METHODOLOGICAL APPENDIX

Implications of the changing nature of work for employment and inequality in Ghana

Carlos Gradín and Simone Schotte*

September 2020

WIDER Working Paper 2020/119

1 Testing for polarization

1.1 Simple test for job and wage polarization

As a simple test for job polarization, we fit a quadratic regression—at the 3-digit occupational level—of the log change in employment share on initial log mean weekly earnings (and its square), testing the significance of the parameters (Goos and Manning 2007; Sebastián 2018a). Specifically, we estimate the following model:

$$\Delta \log(E_{j,t}) = \beta_0 + \beta_1 \log(y_{j,t-1}) + \beta_2 \log(y_{j,t-1})^2, \quad (2)$$

where $\Delta \log(E_{j,t})$ is the change in the log employment share of occupation j between survey wave $t-1$ and t, $\log(y_{j,t-1})$ is the logarithm of the mean labour earnings in occupation j in survey wave $t-1$, and $\log(y_{j,t-1})^2$ is the square of initial log mean labour earnings.

We repeat the same exercise with log change in earnings as the dependent variable (Sebastián 2018b), estimating the following model:

$$\Delta \log(y_{j,t}) = \varphi_0 + \varphi_1 \log(y_{j,t-1}) + \varphi_2 \log(y_{j,t-1})^2, \quad (3)$$

where $\Delta \log(y_{j,t})$ is the change in log mean labour earnings in occupation j between survey wave $t-1$ and t.

Both equations are estimated by weighting each occupation j by its initial employment share to avoid that results are biased by compositional changes in small occupation groups.

1.2 Regression of changes in employment and earnings on the level of routine intensity

* United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki, Finland; corresponding author: schotte@wider.unu.edu

In a next step, we fit a quadratic regression—at the 3-digit occupational level—of the log change in employment share on the initial level of routine intensity (Sebastián 2018a). Again, we repeat the same exercise with log change in earnings as the dependent variable, estimating the following two equations, where RTI_j measures the (time-invariant) routine-task intensity of occupation j:

$$\Delta \log(E_{j,t}) = \delta_0 + \delta_1(RTI_j) + \delta_2(RTI_j)^2, \quad (4)$$

$$\Delta \log(y_{j,t}) = \vartheta_0 + \vartheta_1(RTI_j) + \vartheta_2(RTI_j)^2. \quad (5)$$

2 The role of occupations in inequality

2.1 Shapley decomposition of changes in inequality

One way of exploring the role played by occupations (and the tasks performed by workers) in explaining inequality trends is to quantify the extent to which earnings differentials can be attributed to differences in earnings among workers performing different tasks, i.e. who are employed in different occupations (between-occupation inequality), as opposed to differences in earnings among workers performing similar tasks, i.e. who are employed in the same occupation (within-occupation inequality) but differ in other personal or job characteristics, such as skills, experience, geographic location, or formality status.

To investigate these effects, we follow the approach of decomposing overall inequality into between-group and within-group inequality, when groups are defined according to the occupation in which a worker is employed. The mean log deviation is the only inequality measure that can be expressed as the sum of inequality between groups (inequality remaining when all workers of each group are given the average earnings of their occupation) and inequality within groups (inequality remaining once all individual earnings are re-scaled so to remove inequality between occupations).[1] The decomposition of the Gini index, however, does not allow for an exact partition into between-group and within-group components, unless the income ranges of the defined groups are non-overlapping. In the classical decomposition (Bhattacharya and Mahalanobis 1967; Pyatt 1976), there is thus a third residual term that relates to both between-group and within-group inequality. Furthermore, even if the subgroup income ranges do not overlap, the within-group term weighs each group based on the product of the share of workers in the occupation and the share of

