

Hansen, Jörgen; Wahlberg, Roger

**Working Paper**

## Endogenous Schooling and the Distribution of the Gender Wage Gap

IZA Discussion Papers, No. 78

**Provided in Cooperation with:**

IZA – Institute of Labor Economics

*Suggested Citation:* Hansen, Jörgen; Wahlberg, Roger (1999) : Endogenous Schooling and the Distribution of the Gender Wage Gap, IZA Discussion Papers, No. 78, Institute for the Study of Labor (IZA), Bonn

This Version is available at:

<https://hdl.handle.net/10419/20912>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

IZA DP No. 78

## **Endogenous Schooling and the Distribution of the Gender Wage Gap**

Jörgen Hansen  
Roger Wahlberg

December 1999

# Endogenous Schooling and the Distribution of the Gender Wage Gap

**Jörgen Hansen**

*IZA, Bonn, Germany*

**Roger Wahlberg**

*Department of Economics, Göteborg University, Sweden*

Discussion Paper No. 78

December 1999

IZA

P.O. Box 7240

D-53072 Bonn

Germany

Tel.: +49-228-3894-0

Fax: +49-228-3894-210

Email: [iza@iza.org](mailto:iza@iza.org)

This Discussion Paper is issued within the framework of IZA's research area *General Labor Economics*. Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent, nonprofit limited liability company (Gesellschaft mit beschränkter Haftung) supported by the Deutsche Post AG. The center is associated with the University of Bonn and offers a stimulating research environment through its research networks, research support, and visitors and doctoral programs. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public. The current research program deals with (1) mobility and flexibility of labor markets, (2) internationalization of labor markets and European integration, (3) the welfare state and labor markets, (4) labor markets in transition, (5) the future of work, (6) project evaluation and (7) general labor economics.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

## **ABSTRACT**

### **Endogenous Schooling and the Distribution of the Gender Wage Gap<sup>\*</sup>**

Previous studies on gender wage discrimination have relied on OLS when estimating the wage equations. However, there exists a number of recent studies, devoted to estimating the return to education, that have shown that OLS may produce biased estimates for a number of reasons. Consequently, if we neglect this potential bias in OLS when estimating the gender wage gap, there is every reason to believe that the estimated wage gaps will also be biased. We first estimate wage equations using instrumental variables procedures and panel data, and then investigate the distributions of wage discrimination experienced. We find overwhelming evidence that OLS seriously overestimate the unexplained gender wage gap.

JEL Classification: J3, J7

Keywords: Return to education, endogenous schooling, panel data, individual effects, distribution of discrimination

Jürgen Hansen  
IZA  
P.O. Box 7240  
53072 Bonn  
Germany  
Tel: +49 228 38 94 223  
Fax: +49 228 38 94 210  
Email: Hansen@iza.org

---

<sup>\*</sup> We are thankful to Christian Belzil, Alan Duncan, Peter Ericson, Lennart Flood, Almas Heshmati, Anders Klevmarken, Thomas Lemieux, and Claude Montmarquette for many helpful comments on earlier drafts of this paper.

# 1 INTRODUCTION

Estimating the return to education and the gender wage gap have been the target for labor economists for decades. In most empirical work, the wage functions are derived from the Mincer equation (Mincer (1974)), relating the logarithm of earnings (or wages) linearly to years of education.<sup>1</sup> For some time, labor economists have argued that estimating this relationship by ordinary least squares (OLS) produce biased estimates for at least three reasons. First, the OLS estimates may suffer from "ability bias". Specifically, if schooling decisions are positively correlated with ability and ability only enters the wage equation through an additive error term, then it can be shown that OLS yields estimates that are upward biased. Second, if the measure of educational attainment (usually years of education) is measured with error, this will produce a negative bias in the OLS estimates of the return to education (see Ashenfelter and Krueger (1994) and Griliches (1977)). Third, schooling attainment can be thought to be a function of ability and subjective discount rates. If these two arguments are correlated (presumably negative), this will also bias the OLS estimates (see Card (1995) and Belzil and Hansen (1998)).

Despite these recognitions of the inability of OLS to obtain consistent estimates in wage regressions, OLS have been the predominant estimator in wage discrimination studies. This is surprising. Arguably, the bias in OLS will not only affect the wage parameters (such as the return to education), but also the wage discrimination itself. The heavy reliance on OLS in the previous literature on gender wage discrimination motivates us to study what effect treating schooling as endogenous will have on the unexplained gender wage gap.

Consequently, our first problem is how to obtain consistent estimates of the wage regressions. One way to reduce the problem of ability bias is to include proxy variables for labor market ability in the wage equation. Typical variables that have served as proxies for ability in previous literature are: scores from IQ-tests (see for instance Blackburn and Neumark (1993) and Griliches (1977)), and information about family background (see Blackburn

---

<sup>1</sup>Recent exceptions are Belzil and Hansen (1998) who estimates the return to education from a structural, dynamic programming model.

and Neumark (1995) and Lam and Schoeni (1993)). A common result from all those studies was that when proxy variables were included in the wage equation, the estimated return to education fell. However, even if these procedures may reduce the ability bias, we still face problems with bias due to measurement error and subjective discounting. To overcome these hurdles we may use instrumental variable techniques. The problem then reduces to finding suitable instruments. One approach is to use exogenous influences, that affects schooling attainment. For instance, information on changes in the minimum school leaving age has been used (see Harmon and Walker (1995)). Card (1993) used data on proximity to colleges as instrument, with the motivation that the cost of attending college is lower if there exists a college close to where the students live. The results from those two studies indicated that OLS underestimate the true return to education substantially.

