Wage gap decomposition models:
A methodological contribution

Magno Rogério Gomes*
Solange de Cássia Inforzato de Souza**
Gabriela Gomes Mantovani***
Vanessa Fortunato de Paiva****

Abstract
This article aims to introduce a methodological contribution to the wage decomposition, from the traditional decomposition of Oaxaca–Blinder. Specifically, we review the usual wage decomposition models and proposes a new technique that do not choose an advantage/disadvantage group in advance, a solution to the index problem, one of the most present criticisms in the decomposition salary of Oaxaca–Blinder. It is based on the analysis of the current wages of the analyzed groups, and not on the impacts of the income that the individuals should have, as usually appears in the models of decomposition. It applies the gender wage decomposition procedures in the Brazilian labor market, under the different approaches, from the microdata of the PNAD of 2015, which confirm the wage differences and discrimination between men and women. Finally, this article contributes to the decision on empirical strategies in the field of wage inequalities and, particularly, their interpretation.

Keywords: Wage decomposition, gender discrimination, Oaxaca–Blinder
JEL Codes: C40, J31, J71

1 Introduction
Wage inequalities in the labor market are a relevant issue in the field of economic sciences and it motivates theoretical and empirical studies in Latin America and developed countries. Brazil, in particular, is a country of significant economic
positioning in the world economy which, despite the intense advances made in previous decades, maintains alarming levels of social and economic inequality. These are historically materialized in the poor income distribution, as well as difficulties in accessing occupations, discriminatory ethnic miscegenation, and gender pay gaps.

The theoretical debate on income inequalities is based on two, not necessarily excluding, positions. The first, related to human capital, emphasizes education and experience for increases in labor income because they contribute to increased productivity, as argued by Schultz (1961), and Becker (1962). In the second position, specialists in discrimination economics and labor market segmentation. From different hypotheses, these researchers point out that there is discrimination whether wages differ between groups, while maintaining individual skills (Becker, 1962), or wage segmentation if jobs characteristics, market structure, process history and social classes, cause these income inequalities, expressed by Doeringer and Piore (1970), Vietorisz and Harrison (1973), and Reich, Gordon, and Edwards (1973).

International and Brazilian empirical findings on wage inequalities incorporate the relevance of productive attributes such as education and experience, as well as sources of inequalities not associated with these elements to explain wage gaps between groups such as gender and race, jobs, economic sectors, regions, and others. Examples are the works by Cobb-Clark and Tan (2010) for Australia; Biltagy (2014) for Egypt; and Fonseca (2017) for Portugal, which demonstrate occupational segregation and wage discrimination in these countries. Recently, in Brazil, Mariano, Costa, Guimarães, and Sousa (2018), Mattei and Baço (2017), and Gomes (2016), among others, confirm the wage differences caused by both productive factors related to human capital and non-productive factors resulting from wage discrimination in the country and its regions.

In the aforementioned studies there is a prevalence of the use of the decomposition procedure of Oaxaca (1973) and Blinder (1973), which gives them great scientific importance. The purpose of this technique is to disaggregate the wage differentials between groups into one part explained by differences in the productive characteristics of the group (education and experience) and another explained by the differences in remuneration of these characteristics between groups, using as a measure of income discrimination or labor market segmentation. To this end, the Oaxaca–Blinder wage decomposition often adopts a labor market advantage group beforehand. This arbitrary choice of the advantage/disadvantage group, the index problem, provokes different interpretations.
and is one of the reasons why researchers direct their criticism of the wage decomposition, since there would be no specific reason to assume the coefficients of a particular group such as the one discriminated by the labor market.

In this sense, other wage decompositions have been proposed and dealt with in the economic literature. Reimers (1983) and Cotton (1988), for example, proposed using the means of the study group coefficients as the non-discriminant coefficient; Juhn, Murphy, and Pierce (1993) used a wage decomposition in the wage distribution percentiles; and DiNardo, Fortin, and Lemieux (1996), Autor, Lawrence, and Melissa (2005), and Machado and Mata (2005) performed the wage decomposition for the entire wage distribution based on the wage density function. However, these other techniques do not particularly address the index problem and present computational difficulties that may do research unfeasible. In this paper, the main focus is to concentrate on the differences in wage averages, as proposed by Oaxaca–Blinder, and to present a possible solution to the problem of early selection of an advantageous group.

Given the above, this article aims to introduce a methodological contribution to improve the wage decomposition from the traditional Oaxaca–Blinder decomposition, aiming at the solution to the problem of early choice of the advantage/disadvantage group (index problem) in the labor market. Specifically, we intend to review the wage decomposition models that maintain the index problem and the ways of interpreting the results obtained by the usual procedure to, finally, propose a new technique that does not require the prior choice of an advantage/disadvantage group to perform it. Illustrates the use of the decomposition procedure in measuring gender pay gaps and discrimination through the 2015 PNAD microdata. This is intended to contribute to the improvement of the Oaxaca–Blinder wage decomposition model and the understanding of the reality of the Brazilian labor market.

In addition to this introduction, this article covers three more sections. The second section presents theoretical and empirical approaches to wage inequalities and the use of the Oaxaca–Blinder procedure. The third addresses the usual methodology of Oaxaca–Blinder decomposition as an empirical strategy and its alternative versions of measurement and interpretation. The fourth section describes the wage decomposition model and the proposal to solve the index problem, as well as applying it to the Brazilian labor market. In the end of the article, there are the final remarks.
2 Salary inequalities, theories and methods: A literature review

Wage differentials are treated in the economic literature by the compensatory wage differential theories, human capital, discrimination and segmentation (or dual market). The compensatory wage differential theory defends the importance of working conditions in the wage differentiation between workers. The theory of human capital, from the point of view of labor supply, has as exponents Mincer (1958), Becker (1962), and Schultz (1961), who affirms the existence of a positive relationship between education and personal income, that is, education is seen as an investment for the worker to improve and/or increase his productivity and remuneration.

The economic theory of discrimination argues that workers are treated in different ways based on some unimportant criterion because of the groups to which they belong (Becker, 1962; Becker & Chiswick, 1966) either by statistical discrimination, discrimination by employer, employee or by customer. That is, according to Loureiro (2003), discrimination occurs when a group of individuals have equal skills, training, education, productivity and experience, but receive different salaries and treatments due to their skin color, gender, economic and social condition, where they live, physical appearance, without these personal characteristics impacting worker productivity.

In the theory of labor market segmentation (or dual market), the wage gap occurs when workers are treated differently due to job characteristics. In this case, the explanation of wage gaps occurs by allocative adjustment, in which the personal attributes of individuals determine which market the worker will be allocated (Doeringer & Piore, 1970), by technological dualism (Vietorisz & Harrison, 1973), or by differences of social classes, which intensify the behavioral and personal differences of the workforce (Reich et al., 1973). The three strands of segmentation theory complement each other and keep the perspective of demand and work.

There are several empirical evidences of wage differences in the labor market and the contributions of different factors to explain such inequalities, related to human capital and gender and race issues, economic sectors or geographical regions. In most of the research, the Oaxaca–Blinder wage decomposition technique is applied, with some of the results described below.

