This document is published in:

Journal of Applied Econometrics (2008), 23(4), 463–485.

DOI:10.1002/jae.1006

A NONPARAMETRIC DECOMPOSITION OF THE MEXICAN AMERICAN AVERAGE WAGE GAP

RICARDO MORA*
Universidad Carlos III de Madrid, Madrid, Spain

SUMMARY

This paper shows that average wage gap decompositions between any two groups of workers can be carried out using nonparametric wage structures. It also proposes an algorithm to correct for sample selection in nonparametric models known as tree structures. This paper studies the wage gap between third-generation Mexican American and non-Hispanic white workers in the southwest. It is shown that the decomposition heavily depends on functional assumptions, and that different aproaches to flexibility may render sufficiently good and similar results.

1. INTRODUCTION

Data from the US Current Population Survey shows that in 1995 Mexican male workers in the USA earned an average 36.5% lower hourly wage than non-Hispanic white workers. Even for Mexican Americans born in the USA, the wage gap still averaged 19.4%. This paper follows a long tradition in trying to explain the sources of this differential. Its main contribution and novelty lie in proposing a decomposition of the wage gap which is less dependent on functional form assumptions.

In the literature on wage discrimination, observed wage differentials between two groups of workers are decomposed into several components. The first contributions to average wage decompositions by Blinder (1973) and Oaxaca (1973) propose a methodology to identify a component of the wage differential which reflects the effect on the wage gap of human capital differences between the groups. The rest of the wage gap can then be interpreted as discrimination, since it provides an estimation of that part of the wage gap which arises because workers in at least one of the groups obtain wages which are different from those their human capital would normally comand. Several models on labor market discrimination, *inter alia*, Thurow (1969), Bergmann (1971), and Madden (1975), suggest that this effect should be further decomposed into two components: the cost imposed to the minority and the benefit obtained by the majority. Another source for the observed wage gap arises because the sample of workers in both groups is not random. Reimers (1983) forcefully argues that if unobserved factors in the decision to participate in the labor market are related to unobserved productivity differentials, then average wages across groups will also differ due to a sample selection effect.

It has long been recognized that the implementation of wage gap decompositions is not without difficulty. First, the variable specification must be complete in the sense that all systematic effects

* Correspondence to: Ricardo Mora, Dpto. Econom´ıa, Universidad Carlos III de Madrid, 126, 28903 Getafe, Spain.

E-mail: ricmora@eco.uc3m.es

of human capital on productivity are conveniently measured. Second, sample selection must be adequately addressed, if only to assess the importance of the participation component and to obtain consistent estimates. Third, decomposing the discrimination effect into favoritism and discrimination entails an index number problem as it requires a non-testable assumption on the characteristics of the wage structure without discrimination. Finally, assigning the sample selection effect to either discrimination or productivity differentials also requires further assumptions on the participation decision determinants, as shown by Neuman and Oaxaca (2004a).

Yet, in spite of the abundant empirical literature developed in the last 30 years, the role of functional form specification seems to have been overlooked. The basic decomposition between a *human-capital* and a *discrimination* term assumes linearity in the wage structure and ordinary least squares (OLS) estimates with a constant term. More complex decompositions still rely on at least a linear parametric structure for the wage equations. In what constitutes the main methodological contribution of this paper, the Blinder–Oaxaca decomposition is generalized to nonparametric wage structures with additive error terms. It will be argued that this decomposition can be carried out with any nonlinear or nonparametric multivariate model and the usefulness of each model will likely depend on each particular case. In the empirical application of the paper a particular model based on tree structures is proposed and estimated. Models with tree structures have three appealing properties in the context of wage equations. First, the Mincer-like wage equations usually implemented for each group in studies on wage discrimination are a particular case of a tree wage structure. Second, the estimated wage structures—the terminal nodes in the trees—can be interpreted as local labor markets. Finally, in what constitutes the second methodological contribution of the paper, it is shown that the sample selection correction can be implemented and that identification of the parameters in the wage equations follows from restrictions which can be interpreted as exclusion restrictions.

The main empirical contribution of this paper is to assess with some highly nonlinear wage equations and with tree wage structures the consistency of wage gap decompositions between Mexican American and non-Hispanic workers obtained from simple functional specifications. In particular, wage gaps observed in a sample from 1994 to 2002 between third-generation Mexican American and non-Hispanic white male workers in the four border states with Mexico (i.e., California, Arizona, New Mexico and Texas) are decomposed using different functional specifications ranging from a very simple Mincer-like equation to highly nonlinear wage equations and tree wage structures.

In the past, substantial efforts have been made to study wage differentials between Hispanics and Mexicans with respect to non-Hispanic white workers. Early results on wage differentials amongst Hispanic and white non-Hispanic can be found in Fogel (1966) and Poston and Alvirez (1973). More recently, Reimers (1983), Verdugo (1992), Cotton (1993), and Trejo (1997) have studied the wage gap between black, Hispanic or Mexican, and non-Hispanic white workers. The second and fourth contributions focus on Mexican American workers. All studies estimate linear wage functions separately for each group including second-order terms for experience, and then carry out Blinder–Oaxaca wage gap decompositions.

As pointed out by Trejo (1997), there is no general consensus on the economic prospects for Mexican American workers. The difficulty in reaching generally accepted results derives in part from the complexities of the processes of assimilation of Mexican Americans in American society. The large inflows of immigrants from Mexico with very low human capital during the 1980s and 1990s may bias time series studies on the evolution of the wage gap. First, there is an obvious danger of capturing a pessimistic picture of the Mexican workers' future in the American labor

market by including all recent legal immigrants, a group with reported low skills. Unreported illegal immigration, on the other hand, is also bound to bias analysis of the trends for legal workers. Not only do we have the short-term effect of the large inflows of low-skilled workers in the labor market, but also there are other unexpected effects. For example, Pagan and Davila (1996) have found that the Immigration Reform and Control Act of 1986 reduced the true wages of male natives most likely to be mistaken as unauthorized, that is, those with low human capital in terms of education and English proficiency.

This paper closely follows Trejo (1997) by focusing on third-generation workers. This group includes all workers born in the USA with at least one parent also born in the USA.[1] Since third-generation Mexican Americans have had ample time to adapt to the US labor market, short-term adjustments to large flows of Hispanic immigration in the labor market should not constitute a serious problem. It will also be argued that English proficiency for third-generation workers is the same for Mexican Americans and other workers.

A number of interesting conclusions stem from the application of multivariate nonparametric techniques to the wage gap between Mexican American and white male workers. First, regarding observed productivity differentials, results are in line with those previously found in the literature. If Mexican American workers had the same observed characteristics as white non-Hispanic workers, then the wage gap would decrease between 14 and 22 points, depending on the estimated functional form of the wage equation. Second, similar models in terms of goodness of fit lead to quite different estimates of the discrimination component of the average wage gap decomposition. This result hinges not only on the differences in variable specification, as is usually claimed in the literature, but also on the differences in functional form.

The rest of the paper is organized as follows. Section 2 presents the standard decompositions carried out in the literature, and extends them to the case of nonparametric structures. Section 3 describes the data and presents results of the estimations and decompositions. The paper ends with a summary of the findings and briefly discusses their implications for efforts to improve the economic situation of Mexicans in the US.

2. DECOMPOSITION OF AVERAGE WAGE DIFFERENTIALS

2.1. Linear Wage Structures

Under the linear parametric approach, the average wage gap between non-Hispanic white male workers (from now on indexed by W) and Mexican American male workers (indexed by M) can be decomposed between a *human-capital* and a *discrimination* component after fitting a linear wage function for each group. Let w_i be the logarithm of the wage for male worker i and let x_i be a column vector of worker i's observed characteristics which also includes a constant term. Then, the wage equation for each group can be conveniently expressed as

$$w_i = \sum_{j=W,M} (x_i' b_j) \cdot I\{i \in j\} + e_i \tag{1}$$

[1] That is, at least by one line of descent, it is the third generation. This is a less restrictive condition than in Trejo (1997), where third-generation workers are those with both parents born in the USA. In practice, only 22% of Mexican workers with one parent born in the USA do not have both parents born in the USA.

where the term e_i can be interpreted as unobserved productivity and the indicator function $I\{i \in j\}$ takes the value 1 if worker i belongs to group j, $j = W, M$, and 0 otherwise. The wage gap between workers in group W and workers in group M is defined as

$$\overline{w}_W - \overline{w}_M = (\overline{x}'_W b_W + \overline{e}_W) - (\overline{x}'_M b_M + \overline{e}_M) \tag{2}$$

where the bar superscript stands for the average operator and $\overline{x}_j$ and $\overline{e}_j$ denote the average values within the sample of group j workers of the vector of observed characteristics and unobserved productivities, respectively. Assuming that wages would also follow a linear structure in the nondiscriminatory case, $b(x) = x'b$, then the wage gap $\overline{w}_W - \overline{w}_M$ can be decomposed into four components:

$$\overline{w}_W - \overline{w}_M = (\overline{x}'_W - \overline{x}'_M)b + \overline{x}'_W(b_W - b) + \overline{x}'_M(b - b_M) + \overline{e}_W - \overline{e}_M \tag{3}$$

The first component in the right-hand side of the equation, $(\overline{x}'_W - \overline{x}'_M)b$, measures the effect of different average characteristics and, thus, it measures the effect of observed productivity differentials. The following two components capture differences in the wage premia assigned to each group. Following Oaxaca (1973), if $\overline{x}'_W(b_W - b)$ is positive, we can say that group W benefits from favoritism, while a positive sign in $\overline{x}'_M(b - b_M)$ would indicate that group M suffers discrimination. Finally, the last component $\overline{e}_W - \overline{e}_M$ reflects differences across groups between averaged unobserved productivity differentials.