[1] Among the indices verifying the usual inequality properties, only the Generalized Family of Entropy indices can be decomposed into the sum of inequality between groups and a weighted sum of group inequality (with weights depending on population and income shares), property known as additive decomposability (see Shorrocks 1984). Among the members of this family, only the Mean Log Deviation and the Theil index verify that the weights of the within-group term add up to one, and therefore the sum of weights is independent of the distribution. Only in the case of the Mean Log Deviation are the weights the population shares and then are independent of between-group inequalities (making the index path independent, with inequality between groups being the inequality remaining when all workers of each group are given the average earnings of their occupation, and inequality within groups being the inequality remaining once all earnings are re-scaled so to remove inequality within each occupation). See, for instance, the formal discussion in Chakravarty (2009).

* United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki, Finland; corresponding author: schotte@wider.unu.edu

earnings held by workers in that occupation, and is thus not independent of between-group inequalities.

Formally, the decomposition can be expressed as follows: Let y denote individual earnings and $G(y)$ denote the Gini index. Further, let $y_b = (m^1, \dots, m^J)$ be a vector in which the earnings of all workers in occupation $j = 1 \dots J$ are replaced by the average earnings in that occupation, m^j. That is, within-occupation inequality is removed, and only between-occupation inequality remains. Let $y_w = (y^1 \frac{m}{m^1} \dots y^J \frac{m}{m^J})$ be a vector in which the earnings of all workers are re-scaled so that all occupations have the same average earnings m. That is, between-occupation inequality is removed, and only within-occupation inequality remains. The Gini estimated on the first vector, $G(y_b)$, has been widely used as a measure of between-group inequality. However, for the reasons discussed above, this term might be understating the actual contribution of inequalities between groups to total inequality, and the residual term, $G(y) - G(y_b)$, cannot immediately be interpreted as the contribution of within-group inequality. In fact, one could follow an alternative approach and directly estimate the contribution of inequality within groups as $G(y_w)$, and use the residual term, $G(y) - G(y_w)$, to approximate the between-group contribution. That is, there are two plausible distinct estimates for the between-group contribution, $G(y_b)$ and $G(y) - G(y_w)$, and two for the within-group, $G(y_w)$ and $G(y) - G(y_b)$, which means that the contributions are path dependent. As mentioned above, only the Mean Log Deviation, M, verifies that $M(y) = M(y_b) + M(y_w)$, with therefore the contributions being independent of the path followed to estimate the contributions:

$$M(y_w) = M(y) - M(y_b) \text{ and } M(y_b) = M(y) - M(y_w)., \quad (1)$$

In other cases, like the Gini index, to obtain a path independent decomposition of inequality, one can follow the Shapley decomposition, in which the contribution of each term is estimated as the average of the two possible contributions, so that the between-group G_B and within-group G_W contributions add up to total inequality:

$$G = G_B + G_W;$$
$$\text{with } G_B = \frac{1}{2}[G(y_b) + G - G(y_w)] \; and \; G_W = \frac{1}{2}[G(y_w) + G - G(y_b)], \quad (2)$$

Note that in general G_B will be lower than $G(y_b)$, but will likely better reflect the between-group nature of inequality.

As a result, changes in inequality over time can also be decomposed into the sum of the contribution of each term:

$$\Delta G = \Delta G_B + \Delta G_W, \quad (3)$$

2.2 Shapley decomposition of changes in between-group inequality: the contribution of workers shares and average earnings

The contribution of inequalities between occupations to explain the trend in overall inequality ΔG_B may come from two different channels. First, changes in the structure of employment can affect inequality trends. For example, if middle-income occupations decrease in size and low- and high-income groups expand, while the earnings differences between occupations remain stable, overall inequality will rise. Second, changes in the earnings gap between occupations may also impact the overall distribution of earnings. If, for example, incomes grow faster in high-paying occupations than in low-paying occupations, while the structure of employment remains unchanged, this will

result in an increase in overall earnings inequality. To disentangle whether changes in employment structure or changes in average earnings are driving the trend in inequality between occupations, we repeat the analysis with counterfactual distributions in which either the occupational shares or the occupational mean earnings are kept constant.