When panel data is available, we don't have to rely on the cross-sectional approaches outlined above. First, we could assume that ability is an individual specific, time-invariant effect and estimate the parameters with a fixed-effect estimator. This would produce unbiased estimates of the return to education. However, this rate can only be identified from individuals that return to school.<sup>2</sup> Second, since it is difficult to find suitable instruments, we could adopt a procedure proposed by Hausman and Taylor (1981) (hereafter HT). They derived a two-stage least-squares random effects estimator, in which identification is achieved through the use of individual means over time of the included regressors in the primary equation. The main advantage of this procedure is that it *does not require any exclusion restrictions* in order to identify the system, as opposed to the cross-sectional procedures.

Given a set of consistent estimates of the parameters in the wage equation, we are able to assess the proportion of the observed gender wage gap that is attributable to discrimination. To do so, we adopt a methodology suggested by Jenkins (1994) which is based on the distribution of wage differences. This approach contrast the standard practice to decompose the *average* wage gap into explained and unexplained parts (Oaxaca and Ransom (1994), Blinder (1973) and Oaxaca (1973)). As Jenkins argues "*the measurement and*

---

<sup>2</sup>According to our sample, very few individuals return to acquire more schooling (see the data section). The estimated return would then reflect the marginal return of these people, and would not necessarily coincide with the average return in the sample.

*analysis of earnings discrimination should take into account the complete distribution of discrimination experienced.*” That is, focusing on differences in averages does not provide a complete picture of the discrimination faced by women and this measure of discrimination may be consistent with very different distributions of discrimination. To exemplify, suppose that the unexplained part of the average wage gap is found to be 10%. From this measure we are not able to infer whether all women in the sample faces a wage discrimination of 10%, or if half the women experience a discrimination of 20%, and the other half does not experience any discrimination at all. Clearly, relying only on differences in averages can be quite misleading.

Another very attractive feature of the distributional approach is the ability to examine the sources of variations in discrimination. For instance, we can use this framework to investigate if wage discrimination between men and women differ both across different individual characteristics (such as educational attainment) and over time.

The outline of the paper is as follows. In the following section, the empirical framework is discussed. We briefly outline how we estimate the wage equations and describe different measures of discrimination. Section 3 provides a description of the data. Section 4 presents the results, with focus on the size and composition of the discrimination. In section 5 we test the robustness of our results to various potential model misspecifications. The paper ends with conclusions and discussion in Section 6.

## 2 EMPIRICAL FRAMEWORK

In this section, we will first briefly describe how we obtain consistent estimates of a reduced form wage equation, accounting for endogenous schooling. After this, we will describe various measures of wage discrimination that has been used in the literature, both the traditional approach based on decomposition of differences in average wages, and newer approaches based on the whole distribution of wage differences.

## 2.1 ESTIMATING WAGE EQUATIONS WITH ENDOGENOUS SCHOOLING

The empirical relationship to be estimated in this paper follows a traditional Mincer (1974) equation, that relates the logarithm of hourly wages ( $LNW_{it}$ ) to a vector of time-invariant characteristics ( $Z_i$ ) as well as to a vector of time-varying characteristics ( $X_{it}$ ). Formally, the equations to be estimated are:

$$LNW_{it} = \alpha^w + \gamma^{w'} Z_i + \beta^{w'} X_{it} + \mu_i^w + \varepsilon_{it}^w \quad \forall i \in W \quad (1a)$$

and

$$LNW_{it} = \alpha^m + \gamma^{m'} Z_i + \beta^{m'} X_{it} + \mu_i^m + \varepsilon_{it}^m \quad \forall i \in M \quad (1b)$$

where  $W$  denotes the set of women in the sample and  $M$  denotes the set of men. Further,  $i = 1, \dots, N^w$  for women,  $i = 1, \dots, N^m$  for men, and  $t = 1, \dots, T_i$ . Subindex  $i$  denotes individuals and subindex  $t$  denotes time-periods. The last two terms of the above relationships concerns the error structure of the model. Unobserved, time-invariant individual effects (such as, ability, motivation, etc.) are assumed to be captured in  $\mu_i$ , while  $\varepsilon_{it}$  is the remainder disturbance, reflecting effects of unobservable variables that vary both across individuals and over time. It is assumed that: (1) the sequence  $\{\varepsilon_{it}\}$  consists of normal i.i.d. random variables with mean zero and variance  $\sigma_\varepsilon^2$ ; (2)  $\varepsilon_{it}$  and  $\mu_i$  are mutually independent; and finally, (3)  $\mu_i$  is normally distributed with mean zero and a homoscedastic variance  $\sigma_\mu^2$ . As argued in the previous section, there exists compelling evidence that the included human capital variables (especially years of education) are correlated with unobserved ability ( $\mu_i$ ).<sup>3</sup> When applying instrumental variables techniques in studies using cross-sectional data, one problem is to find exogenous variables that can be used as instruments, and at the same time, be legitimately excluded from the wage equation. When panel data information is available, individual means over time of all included regressors can serve as valid instruments (see HT).<sup>4</sup> To obtain the estimates of equations (1a) and (1b), we

---

<sup>3</sup>It is assumed throughout this paper that the stochastic term  $\varepsilon_{it}$  is uncorrelated with all included covariates.