The decomposition in equation (3) cannot be computed without estimation because b, b_m, and b_w are unknown vectors. Furthermore, estimation of equation (1) only provides estimates for b_M and b_W, while b, the non-discriminatory wage structure, is not identified without further assumptions. In this respect, extending the employer discrimination model of Arrow (1972) and Becker (1957), Neumark (1988) shows that if employers only care about the proportion of workers from group M and from group W within each type of labor x, the nondiscriminatory wage is a weighted average of each group's wage. More specifically, let $\alpha(x)$ be the proportion of group M workers within the total number of type x workers, then:

$$b(x) = \alpha(x)b_M(x) + (1 - \alpha(x))b_W(x) \tag{4}$$

where $b_M(x)$ and $b_W(x)$ are the wages for type x workers who belong to group M and group W, respectively. Equation (4) shows that the nondiscriminatory wage structure will, in general, not be linear in the workers' characteristics even if the actual wage structures, $b_M(x)$ and $b_W(x)$, are. In order to obtain a linear approximation to the Neumark nondiscriminatory structure to carry out the decomposition in equation (3), Neumark (1988) suggests fitting, based on a weighted least squares criterion, the estimated marginal productivity $\widehat{b(x)}$ on the workers' characteristics, X. He then shows that this procedure is equivalent to the linear estimation of the nondiscriminatory wage structure by using a pool regression with all individual observations.[2]

The Neumark pool decomposition, as usually named, identifies b, b_M, and b_W as the OLS estimates obtained from the pool sample, the sample of group M workers, and the sample of

[2] Oaxaca and Ransom (1994) show that Neumark's pooled decomposition is a generalization of previous proposals by Oaxaca (1973), Reimers (1983), and Cotton (1988).

group W workers, respectively. Since a constant term is included in the OLS estimates, the fourth component in equation (3) is identified to be zero.

If unobserved factors in the decision to participate in the labor market are related to unobserved productivity, then average wages across groups may also differ due to a sample selection effect. The standard two-stage Heckman procedure has been advocated by Reimers (1983) and Neuman and Oaxaca (2004a, 2004b). Model (3) is extended into a two-equation linear model of participation and wage determination among employed workers:

$$w_i = \sum_{j=W,M} (x_i' b_j) \cdot I\{i \in j\} + e_i \tag{5}$$

$$y_i = \sum_{j=W,M} (\gamma_j' z_i) \cdot I\{i \in j\} + u_i \tag{6}$$

The latent variable y_i is related to the decision to participate in the labor market so that if $y_i > 0$, then the individual works and earns a wage w_i. z_i is a vector of the determinants of the participation decision and γ_j are the associated parameter vectors. e_i and u_i are i.i.d. error terms that are assumed to follow a bivariate normal distribution with $\text{var}(e_i|j) = \sigma_{ej}$, $\text{var}(u_i|j) = \sigma_{uj}$, and $\text{cov}(e_i, u_i|j) = p_j$, $j = M, W$. Under these conditions, it is well known (see Heckman, 1979) that the expected value of the observed wage is

$$E[w_i|y_i > 0] = \sum_{j=W,M} (x_i' b_j + \rho_j \sigma_{ej} \lambda_j) \cdot I\{i \in j\} \tag{7}$$

where λ_j is the inverse of Mills' ratio associated with group j. No exclusion restriction is strictly necessary to identify the selection bias in the bivariate normal case, as identification is ensured by the nonlinearity of the inverse of Mills' ratio. Reimers (1983) applies to each group Heckman's (1979) two-stage procedure to decompose the observed average wage gap controlling for the average effect of different selectivity bias:[3]

$$\overline{w}_W - \overline{w}_M = (\overline{x}_W' - \overline{x}_M')\widehat{b} + \overline{x}_W'(\widehat{b}_W - \widehat{b}) + \overline{x}_M'(\widehat{b} - \widehat{b}_M) + \widehat{c}_W \overline{\lambda}_W - \widehat{c}_M \overline{\lambda}_M \tag{8}$$

where again $\widehat{b}$ is identified by two-stage estimation on the pool sample.

2.2. Nonparametric Wage Structures

Reasonable linear approximations of the true wage structure may still result in important biases in the estimation of components in average wage gap decompositions. To illustrate this with a simple numerical example, suppose that there are two groups of workers, W and M. Assume that log wages w_i are a nonlinear function of a variable x_i which only takes three values, $x_i = 0, 1, 2$:

$$w_i = 1 \cdot I(x_i < 2) + 6 \cdot I(x_i = 2, i \in W) + 5 \cdot I(x_i = 2, i \in M)$$

Figure 1 shows the wages for both groups of workers. Assume that there are four group M workers and four group W workers. In each group there is at least one worker with $x_i = 0, 1, 2$.

[3] Neuman and Oaxaca (2004a) further propose alternative ways to decompose the sample selection effect, $\widehat{c}_W \overline{\lambda}_W - \widehat{c}_M \overline{\lambda}_M$, in order to isolate differences in wages which stem from discrimination at the participation decision.

Two group M workers have $x_i = 0$, while two group W workers have $x_i = 1$. Since workers' characteristics in the sample only differ across groups for $x_i < 2$—values for which wages are the same across groups—the variable x_i actually does *not* contribute to the average wage differential and the wage gap should be uniquely attributed to discrimination against type M workers—or favoritism for W workers—with $x_i = 2$. However, a first-order linear decomposition of the wage gap (see Figure 1) will erroneously conclude that the productivity differentials are large and against group W workers with $x_i = 0, 1$. Overall, discrimination against M workers will be heavily underestimated in this example. Of course, the bias arises from a well-known functional form problem in the wage equations, and will persist in the decompositions of wage gaps regardless of the chosen nondiscriminatory wage structure.[4]

It is straightforward to generalize wage gap decompositions to nonparametric wage structures with additive error terms. Assume that the wage for worker i takes the following additive form:

$$w_i = \sum_{j=W,M} b_j(x_i) \cdot I\{i \in j\} + e_i \tag{9}$$

and let $b(x)$ be the nondiscriminatory wage structure. The average wage gap between the two groups is

$$\overline{w}_W - \overline{w}_M = \sum_{i \in W} N_W^{-1} b_W(x_i) - \sum_{i \in M} N_M^{-1} b_M(x_i) + \sum_{i \in W} N_W^{-1} e_i - \sum_{i \in M} N_M^{-1} e_i \tag{10}$$

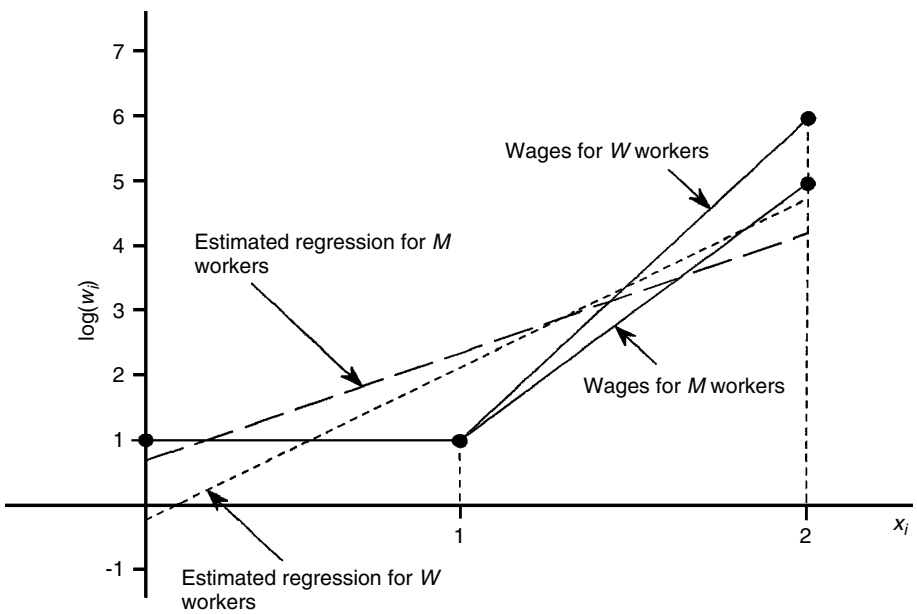

Figure 1. Example of bias in the decomposition of the average wage gap due to functional specification error

[4] If we assumed that group W wages would prevail without discrimination, then the productivity differentials effect as measured by the linear model would be 0.625. The figure goes down to 0.455 if group M workers were those of the nondiscriminatory wage structure.

where N_M and N_W are the number of observations in the sample for group M and group W workers, respectively. As in the parametric case, we can decompose the observed average differences into four components:

$$\begin{aligned}\overline{w}_W - \overline{w}_M = &\left(\sum_{i\in W} N_W^{-1}(b_W(x_i) - b(x_i))\right) + \left(\sum_{i\in M} N_m^{-1}(b(x_i - b_m x_i))\right) \\ &+ \left(\sum_{i\in W} N_W^{-1} b(x_i) - \sum_{i\in M} N_M^{-1} b(x_i)\right) + \left(\sum_{i\in W} N_W^{-1} e_i - \sum_{i\in M} N_M^{-1} e_i\right)\end{aligned} \tag{11}$$

Interpretation of these terms is similar to that of the linear parametric specification without sample selection bias. As in equation (3), the first term on the right-hand side of equation (11) reflects favoritism for group W workers, while the second component measures discrimination against group M workers. The third term is the wage gap that would exist in the absence of discrimination and favoritism provided that the individuals' average of unobserved productivities was similar across groups in the sample. Finally, the fourth term is the average effect of these unobserved productivities in the wage gap.