Let $\Delta G_{bm} = f(m_0, m_1; e_0)$ be the change in ΔG_B in the counterfactual situation in which occupational employment shares e_t are held constant over two periods t=0 and t=1. Since only the mean earnings across occupations m_t are allowed to change, one can say that the only between-occupation inequality change was driven by changes in average earnings. Similarly, one can define $\Delta G_{be} = f(m_0; e_0, e_1)$ as the corresponding change in ΔG_B when the mean earnings by occupation are held constant, instead, and only employment shares change, explaining, therefore, the change in the contribution of between-occupation inequality.

Similarly to what was done before, one can then define the Shapley contribution of these two elements (employment shares and mean earnings) to the change in inequality between occupations, respectively ΔG_{BE} ad ΔG_{BM}, as given by:

$$\begin{aligned} \Delta G_B &= \Delta G_{BE} + \Delta G_{BM} \\ \Delta G_{BE} &= \frac{1}{2}[\Delta G_{be} + \Delta G_b - \Delta G_{bm}] \\ \Delta G_{BM} &= \frac{1}{2}[\Delta G_{bm} + \Delta G_b - \Delta G_{be}]. \end{aligned} \tag{4}$$

Where a higher contribution of ΔG_{BE} indicates that inequality between occupations is changing due to a composition effect (changes in employment shares) while a higher contribution of ΔG_{BM} indicates that changes in inequality between occupations is driven by changes in the structure of average earnings.

2.3 RTI and inequality, the concentration index

The previous decompositions help to identify if a significant share of inequality or its trend is explained by differences in earnings across occupations. Inequality between occupations is the result of their specific characteristics, such as the skills required for the job, for instance, or the nature of the tasks performed by workers, the main interest of this project. To further explore the relevance of the task composition of occupations in explaining trends in inequality, we calculate, in addition, the routine-task intensity (RTI) concentration index for the distribution of average earnings by occupations. Here occupations are sorted by their RTI, while for the conventional measure of inequality between occupations, $G(y_b)$, occupations are sorted by their average earnings.

The Gini index is, in general, geometrically defined as twice the area between the Lorenz curve and the 45º line. If workers are assigned the average earnings in their occupation, one obtains the Gini index of inequality between occupations. The Gini concentration index is defined in similar terms as twice the area between the concentration curve (see Kakwani 1980) and the diagonal. The concentration curve is a generalization of the Lorenz curve in which the variable that is accumulated in the abscissa is not necessarily sorted by the variable accumulated in the ordinate. While the Lorenz curve of the distribution of earnings between occupations plots the cumulative distribution of occupation earnings for each cumulative proportion of employment, with occupations sorted by earnings, the concentration curve does the same but with workers sorted by another outcome. In this case we will use the inverted RTI measure to sort occupations from

highest to lowest routine tasks intensity. One would in general expect least routine tasks to have higher earnings, although this relationship is not necessarily monotonic in practice. The concentration curve is non-decreasing in the proportion of workers but, unlike the Lorenz curve, is not necessarily convex and may fall above the diagonal. The Gini inequality and concentration indices are identical when there is perfect correlation between earnings and the inverted RTI, that is when least routine occupations tend to have higher average earnings. Therefore, the ratio between the concentration and the inequality indices is a measure of the association between RTI and average earnings (based on the Gini metrics) and gives an idea of to what extent between-occupation inequalities are related to their level of RTI or, alternatively, to other factors (like skills, for example).