<sup>4</sup>When the panel is balanced, there exists even more efficient instruments (see Breusch, Mizon and Schmidt, 1989, and Amemiya and MaCurdy, 1986).



modify the HT approach to allow for unbalanced panel following Rosholm and Smith (1996).

## 2.2 MEASURING WAGE DISCRIMINATION

As was argued in the introduction, there exists two different ways to measure wage discrimination. The traditional approach uses the regression estimates and average observed characteristics ( $\bar{Z}$  and  $\bar{X}$ ) to partition the average (log) wage gap into explained and unexplained parts. Formally, we can write this as:

$$\overline{LNW}^m - \overline{LNW}^w = (\bar{Z}^m - \bar{Z}^w)\hat{\gamma}^m + (\bar{X}^m - \bar{X}^w)\hat{\beta}^m + \bar{Z}^w(\hat{\gamma}^m - \hat{\gamma}^w) + \bar{X}^w(\hat{\beta}^m - \hat{\beta}^w) \quad (2)$$

where the first two terms represents the part of the observed average wage gap that is attributable to differences in characteristics, and the last two terms summarizes the discrimination. In the above equation we have implicitly assumed the male parameter set as the non-discriminatory (or reference) parameter set, that is, the return to observables women would face in the absence of discrimination. While this decomposition is attractive in its simplicity, it is not capable of describing distribution of discrimination.

Instead of relying on differences in averages, it is more appealing to compare the whole distributions of predicted wages. As a starting point, we need to predict wages for women under two different regimes, one using the set of parameters from the regression based on the female sample, and one using the male parameter set. Specifically, we have:

$$\hat{y}_i = \exp(\hat{\alpha}^w + \hat{\gamma}^{w'}Z_i + \hat{\beta}^{w'}X_{it} + \frac{1}{2}(\hat{\sigma}_{\mu w}^2 + \hat{\sigma}_{\varepsilon w}^2)) \quad \forall i \in W \quad (3a)$$

and

$$\hat{r}_i = \exp(\hat{\alpha}^m + \hat{\gamma}^{m'}Z_i + \hat{\beta}^{m'}X_{it} + \frac{1}{2}(\hat{\sigma}_{\mu m}^2 + \hat{\sigma}_{\varepsilon m}^2)) \quad \forall i \in W \quad (3b)$$

where  $\hat{r}_i$  can be thought of as the reference wage, that is, the wage woman  $i$  would receive in the absence of discrimination.<sup>5</sup> The problem in this framework is how to summarize all the information contained in these two distributions. Jenkins (1994) proposed a number of measures for analyzing the joint distribution of wage differences. The first of these is related to the Gini coefficient, and is based on the area between the Generalized Lorenz Curve (GLC) and the Generalized Concentration Curve (GCC).<sup>6</sup> In the case of no discrimination, these two curves will coincide and the larger the area between the curves is, the more discrimination there is. Formally, this index ( $C$ ) is defined as:

$$C = \left(1 + \frac{1}{2N^w}\right) \left(\frac{\bar{r} - \bar{y}}{\bar{y}}\right) - \left(\frac{1}{N^w}\right)^2 \sum_{i \in W} i \left(\frac{\hat{r}_i - \hat{y}_i}{\bar{y}}\right) \quad (4)$$

where  $\bar{r}$  and  $\bar{y}$  denotes average of respective distribution. This index incorporates the differences in means as in traditional decomposition approaches but also includes a term which is the predicted wage gap weighted by a woman's rank in the predicted wage distribution.

The second measure of wage discrimination suggested by Jenkins (1994) is related to  $C$  but introduces a new parameter ( $\alpha$ ) that incorporates the degree of aversion against discrimination. The index is defined as:

$$J_\alpha = \sum_{i \in W} \omega_i (1 - d_i^{-\alpha}) = 1 - \sum_{i \in W} \omega_i d_i^{-\alpha} \quad (5)$$

where

$$\omega_i = \frac{\hat{y}_i}{N^w \bar{y}}$$

---

<sup>5</sup>These two measures are similar to two of the measures presented in Jenkins (1994). However, since the wage rate is log-normal distributed, we need to include the estimates of the variances within the exponential brackets. Further, by assumption, the two variance components are independent, hence,  $\text{Var}(\mu_i + \epsilon_{it}) = \text{Var}(\mu_i) + \text{Var}(\epsilon_{it})$ .

<sup>6</sup>GLC is obtained by ordering women in ascending order of observed wage and plot the cumulative wage per capita against cumulative sample share. GCC is obtained in a similar fashion except that observed wages are replaced by the predicted reference wages. However, the ordering remains the same as for GLC. For more details, see Jenkins (1994).

$$d_i = 1 + \frac{|\hat{r}_i - \hat{y}_i|}{\bar{r}}$$

$$\alpha > 0$$

The variable  $d_i$  is a normalized wage gap (the gap for a woman relative to the mean of the reference distribution) and  $\omega_i$  is woman i:s wage share.  $J_\alpha$  allows us to aggregate the wage gaps in different ways using different values of  $\alpha$ . Higher values for  $\alpha$  corresponds to greater degree of discrimination aversion. The aversion parameter can be thought to represent the increase in wages required to compensate a women for a small increase in the wage gap.

In the result section below, we will report the values of these indices using both OLS and the HT approach to estimate the wage regression parameters. These indices will also be compared with more conventional measures of discrimination based on differences in averages.

