A new outlook on identification in detailed wage decompositions*

María Arrazola†

José de Hevia‡

†*Department of Economic Analysis, Universidad Rey Juan Carlos, Paseo Artilleros s/n 28032 Madrid, Spain (e-mail: maria.arrazola@urjc.es)*

‡*Department of Economic Analysis, Universidad Rey Juan Carlos, Paseo Artilleros s/n 28032 Madrid, Spain (e-mail: jose.dehevia@urjc.es)*

Abstract

Some articles have shown that the conventional Oaxaca's wage decomposition methodology poses a serious problem when categorical variables are included: the contribution of each dummy variable to the wage discrimination may vary. This article shows that this problem originates from an inadequate interpretation of the models and of the contributions of dummy variables to discrimination, and that with an adequate interpretation of the model the problem is easily solved. We also show that the proposal made by Gardeazabal and Ugidos (2004) for the solution of this problem, widely employed in empirical literature**,** contains some important drawbacks.

JEL: J31, J71

Key words: wage gap, Oaxaca's decomposition, gender differentials

Word count: 3585

* Financial support from URJC-CM (2008-CSH-3467) is gratefully acknowledged

I. Introduction

In the context of the analysis of wage differentials between different socioeconomic groups Oaxaca (1973) and Blinder (1973) proposed a decomposition of those differences into two components, one related to the differences in the characteristics of the individuals, and the other to the differences in the effects on the wages of those characteristics. This component is used as an estimate of wage discrimination. That proposal allows the calculation of the separate contribution to wage discrimination of each of the variables considered in the analysis. However, when the variables are categorical, it is not clear how their contribution to discrimination should be calculated. Oaxaca and Ransom (1999) show that "conventional decomposition methodology cannot identify the separate contributions of dummy variables to the wage decomposition, because it is only possible to estimate the relative effects of a dummy variable. So, the discrimination component is not invariant to the choice of the "left-out" reference group". To solve this problem, Gardeazabal and Ugidos (2004) suggest the incorporation into the analysis of a restriction on the parameters of the wage equation.

There are many empirical articles quoting the problem posed in Oaxaca and Ramson (1999) and those which use the solution proposed in Gardeazabal and Ugidos (2004) (some recent examples can be found in, for instance, Fortin, 2008, Simón et al. , 2008, Gang et al., 2008, Reagan and Oaxaca, 2009 and many others). For that reason, this article aims to clarify and solve the problem in the Oaxaca's decomposition when categorical variables are used. We suggest that at the back of the confusion in the literature with respect to the calculation of the contributions of categorical variables to discrimination, the problem is not one of identification, as proposed by Oaxaca and Ransom (1999) and Gardeazabal and Ugidos (2004), but, rather, it lies in an inadequate

interpretation of the models used, and of those contributions. We show that the problem is easily solved with an adequate interpretation of the model. Additionally, it will be seen that the proposal of Gardeazabal and Ugidos (2004) contains important drawbacks.

In Section II we discuss the problem of calculating the contributions to the discrimination component when categorical variables are employed. Section III discusses the Gardeazabal and Ugidos (2004) proposition. Section IV discusses the problem when several groups of categorical variables are considered. Section V is an illustration of the problem with an application to Spanish data. Section VI concludes the article.

II. Conceptual framework

Conceptual framework is habitual in the literature. As in Gardeazabal and Ugidos (2004) a linear regression model with a single set of dummy variables for *J* different educational levels has been considered:

$$w_g = \beta_{0g} + \sum_{j=1}^{J} \beta_{jg} D_{jg} + u_g \qquad (1)$$

where w_g is the log wage of a person belonging to group g, β_{0g} and β_{jg} are parameters, D_{jg} is a variable that takes the value 1 when the individual has studies in category *j* and 0 otherwise, and u_g is a disturbance term with $E\lfloor u_g / D_g \rfloor = 0$.

