

$$W_{gR} = X_g \alpha_g + e_{gR} \tag{3}$$

A standard approach is the Heckman's model. The household's decision to participate in the regular wage sector is treated as binary and specified as a function of observed covariates and a stochastic error. Selection is accounted for by controlling the correlation between unobservables that affect participation and wage. This would account, for instance, for the decisions of people with higher than average productivities who may be more likely to participate in the regular wage market and, as a result of their higher productivity, also earn higher than average wages resulting in inconsistent parameter estimates (see Borjas, 1987).

However, this approach is overly restrictive since it ignores alternative sources and layers of selectivity.

We address the concerns with multiple layers of selectivity separately for men and women by dividing the population into 3 sectors: i. regular wage employment (R); ii. self-employed and casual wage employed (S); and iii; out of the labor force and unemployed (O).[1] We follow a multinomial selection approach (Lee, 1983) that specifies an individuals' discrete choice of labor market sector j based on a random utility model $U_{gj} = z_g\beta_g + \mu_{gj}$, $j = R, S, O$. Sector choice is based on the maximum utility she/he derives from the j alternative classes. Wages are only observed in our analysis when the regular wage employment (R) is chosen which happens when the choice R provides the maximum utility among the j alternatives. Specifying the choice model as a multinomial logit, selection due to each labor market alternative in equation (3) can be accounted for by augmenting equation (3) with a series of conditional expectations as additional explanatory variables.

These selection correction terms represent expectations of the error term in equation (3), e_{gR}, conditional on each alternative being chosen, $E_j(e_{gR}|\mu_{gj})$. They are computed using a variant of the Dubin and McFadden (1984) model following the approach of Bourguignon et al. (2007). To compute the expectations, an individual's labor market sector choice probabilities, λ_j, conditional on the explanatory variables z , are first estimated under the assumption that the e_{gR} (equation 3) are i.i.d Gumbel distributed (McFadden, 1973). The conditional means of e_{gR} are then calculated and used as selection variables in equation (3). Integrals in the conditional expectations, $m(\lambda_j)$, are computed using numerical quadrature since they have no closed form analytical expressions. The selectivity corrected wage earning function is thus:

$$W_{gR} = X_g\alpha_g + \sigma\Big(\rho_R\big[m(\lambda_R)\big] + \rho_S\big[m(\lambda_S)\frac{\lambda_S}{\lambda_S - 1}\big] + \rho_O\big[m(\lambda_O)\frac{\lambda_O}{\lambda_O - 1}\big]\Big) + \varepsilon_{gR} \qquad (4)$$

where the terms in the square brackets are the selection correction terms. Selection effects related to each sector j are captured by coefficients, $\sigma\rho_j$ where σ is the standard deviation of

[1] An alternative is to consider a more disaggregated classification a). unemployed; b) out of the labor force; c) informal or casual wage jobs; d) self-employed and e) regular (or formal) wage jobs. The disaggregation can be important for identifying different types of selectivity since there are significant differences between subgroups of workers engaged outside of the regular wage sector, for instance. Self-employment in India is shown to be a more productive and dynamic sector on average than casual wage work which involves mostly manual labor with rudimentary production modes (Banerjee and Duflo, 2008). Consequently, the nature of selection into and out of the different nonwage subsectors can be different. However, keeping with the scope of our analysis and following the comment above we focus on selection between the three aggregate sectors (R,S,0) and provide the results of more disaggregated selectivity models upon request.

the errors in equation (3). The model is identified by excluding at least one of the variables in X from Z, and by the nonlinearity of functional forms used.

Given the above selection model, our approach can be described in two steps.

Step 1: This step simply allows us to see if selection is present in the male and female regular wage equations. To this end, we estimate equation (4). We identify the selection equation using a dummy variable for landownership, marital status, number of children in the household, number of adults in the household, whether the individual is the head of the household, and whether they are recipients of ration cards that are based on pre-determined poverty status. We expect these variables to affect participation in the j-th labor market but not wages. Given the two step selection approach (of computing the conditional expectations and then using them as selection correction variables in equation 4) we correct the standard errors by bootstrapping equation (4).