If we further assume Neumark's nonparametric wage structure, i.e., $b(x) = \alpha(x)b_M(x) + (1 - \alpha(x))b_W(x)$, then:

$$\begin{aligned}\overline{w}_W - \overline{w}_M = &N_W^{-1}\sum_{i\in W}\alpha(x_i)(b_W(x_i) - b_M(x_i)) \\ &+ N_m^{-1}\sum_{i\in M}(1 - \alpha(x_i))(b_W(x_i) - b_M(x_i)) \\ &+ \left(N_W^{-1}\sum_{i\in W}(\alpha(x_i)b_M(x_i) + (1 - \alpha(x_i))b_W(x_i))\right. \\ &\left.-N_M^{-1}\sum_{i\in M}(\alpha(x_i)b_M(x_i) + (1 - \alpha(x_i))b_W(x_i))\right) \\ &+ \left(\sum_{i\in W} N_W^{-1} e_i - \sum_{i\in M} N_M^{-1} e_i\right)\end{aligned} \tag{12}$$

As in the parametric case, if unobserved factors in the decision to participate in the labor market are related to unobserved productivity, then average wages across groups may also differ due to a sample selection effect. Unlike the parametric case, exclusion restrictions are typically crucial to identify the selection bias in nonparametric models (see, for instance, the survey in Vella, 1998). In the following, a model based on classification and regression trees is presented and sufficient conditions are obtained to identify the wage parameters in the presence of sample selection bias.

In equation (5), the existence of two groups of workers, workers from group M and workers from group W, both for the wage and the participation equations was implicity assumed. In contrast,

in models following tree structures, the number of groups is unknown and must be estimated together with the parameters which enter the equations linearly. To present the tree model, it is useful to partition the variable set into seven elements: $\{w_i, y_i, s_i, x_i, z_i, e_i, u_i\}$. The first two elements, w_i and y_i, simply denote, as before, the dependent variable in the wage equation and the latent participation index. As in equation (5), if $y_i > 0$, then i participates in the labor market.

The third element, s_i, is a column vector which consists of all (possibly real) variables that *potentially* define membership to a group. In the traditional linear approach, s_i only contains the dichotomous variable describing membership to groups M and W. Here, for example, s_i could be a three-dimensional vector including ethnic origin and the geographical coordinates of the place of residence (two real variables). Alternative groups are then defined by the discrete values of the ethnic variable and intervals for the geographical coordinates. The notation $s_i \in g_w \in G_w$ denotes that i belongs to group g_w among the set G_w of different wage groups. Also, if $s_i \in g_y \in G_y$, then i belongs to group g_y among the set G_y of different participation groups. This notation does not exclude the possibility that membership to groups in the wage equations and in the participation equations are defined through two disjoint subsets of variables in s_i. In the three-variable example, it could be that the ethnic variable would define membership to wage equations while the geographical coordinates would define membership to participation equations.

The vector x_i includes all variables that enter linearly into the wage equations, while z_i includes all variables that enter linearly into the participation equations. Finally, e_i and u_i are jointly i.i.d. disturbances with positive covariance. The tree model takes the form

$$w_i = \sum_{g \in Gw} (x_i' b_g) \cdot I\{s_i \in g\} + e_i \tag{13}$$

$$y_i = \sum_{g \in Gy} (z_i' \gamma_g) \cdot I\{s_i \in g\} + u_i \tag{14}$$

It is important to stress that both the sets of coefficients $\{b_g\}_{Gw}, \{\gamma_g\}_{Gy}$ and the partitions $\{G_w, G_y\}$ are unknowns in this model. Under the usual assumptions for the disturbances—for example, normality—numerical methods must be used to find the ML estimates for γ_g in all possible groups defined by s_i. This will usually lead to a computationally intractable problem even for relatively low-dimensional variable vectors. In many studies on wage equations, it has been shown that the assumption of normality is restrictive and, as a result, nonparametric methods have been developed to correct for the selection bias in the wage equations. Unfortunately, in our context, this venue would lead to an even more serious computational problem. A practical solution consists in keeping the normality assumption and assuming piecewise participation equations, for which the ML estimator for γ_g is simply the share of positive responses in group g:

$$w_i = \sum_{g \in Gw} (x_i' b_g) \cdot I\{s_i \in g\} + e_i \tag{15}$$

$$y_i = \sum_{g \in Gy} \gamma_g \cdot I\{s_i \in g\} + u_i \tag{16}$$

Two problems must be overcome before implementation of wage gap decompositions in this framework. First, due to the curse of dimensionality, equations (15) and (16) cannot be estimated with a least squares or maximum likelihood criterion even in relatively low-dimensional analyses.

Under general conditions, however, consistent estimates can be obtained from equations (15) and (16) using partition-based algorithms like regression and classification trees.[5] Second, the model as it stands presents a potential problem of identification. To see this, consider the conditional expectation of observed wages for equation g in a given set of groups G_w:

$$E[w_i|y_i > 0, s_i \in g, x_i] = x_i' b_g + \rho_g \sigma_g \lambda(s_i) \tag{17}$$

Since identification follows from variation of $\lambda(s_i)$ in g, conditions similar to exclusion restrictions are sufficient to ensure it. To see this, split the vector s_i into vectors s_{1i} and vectors s_{2i} such that

$$w_i = \sum_{g \in Gw} (x_i' b_g) \cdot I\{s_{2i} \in g\} + e_i \tag{18}$$

$$y_i = \sum_{j \in G_1} \left\{ \sum_{g \in j} \gamma_{jg} \cdot I\{s_{2i} \in g\} \right\} I\{s_{1i} \in j\} + u_i \tag{19}$$

This is a restricted version of model (15) where any group in the partician G_y is restricted to belong to one of a predetermined set of major groups G_1 which can itself be interpreted as an 'upper-level' partition of G_y. Any group in the participation equations defined through this restriction cannot be replicated in G_w because the variables which define G_1, s_1, are excluded from s_2. Identification follows if γ_{jg} changes across the groups $g \in Gw$ defined by s_2.

Estimation of the model can be carried out through a two-step procedure based on Heckman's (1979) two-step estimator adapted to the recursive splitting algorithms from classification and regression trees. First, a recursive splitting algorithm is employed to estimate the participation model. In the second stage, the inverse of Mills' ratio for each observation is incorporated into vector x_i as an additional variable and a recursive splitting algorithm is employed to estimate the wage equations. Under the assumption of normality, this procedure renders consistent estimates of very flexible functional forms and corrects for sample selection bias.

The assumption of normality can nevertheless be tested and a strategy is available if the null is rejected. When the errors are not normal, the conditional expectation of observed wages for equation g in a given set of groups G_w takes the form (see, for example, Vella, 1998)

$$E[w_i|y_i > 0, s_i \in g, s_{2i} \in g, s_{1i} \in j, x_i] = x_i' b_g + g_g(\gamma_{jg}) \tag{20}$$

This raises two difficulties. First, the estimation of the index γ_{jg} cannot be based on distributional assumptions. Second, identification of $g_g(\gamma_{jg})$ is no longer ensured. Within the framework of

[5] See Breiman *et al*. (1984) for an introduction to regression trees. In regression trees, the explanatory variables used may be both categorical and continuous, and it is this that makes the method useful in contexts that would be too complex to use analysis of covariance, for instance. Under very general conditions, partition-based algorithms like regression trees are consistent (see Stone, 1977). Regression trees are a step-optimal estimation strategy with three algorithms: (a) recursive sample partitioning of the estimation sample with the splitting variables until no further splits are possible—at each step of the splitting, the overall mean square error is minimized; (b) recursive computation of a sequence of encompassing models—at each step, a mean square error-complexity trade-off function is minimized; (c) selection of the least complex model which deviates less than *SE* standard errors from the model with the smallest test sample mean square error among the sequence obtained in (b). The interested reader may consult Durlauf and Johnson (1995), Cotterman and Peracchi (1992), and Tronstad (1995) for economics-oriented applications of regression trees.

tree structures, however, these two problems are solved under the exclusion restrictions already proposed. First, the recursive splitting estimators for γ_{jg} do not require a normality assumption for u_i. Second, a Taylor approximation of $g_g(\gamma_{jg})$ is fully identified as the exclusion restrictions prevent the possibility of a constant function $g_g(\gamma_{jg})$ in the wage equation of any group $g \in Gw$ defined by s_2.