### 3 DATA

The empirical model outlined above is implemented on a sample of labor force participants extracted from the Swedish survey "Household Market and Nonmarket Activities (HUS)". The HUS survey is devoted to the construction of a reliable data base which can be used to study the dynamics of household behavior. The first wave of HUS took place in 1984, and provides detailed information on a random sample of about 2,300 Swedish households. Information was gathered in a personal interview, and the questions cover a broad set of information, such as: family background, education, market work experience, family composition, labor market status at the time of the interview, child care arrangements, wages and incomes, and family composition. The second wave took place in 1986, where the respondents from 1984 were contacted again. Also, a new additional sample was questioned. The participants in 1986 were later contacted in 1988 and in 1991 (with a limited amount of questions) and in 1993. Finally, the most recent wave took place in 1996.<sup>7</sup>

---

<sup>7</sup>For more details on coding procedures, see Klevmarken and Olovsson (1993).

The panels used in this paper are composed from the following waves: 1984, 1986, 1993, and 1996. Since we lack information on some of the included variables for the two waves in 1988 and 1991, they are not included in this study. Individuals for whom information concerning the included variables was missing were excluded. The upper age limit was set to 65 years of age. In addition, we excluded individuals who reported that they were self-employed. Further, only those individuals that are working are considered in this study.<sup>8</sup> After these selections, we were left with 545 males, and 472 females.

Another important selection we have to do is to exclude those who have left the panel. There exist a growing literature on how to account for the potential bias this introduces (for a recent survey, see Vella (1998)). It is however beyond the scope of this paper to correct for this problem. In Tables 1 and 2, we present sample descriptives for both those individuals included in our sample but also for those who attrited from the panel (in other word, those who only participated in one survey). From these tables we can conclude that there does not seem to be any systematic difference between those who stayed in the panel and those who left. Those individuals who left the panel have on average slightly lower work experience and wages, but apart from that the samples seem to be similar. We take this as an indication that attrition from this panel does not introduce any serious bias of our estimates.<sup>9</sup>

Information on wages and incomes comes from two sources, and all refers to the previous year. For about two thirds of the sample, this information was collected from the Swedish tax registers, which ensures a high degree of accuracy. For the remaining one third, this information was collected during the interview. The wage measure used in this paper was constructed as the ratio of the individuals annual labor income to their annual hours of work.<sup>10</sup> Hours of work includes both hours spent on extra work and/or overtime.

The education and market work experience variables are the reported actual years of each activity. The first time the individuals participates in the

---

<sup>8</sup>This may introduce selection bias that should be controled for during estimation. We will discuss this later in section 5.

<sup>9</sup>We have also estimated the wage equations with OLS for those who have attrited. For details, see section 5.

<sup>10</sup>The hourly wage rate was deflated by consumer price index using 1983 as the base year.

survey, they are asked about how many years of schooling and market work experience they have had. In subsequent surveys, individuals are asked to report their labor force status month by month since they last participated. Using this information, we can identify possible changes in both years of education and market work experience between the survey times.

The remaining covariates in the model are defined as follows: four binary variables concerning the educational status of the parents corresponding to medium schooling (essentially high school equivalent education and vocational schooling), and to high schooling (a university degree or at least some university education); we also include one binary variable indicating whether the person's mother worked or not while growing up; two binary variables concerning the work status of the father while growing up (the first is equal to one if the father was farmer, and the second is equal to one if the father was a blue-collar; hence, the excluded category is those whose father was a white-collar), a binary variable that is equal to one if both parents were born outside Sweden, two binary variables identifying the area of residence (medium-sized city or countryside), and a binary variable which is equal one if the person is married. In addition to these variables, time dummies are included to capture macro-economic shocks and general wage inflation. Finally, we also included two age dummies to capture possible cohort effects both on wages and on schooling.

## 4 EMPIRICAL RESULTS

In this section, we first discuss the estimated coefficients from the wage equations. Special attention is devoted to determine the effect of ability and endogeneity on the returns to human capital. The size and composition of the gender discrimination are discussed in sections 4.2 and 4.3.

### 4.1 WAGE ESTIMATES AND RETURNS TO HUMAN CAPITAL

The results from the wage regressions for males are presented in Table 3, and the female results are in Table 4. The first column in these tables refers to estimates obtained by OLS and in the second column, the HT estimates are reported.

Family background variables (dummy variables indicating parents' education, parents' work status, and whether both parents were born outside Sweden or not) are included because they are believed to capture some of the effects of pre-school investments, and hence ability on the labor market. For instance, if both parents were born outside Sweden, this may affect the child's possibility to profit from education. Similarly, the level of parents' education and their working status may be thought of as indicators of intellectual stimulus received from the parents. However, most of the effects of family background on wages are insignificant both for males and for females, even if there seems to exist a positive relationship between parents' education and wages. One exception is the immigration status of the parents. If both parents were born outside Sweden, this has a significant and negative effect on male wages, but not on female wages. The estimates in Tables 3 and 4 shows that persons whose fathers were blue-collar have lower wages, a result that holds for both males and females. Further, individuals residing outside urban areas (Stockholm, Göteborg, and Malmö) have lower wages, especially those living in the countryside. Concerning the effect of marriage on wages, we find a surprising result. Being married or cohabiting does not seem to influence male wages, but for females there is positive and significant effect. Married women's wages are about 6% higher than wages for single women. This is indeed surprising, as one would expect married women's wages to be lower than those of single women because of work interruptions due to child bearing. Our result may then indicate that the depreciation rate of human capital is quite low in Sweden. The marriage coefficient for males is also contradicting most previous work, which has documented a positive effect of marriage on male wages. Finally, the results also reveal that there has been an increase in real wages over the time of about 10% for both males and females.