Internationally, Cobb-Clark and Tan (2010) found a significant effect of
non-cognitive skills on the likelihood of employment in Australia’s labor market from 2001 to 2006, and although women earn lower wages than men being employed in the same occupation, characterizing segregation in the labor market. Biltagy (2014) estimated that the gender pay gap in Egypt in 2006 is largely due to discrimination against women in the labor market. Fonseca (2017) found a wage gap that favors men in 2007 and 2012 in Portugal, also mostly explained by discriminatory practices, despite the reduction in wage discrimination, due to better women’s productive endowments.

In Brazil, uses of Oaxaca Blinder decomposition can be identified. In summary, from a gender perspective, for example, Matos and Machado (2006) found the presence of color discrimination between white men and black men, and the worst situation of black women in terms of income, according to 1987 and 2001 PNAD. Carvalho, Néri, and Silva (2006) confirmed gender and color discrimination for 2003, and with more recent data, Souza (2015) and Maia (2015) affirmed the unfavorable wage gap for women and the most pronounced wage gap between white men and non-white women.

Under another scientific methodology for the study of wage inequalities, Mariano et al. (2018) used the Ñopo pairing with 2014 PNAD data for metropolitan regions of Brazil and found a wage gap in favor of men and whites.

In addition to studies for the country as a whole, regional delimitations and labor market insertions using the Oaxaca Blinder methodology do not contradict the research cited and point to new data. Souza (2015), for Bahia in 2013, show that discrimination against women is lower in the insertion of the first job, being more intense among white individuals in the formal market. Gomes (2016), for the same year, confirmed that the Northeast region exhibited the lowest gender discrimination against women when compared to the South region. Mattei and Baço (2017), for the formal market of the state of Rio Grande do Sul, in the 2000, 2007 and 2014, found that gender pay gaps narrowed over the reporting period and this fall is associated with reduced pay discrimination against women.

In sum, the surveys listed above have made it possible to detect wage differences between demographic groups, especially by gender, and have shown the use of the traditional Oaxaca–Blinder wage decomposition. Nevertheless, there are criticisms of this method arising from the index problem, translated by the early choice of a group in advantage or disadvantage and the admission of the group that suffers positive or negative discrimination in the labor market.

---

1 Non-cognitive skills are related to the five major personality traits: extroversion, emotional stability, pleasantness, conscientiousness, and openness to new experiences.
This article aims to improve the Oaxaca–Blinder procedure, proposing a way to solve this problem, as well as to facilitate counterfactual interpretations.

3 Wage decomposition as empirical strategy

As exponents of wage discrimination studies, Ronald Oaxaca and Alan Blinder developed a counterfactual technique of decomposition of wage differences using wage determination equations, mincerian equations, set forth in the authors’ work on wage differences between men and women in the urban region of the United States in 1967, both published in 1973 (Oaxaca, 1973; Blinder, 1973). Oaxaca and Blinder perform a decomposition in which the independent variables of the equation are divided into two portions: one portion corresponds to the wage differentials resulting from the endowments of individuals (productive characteristics) and the other is the proxy of discrimination, when similar characteristics are valued in different ways.

Other wage decompositions present in the literature result from the idea that there is a non-discriminatory coefficient vector that should be used to determine the contribution of worker or job characteristics to wage differences, as pointed out by Jann (2008). However, the early choice of an advantage/disadvantage group and the admission of the group experiencing positive or negative discrimination remain and lead to different interpretations.

Reimers (1983) and Cotton (1988) propose the use of the average of the estimated coefficients of the two groups as non-discriminant coefficient. Reimers (1983) suggests the use of means of the two parameters $\hat{\beta}^* = 0.5\hat{\beta}_H + 0.5\hat{\beta}_M$, where $\hat{\beta}^*$ corresponds to the non-discriminant estimator, $\hat{\beta}_H$ the estimated coefficient of men, and $\hat{\beta}_M$ the estimated coefficient of women. Cotton (1988) suggests a more robust average with the weight of group distributions. That is,

$$\hat{\beta}^* = \frac{n_H}{n_H + n_M} \hat{\beta}_H + \frac{n_M}{n_H + n_M} \hat{\beta}_M,$$

where $n_M$ is the women’s group sample number, $n_H$ is the sample number of men. While using parameter averages or weighting coefficients, these choices are somewhat arbitrary.

Following the same idea as Reimers (1983) and Cotton (1988), Oaxaca and Ransom (1994) present a weighted matrix decomposition in the following model:

$$D = \{E(X_H) - E(X_M)\}'\{W\beta_H + (I - W)\beta_M\} + \{(I - W)'E(X_H) + W'E(X_M)\}'(\beta_H - \beta_M),$$
where $W$ is the relative weight matrix of the men’s group coefficients, and $I$ is the identity matrix. The estimated matrix of $W$ can be found from:

$$\hat{W} = W = (X_M'X_M + X_M'X_M)^{-1}X_M'X_M,$$

where $X$ are the observed values of the matrix and is equivalent to the use of the pooled model coefficients of the studied groups:

$$W = \text{diag}(b - b_M) \times \text{diag}(b - b_M)^{-1},$$

where $\beta$ the pooled model coefficient.

Oaxaca and Ransom (1994) and Neumark (1988) also suggest including a group indicator as an additional covariate so that no transfers of the unexplained part of the differential to the explained component occur. In both procedures there is an important methodological advance, but there is still the difficulty of identifying the market salary (non-discriminant).

Researchers have also used other decomposition techniques that generally emphasize decomposition in data distributions. Juhn et al. (1993) in a study of gender pay gaps for the United States between 1963 and 1989, present an alternative of wage decomposition in the wage distribution percentiles. Their method starts from a simple wage equation, $\hat{Y}_{it} = X_{it}\beta_t + u_{it}$, where $\hat{Y}_{it}$ is the wage ln for each individual $i$ in the year $t$; $X_{it}$ is the vector of individual characteristics of individuals; and $u_{it}$ is the value of unobserved wages (residue). The authors put the residue with a function $u_{it} = F_t^{-1}(\theta_{it} \mid X_{it})$, where $\theta_{it}$ corresponds to the percentile of the individuals of the residual distribution and $F_t(\cdot)$ the residual distribution function of the wage equation. Therefore, $F_t^{-1}(\cdot \mid X_{it})$ is the inverse of the cumulative distribution of waste given the characteristics of workers in year $t$. The authors break down the wage gap into three parts: differences attributed to the individual characteristics of workers, the part explained by changes in characteristics prices (betas), and the part of the differential due to residual distribution, given by

$$Y_{it} = X_{it}\bar{\beta} + X_{it}(\beta_t - \bar{\beta}) + \bar{F}^{-1}(q_{it} \mid X_{it}) + [F_t^{-1}(q_{it} \mid X_{it}) - \bar{F}^{-1}(q_{it} \mid X_{it})],$$

where $\bar{\beta}$ is the sum of the observable prices for the whole period and $\bar{F}^{-1}(\cdot \mid X_{it})$ the average cumulative distribution.

For Juhn et al. (1993), the main advantage over standard wage decomposition is that this method allows us to see how changes in composition affect the entire
wage distribution and not just its variation. It is possible to determine how changes in individual distributions affect other measures of inequality, such as interquartile range or ninetieth percentile differential, or how the effects are different for above and below average differences. However, the problem of early choosing a group with an advantage or disadvantage in wage decomposition remains.