A nonparametric version of equation (8) can be carried out for model (18) as the selection bias in model (18) also enters additively in the expected wage equations for the workers. Let $\tilde{s}_{2i}$ be the vector formed by the remaining elements of vector s_{2i} after excluding the dichotomous variable which identifies W and M workers and let $b(x_i, s_{2i}) = \sum_{g \in Gw}(x_i' b_g) \cdot I\{s_{2i} \in g\}$ be the expected wage of i in the population, i.e., the wage structure for i. Following Neumark, the nondiscriminatory wage structure is defined as the weighted average of the wage structures of W and M workers:

$$b(x_i, \tilde{s}_{2i}) = \sum_{i \in W} \alpha(x_i, \tilde{s}_{2i}) b(x_i, s_{2i}) - \sum_{i \in M} (1 - \alpha(x_i, \tilde{s}_{2i})) b(x_i, s_{2i}) \tag{21}$$

Then, equation (11) takes the form

$$\begin{aligned} \overline{w}_W - \overline{w}_M = & \left(\sum_{i \in W} N_W^{-1} (b(x_i, s_{2i}) - b(x_i, \tilde{s}_{2i})) \right) \\ & + \left(\sum_{i \in M} N_M^{-1} (b(x_i, \tilde{s}_{2i}) - b(x_i, s_{2i})) \right) + \\ & + \left(\sum_{i \in W} N_W^{-1} b(x_i, \tilde{s}_{2i}) - \sum_{i \in M} N_M^{-1} b(x_i, \tilde{s}_{2i}) \right) \\ & + \left(\sum_{i \in W} N_W^{-1} c(s_i) \lambda(s_i) - \sum_{i \in M} N_M^{-1} c(s_i) \lambda(s_i) \right) \\ & \left(\sum_{i \in W} N_W^{-1} e_i - \sum_{i \in M} N_M^{-1} e_i \right) \end{aligned} \tag{22}$$

where $c(s_i) = \sum_{j \in Gy} \sum_{g \in j} I\{s_{2i} \in g, s_{1i} \in j\} \rho_{jg} \sigma_{jg}$ under the assumption of normality.

3. DATA, ESTIMATION TECHNIQUES, AND RESULTS

3.1. Dataset and variables

This section aims to study the sensitivity to different functional form specifications of wage gap decompositions between third-generation Mexican American and non-Hispanic male workers in the border states with Mexico. The data used correspond to extracts of the Merged Outgoing Rotation Groups file of the Current Population Survey (CPS) prepared by the NBER for the 1994–2002 period. In this study, a male worker is said to be Mexican American if he reports to

be either Mexican American, Mexicano, or Chicano in the CPS's variable 'ethnic affiliation'. A male worker is said to be non-Hispanic white if he reports to be 'white' when asked about his race and does not report to belong to any of the available Hispanic groups.[6] From 1994 onwards the CPS dataset provides information on the respondent's and his parents' birthplace, so that it is possible to identify third-generation Mexican Americans, i.e., workers born in the USA with at least one parent also born in the USA.[7]

The restriction of the sample to third-generation males in the four border states aims to analyze sensitivity to functional form in wage differentials with respect to a homogeneous minority group. Hispanics are a very heterogeneous ethnic group and previous research on their wage differentials has already acknowledged this fact by having usually reported wage equations estimates for each of the subgroups. Within the Hispanic population, Mexican Americans are the largest single group in the USA, thus constituting a good choice for analysis in terms of sample size. In addition, the presence of Mexican Americans in the USA has been longer than that of any other Hispanic group. Almost two-thirds of them live in the four border states with Mexico (i.e., California, Arizona, New Mexico, and Texas), where they represent an important share of the total population. Using the CPS data, it can be seen that the ratio of Mexican Americans to non-Hispanic whites exceeds 0.30 only in these four states. The highest value for this ratio elsewhere is about half that number (0.156 in Nevada). Moreover, the increase in the numbers of Mexican Americans in other parts of the country has been a relatively recent phenomenon which is closely related to the increase of the presence of all Hispanics in all major US Metropolitan areas. As a consequence, the ratio of third-generation Mexican Americans to third-generation non-Hispanic whites in the border states is 0.17, while for the non-border states the same ratio is only 0.0086, reaching its highest value of 0.06 in Colorado. Thus, third-generation Mexican Americans in the border states constitute a relatively homogeneous group clearly differentiated from other Hispanics.

The study of third-generation workers as opposed to all workers presents two addittional advantages. First, it is possible to identify a sample of workers, both Mexican American and non-Hispanic, with a prolonged exposure to the US labor market. Important issues for the overall population of Hispanics and Mexican Americans, such as controlling for the effect of being mistaken as an illegal immigrant, or the effect of insufficient time to adapt to the US labor market, are therefore likely to be minor problems for the third generation.

Second, it can also be argued that the level of English proficiency—a variable not usually reported in the CPS—for the third generation is not likely to be an important factor in wage regressions. Controlling for the level of English proficiency is usually done with self-assessed reports. Datasets which contain information on self-assessed English proficiency suggest that English proficiency for third-generation workers can be assumed to be the same for Mexican Americans and other workers. Table I summarizes data from the 1990 Decennial Census from the Bureau of the Census. It is shown that the longer a worker has lived in the USA, the better he claims his English is. The majority of non-Hispanic workers who were born in the USA could only speak English. In the case of Mexican Americans, although the majority could speak Spanish, almost 90% of those born in the USA reported to speak only English or to speak English 'very well'. Therefore, for a third-generation Mexican American

[6] These are: Mexican American, Chicano, Mexicano, Puerto Rican, Cuban, Central or South American, Other Spanish.

[7] That is, at least by one line of descent, it is the third generation. This is a less restrictive condition than the one in Trejo (1997), where third-generation workers are those with both parents born in the USA. Nevertheless, both definitions are highly correlated for Mexican Americans, as only 22% of Mexican Americans that have one parent born in the USA do not have both parents born in the USA.

the probability that at least one parent will have a good command of English is greater than 99.998%.[8]

These percentages do not account for measurement error in the English proficiency variable. Bilingual speakers, such as the majority of third-generation Mexican Americans, may systematically judge language skills differently from monolingual speakers. In the only study directly addressing this issue, Dustmann and van Soest (2001) report important differences in wage equation estimates after controlling for this problem using data on first-generation immigrants from the German Socio-Economic Panel. To sum up, by restricting the sample to third-generation workers it can be expected that estimation biases derived from not controlling correctly by the level of English proficiency are minimized.

The sample only includes males from 25 to 62 years of age, so that the retirement decision does not result in sample selection bias across ethnic populations. Finally, those observations in the 1.67% upper tail of the yearly income distribution were also excluded in the results presented below as most of these observations were top-coded in the variable on earnings.[9] The number

Table I. English proficiency for Mexican American workers for pay[a,b]

Year of entry	Only English	Very well	Well	Not well	Not at all
1987–1990	3.53	13.60	11.43	30.99	40.45
	(15.00)	(26.59)	(31.82)	(23.09)	(3.50)
1985–1986	3.64	8.90	19.00	39.28	21.19
	(13.23)	(35.32)	(28.82)	(19.97)	(2.67)
1982–1984	3.40	14.72	22.66	41.81	17.41
	(15.14)	(35.51)	(32.80)	(14.16)	(2.39)
1980–1981	3.97	16.49	24.70	38.67	16.17
	(10.55)	(37.04)	(33.61)	(17.06)	(1.74)
1975–1979	3.56	19.44	30.31	34.00	12.68
	(14.09)	(43.02)	(32.30)	(9.60)	(1.00)
1970–1974	3.94	23.98	30.73	31.52	9.84
	(21.58)	(46.21)	(24.39)	(7.04)	(0.77)
1965–1969	3.23	31.12	29.41	27.50	8.73
	(38.11)	(41.97)	(14.40)	(4.95)	(0.57)
1960–1964	3.69	34.84	31.56	23.22	6.70
	(55.22)	(32.30)	(10.32)	(1.93)	(0.22)
1950–1959	8.60	39.41	27.47	19.15	5.36
	(63.61)	(27.01)	(7.81)	(1.43)	(0.15)
Before 1950	9.26	41.01	26.29	18.00	5.44
	(67.92)	(21.64)	(8.73)	(1.71)	(0.00)
Born US	37.51	49.02	9.95	3.12	0.40
	(95.46)	(3.58)	(0.64)	(0.30)	(0.01)

[a] Weighted tabulations from a 1% random sample of the 1990 Decennial Census from the Bureau of the Census. Data refer to male employed workers for pay.
[b] Percentage population estimates. Values for non-Hispanic white workers in parentheses.