As the explanatory variables of (1) are not continuous, the parameters do not receive their usual interpretation: neither is β_{0g} an intercept in the traditional sense[1],

[1]In regression models with continuous explanatory variables the intercept is interpreted in two ways: (1) It fits the mean of the endogenous variable to the mean of the exogenous ones. (2) This is the expected value of the endogenous variable when all the

nor are the β_{jg}´s slope coefficients[2]. The parameters β_{0g} and β_{jg}´s together model *J* intercepts**,** one for each category *j*. Those parameters represent the mean value of the endogenous variable for *J* different subpopulations. The average wage for each *j* education level has two components: one common to all the levels, (β_{0g}), and another specific to the *j*-th education level, (β_{jg}). Starting from (1) and taking into account that $\sum_{j=1}^{J} D_{jg} = 1$, yields:

$$w_g = \beta_{0g} \sum_{j=1}^{J} D_{jg} + \sum_{j=1}^{J} \beta_{jg} D_{jg} + u_g \Rightarrow w_g = \sum_{j=1}^{J} (\beta_{0g} + \beta_{jg}) D_{jg} + u_g$$

So that:

$$E[w_g / D_{jg} = 1] = \beta_{0g} + \beta_{jg} \quad \forall j = 1,2,...,J$$

There is no single intercept in (1), but there really are *J* intercepts, one for each education level, each of them with the two components mentioned. Note that β_{0g} is part of the effect of each of the educational levels on the wages and not an intercept with an independent interpretation.

Neither of the two components of the average of each subpopulation can be

---

exogenous variables are simultaneously cancelled out. Neither of these two interpretations is suitable in this case since the dummy variables are capable of fitting the mean of the endogenous variable to the mean of the explanatory variables, without needing β_{0g}, and, because, it is impossible for all the dummy variables to be cancelled out simultaneously.

[2] As the dummy variables are not continuous, the β_{jg}´s cannot be interpreted as slopes of a regression line.

estimated separately since in (1) there is exact multicollinearity. However, for a study of wage differentials among social subgroups, a separate identification of β_{0g} and the β_{jg}´s is not necessary. What is relevant is not which part of the wages is common to all the individuals, and which part is specific to each subpopulation, but whether the wages are different between subpopulations, to find out the magnitude of the differences, and, in the case of the wage discrimination analysis, to determine to what extent belonging to a subpopulation implies suffering wage discrimination.

If the idea of estimating separately β_{0g} from the β_{jg}´s is abandoned, and it is assumed that what is relevant is to approximate the parameters δ_{jg} such as:

$$E\left[w_g / D_{jg} = 1\right] = \beta_{0g} + \beta_{jg} = \delta_{jg} \quad \forall j = 1,2,..., J$$

gives:

$$w_g = \sum_{j=1}^{J} \delta_{jg} D_{jg} + u_g \tag{2}$$

in which there is no exact multicollinearity problem. Taking into account that $\sum_{j=1}^{J} D_{jg} = 1$, (2) could be rewritten in many different ways. For example, taking $D_{1g} = 1 - \sum_{j=2}^{J} D_{jg}$, (2) becomes:

$$w_g = \gamma_{1g} + \sum_{j=2}^{J} \gamma_{jg} D_{jg} + u_g \tag{3}$$

with

$$\gamma_{1g} = \delta_{1g}$$

$$\gamma_{1g} + \gamma_{jg} = \delta_{jg}$$

There are different models analogous to (3) depending on the left-out category. Models like (3) are those which are usually employed in the analysis of wage

discrimination, and in which Oaxaca and Ransom (1999) and Gardeazabal and Ugidos (2004) point out that the conventional decomposition methodology cannot identify the separate contribution of dummy variables to wage discrimination.