The method proposed by DiNardo et al. (1996) is a generalized semiparametric decomposition from the Oaxaca decomposition, in which the authors emphasize changes in the wage distribution density. They make a counterfactual analysis and answer, for example, what would be the distribution of workers’ wages in 1988 if they had the same characteristics (productive and non-productive) of 1979. The technique implemented has advantages over traditional Oaxaca decomposition, such as not only using averages, but wage density. On the other hand, as in Oaxaca’s usual decomposition, the proposal ignores the effects of the general equilibrium and depends on the ordering of the explanatory factors, limitations that do not preclude showing the significant importance of the various explanatory factors, according to the authors. Despite these merits, researchers do not focus on the index problem.

Following the essence of DiNardo et al. (1996), Machado and Mata (2005) also extend the Oaxaca–Blinder decomposition and propose a decomposition for the entire wage distribution. The method is based on estimates of the wage density function over a given period, implicit in the counterfactual distributions of the observed attributes. It was applied to Portuguese wages from 1986 to 1995.

The counterfactual nature of the proposed method requires the estimation of the wage distribution conditional on the variables of interest through quantile regressions, i.e., models for conditional wage distribution quantiles are estimated. These quantile regressions capture the impacts of changes in covariates on the conditional wage distribution. The authors’ main contribution is to marginalize the estimated conditional distribution, using different scenarios for the distribution of worker attributes. In short, the method of Machado and Mata (2005) assesses the impact of wages on all wage density without imposing a functional relationship between other variables, however, it is indifferent to the implications

---

2 The major advantage of this over the more standard variance accounting framework is that it allows us to look at how composition changes have affected the entire wage distribution and not just the variance. We can determine how changes in the distribution of observables have affected other inequality measures such as the inter-quartile range or the ninetieth-tenth percentile differential or how the effects have been different for inequality above and below the mean (Juhn et al., 1993, p.428).
of arbitrary choice of the group that is favored or not in the labor market.

**Firpo, Fortin, and Lemieux (2009)** propose a decomposition method beyond the means and implement a new regression method to estimate the effect of variables on the unconditional quantiles of a result variable. The model uses the regressed influence function (RIF) consisting of three different regression estimators, based on an OLS standard regression (RIF-OLS), where the refocused influence function is the dependent variable, a Logit regression (RIF-Logit) and a non-parametric Logit Regression (RIF-NP). As mentioned, it has an advantage because it surpasses the analysis of averages, but its limitation lies in the assumption that the covariations of the variables are independent, admitted by the authors.

According to **Fortin, Lemieux, and Firpo (2011)**, it is important to employ a salary decomposition with data aggregation. The authors' idea does not involve parametric assumptions, besides allowing to compute the effects of covariates individually. As written, the refocused influence function (RIF) is used for interest distribution statistics. Once the RIF regression is estimated, the estimated coefficients can be used to perform the decomposition, detailed in the same way as in the Oaxaca–Blinder decomposition pattern. The disadvantage of this approach is that RIF regression coefficients only provide a local approximation for the effect of changes in the distribution of a covariate of statistical distribution of interest. But it is important because it provides one of the few options available to calculate detailed decomposition for different distribution parameters. In the aforementioned works, scholars broaden Oaxaca’s decomposition, but there is no focus on the problem of prior choice of demographic advantage or disadvantage.

**Ñopo (2008)** develops a nonparametric matching method to estimate the gender pay gap. The objective is to re-sample all women without replacement and combine each observation with a synthetic man with the same observable characteristics and a salary obtained by averaging all men with the same characteristics \((x)\). Thus, one woman is first selected from the sample (without replacement); second, all men who have the same “\(X\)” characteristics as the woman previously selected are selected; third, with all the individuals selected in step 2, a synthetic individual is constructed whose salary is the average of all of them and matches that of the original woman. Then the observations of both individuals (the synthetic man and the woman) are inserted in new samples of paired individuals. The process is repeated (from 1 to 4) until the original female sample is exhausted. The new dataset contains observations of paired women, paired men, matchless women, and unmatched men, where sets of matched men
and women have the same empirical probability distributions for characteristics \( X \). Counterfactual analysis proceeds according to the original configuration of Oaxaca–Blinder, has advantages such as not assuming the assumptions of parametric models and acting on the wage distributions of the studied groups, however, they are also indifferent to the index problem.

As described, researchers have developed new methods of wage decomposition in recent decades, building on the work of Oaxaca (1973) and Blinder (1973) and seeking to go beyond the use of observed averages. These other techniques involve studies on sample distributions and assumptions whose limits and potentialities have been previously exposed.

In summary, from attempts to improve wage decomposition, Reimers (1983) uses the use of the estimated group coefficients as the non-discriminant coefficient, Cotton (1988) recommends a more robust average with the weight of group distributions, and Oaxaca and Ransom (1994) present an alternative by weighting the decomposition by a weight matrix. Juhn et al. (1993) and Firpo et al. (2009) make the wage decomposition in the wage distribution percentiles, and Machado and Mata (2005) propose an estimate of decomposition in the entire wage distribution in different periods. Other scholars offer their contribution through semiparametric models, such as DiNardo et al. (1996), Fortin et al. (2011) and Nopo (2008). Nevertheless, the papers do not focus on the awkwardness of the anticipation of choosing the advantage or disadvantage group, or the group that suffers positive or negative discrimination.

Given the aforementioned, the next section is dedicated to presenting four ways to perform the traditional Oaxaca–Blinder wage decomposition, which, a priori, determine an advantage/disadvantage group and its different results and interpretations for the Brazilian reality. Finally, as the main contribution of this study, a wage decomposition is proposed in which there is no arbitrary choice of an advantageous group, which is not explored in the known wage decomposition methodologies, starting from a counterfactual interpretation. Not competing with each other, the models shown help the researcher’s decision on the empirical strategy to be adopted.

4 Solution to the index problem
To perform wage decompositions, it is necessary to estimate the wages of the groups under analysis and to correct the sample selection bias. Wages are estimated using the semilogarithmic equation developed by Mincer (1974) and is used in studies that aim to measure wage returns given the productive and
Wage gap decomposition models: A methodological contribution

non-productive characteristics of individuals, such as individuals with different occupations, race or gender.

In this paper, the application of the decomposition technique will be for the groups of men and women and the original equation will extend because, besides presenting productive variables (human capital) as determinants of men’s and women’s wages, it will also add other characteristics present in the theoretical and empirical economic literature:

\[
\ln(W_i) = \beta_0 + \beta_1 ESC_i + \beta_2 EX_i + \beta_3 EX_i^2 + \beta_4 B_i + \beta_5 F_i + \beta_6 SET_i \\
+ \beta_7 OCUP_i + \beta_8 URB_i + \beta_9 UF_i + \beta_{10} IMR_i + \mu_i, \quad (1)
\]

where \(\ln(W_i)\) corresponds to the natural logarithm of income adjusted by the working hours of a given group; \(ESC_i\) regards the years of education of the worker; \(EX_i\) the experience proxy; \(EX_i^2\) the squared experience; and \(B_i\) the color, separated into white and non-white (brown and black).