[8] Trejo (1997) estimates returns to English proficiency based on self-assesments using data from the November 1979 and November 1989 Current Population Survey for all Mexican workers. He does not report separate results for second- and higher-generation workers, but finds that the difference in returns with respect to only-English speakers for workers who speak English at least well is either insignificant (in 1979) or only marginally significant (in 1989).
[9] The benchmark 1.67 is the percentage of top-coded observations in 1997. In all other years, the percentage was lower than 1.5%. In order to gauge the importance of this truncation, in results not presented here, these individuals were included in the participation model as not-participating. The results on the participation model remained the same. Including top-coded

of observations for the entire sample is 75,949. This includes workers for pay, self-employed, unemployed, and other. Only the first group, a total of 53,865 observations (70.92%), enters into the wage sample. The rest of the observations, 22,084, which include those self-employed (10,850 observations, or 49.13%), those unemployed (11.55%), and other (39.32%), are pooled together in the analysis as nonparticipants in the working-for-pay market.

Wages are computed as the logarithm of weekly hourly wages, deflated by wage inflation. Human capital variables include age, years of education, and a dummy variable for vocational training. In addition to these human capital variables, information on the degree of Hispanic presence sorrounding the worker is also taken into account in an attempt to capture segregation effects on wages. This information includes the historical percentage of Hispanics in the metropolitan area of residence, the historical percentage of Hispanics in the occupation, and the historical percentage of Hispanics in the industry to which the job belongs. Business cycle effects are controlled for by introducing dummies for the month and year of the sample.

In the participation equations some variables related to the relatives of the workers were constructed using an algorithm to merge workers with their parents/wives living in the same household. These variables are the age and education of the spouse and whether the worker lived with his parents. Also in the participation equations, codes for the states and whether the observation was prior to 1997 were included to account for legislative changes at state level on social benefits for children.

Table II shows means and definitions of the variables in the total and wage sample. On average, third-generation Mexicans have lower wages, are younger, less educated than non-Hispanic whites, and tend to live in areas and work in occupations and industries with a traditionally higher Hispanic presence. A larger share of them live with at least one of their parents and their wives have lower levels of education and are younger than the wives of non-Hispanic workers.

4. RESULTS

4.1. Participation Models

The wage sample includes only employees working for pay. Around 76.1% third-generation Mexican Americans earn a wage. In contrast, the figure for non-Hispanic whites is only 70.1%. In this section, several participation models are estimated to account for these figures. The purpose of the exercise is to compute the inverse of Mills' ratio, λ_i, and the index, γ_{jg}, to correct for incidental truncation in the wage equations.

Several functional forms are considered. All models can be interpreted as random utility models where the error term difference follows a normal distribution. The set of control variables includes the respondent's years of education, whether he holds a vocational degree, his age, veteran status, whether the respondent lives with at least one of his parents, the percentage of Hispanics in the Metropolitan Area in 1992 and 1993, and state and time dummies.

For the purpose of comparability among the models, in the results presented the wife's years of education and the wife's age variables are each simplified into four categories, with values ranging from 1 to 4, in the following way. *Wife's Ag*. takes value 1 if no wife is present, value 2 if her age

observations in the wage equations by assuming a gamma distribution for the upper tail—admittedly a very parametric solution—led to qualitatively similar results for the wage gap decompositions. Nevertheless, inferences on the entire population from the results presented here should be made with caution.

Table II. Variable definitions[a] and mean values

No. of observations:	White non-Hispanics		Mexican Americans	
	All	Wage sample	All	Wage sample
	65,026	45,558	10,923	8,307
Wages	—	2.65	—	2.35
Age	42.17	40.93	39.31	38.54
Education	14.23	14.31	12.35	12.57
Vocational Degree	0.09	0.10	0.08	0.08
Hispanic Area	20.90	20.63	27.08	26.83
Hispanic Occupation	23.02	22.52	29.94	29.59
Hispanic Industry	20.23	19.23	22.42	21.67
Veteran Status	0.26	0.25	0.19	0.19
Marital Status	0.65	0.66	0.62	0.65
Parents	0.05	0.04	0.13	0.11
Wife's Education	14.06	14.07	12.15	12.32
Wife's Age	41.71	40.60	39.16	38.30

[a] *Wages*: logarithms of earnings per week divided by hours per week at the job deflated by annual wage inflation. *Age*: age of respondent. *Education*: number of years of education. *Vocational degree*: 1 if attended vocational degree in college, 0 otherwise. *Hispanic Area*: average 92/93 percentage of Hispanic population in the Metropolitan Statistical Area FIPS code or the state if MSA code not available. *Hispanic Occupation*: average 92/93 percentage of Hispanic population in the 2-digit Detail Occupation Recode from the 1980 Census. *Hispanic Industry*: average 92/93 percentage of Hispanic population in the NBER 2-digit Detailed Industry Classification. *Veteran Status*: 1 if veteran, 0 otherwise. *Marital Status*: 1 if wife present, 0 otherwise. *Parents*: 1 if lives with at least one parent, 0 otherwise. *Wife's Education*: wife's years of education if $Marital\ Status = 1$. *Wife's Age*: wife's age if $Marital\ Status = 1$.

lies within [0, 35), value 3 if it lies within [35, 45) and value 4 if it is larger than 45. *Wife's Ed.* takes value 1 if no wife is present, value 2 if her years of education lie within [0, 12), value 3 if it lies within [12, 14) and value 4 if it is larger than 14. This simplification will prove useful in the identification of the wage equations in the nonparametric models. On the other hand, the results for the marginal effects in the two parametric equations are equivalent to the results obtained from the original variables and the goodness-of-fit measures are invariant within a four-digit accuracy.

Table III reports results from a parametric model where the respondent's age and education enter quadratically into the index function. Two equations are estimated: one for Mexican Americans and one for non-Hispanic whites. Most parameters are significant in the two equations, although the standard errors for the coefficients in the Mexican American equation are larger than in the non-Hispanic whites equation, probably reflecting the smaller sample size in the first group. Some of the estimates are significant only in one of the two equations. *Hispanic Area*, for example, is strongly significant only in the Mexican American sample.

Given the nonlinearity in both the normal distribution and the index function formulation and the fact that some of the dummy variables are interdependent, the estimated coefficients provide little insight as to the contribution of each variable to the probability to participate. To present some evidence on these contributions, marginal effects are computed for each variable as the average change across individuals in the predicted probability after one unit increases from each possible value taking into account the dependency among the variables. For example, in order to compute the marginal effects for the dummy variable *California*, all other state dummies are set to zero.

Table III. Participation results: probit model[a]

	Non-Hispanic whites			Mexican American		
	Coeff.	SE	dF/dx	Coeff.	SE	dF/dx
Age	0.034	0.0054	−0.0122	0.0554	0.0138	−0.0091
(*Age*^{2})/100	−0.070	0.0053		−0.0731	0.0139	
Education	0.055	0.0097	0.0112	0.1533	0.0230	0.0269
(*AgexEd*)/100	−0.055	0.0209		−0.161	0.0496	
Vocational Degree	0.1438	0.0189	0.0459	0.1157	0.0558	0.0320
Wife's Ag.	0.0850	0.0076	0.0269	0.0630	0.0242	0.0167
Wife's Ed.	−0.0343	0.0073	−0.0114	0.0100	0.0040	0.0029
Parents	−0.4112	0.0235	−0.1514	−0.4031	0.0438	−0.1260
Veteran Status	0.1864	0.0132	0.0653	0.0764	0.0383	0.0215
California	−0.0622	0.0199	−0.0225	−0.0141	0.0472	−0.0042
Arizona	0.0344	0.0242	0.0123	0.1488	0.0647	0.0423
Texas	0.0887	0.0206	0.0313	0.1356	0.0459	0.0382
Hispanic Area	0.0664	0.0487	0.0237	1.0160	0.1128	0.2883
Constant	−0.3771	0.1760		−2.1237	0.4124	
Pseudo-R^2	0.0487			0.0797		
No. of observations	65,026			10,923		
λ	0.5934			0.3742		

[a] Time variables were also included in the equations. dF/dx is the average of each individual's marginal effects. For each variable, these are computed as the average change in the predicted probability after one unit increase from each possible value. Variable dependence such as in *Age* and *Age*^{2} and the state dummies is taken into account. Pseudo-$R^2 = 1 - \frac{L_1}{L_0}$, where L_0 is the log-likelihood of the model only with a constant. λ is the mean of inverse of Mills' ratio. *Wife's Ag.* and *Wife's Ed.* take values between 1 and 4 according to *Wife's Age* and *Wife's Education* (see main text). See Table II for all other variable definitions.