3 Reweighting/RIF regression decomposition of changes in the distribution of earnings

In a final step, we follow the estimation methodology developed by Firpo et al. (2007, 2009) to quantify the role played by different variables at different points of the earnings distribution in determining inequality trends over time. The original approach presents an extension of the Oaxaca-Blinder decomposition method, where each variable's contribution to the change in earnings is decomposed into a 'composition' effect and a 'wage structure' effect at each percentile of the earnings distribution. The estimation method has been expanded by Fortin et al. (2011) to different functionals of the earnings distribution, including measures of dispersion such as the Gini coefficient (see also Firpo et al. 2011, 2018 for an extensive discussion), which makes this approach particularly useful in the context of our analysis. It allows us to quantify the extent to which intertemporal changes in inequality can be attributed to changes in the distributions of certain worker characteristics and changes in the labour market remuneration of these characteristics.

Technically, the estimation strategy is performed in two stages. At the first stage, distributional changes between the initial and final periods are divided into an aggregate earnings structure effect and an aggregate composition effect, using a reweighting method approach, based on a semi-parametric propensity score procedure. In the second stage, the two components are further subdivided into the contribution of different sets of explanatory variables using the recentred influence function (RIF) regressions.

In general, any distributional parameter (earnings quantile, Gini index, interquantile ratio, etc.) can be written as a functional $v(F_y)$ of the cumulative distribution of earnings $F_y(y)$. The first part of the decomposition consists of dividing the overall change in the distributional measure between two years ($t = 0$ and $t = 1$) into a composition effect driven by changes in the distribution of the workers characteristics, X, observed in both periods—such as the increasing level of education, lower average RTI, or a higher share of workers from a specific demographic group—and an earnings structure effect that reflects how the conditional distribution of earnings $F(y|X)$ of workers with those given characteristics changes over time (the effect of changes in the returns to characteristics).

Let $v\left(F_{y_s|t}\right)$ be the distributional measure of interest when workers in year t obtain earnings under the earnings structure prevailing of year s. Then $v\left(F_{y_0|t=0}\right)$ and $v\left(F_{y_1|t=1}\right)$ are the observed measures in, respectively, the initial and final periods, while $v\left(F_{y_0|t=1}\right)$ reflects the counterfactual measure that would have prevailed if workers in the final period obtained their earnings under the earnings structure of the initial period (which is never observed). By adding and subtracting this

counterfactual measure to the overall difference over time Δ_o^v, we get the first aggregate decomposition:

$$\begin{aligned}\Delta_o^v &= v\left(F_{y_1|t=1}\right) - v\left(F_{y_0|t=0}\right) \\ &= \left[v\left(F_{y_1|t=1}\right) - v\left(F_{y_0|t=1}\right)\right] + \left[v\left(F_{y_0|t=1}\right) - v\left(F_{y_0|t=0}\right)\right] \\ &= \Delta_S^v \qquad + \qquad \Delta_X^v.\end{aligned}$$

where Δ_S^v is the earnings structure effect and Δ_X^v is the composition effect.

In this context, the counterfactual distribution can be estimated using reweighting, a method that provides a consistent estimate of the counterfactual distribution under the ignorability assumption. This procedure consists of replacing the marginal distribution of X in the initial year with the marginal distribution of X in the final year. This is done by multiplying the sampling weight of each initial observation by a reweighting factor $\psi(X)$, estimated with propensity score reweighting. Applying the Bayes' rule, this factor can be expressed as:

$$\psi(X) = \frac{\Pr(X|t=1)}{\Pr(X|t=0)} = \frac{\Pr(t=1|X)}{\Pr(t=0|X)} \frac{\Pr(t=0)}{\Pr(t=1)}.$$

In practical terms, we estimate the reweighting factor using a simple reweighting procedure. Using the pooled sample of workers of both periods, we estimate the probability of each worker being observed in the final year conditional on a set of covariates ($\Pr(t=1|X)$) and its complement ($\Pr(t=0|X) = 1 - \Pr(t=1|X)$). This is done using a logit model in which the dependent variable is a year dummy and the explanatory variables are the key covariates and relevant interactions. The predicted probability scores are then used to reweight the initial sample so that it resembles the final sample in terms of average worker characteristics. That is, we increase the weight of workers in the initial year with characteristics that were more common in the end year, and similarly reduce the weight of those with characteristics that were less common. The reweighted distribution reflects the counterfactual that combines the final marginal distribution of characteristics and the initial conditional earnings distribution, $F_{y_0|t=1}$. We estimate the distributional measure of interest v for this counterfactual distribution, which is used in the aggregate decomposition described above.