The same as in (1), in (2) and (3) J intercepts are modeled, although they are represented differently in each model. There is no multicollinearity either in (2) or in (3), and, therefore, these models can be estimated without any problem. However, the interpretation of the parameters in each of the models is different. Whereas in (2) the effect on the expected wages of possessing a *j*-th education level is $E\left[w_g \middle| D_{jg} = 1\right] = \delta_{jg}$, in (3) it is $E\left[w_g \middle| D_{1g} = 1\right] = \gamma_{1g}$ if *j*=1 and $E\left[w_g \middle| D_{jg} = 1\right] = \gamma_{1g} + \gamma_{jg}$ if *j*=2, 3, …, *J* . It would not be correct to assign, for instance, in (3), only γ_{jg} if *j*=2, 3, …, *J* to the *j*-th group. When in models like (3) the effect of the educational level on the wages is assigned, attention should not solely be paid to the parameter linked to each dummy variable, because that parameter only includes part of that effect. The value of the parameter linked to each dummy variable changes with the left-out category because the interpretation of each of those parameters varies, but, in any case, the total effect on wages of belonging to a group with a *j* education level does not change. The origin of the confusion existing in the literature is, precisely, to consider that the effect of each educational level on the wages is only measured with the parameter linked to each dummy variable.

To find out the total effect of each dummy variable, it is enough to consider, in (3), that $\sum_{j=1}^{J} D_{jg} = 1$ and obtain a model similar to (2):

$$w_g = \gamma_{1g} \sum_{j=1}^{J} D_{jg} + \sum_{j=2}^{J} \gamma_{jg} D_{jg} + u_g \Rightarrow w_g = \gamma_{1g} D_{1g} + \sum_{j=2}^{J} \left(\gamma_{1g} + \gamma_{jg}\right) D_{jg} + u_g \qquad (4)$$

In the context of the decomposition of the difference in the average wages

between different groups proposed by Oaxaca (1973), if it is wished to prevent the problem of the variation in the contributions of dummy variables to discrimination, (2) or (4) should be used in calculations. However, (3) is generally employed, in which it is a mistake to assign the parameter linked to the dummy variable as the total effect of each education level.

Taking as an example the analysis of the wage differentials between men and women, and, starting from (3) and from the OLS estimates, the decomposition of the wage differences usually carried out is[3]:

$$\overline{w}_m - \overline{w}_f = \underbrace{\hat{\gamma}_{1m} - \hat{\gamma}_{1f} + \sum_{j=2}^{J}\left(\hat{\gamma}_{jm} - \hat{\gamma}_{jf}\right)\overline{D}_{jf}}_{Discri\,\min ation} + \underbrace{\sum_{j=2}^{J}\hat{\gamma}_{jm}\left(\overline{D}_{jf} - \overline{D}_{jm}\right)}_{Characteristics} \qquad (5)$$

where *g*= *m*(male), *f*(female), $\overline{w}_g$, are the sample averages of log wages, $\hat{\gamma}_{jg}$ are OLS estimates, and $\overline{D}_{jf}$ is the average value of a dummy variable.

In (5), what is usually done in the literature is to consider that $\left(\hat{\gamma}_{1m} - \hat{\gamma}_{1f}\right)$ is the contribution of the intercept to wage discrimination, and $\left(\hat{\gamma}_{jm} - \hat{\gamma}_{jf}\right)\overline{D}_{jf}$ is the contribution of each dummy variable to discrimination. In this context, Gardeazabal and Ugidos (2004) say that "the contribution of each variable to discrimination is not invariant to the left-out category" but, this result is held up by an incorrect interpretation of the parameters of (3), which does not take into account that the γ_{jg} only includes the differences in average wages with respect to the left-out category.

Given that $\sum_{j=1}^{J}\overline{D}_{jf} = 1$, it is obtained, from (5)**,** that:

[3]We present the descomposition based on the assumption that the estimated male wage structure is the nondiscriminatory standard.