Another interesting result from Tables 3 and 4 is the quite large difference in returns to human capital between genders. For males, an additional year of schooling implies a wage gain of about 3.5% (in the OLS case), while this figure is only 2.9% for females. Regarding market work experience the gender difference is even larger, an increase with one percent raise male wages with about 0.2% and female wages with only 0.1%. Allowing for schooling to be endogenous has a dramatic effect on the estimated returns. For males, the return increases to over 6%, an increase of 80%. This result is in line

with much of the previous work on endogenous schooling and the return to education, which have typically found that OLS underestimates the true return. For females however, the return decreases to 0.7%, a drop of over 75%. This result is surprising but confirm that the return differs substantially between gender. However, the female return using HT is estimated without precision, and it is not significantly different from the OLS estimate. We will return later in the next section to check the robustness of this result, especially concerning the identifying assumptions made.

Finally, we note that the values of the Hausman test statistics ( $\widehat{TEST}$ ) indicates that instrumentation of the schooling variable is sufficient to remove any correlation between the individual specific terms (ability) and the remaining regressors, as the null hypothesis of zero correlation is not rejected for males ( $\widehat{TEST} = 0.079$ ) or for females ( $\widehat{TEST} = 0.490$ ).<sup>11</sup>

## 4.2 THE SIZE OF THE GENDER WAGE GAP

In Tables 5 we present summary statistics for two different distributions of discrimination. The first set of rows refers to the raw wage gap (defined as  $\hat{r}_i - \hat{g}_i$ ), while the second set of rows refer to the distribution of wage differentials expressed as percentages. Concerning the raw wage gap, we note that for 3% of the women in our sample they would actually be worse off if they faced the male return to observables. However, a vast majority would receive higher wages if they faced the same returns as men. When we treat schooling as endogenous, this proportion is somewhat lower (86% instead of 97%). We further note that the raw wage gap differs substantially between the two estimation approaches. The average wage gap is 6.5 SEK when estimating with OLS but only 4.6 SEK when estimating with HT, a difference of about 40%. That is, the results indicate that OLS substantially overestimates the true gender wage discrimination.

In the second set of rows in Table 5, we again see that the unexplained wage differential is much higher in the OLS case. Moreover, we observe a wide range of the differentials. In the OLS case, the mean unexplained wage differential is about 14%. However, a closer look at the distribution reveals

---

<sup>11</sup>The associated critical values for this test statistic are: 26.0 at the 10%-level and 28.9 at the 5%-level.

that for some women the differential is as high as 43% while for some in the lower end of the distribution, the differential is negative. These results clearly show that there is much more information available than what a traditional decomposition of difference in averages would reveal. Concerning the differentials when using HT, we find the same pattern as for OLS, but with even higher variation.<sup>12</sup>

It is also informative to illustrate the complete distribution of discrimination graphically using the Generalized Lorenz Curve (GLC) and the Generalized Concentration Curve (GCC). In Figure 1 we have drawn the GLC for both the OLS case (GLCO) and for the HT case (GLCHT). In addition, we have also included the GCC for both cases GCCO (OLS) and GCCHT (HT). The first thing we note is that the GCC lies everywhere above the GLC. This is true for both OLS and for HT. This indicates that wage discrimination is present over the whole distribution.<sup>13</sup> Further, the curves in Figure 1 confirm the difference in discrimination between OLS and HT since the curves associated to OLS lies everywhere above that of HT. This strengthens our conclusion that OLS tend to overestimate the unexplained gender wage gap.

In Table 6 we present values of the various indices described in section 2.2. We also report the percentage difference between  $\hat{r}$  and  $\hat{y}$ , evaluated at sample means. We refer to this measure as  $Df$ . This measure is essentially identical to the traditional measure based on differences in average wages (equation 2). In the first column, we report the values when the wage equations were estimated with OLS and in the last two when we used HT. Again, the results are consistent with what we have seen above. OLS yields substantially higher measures of discrimination, regardless of which index we use ( $C$  or  $J$ ) and which value on the discrimination aversion parameter ( $\alpha$ ) we use. The results in the third column are based on a version of our original HT approach. The difference is that we also use indicators for minimum school leaving ages as instruments, in addition to the averages across time of the included variables

---

<sup>12</sup>Our OLS results are similar to results from previous Swedish studies, who have typically found discrimination to be around 10%-15% (see Andersson (1995) and Palme and Wright (1992)).

<sup>13</sup>Note that this does not conflict with the results in Table 5 where a small share of the sample had negative raw wage gaps, since GLC is defined as  $\sum_{i=1}^k \hat{y}_i / N^w$  against  $p = k / N^w$  and GCC is defined as  $\sum_{i=1}^k \hat{r}_i / N^w$  against  $p$ .



in the wage equation.<sup>14</sup> This serves as a validity and robustness check of our original HT results. As can be seen, the values of the different indices do not differ much between the two HT approaches. The index values using the extra information on school leaving age implies slightly higher discrimination than our original HT values. However, there is still a substantial difference between the new HT values and the OLS values.