The binary variable \(F_i\) represents the formality of employment; the \(OCUP_i\) vector the dummies for the occupations of each individual, i.e., Leaders (directors and managers), Science and Arts Professionals (SAPs), Technical and Operational. The vector \(SET_i\) indicates the sectors of the economy, which are commerce, industry, services and agriculture, the latter being the base variable. The variable \(URB_i\) represents the urban or rural area and the vector \(UF_i\) the macroregions of Brazil: South, Southeast, Midwest, North and Northeast region as the base region.

The \(IMR_i\) corresponds to the inverse Mills ratio (IMR), parameter estimated by the Probit model from a labor market participation equation. After estimating the participation equation, we calculate the IMR, a decreasing monotone function of a probability,

\[
IMR = \frac{\phi(z_i)}{1 - \varphi(z_i)},
\]

where \(\phi\) represents a density function of a standard normal variable and \(\varphi\) represents a distribution function of a standard normal variable. This variable is

---

3 This study considered employed workers aged 14 years and over, as it is the minimum legal age to start working hours in Brazil, starting from PNAD (2015). Equation (1) will be estimated for the studied groups, men and women, but can be used for a range of studies that seek to estimate differences between two groups.

4 The experience proxy is generated from the individual’s age minus the age at which he or she started working.

5 By formal, it is considered employees with a formal contract, military, statutory civil servants, domestic workers with a formal contract, self-employed and formal employers.
used in the wage determination equations to correct a possible sample selection bias, due to the non-observation of labor supply. According to Heckman (1979), a sample selection bias can occur for two reasons. The first deals with the self-selection of the individuals, and the second, with the sample selection decisions by the researchers. Lastly, $\mu_i$ represents the random error term (see Appendix).

Following, we describe the wage decomposition models, their interpretations and applications for the Brazilian case, based on data from PNAD—National Household Sample Survey—2015. The models allow the detailing of the differences that are explained or not by the individual attributes, but they present a problem of identification, that is, the invariance of the decomposition results due to the choice of the variables to be omitted in the sets of categorical variables used in the regressions to determine the wage values. To correct the problem, we use the standardization process exposed by Yun (2005).

The results of the salary decompositions presented (models 1, 2, 3, 4, 5) are shown and summarized in Table 2.

**Model 1 — Assumes a negative discrimination hypothesis against women: men’s salary as the market salary** Suppose that $H$ represents men and $M$ women, the researcher may assume that discrimination against women exists and the salary paid to men is the market salary (index problem) in the absence of discrimination, that is, it would allow negative discrimination against women.

This model follows the usual Oaxaca–Blinder (1973) decomposition, also presented in Jann (2008). It expresses the differences in average income between the two groups studied, in which part of this gap can be explained by differences in the characteristics of individuals, and the other part cannot be explained (differences in coefficients) that is attributed to discrimination when comparing the wages of men and women or whites and nonwhites.

Starting from the mincerian equations of the studied groups we have:

$$Y^H_i = \beta^H_0 + \sum_{j=1}^{n} \beta^H_j X^H_{ji} + u^H_i$$

$$Y^M_i = \beta^M_0 + \sum_{j=1}^{n} \beta^M_j X^M_{ji} + u^M_i,$$

where $Y_i$ corresponds to the natural logarithm of the hourly wage; $\beta_j$ is the coefficient vector; $X_{ji}$ is the matrix of explanatory variables of the model; and $u_i$ represents the error term vector. The superscript $H$ represents the group of
Individual salary averages are calculated using the OLS and can be written as

\[
\bar{Y}^H = \beta_0^H + \sum \bar{X}^H \beta_j^H \quad (2')
\]

\[
\bar{Y}^M = \beta_0^M + \sum \bar{X}^M \beta_j^M \quad (3')
\]

From equations (2') and (3'), and indexing male salary as “minuend” and female salary as “subtracting”, i.e., subtracting (3') from (2') we have:

\[
(\bar{Y}^H - \bar{Y}^M) = (\beta_0^H + \sum \bar{X}^H \beta_j^H) - (\beta_0^M + \sum \bar{X}^M \beta_j^M) \quad (4)
\]

As noted earlier, one observation is made by Jann (2008): it is considered that there is negative discrimination against women, and thus the results of decomposition are represented as follows:

\[
\beta_j^M = \beta_j^H - (\beta_j^H - \beta_j^M), \quad \text{with } \Delta \beta = (\beta_j^H - \beta_j^M) \quad (\text{negative discrimination}),
\]

\[
(\bar{Y}^H - \bar{Y}^M) = (\beta_0^H - \beta_0^M) + \sum \bar{X}^H \beta_j^H - \sum \bar{X}^M (\beta_j^H - \Delta \beta), \quad (5)
\]

\[
(\bar{Y}^H - \bar{Y}^M) = (\beta_0^H - \beta_0^M) + \sum \bar{X}^H \beta_j^H - \sum \bar{X}^M \beta_j^H + \sum \bar{X}^M \Delta \beta, \quad (6)
\]

\[
(\bar{Y}^H - \bar{Y}^M) = (\beta_0^H - \beta_0^M) + \sum \beta_j^H (\bar{X}^H - \bar{X}^M) + \sum \bar{X}^M \Delta \beta, \quad (7)
\]

\[
\sum \beta_j^M (\bar{X}^H - \bar{X}^M) = \text{Part explained by differences in appropriations} = E_1,
\]

\[
(\beta_0^H - \beta_0^M) + \sum \bar{X}^M (\beta_j^H - \beta_j^M) = \text{Unexplained part (discrimination proxy)} = D_1.
\]

The application of Model 1, which assumes the hypothesis of negative discrimination against women, as shown in Table 2, is analyzed as follows: the total wage gap between men and women was 0.0610 of the ln of the hourly wage of men and women. As a percentage of women’s wages, the difference was 6.29, that is, for their wages to match men’s wages, their earnings should increase by 6.29%.

The unexplained part (discrimination proxy) corresponds to the differences in the coefficients applied to the characteristics of women. If there were no gender discrimination, women’s earnings would have to increase by 25.43%. Regarding the explained part of the model, differences in characteristics between men and women (H − M) are valued by the coefficients of men, considered the market
wage, without discrimination.

The interpretation is that if men had the same characteristics as women, their incomes should be higher. In percentage terms, men’s current earnings correspond to 84.74% of the earnings they would have if they had the characteristics of women, which shows that women have, on average, their characteristics superior to those of men. As the explained part was negative, this percentage can also be seen as a buffer to reduce gender pay gaps.

Model 2 – Assumes a positive discrimination hypothesis in favor of men: women’s salary as the market’s salary From the subtraction of equations (3’) and (2’) and indexing male salary as “minuend” and female salary as “subtracting”, positive discrimination in favor of males is considered, according to Jann (2008). In other words, it is hypothetically assumed that the market wages would be those paid to women and that men would have a higher remuneration than women, due to the fact that entrepreneurs prefer male labor.