$$\overline{w}_m - \overline{w}_f = \underbrace{\left(\widehat{\gamma}_{1m} - \widehat{\gamma}_{1f}\right)\left(\sum_{j=1}^{J} \overline{D}_{jf}\right) + \sum_{j=2}^{J}\left(\widehat{\gamma}_{jm} - \widehat{\gamma}_{jf}\right)\overline{D}_{jf}}_{Discri\min ation} + \underbrace{\sum_{j=2}^{J} \widehat{\gamma}_{jm}\left(\overline{D}_{jf} - \overline{D}_{jm}\right)}_{Characteristics} \Rightarrow$$

$$\overline{w}_m - \overline{w}_f = \underbrace{\left(\widehat{\gamma}_{1m} - \widehat{\gamma}_{1f}\right)\overline{D}_{1f} + \sum_{j=2}^{J}\left(\widehat{\gamma}_{1m} + \widehat{\gamma}_{jm} - \left(\widehat{\gamma}_{1f} + \widehat{\gamma}_{jf}\right)\right)\overline{D}_{jf}}_{Discri\min ation} + \underbrace{\sum_{j=2}^{J} \widehat{\gamma}_{jm}\left(\overline{D}_{jf} - \overline{D}_{jm}\right)}_{Characteristics} \quad (6)$$

In (6), that is the decomposition of wage differentials obtained from (4)**,** the contribution of each variable to discrimination is invariant to the left-out category. It has been demonstrated that, with a simple reparametrization of models like (3) in models like (4), the problem of variation in contributions can be avoided. The effect on average wages of having a j-th education level, explicitly included in (2) and (4), should not be confused with the parameters linked to the dummy variables in (3). This confusion is responsible for the contribution of each dummy variable to discrimination made in many works not being invariant to the left-out category.

III. Comments on the Gardeazabal and Ugidos proposal

To avoid the problem of the variation in the contribution to discrimination of each dummy variable, Gardeazabal and Ugidos (2004) propose to estimate (1) by assuming that $\sum_{j=1}^{J} \beta_{jg} = 0$. Starting from this assumption, they suggest the following decomposition:

$$\overline{w}_m - \overline{w}_f = \underbrace{\left(\hat{\beta}_{0m} - \hat{\beta}_{0f}\right) + \sum_{j=1}^{J}\left(\hat{\beta}_{jm} - \hat{\beta}_{jf}\right)\overline{D}_{jf}}_{Discri\min ation} + \underbrace{\sum_{j=1}^{J} \hat{\beta}_{jm}\left(\overline{D}_{jf} - \overline{D}_{jm}\right)}_{Characteristics} \quad (7)$$

We have shown that, in order to obtain invariant contributions to the discrimination component, it is not necessary to impose any restriction on the parameters of the model, but it is sufficient to reparametrize the model adequately. The

restriction on the parameters proposed in Gardeazabal and Ugidos (2004) is only needed if it is wished to identify the β_{0g} and β_{jg} separately but this is not necessary in most of the empirical applications because what is of interest is whether the wages are different between subpopulations. As well as being unnecessary, the proposal of Gardeazabal and Ugidos (2004) has three other drawbacks: difficulties in interpreting the results, arbitrariness of the restriction imposed, and possible biases in the results.

With respect to the first problem, it should be pointed out that Gardeazabal and Ugidos (2004) interpret that $\left(\hat{\beta}_{0m} - \hat{\beta}_{0f}\right)$ is the contribution of the intercept to the discrimination component, and that $\left(\hat{\beta}_{jm} - \hat{\beta}_{jf}\right)\overline{D}_{jf}$ is the contribution of each dummy variable to that component. However, this interpretation is not adequate. It can be observed that, in their context, $\hat{\beta}_{0g} = \dfrac{\sum_{j=1}^{J} \hat{\delta}_{jg}}{J}$ and it should be interpreted as an average of the wage averages per educational level, and the $\hat{\beta}_{jg}$ as differences with respect to that average. Neither $\left(\hat{\beta}_{0m} - \hat{\beta}_{0f}\right)$, nor the $\left(\hat{\beta}_{jm} - \hat{\beta}_{jf}\right)\overline{D}_{jf}$ are easy to interpret. They are, respectively, the contribution to the discrimination component of a difference in averages of wage averages, and the contribution of a difference in the differences in those averages.