As a last check of the validity of the difference between OLS and HT in terms of discrimination, we have calculated Generalized Lorenz (GL) and Inverse Generalized Lorenz (IGL) ordinates for the absolute raw wage gap ( $|\hat{r} - \hat{y}|$ ). The results are presented in Table 7. These results serve as discrimination dominance check and as can be seen, the ordinates for OLS lies everywhere above those of HT. This is true both regarding the GL ordinates as well as for the IGL ordinates. The results presented in Tables 5 to 7 have shown overwhelming evidence that OLS may seriously overestimate the unexplained gender wage gap.

### 4.3 THE COMPOSITION OF THE GENDER WAGE GAP

In Table 8 we have decomposed the distributional index by different sample subgroups. The objective is to investigate which groups of women are worse off in terms of wage discrimination. In doing that, we focus on two measures only: the traditional measure ( $Df$ ) and the  $J_\alpha$ -index (using  $\alpha = 0.25$ ). Further, since we have clearly shown the difference between OLS and HT, we only use wages predicted using HT.

As a reference we include the values of  $Df$  and  $J$  for the whole sample. As can be seen, Table 8 contains some very interesting results. First, gender wage discrimination differs substantially across educational groups. Among the low educated, the unexplained wage differential is about 8%. However, among those with a college degree, this differential is almost 14%. This is clearly a result of the very low return to education we have seen for females.

An interesting result is the larger discrimination (15%) among women whose mothers are highly educated. An intuitive explanation for this is the in-

---

<sup>14</sup>Using this type of information as potential instruments was suggested by Harmon and Walker (1995). Sweden has experienced a number of exogenous changes in the educational system over time, and it is natural to use this information here.

tergenerational transfer of human capital from mothers to daughters, which is confirmed in our first stage regression results (not reported here to save space). This implies that daughters of highly educated mothers have on average higher education, and as we have seen, face higher discrimination.

Another interesting question is whether wage discrimination is a cohort related problem. We may expect that women from older cohorts are more discriminated in terms of wages. This expectation is also confirmed in Table 8, where we see that the average unexplained wage differential among young women (younger than 35) is only around 5%, while that of older women is twice as high (around 10%). We can also see that wage discrimination differs substantially across different values of work experience. The average unexplained wage gap among those women with more than 25 years of work experience is over 20%. This fact has at least two explanations: First, there is a large difference in return to work experience between men and women. Second, the strong correlation between age and work experience implies that part of this increase in discrimination can be attributed to age differences.

Finally, we see that there seems to be a time trend in the gender wage gap. In 1983, the average unexplained wage gap was only 5.5% while in 1995, this had increased to 16.6%. A possible explanation for this is the increased wage dispersion in Sweden since the beginning of the 1980s (see Edin and Holmlund (1995)).

## 5 ROBUSTNESS OF THE RESULTS

In an attempt to explore the robustness of our HT results towards possible over-identification, the HT estimates were also obtained using different instrument sets. The results from this sensitivity analysis is found in Tables A1 and A2 in appendix. In the first column of these tables we present the wage equation estimates when the time dummy and the age dummy variables were excluded from our set of instruments. In the second column, we report estimates when the only time-varying variables included among the instruments were the time dummy and the age dummy variables. Hence, in this case, identification hinges on variations in participation rates in the panel, while in the first case identification is mainly achieved through variations in work experience. The third column shows results when we included indicators for

exogenous changes in the Swedish educational system. We have already seen that this did not change the discrimination indices much, and as can be seen from Tables A1 and A2, the estimates are similar to those reported in Tables 3 and 4. This is true regardless of which IV-set used and we take this result as an indication that over-identification is not a problem.

Whenever we use panel data for estimation purposes, we have to be concerned with non-random attrition from the panel. In HUS, there has been a quite substantial drop out of the panel and it is essential to check if this may affect our results. In order to evaluate if those who left the panel are systematically different from those who stayed, we estimated wage regressions only on those who left the panel. The results from these OLS regressions are shown in the first column of Tables A3 (males) and A4 (females). Among those estimates that are significant, the results are virtually identical to those reported in Tables 3 and 4. We take this, together with the similarities in the data that was described in section 3, as an indication that attrition is not a major problem in this study.

As a final check of the robustness of our results, we have estimated versions of Heckman's two-step method to investigate if the fact that we focus only on working individuals imposes any bias. However, the participation rates in market work in Sweden were very high in the 1980's and 1990's, both for men and women. This fact may reduce the importance of adjusting for this type of selection. The results are shown in column two of Tables A3 (males) and A4 (females). Again, the estimates are very similar to the ones reported in Tables 3 and 4, and further, the selection correction terms are not significant at any conventional level.

## 6 CONCLUSIONS

The purpose of this paper has been to study how the unexplained wage gap is affected by treating years of education in a traditional wage equation as endogenous. There exists a number of recent studies that have documented a significant difference between the OLS estimates and the IV estimates of the return to education. However, it has been proven difficult to find legitimately instruments. Using panel information on a sample of Swedish labor

force participants, we adopted a random effects instrumental variable estimator derived by Hausman and Taylor (1981) that uses variation over time to identify the system of equations. For males, the return to education is significantly higher when we use IV (6.3%) than when we use OLS (3.5%). This tendency of OLS to underestimate the return to education has also been found in previous work. However, for females we found no significant difference between the OLS estimate and the IV estimate. In contrast to the results for males, the IV estimate of the return to education was lower than the corresponding OLS estimate.