Average marginal effects show that years of education, both for the respondent and his wife, and the Hispanic presence have larger effects on the probability to participate in the Mexican American sample. In contrast, holding a vocational degree, the wife's age, the presence of one of the parents at home, and veteran status have larger average impacts in the non-Hispanic sample. Admittedly, interpretation of these results is complicated by the fact that these equations are pooling selection from several decisions, i.e., the decision to work, the decision to look for a paid job, and the decision to accept a paid job.[10] However, it can be stressed from the results that there are significant differences among the two groups in the truncation process. These differences can potentially lead to differences in average wages arising from selection bias if the unobserved components are correlated. Correction for this selection bias can be carried out by construction of the inverse of Mills' ratio, λ_i, for all workers. The average value of this variable is larger in the non-Hispanic white sample, 0.5934 versus 0.3742, reflecting the fact that participation is large in both samples and slightly larger in the Mexican American sample.

Table IV reports marginal effects, goodness-of-fit measures, and descriptive statistics for λ_i for several participation models. The first two columns show detailed results for the parametric especification already reported in Table III.

[10] For convenience, the focus here is on correcting for incidental truncation in the wage equations. For that goal, pooling the nonparticipating alternatives does not seem too distorting, especially in the nonparametric models. However, as already stated, discrimination may also arise in the decisions to participate. If the goal were to asses the degree of that discrimination, then the difference between the nonparticipating alternatives, i.e., inactivity, self-employment, and unemployment, should be made explicit. This is a very interesting problem which is beyond the scope of this paper.

Table IV. Participation results in three models[a]

	Quadratic		5th order		Tree	
	NHW	MA	NHW	MA	NHW	MA
Marginal effects						
Age	−0.0122	−0.0122	−0.0129	−0.0116	−0.0006	−0.0006
Education	0.0112	0.0097	0.0139	0.0164	0.0003	0.0003
Vocational Degree	0.0459	0.0320	0.0391	0.0204	0	0
Wife's Ag.	0.0269	0.0167	0.0227	0.0132	−0.0242	−0.0180
Wife's Ed.	−0.0114	0.0029	−0.0100	0.0028	−0.0071	−0.0055
Parents	−0.1514	−0.1260	−0.1354	−0.1023	−0.0122	−0.0140
Veteran Status	0.0653	0.0215	0.0649	0.0215	0.0000	−0.0000
California	−0.0225	−0.0042	−0.0258	−0.0039	−0.0137	−0.0101
Arizona	0.0123	0.0423	0.0093	0.0412	0.0136	0.0101
Texas	0.0313	0.0382	0.0271	0.0385	0.0120	0.0178
Hispanic Area	0.0237	0.2883	0.2221	0.5187	0.0246	0.0316
Pseudo-R^2	0.0487	0.0797	0.0816	0.1299	0.1900	0.2968
No. of observations	65,026	10,923	65,026	10,923	65,026	10,923
λ: Average	0.5934	0.3742	0.5606	0.3526	0.5718	0.5936
λ: SD	0.1912	0.1673	0.2198	0.2000	0.6661	0.7517
λ: Minimum	0.1589	0.0423	0.0782	0.0216	0.0957	0.0898
λ: Maximum	1.3997	1.5555	1.8956	1.6981	3.1804	3.1804

[a] See Table III for the computation of marginal effects, Pseudo-R^2, and λ. 'Quadratic' refers to the model in Table III. '5th order' refers to a parametric model with terms to the 5th order in *Age, Education*, and *Hispanic*. 'Tree' refers to model (23) with G_1 defined by state dummies, *Wife's Ag., Wife's Ed., Parents*, and ($Year < 1997$). Finally, 'NHW' refers to the non-Hispanic white sample and 'MA' to the Mexican American sample.

The third and the fourth columns report results for a highly nonlinear parametric specification of the index function (including terms up to the fifth order in the human capital variables and *Hispanic Area*). All coefficients of the variables shown in the table were significant at the 99% confidence interval except for the *Age* variables.[11] Significance of *Hispanic Area* for all parameters turned out to be very high even for the non-Hispanic white sample. The pseudo-R^2 increased by 67.56% in the non-Hispanic white sample and by 62.97% in the Mexican American sample. This large increase in fit in this flexible parametric model, however, hardly translates into large increases either in the average marginal effects or in the average for λ_i in the two samples. But it does have an important effect both on the distribution of the individuals' marginal effects (not shown in the table) and the distribution of the inverse Mills' ratio. In particular, both the standard deviation and the spread of λ_i significantly increase in the two samples using this more flexible model.

The last two columns in Table IV show the results of the estimation of the participation equation in model (18):

$$y_i = \sum_{j \in G_1} \left\{ \sum_{g \in j} \gamma_{jg} \cdot I\{s_{2i} \in g\} \right\} I\{s_{1i} \in j\} + u_i \tag{23}$$

[11] Dropping the fifth-power term for *Age* led to significant estimates for the remaining coefficients. The marginal effect results, however, remained invariant. Including a sixth-order term led to multicollineality problems in *Age* and the included coefficients for the three variables became nonsignificant.

The partition G_1 is obtained by the interaction of all value combinations in the sample for the four state dummies, *Wife's Ag., Wife's Ed., Parents*, and a dummy variable for ($Year < 1997$) which aims to capture the effects on participation brought about by the change in legislation for children's social benefits. The grid obtained is composed of 88 cells (out of 256 potential combinations). A classification tree algorithm is carried out within each of these cells using as splitting variables (i.e., as s_{2i}) *Age, Education, Vocational Degree, Veteran Status, Ethnic* (i.e., 0 if non-Hispanic white, 1 if Mexican American), *Time*, and *Hispanic Area* and taking 2 as the *SE* rule.[12]

The pseudo-R^2 rises to 0.1900 in the non-Hispanic white sample and to 0.2968 in the Mexican American sample. A higher fit for the overall sample is an outcome of the algorithm and should not be surprising. A more honest goodness-of-fit measure can be computed with test samples, leading to values closer to the 5th order model. However, the reported pseudo-R^2 is useful in the sense that it shows the flexibility of the tree structure to account for variability within the sample. As in the flexible parametric model, this enhanced flexibility again does not translate into larger average effects. In fact, the average marginal effect for *Hispanic Area* (0.0246 and 0.0316 for the non-Hispanic white and the Mexican American sample, respectively) is lower than the average marginal effect in the flexible parametric model. Nevertheless, the greater flexibility does translate into the distribution of the marginal effects (not shown in the table). As an illustration, the lower and upper bounds of the individual marginal effect of *Hispanic Area* for the tree structure are -0.3074 and 0.4002 for both samples. In contrast, the lower and upper limits in the flexible parametric model are 0.0776 and 0.2356 for the non-Hispanic white sample and 0.1719 and 0.6502 for the Mexican American sample. As shown in Table IV, as a result of this greater variation in the individual effects, the distribution of the inverse of the Mills' ratio also increases both in terms of its standard deviation and spread.

4.2. Wage Equations

In this section, different wage equations are estimated in order to carry out in the following section wage gap decompositions between Mexican Americans and their non-Hispanic white counterparts.

As in the previous section, several functional forms are considered. The set of control variables includes the respondent's years of education, whether he holds a vocational degree, his age, veteran status, the percentage of Hispanics in the Metropolitan Area in 1992 and 1993, the percentage in 1992 and 1993 of Hispanics in the industry where he works, the percentage in 1992 and 1993 of Hispanics in the occupation where he works, a time trend, and the correction for incidental truncation.

Table V reports results from a parametric model where the respondent's age and education enter quadratically into the wage equation. Two equations are estimated: one for Mexican Americans and one for non-Hispanic whites.

The estimates are obtained using the Heckman two-stage procedure with the appropiate standard errors. Results replicate the basic outcome in the literature. In short, returns to age and education for third-generation Mexican Americans are not significantly different from those for the non-Hispanic white workers. Although the point estimates for *Vocational Degree* are rather different, none of the estimates is estimated with accuracy. Thus, the coefficient is not significantly different from zero in

[12] It is noteworthy to stress that several of the splitting variables, such as *Hispanic Area*, take more than 45 different values. Maximum likelihood in this multivariate computer-intensive setup is therefore computationally intractable.

Table V. Wage equations results: quadratic model[a]

	Non-Hispanic whites			Mexican American		
	Coeff.	SE	dF/dx	Coeff.	SE	dF/dx
Age	0.0521	0.0028	0.0107	0.0357	0.0065	0.0105
(*Age*^{2})/100	−0.0547	0.0030		−0.0271	0.0067	
Education	0.0303	0.0047	0.0482	0.0585	0.0107	0.0537
(*AgexEd*)/100	−0.0438	0.0106		−0.0125	0.0232	
Vocational Degree	0.0009	0.0087	0.0009	0.0413	0.0216	0.0413
Hispanic Area	−0.2365	0.0619	−0.2365	−0.7237	0.1125	−0.7237
Hispanic Occupation	−0.0075	0.0006	−0.0075	−0.0011	0.0014	−0.0011
Hispanic Industry	−0.0019	0.0002	−0.0019	−0.0020	0.0005	−0.0020
Veteran Status	−0.0380	0.0071	−0.0380	−0.0037	0.0161	−0.0037
Trend	−0.0008	0.0001	−0.0008	−0.0003	0.0002	−0.0003
Constant	1.2203	0.0884		1.1495	0.1938	
R^2	0.2010			0.2063		
No. of observations	45,558			8,310		
λ	−0.2556	0.0335		−0.3179	0.0494	
ρ	−0.4696			−0.5832		

[a] Time variables were also included in the equations. dF/dx is the average of each individual's marginal effects. λ is the coefficient for the inverse of Mills' ratio and ρ is the coefficient for the correlation between the disturbances. Estimates are obtained using the Heckman two-stage procedure in STATA.

any of the equations. Marginal effects are also presented to show the marginal contribution of each of the main control variables to wages. On average, the marginal effect of education and age are very similar between the two populations. If anything, returns to education are larger in the Mexican American sample. *Veteran Status* and the variables related to the presence of Hispanics have a negative effect on wages both for non-Hispanic whites and for Mexican Americans. Interestingly, while *Veteran Status* seems to depress wages for non-Hispanic whites more strongly than for Mexican American workers—for whom the estimate is not significant—the opposite occurs with respect to *Hispanic Industry* and *Hispanic Area*.