Step 2: This step allows us to re-examine the decomposition exercises of the previous section. Specifically, it allows us to evaluate the robustness of the human capital variables in explaining the gender wage gap once we control for selection bias by augmenting the decomposition models with the conditional expectations described earlier. We treat the selection variables for men and women as attributes of men and women and introduce them into the Unconditional Quantile decompositions based on RIF regressions as additional explanatory variables. Again we bootstrap the standard errors of the decomposition to account for the two-step estimation.

References

Agrawal, T. (2012). Returns to education in India: Some recent evidence. Technical report, Working Paper n. WP-2011-017. Indira Gandhi Research Institute for Development.

Banerjee, A. V. and E. Duflo (2008). What is middle class about the middle classes around the world? *Journal of Economic Perspectives 22*(2), 3–41A.

Becker, G. (1975). *Human capital*. Chicago: University of Chicago Press.

Blau, F. D. and L. M. Kahn (1994). Rising wage inequality and the us gender gap. *American Economic Review 84*(2), 23–28.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human Resources 8*(4), 436–455.

Bollinger, C., J. P. Ziliak, and K. R. Troske (2011). Down from the mountain: Skill upgrading and wages in Appalachia. *Journal of Labor Economics 29*(4), 819–857.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *American Economic Review*, 531–553.

Bourguignon, F., M. Fournier, and M. Gurgand (2007). Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons. *Journal of Economic Surveys 21*(1), 174–205.

Dahl, C. M., D. Le Maire, and J. R. Munch (2013). Wage dispersion and decentralization of wage bargaining. *Journal of Labor Economics 31*(3), 501–533.

Deschenes, O. (2001). Unobserved ability, comparative advantage and the rising return to education in the United States: A cohort-based approach. Technical report, Princeton University Industrial Relations Section Working Paper No. 465.

Dubin, J. A. and D. L. McFadden (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 345–362.

Fasih, T., G. Kingdon, H. A. Patrinos, C. Sakellariou, and M. Soderbom (2012). Heterogeneous returns to education in the labor market. Technical report, Policy Research Working Paper 6170. World Bank.

Katz, L. F. and K. M. Murphy (1992). Changes in relative wages, 1963–1987: supply and demand factors. *Quarterly Journal of Economics 107*(1), 35–78.

Kavuma, S. N., O. Morrissey, and R. Upward (2015). Private returns to education for wage-employees and the self-employed in Uganda. *WIDER Working Paper.*

Kingdon, G. and M. Söderbom (2008). Education, skills, and labor market outcomes: Evidence from Ghana. Technical report, Education Working Paper Series. Numer 12. World Bank.

Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica*, 507–512.

Lemieux, T. (2006). The mincer equation thirty years after schooling, experience, and earnings. In *Jacob Mincer – A Pioneer of Modern Labor Economics*, pp. 127–145. Springer.

McFadden, D. (1973). *Conditional logit analysis of qualitative choice behavior.* Frontiers in Econometrics. New York: Academic Press.

Mincer, J. (1997). Changes in wage inequality, 1970 – 1990. *Research in Labor Economics 16*, 1–18.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review 14*(3), 693–709.

Rankin, N., J. Sandefur, and F. Teal (2010). Learning and earnings in Africa: Where are the returns to education high. Technical report, Oxford: Centre for the Study of African Economies.

Royston, P. and W. Sauerbrei (2004). A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine 23*(16), 2509–2525.

Royston, P. and W. Sauerbrei (2008). *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*, Volume 777. John Wiley & Sons.

Söderbom, M., F. Teal, A. Wambugu, and G. Kahyarara (2006). The dynamics of returns to education in Kenyan and Tanzanian manufacturing. *Oxford Bulletin of Economics and Statistics 68*(3), 261–288.