The proposal of Gardeazabal and Ugidos ignores the fact that the intercept contains the effects of all the dummy variables. Their proposal does not provide the total effect on discrimination of belonging to the *j*-th subpopulation, which is what is relevant from an economic point of view. To correctly calculate the contribution of each dummy variable to discrimination, one only needs to take into account, in (7), that $\sum_{j=1}^{J} \overline{D}_{jf} = 1$ and to observe that the contribution attributed to the intercept is part of the

contribution for having a *j*-th educational level. This leads us to an expression similar to (6):

$$\overline{w}_m - \overline{w}_f = \underbrace{\sum_{j=1}^{J}\left(\hat{\beta}_{0m} + \hat{\beta}_{jm} - \left(\hat{\beta}_{jf} + \hat{\beta}_{0f}\right)\right)\overline{D}_{jf}}_{Discri\,\min ation} + \underbrace{\sum_{j=1}^{J}\hat{\beta}_{jm}\left(\overline{D}_{jf} - \overline{D}_{jm}\right)}_{Characteristics}$$

Additionally, and contrary to what Gardeazabal and Ugidos (2004) suggest, it is inadequate to consider that the global contribution of the dummy variables to the discrimination component is the sum of the contributions of each dummy variable, ignoring the intercept.

With regard to the second drawback, if instead of imposing $\sum_{j=1}^{J}\beta_{jg} = 0$ in (1), any other arbitrary restriction like $\sum_{j=1}^{J}\beta_{jg} = \mu$ is imposed, the results obtained in the decomposition of the discrimination component are exactly the same. This conditions any economic interpretation of (1) and (7). In fact, starting from (1):

$$E\left[w_g / D_{jg} = 1\right] = \beta_{0g} + \beta_{jg} = \delta_{jg} \quad \forall j = 1,2,...,J \Rightarrow$$

$$\sum_{j=1}^{J} E\left[w_g / D_{jg} = 1\right] = \sum_{j=1}^{J}\left(\beta_{0g} + \beta_{jg}\right) = \sum_{j=1}^{J}\delta_{jg} \quad \Rightarrow$$

$$J\beta_{0g} + \sum_{j=1}^{J}\beta_{jg} = \sum_{j=1}^{J}\delta_{jg} \Rightarrow$$

If $\sum_{j=1}^{J}\beta_{jg} = \mu$:

$$\beta_{0g} = \frac{\sum_{j=1}^{J} \delta_{jg}}{J} - \frac{\sum_{j=1}^{J} \beta_{jg}}{J} = \frac{\sum_{j=1}^{J} \delta_{jg}}{J} - \frac{\mu}{J}$$

and

$$\beta_{jg} = \delta_{jg} - \left(\frac{\sum_{j=1}^{J} \delta_{jg}}{J} - \frac{\mu}{J} \right)$$

yields:

$$\beta_{0m} - \beta_{0f} = \left(\frac{\sum_{j=1}^{J} \delta_{jm}}{J} - \frac{\mu}{J} \right) - \left(\frac{\sum_{j=1}^{J} \delta_{jf}}{J} - \frac{\mu}{J} \right) = \frac{\sum_{j=1}^{J} \delta_{jm} - \sum_{j=1}^{J} \delta_{jf}}{J}$$

and

$$\beta_{jm} - \beta_{jm} = \left[\delta_{jm} - \left(\frac{\sum_{j=1}^{J} \delta_{jm}}{J} - \frac{\mu}{J} \right) \right] - \left[\delta_{jf} - \left(\frac{\sum_{j=1}^{J} \delta_{jf}}{J} - \frac{\mu}{J} \right) \right] = \left(\delta_{jm} - \frac{\sum_{j=1}^{J} \delta_{jm}}{J} \right) - \left(\delta_{jf} - \frac{\sum_{j=1}^{J} \delta_{jf}}{J} \right)$$

The decomposition of the discrimination component is always the same regardless of the value of μ. This result is maintained in terms of the OLS estimates.