In so far as segregation is the result of lack of success in adapting to mainstream society, a higher than average Hispanic presence in the area (i.e., segregation) is also a measure of integration failure, and therefore provides useful information on the individual's human capital. Results presented in Table V are compatible with this interpretation.

As expected, the correction for selection bias leads to higher standard errors. For the sample of non-Hispanic white workers, point estimates remained identical up to the third digit except for *Veteran Status*, which increased in absolute value after the corretion from −0.0126 to −0.0380. For the sample of Mexican Americans, controlling for incidental truncation changed the coefficient of *Hispanic Area* from −0.5851 to −0.7237. The sign of the estimate for the λ_i coefficient is negative and significant. A negative correlation between the error terms show that a higher propensity to participate is associated with a lower wage. One could argue that this is more so for the subsample of Mexican American workers, where the coefficient for λ_i is larger in absolute value. However, although the estimates are significantly different from zero, they both fall within the confidence intervals of the other coefficient at the usual levels of significance.

Table VI shows some results for three alternative modelizations of the wage structure. The first two columns correspond to the third and sixth columns of the previous table. The third and the fourth columns show results for a flexible parametric model with terms to the fourth order in *Age,*

Table VI. Wage equations results in three models[a]

	Quadratic		4th order		Tree 1	
	NHW	MA	NHW	MA	NHW	MA
Marginal effects:						
Age	0.0107	0.0105	0.0149	0.0121	0.2447	0
Education	0.0482	0.0537	0.0684	0.0458	0.0273	0.0181
Hispanic Area	−0.2365	−0.7237	−4.3380	−2.6847	−0.3697	−0.3605
Hispanic Occupation	−0.0075	−0.0011	−0.0418	−0.0179	−0.0256	−0.0230
Hispanic Industry	−0.0019	−0.0020	−0.0026	−0.0025	−0.0210	−0.0239
R^2	0.2010	0.2063	0.2131	0.2208	0.2226	0.3512

[a] See Table III for the computation of marginal effects. 'Quadratic' refers to the model in Table V. '4th order' refers to a parametric model with *Education, Hispanic Area, Hispanic Occupation* terms to the 4th order in *Age* and *Hispanic Industry*. 'Tree 1' refers to model (18) with correction in the wage equations carried out with the scores computed from 'Tree' (see Table IV) entering into the wage equations to the fourth power. All control variables are included as s_{2i}. The model is separatelly estimated for non-Hispanic whites and Mexican Americans. The *SE* rule to select the size of the tree is 0.

Education, Hispanic Area, Hispanic Occupation, and *Hispanic Industry*. Naturally, the R^2 must increase (in fact, also the adjusted-R^2 increased) as more variables are poured into the regressions. Nevertheless, only the marginal effect for *Hispanic Area* changes noticeably.

Finally, the last two columns present results for the estimation of a tree model both for the non-Hispanic white workers and for the Mexican American workers. 'Tree 1' refers to model (18) with correction in the wage equations carried out with the index functions computed from 'Tree' (see Table IV). These indexes, γ_{jg}, enter into the wage equations up to the fourth power. All control variables—*Age, Education, Vocational Degree, Veteran Status, Time, Hispanic Area, Hispanic Industry* and *Hispanic Occupation*—are included in vector s_{2i}. On the other hand, the *Wife's Educational Category*, her age category, parents, the state code, and a time dummy are included in vector s_{1i}. These exclusion restrictions guarantee the identification of the model. The model is estimated by splitting the sample into a learning and a test sample. The test sample, whose size is aproximately one-third of the overall size of the dataset, is used to select the complexity of the tree structure. As in the other two models in Table VI, the model is separately estimated for non-Hispanic whites and Mexican American workers.

The tree structure obtains the highest fit of the data. This is somewhat surprising, as the model turns out to have only 38 terminal groups. It is possible to carry out a test for the normality of the error terms within each of these groups. Following Vella (1998), the tests are conducted via artificial regressions in which the third and fourth sample moments of the residuals are regressed against an intercept and the scores from the probit specification. For the Mexican American sample, the null was rejected in all cases. For the non-Hispanic white subsample, the null was rejected using the fourth sample moments in all terminal nodes and using the third sample moment in all but five cases.

To sum up, higher flexibility of the models results in higher fits. As already stressed, this should not come as a surprise, but two points merit consideration here. First, when looking at more honest measures of goodness of fit, such as test sample, the differences among the different models do decrease. This last remark is in a way surprising because specifications are very different from each other. The tree model is a piecewise function that resembles a multivariate histogram. The quadratic model consists of two parsimonious parametric equations. Thus, although all models

seem to explain data variability equally well, the way they do it is different by construction. Finally, evidence against the assumption of normality is found in a flexible nonparametric tree structure.

4.3. Wage Decompositions

Once the wage structure is estimated, decompositions can be easily carried out by substituting the estimates in equation (11). In the following, the decompositions shown are based on the assumption that the nondiscriminatory wage structure is the weighted average within each type of worker of the majority and the minority groups' wage structures, as proposed by Neumark (1988).[13]

Table VII presents seven different decompositions of the wage gap as computed from the test sample.[14] All decompositions correct for selectivity bias. However, the first two use Heckman's two-stage parametric procedure, while the others employ a more general, at most fourth-order polynomial for the index functions in the participation equations.[15]

'Quadratic with λ' presents the decomposition for the quadratic model (Table V) using Heckman's two-stage procedure with the probit formulation from Table III. '4th order with λ'

Table VII. Average wage gap decompositions: test sample results[a]

	1	2	3	4	5	6
Quadratic with λ	30.34	1.52	9.22	19.37	0.27	−0.04
4th order with λ	30.34	2.13	12.08	20.32	−4.38	0.19
Quadratic	30.34	1.50	8.86	19.77	0.47	0.03
4th order	30.34	1.85	9.85	20.60	−1.97	0.26
Tree 1	30.34	1.29	6.31	15.19	6.68	0.87
Tree 2	30.34	1.25	8.99	21.61	−1.00	−0.22
Tree 3	30.34	3.70	20.84	14.05	−9.38	1.14
Tree 4	30.34	0	0	22.24	−4.15	12.26

[a] Decomposition taking Neumark's non-discrimination wages. 1: Average wage gap; 2: favoritism for NHW; 3: discrimination against MA; 4: observed productivity differentials; 5: incidental truncation effect; 6: error effect. 'Quadratic with λ' presents the decomposition for the quadratic model (Table V) using Heckman's two-stage procedure with the probit formulation from Table III. '4th order with λ' refers to the '4th order' model in Table VI using the '5th order' probit model. 'Quadratic' refers to the quadratic model using a 4th order polynomial with the propensity scores to correct for selection. The same correction method is applied to '4th Order' and the tree models. Tree 1 is defined in Table VI. Tree 2 includes a quadratic human capital specification in vector x_i and $SE = 2$. Tree 3 variable specification matches that of Tree 1 but $SE = 2$. Tree 4 is estimated with the pool sample of non-Hispanic whites and Mexican Americans. *Ethnic* enters into s_{2i} and SE = 0.

[13] Given the different sample sizes of the majority and the minority groups, by construction favoritism tends to be smaller than discrimination in the Neumark decomposition.

[14] For brevity, decompositions with the learning sample are not shown. The main feature which distinguishes the results of the learning sample decompositions in all models is that the error component is underestimated. In fact, for the parametric decompositions, this error term is fixed to zero.

[15] The decompositions were also carried out without the correction terms and with an 'at most' quadratic structure in the correction terms to asses the sensitivity of the results to the sample selection problem. When no correction for selection was implemented, there was no pattern as to where the selection component would go. For example, discrimination and favoritism assimilated the selection component in Tree 1, while the error term increased in Tree 2 and the productivity differentials increased in Tree 3. When the number of maximum terms was reduced to 2, the selection component became somewhat volatile, increasing in some models—it accounted for most of the wage gap in Tree 3—and decreasing in others, as in Trees 1 and 2.

refers to the '4th order' model in Table VI using the '5th order' probit model. 'Quadratic' refers to the quadratic model using an 'at most' 4th order polynomial with the propensity scores to correct for selection.[16] The same correction method is applied to '4th order' and the tree models. Tree 1 is defined in Table VI. Tree 2 includes a quadratic human capital specification in vector x_i and $SE = 2$. Tree 3 variable specification matches that of Tree 1 but $SE = 2$. Tree 4 is estimated with the pool sample of non-Hispanic whites and Mexican Americans. *Ethnic* enters into s_{2i}. Therefore, only if this variable outperforms the others in any stage of the recursive splitting algorithm in terms of separating the sample into two more homogeneous subsamples will there be a different wage structure for non-Hispanic whites and Mexican Americans. To ensure that the algorithm does not result in an excessive simplification of the tree structure, the *SE* rule for this model is $SE = 0$.