Thus,

$$\beta^H_j = \beta^M_j + (\beta^H_j - \beta^M_j),$$

being the discriminatory factor $\Delta \beta = (\beta^H_j - \beta^M_j)$,

$$\bar{Y}^H - \bar{Y}^M = (\beta^H_0 - \beta^M_0) + \sum \bar{X}^H \beta^H_j - \sum \bar{X}^M \beta^M_j,$$

(8)

$$\bar{Y}^H - \bar{Y}^M = (\beta^H_0 - \beta^M_0) + \sum \bar{X}^H (\beta^M_j + \Delta \beta) - \sum \bar{X}^M \beta^M_j,$$

(9)

$$\bar{Y}^H - \bar{Y}^M = (\beta^H_0 - \beta^M_0) + \sum \beta^M_j (\bar{X}^H - \bar{X}^M) + \sum \bar{X}^H \Delta \beta,$$

(10)

$$\bar{Y}^H - \bar{Y}^M = \text{Total difference} = G_2,$$

$$\sum \beta^M_j (\bar{X}^H - \bar{X}^M) = \text{Part explained by differences in appropriations} = E_2,$$

$$(\bar{\beta}^H_0 - \bar{\beta}^M_0) + \sum \bar{X}^H (\beta^H_j - \beta^M_j) = \text{Unexplained part (discrimination proxy)} = D_2.$$

Considering, then, a positive discrimination in favor of men, according to Model 2, the counterfactual interpretation of the results is as follows (Table 2). The total wage gap in Brazil remains 0.0610 of the hourly wage of men compared to women; as a percentage of women’s wages, the difference was 6.29, i.e., their earnings should increase by 6.29% to match their wages to men’s wages, an interpretation similar to that given in Model 1.

The unexplained part corresponds to the differences in the coefficients applied to the characteristics of men. In other words, this effect shows the amount, in percentage, that men earn more because they suffer positive discrimination. According to the results of the application for Brazil, it is noted that the value
was positive, that is, that male coefficients are generally higher than female ones, and since the average male characteristics are non-null values, it is evident that the male gender is valued by the market differently (superior) to the female gender. Because of positive discrimination, men have an increase of 0.3442 in ln hourly wages, which corresponds to 41.08%. In other words, men’s salaries average 41.08% higher than they would be if they were valued with women’s coefficients (market salary) (5.28).

Regarding the explained part, it is observed that the differences in characteristics between men and women \((H - M)\) are valued by the coefficients of men (positive discrimination in favor of men). Therefore, it can be said that in a hypothetical scenario where women had the same characteristics as men (women’s coefficients applied to men’s characteristics) their income would decrease by 24.66%. It is noted that the average characteristics of women are mostly higher than men, which is observed by the negative sign. This means that women are more educated than men, on average, and education is a productive factor that is associated with higher wages.

**Model 3 – Assumes a negative discrimination hypothesis against women** (salary of man as the market): Equation of women as “minuend”\(^6\) In this model, subtraction is based on equations \((3')\) and \((2')\), but indexes female salary with “minuend” and male salary as “subtracting”, considering possible negative discrimination against women. Thus,

\[
\beta^M_j = \beta^H_j - (\beta^H_j - \beta^M_j), \quad \Delta \beta = (\beta^H_j - \beta^M_j)
\]

being the discriminatory factor,

\[
(\bar{Y}^M - \bar{Y}^H) = (\beta^M_0 - \beta^H_0) - \sum \bar{X}^H \beta^H_j + \sum \bar{X}^M \beta^M_j,
\]

\[
(\bar{Y}^M - \bar{Y}^H) = (\beta^M_0 - \beta^H_0) - \sum \bar{X}^H \beta^H_j + \sum (\beta^H_j - \Delta \beta) \bar{X}^M,
\]

\[
(\bar{Y}^M - \bar{Y}^H) = (\beta^M_0 - \beta^H_0) + \sum \beta^H_j (\bar{X}^M - \bar{X}^H) - \sum \bar{X}^M \Delta \beta,
\]

\[
(\bar{Y}^M - \bar{Y}^H) = \text{Total difference} = G_3,
\]

\[
\sum \beta^H_j (\bar{X}^M - \bar{X}^H) = \text{Part explained by differences in appropriations} = E_3,
\]

\[
(\beta^M_0 - \beta^H_0) - \sum \bar{X}^M (\beta^H_j - \beta^M_j) = \text{Unexplained part (discrimination proxy)}
\]

\[
= D_3.
\]

In this interpretation, which indexes female salary as “minuend” and male salary as “subtracting”, and maintains the hypothesis of negative discrimination

---

\(^6\) Equations \((3')\) and \((2')\).
against women, the understanding is as follows: In order for men’s wages to match those of women, their wages should be reduced by 5.92% in their incomes. In the explained part, the data show that if men had the same averages of women’s productive and non-productive characteristics, their wages would increase by approximately 18%.

Given the unexplained part, it can be said that due to the existence of discrimination, female incomes are 20.28% lower than they should be in the absence of discrimination, or that female incomes are 79.72% of what their income would have if the market remunerated its characteristics in the same way as the characteristics of men.

Model 4 – Assumes a positive discrimination hypothesis in favor of men (women’s salary as the market): Women’s equation as “minuend”

In this model the subtraction continues from equations (3’) and (2’), indexing female salary with “minuend” and male salary as “subtracting”, i.e. subtracting (2’) from (3’), but assumes possible positive discrimination in favor of men. Thus,

\[ \beta_j^H = \beta_j^M + (\beta_j^H - \beta_j^M), \]

being the discriminatory factor \( \Delta \beta = (\beta_j^H - \beta_j^M) \),

\[ (\bar{Y}^M - \bar{Y}^H) = (\beta_0^M - \beta_0^H) - \sum \bar{X}^H \beta_j^H + \sum \beta_j^M \bar{X}^M, \]  

(14)

\[ (\bar{Y}^M - \bar{Y}^H) = (\beta_0^M - \beta_0^H) - \sum \bar{X}^H (\beta_j^H + \Delta \beta) + \sum \beta_j^M \bar{X}^M, \]  

(15)

\[ (\bar{Y}^M - \bar{Y}^H) = (\beta_0^M - \beta_0^H) + \sum \beta_j^M (\bar{X}^M - \bar{X}^H) - \sum \bar{X}^H \Delta \beta, \]  

(16)

\[ (\bar{Y}^M - \bar{Y}^H) = \text{Total difference} = G_4, \]

\[ \sum \beta_j^M (\bar{X}^M - \bar{X}^H) = \text{Part explained by differences in appropriations} = E_4, \]

\[ (\beta_0^M - \beta_0^H) - \sum \bar{X}^H (\beta_j^H - \beta_j^M) = \text{Unexplained part (discrimination proxy)} = D_4. \]

It is assumed that the market wage is female. Thus, in order for men’s wages to match those of women, their wages should be reduced by −5.92%. The explained part (E4) states that if men were valued equal to women, their income would be 75.33% (5.28) of women’s income (7.01). The unexplained part points out that in a scenario of no positive discrimination, male incomes should fall by −29.12%.

In summary, the wage decomposition models 1, 2, 3 and 4 follow the reasoning of the decompositions presented by Oaxaca (1973) and Jann (2008). However, these models assume a priori that one of the groups experiences positive or
negative discrimination in the labor market. Depending on the choice of equation, such as “minuend” and/or “subtracting”, the results of the wage decomposition will differ and, mainly, the interpretations of wage impacts.