With respect to the third problem, if the parameters of the different subgroups do not exactly fulfill the same restriction, $\sum_{j=1}^{J} \beta_{jg} = \mu \quad \forall g$, the contributions of dummy variables to discrimination calculated imposing the same restriction on all the groups *g*,

are not adequate[4]. Assuming that $\sum_{j=1}^{J}\beta_{jm}=\mu_m$ and $\sum_{j=1}^{J}\beta_{jf}=\mu_f$ with $\mu_m=\mu_f+\pi$ and $\pi\neq 0$. Then:

$$\beta_{0m}-\beta_{0f}=\left(\frac{\sum_{j=1}^{J}\delta_{jm}}{J}-\frac{\mu_m}{J}\right)-\left(\frac{\sum_{j=1}^{J}\delta_{jf}}{J}-\frac{\mu_f}{J}\right)=$$

$$=\left(\frac{\sum_{j=1}^{J}\delta_{jm}-\sum_{j=1}^{J}\delta_{jf}}{J}\right)-\left(\frac{\pi}{J}\right)\neq\left(\frac{\sum_{j=1}^{J}\delta_{jm}-\sum_{j=1}^{J}\delta_{jf}}{J}\right)$$

$$\beta_{jm}-\beta_{jm}=\left[\delta_{jm}-\left(\frac{\sum_{j=1}^{J}\delta_{jm}}{J}-\frac{\mu_m}{J}\right)\right]-\left[\delta_{jf}-\left(\frac{\sum_{j=1}^{J}\delta_{jf}}{J}-\frac{\mu_f}{J}\right)\right]=$$

$$=\left(\delta_{jm}-\frac{\sum_{j=1}^{J}\delta_{jm}}{J}\right)-\left(\delta_{jf}-\frac{\sum_{j=1}^{J}\delta_{jf}}{J}\right)+\left(\frac{\pi}{J}\right)\neq\left(\delta_{jm}-\frac{\sum_{j=1}^{J}\delta_{jm}}{J}\right)-\left(\delta_{jf}-\frac{\sum_{j=1}^{J}\delta_{jf}}{J}\right)$$

The contribution calculated by erroneously assuming that $\mu_m=\mu_f$ is biased. The magnitude and the bias sign depend on the value and sign of π.

IV. Several sets of dummy variables

If we consider, for instance, two groups of dummy variables, we have:

[4] Gardeazabal and Ugidos (2004) indicate that, for the case of men-women wage differences, possibly $\sum_{j=1}^{J}\beta_{jm}\neq\sum_{j=1}^{J}\beta_{jf}$.

$$w_g = \beta_{0g} + \sum_{j=1}^{J} \beta_{Djg} D_{jg} + \sum_{i=1}^{I} \beta_{Zig} Z_{ig} + \sum_{j=1}^{J} \sum_{i=1}^{I} \beta_{jig} D_{jg} Z_{ig} + u_g \quad (8)$$

where w_g is the log wage of a person belonging to group g, β_{0g}, β_{Djg}, β_{Zjg} and β_{ijg} are parameters, D_{jg} is a variable that takes the value 1 when the individual has studies in category *j* and 0 otherwise, Z_{ig} is a variable that takes the value 1 when the individual is in category *i* and 0 otherwise, and u_g is a disturbance term with $E[u_g / D_g, Z_g] = 0$.

The interpretation of the parameters of (8) is similar to that of (1). In (8) there are different *J*x*I* intercepts, one for each subpopulation defined by the values of the dummy variables. Starting from (8), it is impossible to estimate separately all the parameters of the model. The idea of estimating separately the β_{0g} from the β_{Djg}, β_{Zjg} and β_{ijg} can be rejected, assuming that what is relevant is to approximate the parameters δ_{ijg}:

$$w_g = \sum_{j=1}^{J} \sum_{i=1}^{I} (\beta_{0g} + \beta_{Djg} + \beta_{Zig} + \beta_{ijg}) D_{jg} Z_{ig} + u_g \Rightarrow$$

$$w_g = \sum_{j=1}^{J} \sum_{i=1}^{I} \delta_{jig} D_{jg} Z_{ig} + u_g \quad (9)$$