In order to assess the impact of biased wage estimates on the unexplained gender wage gap, we compared the distributions of predicted female wages using both male and female estimates, following Jenkins (1994). The results showed that the estimated wage differential was substantially higher (about 40%) using OLS than when using IV. The estimated wage gap using OLS was around 14%, while it was only 10% using IV. The main explanation for this reduced wage gap is the increased difference in return to human capital in combination with the fact that women have lower accumulated education, and especially work experience, in our sample. Our results were robust both towards different sets of instruments as well as towards a number of different ways to measure wage discrimination. When we investigated the unexplained wage gap for different subgroups of our sample, we found some additional interesting results. First, the degree of wage discrimination differs across age cohorts. Among the youngest women, the unexplained wage gap was found to be around 5% while it was twice as high for older women. We also saw that the wage gap has increased over time from 5.5% in 1983 to over 16% in 1995.

## 7 REFERENCES

1. Amemiya, T. and MaCurdy, T. (1986) "Instrumental-variable estimation of an error-components model," *Econometrica* 54, 869-881.
2. Andersson, P. (1995) "Two essays in labor economics: On labor market discrimination and on the quality of data," Memorandum No. 216, Department of Economics, Göteborg University.

3. Ashenfelter, O. and Kreuger, A. (1994) "Estimates of the economic return to schooling from a new sample of twins," *American Economic Review* 84, 1157-1173.
4. Belzil, C. and Hansen, J. (1998) "Subjective discount rates, unobserved ability, intergenerational transfers and the return to education," Department of Economics, Concordia University.
5. Blackburn, M. and Neumark, D. (1993) "Omitted ability bias and the increase in the return to schooling," *Journal of Labor Economics* 11, 521-543.
6. Blackburn, M. and Neumark, D. (1995) "Are OLS estimates of the return to schooling biased downward? Another look," *Review of Economics and Statistics* 77, 217-229.
7. Blinder, A.S. (1973) "Wage discrimination: Reduced form and structural estimates," *Journal of Human Resources* 18, 437-455.
8. Breusch, T.S., Mizon, G.E., and Schmidt, P. (1989) "Efficient estimation using panel data," *Econometrica* 57, 695-700.
9. Card, D. (1993) "Using geographic variation in college proximity to estimate the return to schooling," NBER Working Paper No. 4483.
10. Card, D. (1995) "Earnings, schooling and ability revisited," *Research in Labor Economics* 14, 23-48.
11. Edin, P.-A. and Holmlund, B. (1995) "The Swedish wage structure: The rise and fall of solidarity wage policy?," In *Differences and changes in wage structures*, ed. R.B. Freeman and L.F. Katz, NBER, University of Chicago Press, Chicago.
12. Griliches, Z. (1977) "Estimating the returns to schooling: Some econometric problems," *Econometrica* 45, 1-22.
13. Harmon, C. and Walker, I. (1995) "Estimates of the economic return to schooling for the United Kingdom," *American Economic Review* 85, 1278-1286.

14. Hausman, J.A. and Taylor, W.E. (1981) "Panel data and unobservable individual effects," *Econometrica* 49, 1377-1398.
15. Jenkins, S.P. (1994) "Earnings discrimination measurement: A distributional approach," *Journal of Econometrics* 61, 81-102.
16. Klevmarken, N.A. and Olovsson, P. (1993) *Household Market and Non-market Activities: Procedures and Codes 1984-1991*, Almquist and Wiksell International, Stockholm.
17. Lam, D. and Schoeni, R.F. (1993) "Effects of family background on earnings and returns to schooling: Evidence from Brazil," *Journal of Political Economy* 101, 710-740.
18. Mincer, J. (1974) *Schooling, Experience, and Earnings*, Columbia University Press, New York.
19. Oaxaca, R. (1973) Male-female wage differentials in urban labor markets, *International Economic Review* 14, 693-709.
20. Oaxaca, R. and Ransom, M.R. (1994) "On discrimination and the decomposition of wage differentials," *Journal of Econometrics* 61, 5-21.
21. Palme, M. and Wright, R. (1992) "Gender discrimination and compensating differentials in Sweden," *Applied Economics* 24, 751-759.
22. Rosholm, M. and Smith, N. (1996) "The Danish gender wage gap in the 1980s: A panel data study," *Oxford Economic Papers* 48, 254-279.
23. Vella, F. (1998) "Estimating models with sample selection bias: A survey," *Journal of Human Resources* 33, 127-169.

Table 1. Descriptive statistics for the male samples.  
Standard deviations in paranthesis.

Variable	Used Sample	Attrited Sample
	Mean	Mean
% fathers medium schooling	20	19
% fathers high schooling	10	14
% mothers medium schooling	19	20
% mothers high schooling	5	8
% fathers were farmer	28	35
% fathers were blue collar	43	37
% mothers never worked	55	53
% parents were immigrants	6	10
Education (years)	11.65 (3.17)	11.40 (3.66)
Experience (years)	24.08 (10.59)	22.67 (12.96)
% residing in medium-sized cities	55	47
% residing in countryside	21	21
% married or cohabiting	88	85
% age less than 35	20	31
% age between 35 and 55	68	52
Wage per hour (1983 SEK)	53.73 (14.71)	49.32 (15.25)
Number of individuals (N)	545	696
Number of observations (NT)	1339	696

Table 2. Descriptive statistics for the female samples.  
Standard deviations in paranthesis.