The following Model 5 materializes the contribution of this article to research in the area of labor economics. It provides a suggestion of wage decomposition in which the researcher does not have to assume that a particular group is discriminated, whether positive or negative. Besides contributing to the correction of the index problem, this wage decomposition also has the advantage in the interpretation of the results, being it simply and objectively.

The proposed wage decomposition and its interpretation start from a counterfactual approach that answers the following questions: what would happen to female (male) incomes if women (male) had the male (female) characteristics? What would happen to women’s (men’s) wages in the absence of discrimination?

Model 5 – Proposed solution to the Index Problem (Impact on Women’s Salaries) In this decomposition one can answer two questions: What would happen to women’s income if women had the masculine characteristics? And what would happen to women’s wages in the absence of discrimination? It is not assumed an advantage or disadvantage salary and therefore represents an alternative for correcting the index problem.

Given the group wage determination equations,

\[
Y_H = \beta_H^0 + \sum \beta_H^j X_H^{ji} + u_H^i, \tag{17}
\]

\[
Y_M = \beta_M^0 + \sum \beta_M^j X_M^{ji} + u_M^i, \tag{18}
\]

\[
E(u_H^i = u_M^i = 0), \tag{19}
\]

where \(Y_i\) corresponds to the natural logarithm of the hourly wage; \(\beta_j\) the coefficient vector; \(X_{ij}\) the matrix of explanatory variables of the model; and \(u_i\) the error terms vector. The superscript \(H\) if the group is male and the superscript \(M\) if the group is female.

Thus, individual salary averages can be written as

\[
\bar{Y}^{MH} = \bar{\beta}_H + \sum \bar{\beta}_H X_{ij}, \tag{20}
\]

\[
\bar{Y}^M = \bar{\beta}_M + \sum \bar{\beta}_M X_{ij}, \tag{21}
\]

where \(\bar{Y}^{MH}\) represents the value of women’s wages if they were valued equal to men’s.
Subtracting (21) from (20) there is
\[ \bar{Y}_{MH} - \bar{Y}_{M} \overset{DT}{=} \left( \hat{\beta}_H - \hat{\beta}_M \right) + \sum \hat{\beta}_H X^H_{ij} - \sum \hat{\beta}_M X^M_{ij}, \] (22)
where DT corresponds to the proxy for gender discrimination in the labor market, i.e. the difference that should be attributed to women’s salary in the absence of gender discrimination; the DE difference corresponding to the statistical differences, wage differences in the group averages, the only difference exclusively due to the fact that women were born as women, which corresponds to the proxy for statistical discrimination cited by Dickinson and Oaxaca (2006).

The other part, DC, refers to the different valuation between men and women in the productive and non-productive characteristics of individuals, which is also a proxy for discrimination. The total decomposition is given by \( DT = DE + DC \),
\[ \bar{Y}_{MH} - \bar{Y}_{M} \overset{DT}{=} \left( \hat{\beta}_H - \hat{\beta}_M \right) + \sum \hat{\beta}_H X^H_{ij} - \sum \hat{\beta}_M X^M_{ij} \] (23)

Thus, in a scenario where women were remunerated similarly to men their incomes would correspond to
\[ \bar{Y}_{MH} = \left\{ \left( \hat{\beta}_H - \hat{\beta}_M \right) + \sum \hat{\beta}_H X^H_{ij} - \sum \hat{\beta}_M X^M_{ij} \right\} + \bar{Y}_{M}. \] (24)

Equation (24) shows that: if \( \hat{\beta}_H = \hat{\beta}_M \) and \( \hat{\beta}_M = \hat{\beta}_H \), then \( \bar{Y}_{MH} = \bar{Y}_{M} \); that is, there would be no significant differences and no wage discrimination. If \( DT > 0 \) there would be discrimination against women.

To identify the portion of the difference corresponding to the explained part of the equation, we simulate what would happen to female incomes if they had the same characteristics as men:
\[ X^M_{ic} = \hat{\beta}_M + \sum \hat{\beta}_M X^H_{ij}, \] (25)
\[ Y^M = \hat{\beta}_M + \sum \hat{\beta}_M Y^M_{ij}, \] (26)
where \( X^M_{ic} \) represents the value of women’s income if they had the same male characteristics.
Starting from this scenario and subtracting (25) from (26) there is
\[
\bar{X}_{IC}^M = \bar{Y}^M - \hat{\beta}_M - \hat{\beta}_M + \sum \hat{\beta}_M \bar{X}_{ij}^H - \sum \hat{\beta}_M \bar{X}_{ij}^M
\]
(27)

In equation (28), it represents the wage differences explained by the differences in the characteristics of individuals. Therefore, in the hypothetical scenario, if women had the characteristics of men, their income would have an aggregate amount per $EC$. $EC > 0$ indicates that men have, on average, their characteristics superior to those of women; $EC < 0$ points out that females have their average characteristics superior to males, according to the next equation:

\[
\bar{X}_{IC}^M = \sum \hat{\beta}_M (\bar{X}_{ij}^H - \bar{X}_{ij}^M).
\]
(28)

Thus, the total difference $G$ is the sum of $DT$ plus $ET$, $G = ET + DT$:

\[
G = \left( \hat{\beta}_H - \hat{\beta}_M \right) + \sum \bar{X}_{ij}^M (\hat{\beta}_H - \hat{\beta}_M) + \sum \hat{\beta}_M (\bar{X}_{ij}^H - \bar{X}_{ij}^M).
\]
(30)

As mentioned earlier, one of the criticisms of models 1, 2, 3, and 4 is the hypothesis that the market pays men higher (discrimination against women) or that the market pays women with market salary and surcharges paid to men are the result of positive discrimination. The arbitrary choice of the type of discrimination, that is, the indexing problem is recurrent in these decompositions, as well as the choice of the equation that will represent the “minuend” and the

\[\text{\footnotesize A similar decomposition is presented by Blinder (1973), but also assumes a pre-discrimination of one of the studied groups.}\]

\[
(\bar{Y}^H - \bar{Y}^L) - \sum_j \beta_j^L (\bar{X}_{ij}^H - \bar{X}_{ij}^L) = \left( \beta_0^H - \beta_0^L \right) + \sum_j \bar{X}_{ij}^L (\beta_j^H - \beta_j^L) + \sum_j (\bar{X}_{ij}^H - \bar{X}_{ij}^L) (\beta_j^H - \beta_j^L).
\]

The superscript $H$ indicates “High-wage” of the group in advantage, and the superscript $L$ indicates “Low-wage” of the group in disadvantage. It also infers the possibility of decomposing a fourth term: $\sum_j (\bar{X}_{ij}^H - \bar{X}_{ij}^L) (\beta_j^H - \beta_j^L)$ that, according to the author, there is no clear economic explanation. This decomposition is also exposed by Jann (2008) and is called threefold.
“subtracting”.

The advantage of Model 5 is that the researcher need not admit that there is negative discrimination against women or that there is positive discrimination in favor of men. There is also no need to choose between “minuend” and “subtracting”, which facilitates interpretations of the estimated results. Thus, Model 5 answers the following question: what would happen to women’s incomes if they were remunerated in the same way as men?