One limitation of this reweighting procedure is that obtaining detailed decompositions of the earnings structure effects is not straightforward. Therefore, these are obtained in the second step, using linear RIF decompositions. The RIF decomposition is similar to the classical Oaxaca-Blinder approach, but with the usual outcome variable, log of earnings, being replaced by the Recentered Influence Function of the target statistic v, $RIF(y; v)$.

The Influence Function of a statistic v, $IF(y; v)$, measures the impact on the statistic of marginally increasing the population mass at a certain point y (i.e. a small contamination) and has an expected value of zero. More formally, if y_ε is a mixture distribution assigning a probability $1 - \varepsilon$ to the original distribution y and ε to a specific point y_i (see Hampel 1974), the IF is obtained as the directional derivative:

$$IF(y_i; v) = \frac{\partial}{\partial \varepsilon} I(y_\varepsilon)|_{\varepsilon=0} \text{ ; with } E_y(IF(y; v) = 0.$$

The $RIF(\mathrm{y}; v)$ is obtained after recentering the IF at the value of the target statistic so that the expectation is v:

$$RIF(y;v) = v + IF(y;v) \text{ with } E_y(RIF(y;v)) = v.$$

The IF and RIF of several distributional measures have already been computed. In particular, the RIF of the τ-th quantile $q_\tau \equiv \inf(y|F(y) \geq \tau)$, is given by:

$$RIF(y_i;q_\tau) = q_\tau + \frac{\tau - \mathbf{1}(y_i \leq q_\tau)}{f_y(q_\tau)}$$

where $\mathbf{1}(y_i \leq q_\tau)$ is an indicator function that takes value 1 if the earnings of worker i falls below the τ-th quantile and 0 otherwise, and f_y is the estimated (non-parametric kernel) density function evaluated at q_τ. The RIF of the Gini index G is given by:

$$RIF(y_i;G) = 2\frac{y_i}{m}\left(\frac{i}{n} - \frac{1+G}{2}\right) + 2\left(\frac{1}{2} - L_i\right)$$

where m is the mean earnings, L_i is the value of the Lorenz curve (the cumulative share of earnings) for the population share with earnings y_i or below.

An OLS regression of the corresponding RIF on X is estimated for the initial and end years, as well as for the counterfactual:

$$RIF(y_{it};v) = \gamma_t X_{it}; t = 0, 1, c.$$

These regressions can be used to decompose the difference in v between two distributions as in conventional Oaxaca-Blinder decompositions, replacing the log of earnings with the corresponding $RIF(y_{it};v)$ estimated for each observation. This is done by adding and subtracting a specific counterfactual that combines the characteristics of the final distribution with the coefficients of the initial one. A RIF decomposition of the distributional change between the counterfactual and final distribution can be used to break the aggregate structural effect previously obtained by reweighting into a pure RIF structural effect $\Delta_{S,p}^v$ (the RIF earnings structure effect in this decomposition) and a RIF reweighing error $\Delta_{S,e}^v$ (the RIF composition effect):

$$\Delta_S^v = (\gamma_1 - \gamma_c)X_{i1} + \gamma_c(X_{i1} - X_{ic})$$
$$= \Delta_{S,p}^v + \Delta_{S,e}^v.$$

Similarly, with a second RIF decomposition for the distributional change between the counterfactual and initial distributions, we can break the reweighting composition effect into a pure RIF composition effect $\Delta_{X,p}^v$ (the RIF composition term in this decomposition) and a specification error $\Delta_{X,e}^v$ (the RIF structural term):

$$\Delta_X^v = \gamma_0(X_{ic} - X_{i0}) + (\gamma_c - \gamma_0)X_{ic}$$
$$= \Delta_{X,p}^v + \Delta_{X,e}^v.$$

The linearity of all these terms, due to the use of linear RIF regressions, makes obtaining the corresponding detailed effects for specific sets of characteristics as straightforward as in the conventional Oaxaca-Blinder approach.