Several remarks stand out inmediately from close inspection of Table VII. First, the error components are relatively small in all decompositions, with the exception of Tree 4. In that case, the splitting algorithm in the learning sample never uses the variable *Ethnic* as a splitting criterion. As a consequence, both discrimination and favoritism are fixed to zero. In the test sample the magnitude of the error term suggests that the more flexible model does fail in capturing the structure of the test sample.[17]

Second, the largest component of the wage gap is almost always the productivity differentials component, ranging from 14.05 logarithmic units to 22.24. Rather interestingly, these two extreme values correspond to two flexible tree structures, and both of them are the worst performers in terms of keeping the unobserved productivity component, the error term, of the test sample low.

Third, the sign of the selection component varies with functional specification. This occurs even in the parametric wage structures with correction à la Heckman. Finally, most of the discrimination terms lie within the [6.31, 12.08] interval, while the favoritism component lies, by construction, in the narrower [1.25, 2.13] interval.

More importantly, no general pattern arises as to whether more functional flexibility gives more or less discrimination. Nevertheless, it is shown that the decomposition heavily depends on functional assumptions, and that different aproaches to flexibility may render sufficiently good and similar results.

5. SOME CONCLUDING REMARKS

This paper shows that average wage gap decompositions between any two groups of workers can be carried out using a nonparametric wage structure. This method does not imply any loss of generality nor pose any additional problems of interpretation. Oaxaca type decompositions are simply generalized to decompositions with differentials that do not have a simple parametric structure.

[16] The selection of the order for the polynomials was based on two algorithms. For the variables entering the vector x_i, the order chosen was that of the variable for which the coefficient turned out to be significant. For the selection correction polynomial, the number of terms varied within the nodes as they were dropped whenever they were not significant. Not performing this algorithm resulted in decompositions which lacked robustness to estimation procedure, especially for Tree 1 and Tree 3.

[17] For this and the other models, several partitions of the sample into a learning sample and a test sample, with varying relative sizes, were carried out. Rather unsurprisingly, given the large size of the dataset, decomposition results were invariant to the first digit.

From the results of the empirical illustration, it can be argued that nonparametric and highly nonlinear functional forms should be carried out to check consistency of the results in any study on wage gap decompositions. The usefulness of each semiparametric and nonparametric model will likely depend on each particular case. However, any flexible proposal must realistically solve the problem of incidental truncation in a way consistent with the flexibility of the wage equations.

In the empirical application of the paper a particular model based on tree structures is proposed and estimated. It is shown that the sample selection correction can be implemented and that identification of the parameters in the wage equations follows from restrictions which can be interpreted as exclusion restrictions.

The main empirical contribution of this paper is to assess with some highly nonlinear wage equations and with tree wage structures the consistency of wage gap decompositions between Mexican American and non-Hispanic workers obtained from simple functional specifications. The results point towards a discrimination component for Mexican Americans which vary in the preferred specifications somewhere between 6 and 12 percentage points—at most a third of the wage gap.

Wage discrimination is a complex process and here I have only looked at it in a partial way. The approach that I have followed is minimalist in the sense that three effects that I have labeled as distinct from discrimination are not necessarily so. First, I have not addressed the choice of residence decisions and therefore my results hinge on the assumption that neighborhood segregation is the result of an assorting algorithm in which human capital differences play a key role. In addition to this, I have only estimated a reduced-form equation of the decision to participate in the for-pay market. This is enough both to test the significance of the effect of selection in the wage gap and to ensure consistency of the estimates. However, it does not help in the decomposition of the wage differential into the productivity and discrimination effect. The problem lies in the fact that self-employment may be lower for Mexican Americans for a variety of reasons, some of them liable to be understood as discrimination. For example, assume that the decision of Mexican Americans to start their own enterprises is affected both by their entrepreneurship and their ability to gather funds coming from mainstream society. If financial markets are incomplete and Mexican Americans have less access to non-market finance, then their self-employment ratios would not only reflect productivity differentials but also an inadequate social network. Finally, the paper confirms the importance of human capital variables. However, accounting is not explaining. It is well known that Hispanics have not been as successful as blacks in closing the educational advantage held by whites, but the economic process that leads to this distressing outcome is largely unknown. There is therefore a need for research on the economic reasons for this failure given the importance of education for the future of Mexican Americans.

ACKNOWLEDGEMENTS

I aknowledge financial support from DGI, grant no. SEJ2006-05710/ECON. I wish to thank seminar participants at the 1999 Latin-American Meeting of the Econometric Society, and Universidad Carlos III de Madrid and especially Pedro Albarrán, Cesar Alonso, Klaus Desmet, Georges Siotis, and Carlos Urrutia for their helpful comments.

REFERENCES

Arrow K. 1972. Some mathematical models of race discrimination in the labor market. In *Racial Discrimination in Economic Life*, Pascal A (ed.). D. C. Heath: Lexington, MA; 187–203.

Becker GS. 1957. *The Economics of Discrimination*. University of Chicago Press: Chicago, IL.

Bergmann BR. 1971. The effect on white incomes of discrimination in employment. *Journal of Political Economy* **79**(2): 294–313.

Blinder AS. 1973. Wage discrimination: reduced form and structural estimates. *Journal of Human Resources* **8**: 436–455.

Breiman L, Friedman JL Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*. Wadsworth: Belmont, CA.

Cotterman R, Peracchi F. 1992. Classification and aggregation: an application to industrial classification in CPS data. *Journal of Applied Econometrics* **7**(1): 31–51.

Cotton J. 1988. On the decomposition of wage differentials. *Review of Economics and Statistics* **70**: 236–243.

Cotton J. 1993. Color or culture? Wage differences among non-Hispanic black males, Hispanic black males and Hispanic white males. *Review of Black Political Economy* **21**(4): 53–67.

Durlauf SN, Johnson PA. 1995. Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics* **10**: 365–384.

Dustmann C, van Soest A. 2001. Language fluency and earnings: estimation with misclassified language indicators. *Review of Economics and Statistics* **83**: 663–674.

Fogel W. 1966. The effect of low educational attainment on incomes: a comparative study of selected ethnic groups. *Journal of Human Resources* 22–40.

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* **47**: 153–162.

Madden JF. 1975. Discrimination: a manifestation of market power? In *Sex, Discrimination and the Division of Labor*, Lloyd CB (ed.). Columbia University Press: New York; 146–174.

Neuman S, Oaxaca RL. 2004a. Wage decompositions with selectivity corrected wage equations: a methodological note. *Journal of Economic Inequality* **2**: 3–10.

Neuman S, Oaxaca RL. 2004b. Wage differentials in the 1990s in Israel: endowments, discrimination, and selectivity. IZA Discussion Paper 1362.

Neumark D. 1988. Employers' discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resources* **23**: 279–295.

Neumark D, Korenman S. 1994. Sources of bias in women's wage equations: results using sibling data. *Journal of Human Resources* **29**(2): 379–405.

Oaxaca R. 1973. Male–female wage differentials in urban labor markets. *International Economic Review* **9**: 693–709.

Oaxaca RL, Ransom MR. 1994. On discrimination and the decomposition of wage differentials. *Journal of Econometrics* **61**(1): 5–21.

Pagan JA, Davila A. 1996. On-the-job training, immigration reform, and the true wages of native male workers. *Industrial Relations* **35**(1): 45–58.

Poston D, Alvirez D. 1973. On the cost of being a Mexican American Worker. *Social Science Quarterly* **53**: 697–709.

Reimers C. 1983. Labor market discrimination against Hispanic and black men. *Review of Economics and Statistics* **65**: 570–579.

Stone CJ. 1977. Consistent nonparametric regression. *Annals of Statistics* **5**: 595–645.

Thurow LC. 1969. *Poverty and Discrimination*. Brookings Institution: Washington, DC.

Trejo S. 1997. Why do Mexican Americans earn low wages? *Journal of Political Economy* **105**(6): 1235–1268.

Tronstad R. 1995. Importance of melon type, size, grade, container, and season in determining melon prices. *Journal of Agricultural and Resource Economics* **20**(1): 32–48.

Vella F. 1998. Estimating models with sample selection bias: a survey. *Journal of Human Resources* **33**(1): 127–169.

Verdugo RR. 1992. Earnings differentials between black, Mexican American, and non-Hispanic white male workers: on the cost of being a minority worker, 1972–1987. *Social Science Quarterly* **73**(3): 663–673.