Taking into account that $\sum_{j=1}^{J} D_{jg} = \sum_{i=1}^{I} Z_{ig} = 1$, (9) could be rewritten in many different ways. For example, considering that $D_{1g} = 1 - \sum_{j=2}^{J} D_{jg}$ and that $Z_{1g} = 1 - \sum_{i=2}^{I} Z_{ig}$, it is obtained that:

$$w_g = \gamma_{11g} + \sum_{j=2}^{J} \gamma_{Djg} D_{jg} + \sum_{i=2}^{I} \gamma_{Zig} Z_{ig} + \sum_{j=2}^{J} \sum_{i=2}^{I} \gamma_{jig} D_{jg} Z_{ig} + u_g \qquad (10)$$

with

$$\delta_{11g} = \gamma_{11g} = \beta_{0g} + \beta_{D1g} + \beta_{Z1g} + \beta_{11g}$$

$$\delta_{jig} = \gamma_{11g} + \gamma_{Djg} + \gamma_{Zig} + \gamma_{jig} = \beta_{0g} + \beta_{Djg} + \beta_{Zig} + \beta_{ijg}$$

Estimating (9) or (10), the different *J*x*I* intercepts of the model could be estimated, but what is not possible is to estimate the parameters of (8).

V. An Illustration for the Spanish Gender Wage Gap

To illustrate the discussion of previous sections, a sample of wage earners from the European Household Panel for Spain for the year 2000 was considered. We had data from 7494 men and 4684 women. Men had, on average, 14.5% higher wages than those of the women.

A wage model has been considered into which 8 dummy variables relative to educational levels were incorporated: no studies (EDU1), primary (EDU2), general lower secondary (EDU3), vocational lower secondary (EDU4), vocational upper secondary (EDU5), general upper secondary (EDU6), short cycle university (EDU7) and long cycle university (EDU8).

The models were estimated by OLS. The results of the decomposition of the discrimination component are shown in Table 1. The columns show the percentage contribution assigned to each variable over the total wage difference. Columns (I) to (VIII) show the contribution assigned to each educational level using (5) making the reference group vary, which is what has customarily been done in the empirical literature and where the problem established in Oaxaca and Ramson (1999) arose. Column (IX) contains the result obtained following Gardeazabal and Ugidos (2004),

and column (X) the result imposing the arbitrary restriction $\sum_{j=1}^{J} \beta_{jg} = -30$. Column (XI) shows the contribution calculated in accordance with (6), which is the decompositon of the wage differences obtained from a model like (4), one of the reparametrizations of (3) that we proposed in section II.

In columns (I) to (VIII), it can be seen that the contributions of the intercept and of each dummy variable vary depending on the left-out variable. The origin of those variations lies in an inadequate assignation of contributions to each dummy variable. It would be incorrect to say, for example, that the contribution of EDU1 is nil when EDU1 is the reference group, but 0.62 when EDU2 is the reference group. Those contributions are either badly calculated, or, rather, badly interpreted, because the contribution assigned to the intercept is part of the contribution from all the dummy variables. When those contributions are correctly assigned using expression (6), the result in column (XI) is obtained.

In columns (IX) and (X), it is seen that the decomposition proposed by Gardeazabal and Ugidos (2004), and that obtained by imposing that $\sum_{j=1}^{J} \beta_{jg} = -30$, lead to the same result, displaying their arbitrary nature and how difficult they are to interpret. The contribution of the intercept contains contributions of all the dummy variables and if those contributions are re-distributed adequately, we obtain the invariant decomposition (column XI). Contrary to what has been suggested by Gardeazabal and Ugidos (2004), it is incorrect to consider that the global contribution of the dummy variables to discrimination is the sum of the contributions of dummy variables, ignoring the intercept. It is also wrong to say that the contribution of education to the discrimination factor is negative (-5.03), and that, therefore, there would be discrimination against men. On the contrary, and as can be seen in column XI, the total

contribution to the discrimination component of education is positive, although the magnitude varies per educational level.