The application to the data from Brazil, as shown in Table 2, determines that the total difference in wages was 0.0610 of the ln of the hourly wage of men and women. In percentual terms of women’s salaries, the difference was 6.29, that is, for their wages to match men’s wages these women’s earnings should increase by 6.29%.

The unexplained part (proxy of discrimination) corresponds to the differences in the coefficients applied to women’s characteristics, i.e., this effect shows what would happen to women’s salary if their characteristics were remunerated equally to male characteristics. According to the results, it can be seen that the value was positive, which shows that the male coefficients, $\beta^M_j$, are generally higher than the females, $\beta^F_j$, and since the average female characteristics are non-null values, it is evident that the male gender is valued by the market differently (superior) to the female gender. Thus, it can be said that in a non-discriminatory scenario, women’s salary should increase by 25.43%, based on their current earnings, i.e. they should receive approximately R$ 8.8 per hour on average.

The part explained is understood as follows: if women had the characteristics of men their wages would reduce by 24.66%, based on their current income, i.e., would receive only $ 5.28 per hour of work.

The proposed Model (5) is also of possible application to men, that is, if the researcher intends to study what would happen to their income if they were valued in the labor market equally to women, as it can be used in research beyond the labor market in studies comparing two distinct groups.

Once again, the syntheses of the studied models and their application to the study of discrimination in Brazil are shown in tables Table 1 and Table 2.

5 Conclusion

This study aimed to propose a wage decomposition model that perfects the Oaxaca–Blinder technique because it dispenses with the anticipated proposition of an advantage/disadvantage group as a reference, present in the traditional decomposition, that is, it deals with the solution to the index problem.
Table 1. Summary of Equations and Wage Decomposition Models.

| Model | Total Difference | Explained | Unexplained |
|-------|------------------|-----------|-------------|
| Lw(1) | \((\bar{Y}^H - \bar{Y}^M)\) \(\sum \beta^H_j (\bar{X}_j^M - \bar{X}_j^H) (\beta^H_0 - \beta^M_0) + \sum \bar{X}^M_j (\beta^H_j - \beta^M_j)\) | | |
| Lw(2) | \((\bar{Y}^H - \bar{Y}^M)\) \(\sum \beta^M_j (\bar{X}_j^H - \bar{X}_j^M) (\beta^H_0 - \beta^M_0) + \sum \bar{X}^H_j (\beta^H_j - \beta^M_j)\) | | |
| Lw(3) | \((\bar{Y}^M - \bar{Y}^H)\) \(\sum \beta^M_j (\bar{X}_j^M - \bar{X}_j^H) (\beta^M_0 - \beta^H_0) + \sum \bar{X}^H_j (\beta^H_j - \beta^M_j)\) | | |
| Lw(4) | \((\bar{Y}^M - \bar{Y}^H)\) \(\sum \beta^M_j (\bar{X}_j^M - \bar{X}_j^H) (\beta^M_0 - \beta^H_0) + \sum \bar{X}^H_j (\beta^H_j - \beta^M_j)\) | | |
| Lw(5) | \((\bar{Y}^{MH} - \bar{Y}^M)\) \(\sum \hat{\beta}_{ij} (\bar{X}_{ij}^H - \bar{X}_{ij}^M) (\hat{\beta}_H - \hat{\beta}_M) + \sum \bar{X}_{ij}^M (\hat{\beta}_H - \hat{\beta}_M)\) | | |

Source: Elaborated by the authors based on the models presented.

Table 2. Results of the usual Salary Decomposition Models and the proposed Model – Brazil, 2015.

| Group | Coefficients | Average Salary hour |
|-------|--------------|---------------------|
| Men   | 2.0084*      | 7.45                |
| Women | 1.9474*      | 7.01                |

| Models | Lw(1) | Lw(2) | Lw(3) | Lw(4) | Lw(5) |
|--------|-------|-------|-------|-------|-------|
| Differences Explained | Coefficients | -0.1656* | -0.2832* | 0.1656* | 0.2832* | -0.2832* |
| Impact % on Salary | -15.26* | -24.66* | 18.01* | 32.73* | -24.66* |

| Gender Discrimination | Coefficients | 0.2266* | 0.3442* | -0.2266* | -0.3442* | 0.2266* |
| Impact % on Salary | 25.43 | 41.08 | -20.28 | -29.12 | 25.43 |

| Total difference | Coefficients | 0.0610* | 0.0610* | -0.0610* | -0.0610* | 0.0610* |
| Impact % on Salary | 6.29 | 6.29 | -5.92 | -5.92 | 6.29 |

Source: Prepared by the authors based on PNAD 2015 data.
Notes: To minimize the problem of heteroscedasticity, common in wage determination models, we used the robust error procedure of White (1980). *significance up to 10%; (without asterisk) not significant.
Some researchers have suggested other wage decompositions, either by using the average of the study group coefficients as the non-discriminant coefficient, sometimes by wage decomposition in the wage distribution percentiles, or by wage decomposition for the entire wage distribution based on the wage density function. These techniques evolved from the traditional Oaxaca wage decomposition, but in some cases remained arbitrary in the choice of weights, and in others indifferent to the *index problem*. In addition, they present operational difficulties due to the need for a large computational capacity. Thus, this work was based on differences in wage averages, as proposed by Oaxaca–Blinder, and offered a solution to the problem of arbitrary choice of an advantage group.

Model 1 assumed the hypothesis of negative discrimination against women, with men’s salary being considered the market salary. Model 2 attributed, a priori, positive discrimination in favor of men and women’s salary as the market salary. Models 3 and 4 address the same hypotheses as the models mentioned above, based on negative discrimination against women and positive discrimination in favor of men, respectively, but show the minuend of the salary equations of females.

The substantial importance of this article lies in the proposition of a counterfactual analysis that does not assume in advance the hypothesis of positive or negative discrimination of a certain group, that is, solves the *index problem* (Model 5). The arbitrary choice of the advantage or disadvantage group is one of the most prevalent criticisms in the traditional wage decomposition of Oaxaca and Blinder. Moreover, it is based on the analysis of the current salaries of the groups analyzed and is not based on the income impacts that individuals should have, as usually shown in the other models presented.

The Oaxaca–Blinder wage decomposition procedure is used to analyze gender wage inequalities in the Brazilian labor market according to different models, which confirms the gender pay gap and gender discrimination in the country. At the same time, it makes possible the real understanding of the *index problem* and the proposed solution.

Thus, it is hoped, with this study, to assist in the decision-making of researchers about the empirical strategy to be used to measure inequalities in the labor market and especially in the interpretation of the results found.

**References**

Autor, D. H., Lawrence, F. K., & Melissa, S. K. (2005). *Rising wage inequality: The role of composition and prices* (Working Paper No. 2096). Harvard Institute of Economic Research.
Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy, 70*(5), 9–49.

Becker, G. S., & Chiswick, B. R. (1966). Education and the distribution of earnings. *American Economic Review, 56*(2), 358–369.

Biltagy, M. (2014). Estimation of gender wage differentials using Oaxaca decomposition technique. *Topics in Middle Eastern and North African Economies, 16*.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources, 8*(4).