The analysis is focused on the effects of interest, that is, the pure RIF composition ($\Delta_{X,p}^v$) and pure RIF earnings structure ($\Delta_{S,p}^v$) effects. Note that these RIF effects add up to the corresponding

reweighting totals only if both error terms are zero. While the reweighting error should be negligible provided we use a rich logit model with interactions among the covariates, and can be generally ignored, the size of the specification error depends on the linearity of the estimated RIF coefficients, that is, by how much they change after reweighting the initial distribution to mimic the final one.[2]

Finally, standard errors are obtained after bootstrapping the entire process (reweighting and RIF decompositions) using a large number of replications (500 in the Ghana case).

References

Bhattacharya, N., and B. Mahalanobis (1967). 'Regional Disparities in Household Consumption in India'. *Journal of the American Statistical Association*, 62(317): 143–61.

Chakravarty, S.R. (2009). *Inequality, Polarization and Poverty: Advances in Distributional Analysis*. New York: Springer.

Firpo, S., N.M. Fortin, and T. Lemieux (2007). 'Decomposing Wage Distributions using Recentered Influence Function Regressions'. Unpublished manuscript, Vancouver: University of British Columbia.

Firpo, S., N.M. Fortin, and T. Lemieux. (2009). 'Unconditional Quantile Regressions'. *Econometrica*, 77(3): 953–73. https://doi.org/10.3982/ECTA6822

Firpo, S., N.M. Fortin, and T. Lemieux (2011). 'Occupational Tasks and Changes in the Wage Structure'. IZA Discussion Paper 5542. Bonn: Institute of Labor Economics (IZA).

Firpo, S., N.M. Fortin, and T. Lemieux (2018). 'Decomposing Wage Distributions using Recentered Influence Function Regressions'. *Econometrics*, 6(2): 1–40.

Fortin, N.M., T. Lemieux, and S. Firpo (2011), 'Decomposition Methods in Economics'. In O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*. Amsterdam: North Holland.

Goos, M., and A. Manning (2007). 'Lousy and Lovely Jobs: The Rising Polarization of Work in Britain'. *The Review of Economics and Statistics*, 89(1): 118–33. http://www.mitpressjournals.org/doi/pdf/10.1162/rest.89.1.118

Hampel, F.R (1974). 'The Influence Curve and Its Role in Robust Estimation'. *Journal of the American Statistical Association*, 60: 383–93. https://doi.org/10.1080/01621459.1974.10482962

Kakwani, N.C. (1980). *Income Inequality and Poverty: Methods of Estimation and Policy Applications*. Report 10092. Oxford: Oxford University Press.

Pyatt, G. (1976). 'On the Interpretation and Disaggregation of Gini Coefficients'. *The Economic Journal*, 86(342): 243–55.

Sebastián, R. (2018a). 'Explaining Job Polarisation in Spain from a Task Perspective'. *SERIEs*, 9: 215–48. https://doi.org/10.1007/s13209-018-0177-1

Sebastián, R. (2018b). 'Technological Change and Employment Polarisation'. Thesis dissertation. Salamanca: University of Salamanca.

Shorrocks, A.F. (1984). 'Inequality Decomposition by Population Subgroups'. *Econometrica*, 52(6): 1369–85. https://doi.org/10.1108/01443581311283475

[2] The specification error, as part of the reweighting composition effect, captures the indirect effect of a change in characteristics, the one that goes through changing the relationship between average characteristics and the target statistic.