VI. Conclusions

With regard to the problem pointed out in Oaxaca and Ransom (1999) related to the decomposition of wage differentials and the use of dummy variables, there are three possible solutions. One is to take into account that, in a model with dummy variables, multiple intercepts are being modeled, and that it is necessary to analyze the differences between those intercepts (one per subpopulation). From this perspective, it is sufficient to make an adequate interpretation of the parameters of the wage equations used in discrimination analysis. The second solution would be to impose any *ad hoc* restriction in the analysis, like the one suggested by Gardeazabal and Ugidos (2004). In this case, the results are conditioned by the restriction imposed, and, it should be borne in mind that the total contribution of each variable to the discrimination factor is not calculated. The third solution could be to abandon a detailed decomposition, knowing that the global discrimination component will not be affected by this problem (see Oaxaca and Ransom, 1999).

References

Blinder, A.S. (1973). 'Wage Discrimination: Reduced Form and Structural Estimates', *Journal of Human Resources*, Vol. 8, pp. 436-455.

Fortin, N.M. (2008). 'The Gender Wage Gap among Young Adults in the United States. The Importance of Money versus People', *Journal of Human Resources*, Vol. 43, pp. 884-918.

Gang, I.N., Sen, K. and Yun, M. S. (2008). 'Poverty in rural India: caste and tribe', *Review of Income and Wealth,* Vol. 54, pp. 50-70.

Gardeazabal, J. and Ugidos, A. (2004). 'More on Identification in detailed wage decompositions', *The Review of Economics and Statistics,* Vol. 86, pp. 1034-1036.

Oaxaca, R.L. (1973). 'Male-Female Wage Differentials in Urban Labor Markets', *International Economic Review,* Vol. 14, pp. 693-709.

Oaxaca, R.L. and Ransom, M. R. (1999). 'Identification in Detailed Wage Decomposition', *The Review of Economics and Statistics,* Vol. 81, pp. 154-157.

Reagan, T.L. and Oaxaca, R. L. (2009). 'Work experience as a source of specification error in earnings models: implications for gender wage decomposition', *Journal of Population Economics,* Vol. 22, pp. 463-499.

Simón, H., Sanromá, E. and Ramos, R. (2008). 'Labour segregation and inmigrant and native-born wage distributions in Spain: an analysis using matched employer-employee data', *Spanish Economic Review,* Vol. 10, pp. 135-168.

TABLE 1

Contribution of Dummy Variables to the Wage Decomposition

	Reference Groups								*GU*	*Arbitrary*	*Invariant*
Contribution	*EDU1* (I)	*EDU2* (II)	*EDU3* (III)	*EDU4* (IV)	*EDU5* (V)	*EDU6* (VI)	*EDU7* (VII)	*EDU8* (VIII)	(IX)	(X)	(XI)
Intercept	207.39	182.13	170.82	130.12	163.97	145.45	103.72	119.51	152.89	152.89	
EDU1	0	0.62	0.90	1.90	1.07	1.52	2.55	2.16	1.34	1.34	5.09
EDU2	-2.92	0	1.30	6.01	2.10	4.24	9.05	7.23	3.38	3.38	21.03
EDU3	-8.37	-2.59	0	9.32	1.57	5.81	15.36	11.74	4.11	4.11	39.10
EDU4	-7.04	-4.74	-3.71	0	-3.09	-1.40	2.41	0.97	-2.08	-2.08	11.86
EDU5	-5.12	-2.14	-0.81	3.99	0	2.18	7.10	5.24	1.31	1.31	19.32
EDU6	-8.22	-4.87	-3.37	2.02	-2.46	0	5.54	3.44	-0.99	-0.99	19.31
EDU7	-15.98	-12.09	-10.34	-4.07	-9.29	-6.42	0	-2.43	-7.58	-7.58	15.98
EDU8	-11.88	-8.46	-6.93	-1.43	-6.01	-3.50	2.13	0	-4.50	-4.50	16.15
EDU1 to EDU8	-59,52	-34,26	-22,97	17,74	-16,11	2,41	44,14	28,35	-5,03	-5,03	147,86
Total	147.86	147.86	147.86	147.86	147.86	147.86	147.86	147.86	147.86	147.86	147.86

Note: EDU1, EDU2, ... , EDU8 represent educational levels. GU means Gardeazabal and Ugidos (2004).