Carvalho, A. P. d., Néri, M., & Silva, D. B. d. N. (2006). *Diferenciais de salários por raça e gênero no Brasil: Aplicação dos procedimentos de Oaxaca e Heckman em pesquisas amostrais complexas*. Rio de Janeiro: IBGE.

Cobb-Clark, D., & Tan, M. (2010). Noncognitive skills, occupational attainment, and relative wages [Forthcoming]. *Labour Economics, 18*(1), 1–13.

Cotton, J. (1988). On the decomposition of wage differentials. *The Review of Economics and Statistics, 70*(2), 236–243.

Dickinson, D., & Oaxaca, R. (2006, September). *Statistical discrimination in labor markets: An experimental analysis* (Working Paper No. 2305). Appalachian State University, IZA DP. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=933035

DiNardo, J., Fortin, N. M., & Lemieux, T. (1996). Labor market institutions and the distribution of wages 1973–1992: A semiparametric approach. *Econometrica, 64*(5), 1002–1044.

Doeringer, P. B., & Piore, M. J. (1970). *Internal labor markets and manpower analysis*.

Firpo, S., Fortin, N., & Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica, 77*(3), 953–973.

Fonseca, L. F. (2017). *Discriminação salarial por género no mercado de trabalho português* (Dissertação). Universidade do Porto, Portugal.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition methods in economics. In *Handbook of labor economics* (pp. 41–102).

Gomes, M. R. (2016). *Diferenças de rendimentos e alocação setorial por gênero no primeiro emprego, reemprego e remanescentes dos trabalhadores: Análise nas regiões Sul e Nordeste* (Dissertação de Mestrado). Centro de Estudos Sociais Aplicados, Universidade Estadual de Londrina, Londrina.

Heckman, J. J. (1979). Sample selection bias as a specification error (with an application to the estimation of labour supply functions). *Econometrica, 47*.

Jann, B. (2008). The Blinder–Oaxaca decomposition for linear regression models. *The Stata Journal, 8*(4), 453–479.
Juhn, C., Murphy, K. M., & Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of Political Economy, 101*(3), 410–442.

Loureiro, P. R. A. (2003). Uma resenha teórica e empírica sobre economia da discriminação. *Revista Brasileira de Economia, 57*(1), 125–157.

Machado, J., & Mata, J. (2005). Counterfactual decompositions of changes in wage distributions using quantile regression. *Journal of Applied Econometrics, 20*(4), 445–465.

Maia, e. a. (2015). A mão de obra feminina no mercado de trabalho brasileiro: Discriminação salarial por gênero. *Ciências Sociais em Perspectiva, 14*(26), 30–53.

Mariano, F. Z., Costa, E. M., Guimarães, D. B., & Sousa, D. T. (2018). Diferenciais de rendimentos entre raças e gêneros, nas regiões metropolitanas, por níveis ocupacionais: Uma análise através do pareamento de Ñopo. *Estudos Econômicos, 48*(1), 137–173.

Matos, R. S., & Machado, A. F. (2006). Diferencial de rendimentos por cor e sexo no Brasil (1987–2001). *Revista Econômica, 8*(1), 5–27.

Mattei, T. F., & Baço, F. M. B. (2017). Desigualdades salariais entre homens e mulheres no mercado de trabalho formal no Estado do Rio Grande do Sul. *Revista Latino Americana de Geografia e Gênero, 8*(2), 143–167.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy, 66*(4), 281–302.

Mincer, J. (1974). *Schooling, experience and earnings*. New York: National Bureau of Economic Research.

Neumark, D. (1988). Employers’ discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resources, 23*(3), 279–295.

Ñopo, H. (2008). Matching as a tool to decompose wage gaps. *The Review of Economics and Statistics, 90*(2), 290–299.

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

Oaxaca, R. L., & Ransom, M. R. (1994). On discrimination and the decomposition of wage differentials. *Journal of Econometrics*(61), 5–21.

Reich, M., Gordon, D. M., & Edwards, R. C. (1973). Dual labor markets: A theory of labor market segmentation. *American Economic Review, 63*(2), 359–365.

Reimers, C. W. (1983). Labor market discrimination against hispanic and black men. *The Review of Economics and Statistics, 65*(4), 570–579.

Schultz, T. W. (1961). Investment in human capital. *American Economic Review, 51*(1), 1–17.

Souza, e. a. (2015). Diferenças salariais por gênero e cor e o impacto da discriminação econômica. *Revista Brasileira de Estudos Regionais e Urbanos, 9*(1), 32–49.
Vietorisz, T., & Harrison, B. (1973). Labor market segmentation: Positive feedback and divergent development. *American Economic Review, 63*(2), 366–376.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838. [http://dx.doi.org/10.2307/1912934](http://dx.doi.org/10.2307/1912934)

Yun, M. (2005). A simple solution to the identification problem in detailed wage decompositions. *Economic Inquiry, 43*(4), 766–772.

**Appendix**

**Table A1.** Results of women and men mincerian regressions and means of individual characteristics.

| Variable               | $\hat{\beta}_H$ | $\hat{\beta}_M$ | $\bar{X}_H$ | $\bar{X}_M$ |
|------------------------|------------------|------------------|-------------|-------------|
| Ln salary hour         | −                | −                | 2.0084      | 1.9474      |
| Years of study         | 0.05046*         | 0.05976*         | 8.3773      | 9.9897      |
| Experience             | 0.02528*         | 0.02342*         | 23.6406     | 21.7828     |
| Experience$^2$         | −0.00028*        | −0.00029*        | 746.4116    | 642.1786    |
| White                  | 0.09953*         | 0.11685*         | 0.4319      | 0.4780      |
| Formal                 | 0.20537*         | 0.20006*         | 0.6013      | 0.6208      |
| Total industry         | 0.25658*         | 0.11156*         | 0.3453      | 0.1308      |
| Commerce               | 0.12587*         | 0.14958*         | 0.2191      | 0.2047      |
| Services               | 0.23900*         | 0.27697*         | 0.2892      | 0.6364      |
| Leaders                | 0.57540*         | 0.58817*         | 0.0535      | 0.0423      |
| PCAs                   | 0.71920*         | 0.57974*         | 0.0528      | 0.1484      |
| Average technician     | 0.31361*         | 0.28757*         | 0.0654      | 0.0811      |
| Urban                  | 0.14798*         | 0.23303*         | 0.8399      | 0.9163      |
| North                  | 0.19787*         | 0.20460*         | 0.0888      | 0.0699      |
| Southeast              | 0.31964*         | 0.31755*         | 0.4173      | 0.4481      |
| South                  | 0.32939*         | 0.32183*         | 0.1539      | 0.1583      |
| Midwest                | 0.38239*         | 0.26923*         | 0.0718      | 0.0679      |
| Federal District       | 0.62411*         | 0.70826*         | 0.0125      | 0.0167      |
| Mills                  | −0.32943*        | 0.65381*         | 0.0553      | 0.1169      |
| Constant               | 0.41522*         | −0.03744*        | −           | −           |

*Source:* Elaborated by the authors from PNAD (2015).

*Notes:* Base sector: agriculture; Base occupation: operational; Base macroregion: Northeast.