Variable	Used Sample	Attrited Sample
	Mean	Mean
% fathers medium schooling	15	20
% fathers high schooling	11	12
% mothers medium schooling	18	22
% mothers high schooling	5	7
% fathers were farmer	34	37
% fathers were blue collar	41	36
% mothers never worked	55	52
% parents were immigrants	9	10
Education (years)	11.46 (3.17)	11.27 (3.36)
Experience (years)	17.37 (9.09)	15.15 (10.04)
% residing in medium-sized cities	53	49
% residing in countryside	20	18
% married or cohabiting	88	85
% age less than 35	19	31
% age between 35 and 55	67	53
Wage per hour (1983 SEK)	43.98 (12.07)	41.91 (12.47)
Number of individuals (N)	472	737
Number of observations (NT)	1141	737



Table 3. Estimated coefficients of wage equations (Males).  
(T-statistics in Parenthesis)

Variable	OLS	HT
Intercept	3.303 (47.04)	2.833 (16.13)
Father medium schooling	-0.019 (1.16)	-0.050 (1.68)
Father high schooling	-0.046 (1.84)	-0.093 (1.98)
Mother medium schooling	0.018 (1.03)	0.013 (0.46)
Mother high schooling	0.114 (3.69)	0.085 (1.61)
Father was farmer	-0.055 (3.07)	-0.047 (1.57)
Father was blue collar	-0.052 (3.13)	-0.030 (1.06)
Mother never worked	-0.008 (0.64)	-0.013 (0.59)
Both parents immigrants	-0.121 (4.80)	-0.081 (1.92)
Education	0.035 (14.96)	0.063 (5.49)
Experience	0.016 (4.29)	0.018 (4.49)
Experience <sup>2</sup> /100	-0.015 (2.17)	-0.011 (1.54)
Resides in medium-sized cities	-0.056 (3.70)	-0.029 (1.35)
Resides in countryside	-0.099 (5.32)	-0.045 (1.59)
Married or cohabiting	0.013 (0.67)	0.011 (0.49)
Year 1986	0.012 (0.73)	0.011 (1.06)
Year 1993	0.102 (5.97)	0.092 (5.59)
Year 1996	0.096 (5.26)	0.080 (3.94)
Age less than 35	-0.038 (0.99)	0.046 (1.38)
Age between 35 and 55	0.026(0.90)	0.067 (2.78)
$\hat{\sigma}_\varepsilon$	0.220	0.133
$\hat{\sigma}_\mu$	-	0.232
Mean of $\theta_i$	-	0.643
$\widehat{\text{TEST}}$	-	0.079
$\overline{\text{R}}^2$	0.338	0.660

Table 4. Estimated coefficients of wage equations (Females).  
(T-statistics in Parenthesis)

Variable	OLS	HT
Intercept	3.246 (57.67)	3.426 (15.34)
Father medium schooling	0.056 (2.59)	0.090 (1.93)
Father high schooling	-0.010 (0.34)	0.023 (0.40)
Mother medium schooling	0.001 (0.58)	0.039 (0.81)
Mother high schooling	0.065 (1.89)	0.119 (1.62)
Father was farmer	-0.040 (1.86)	-0.041 (1.02)
Father was blue collar	-0.031 (1.50)	-0.034 (0.87)
Mother never worked	-0.015 (0.96)	-0.008 (0.28)
Both parents immigrants	-0.019 (0.77)	-0.015 (0.32)
Education	0.029 (11.23)	0.007 (0.35)
Experience	0.010 (3.63)	0.008 (2.08)
Experience <sup>2</sup> /100	-0.012 (2.02)	-0.009 (1.11)
Resides in medium-sized cities	-0.011 (0.66)	0.009 (0.34)
Resides in countryside	-0.039 (1.82)	-0.023 (0.70)
Married or cohabiting	0.060 (2.70)	0.054 (1.61)
Year 1986	0.014 (0.70)	0.019 (1.34)
Year 1993	0.081 (4.02)	0.110 (5.09)
Year 1996	0.046 (2.12)	0.089 (3.40)
Age less than 35	-0.060 (1.91)	0.024 (0.62)
Age between 35 and 55	0.011 (0.48)	0.051 (1.81)
$\hat{\sigma}_\varepsilon$	0.237	0.163
$\hat{\sigma}_\mu$	-	0.293
Mean of $\theta_i$	-	0.627
$\widehat{\text{TEST}}$	-	0.490
$\overline{R}^2$	0.219	0.518

Table 5. Summary statistics for male/female wage distributions.

	OLS	HT
<b>Raw wage gaps</b>		
Mean	6.470	4.610
Median	6.037	3.906
Minimum	-5.098	-7.489
Maximum	24.844	25.269
Std Dev	3.974	4.811
% positive	97%	86%
<b>Differentials (%)</b>		
Mean	14.173	10.188
Median	14.045	9.184
Minimum	-13.771	-17.862
Maximum	42.874	53.444
Std Dev	7.611	10.362

Table 6. Distributional discrimination index estimates.

Index	OLS	HT <sup>1</sup>	HT <sup>2</sup>
$Df$	13.92	9.71	11.368
$C$	0.065	0.045	0.054
$J_\alpha$ $\alpha = 0.25$	0.031	0.025	0.027
$\alpha = 0.5$	0.060	0.050	0.053
$\alpha = 1$	0.116	0.096	0.102
$\alpha = 2$	0.216	0.177	0.189
$\alpha = 5$	0.438	0.362	0.386
$\alpha = 10$	0.653	0.547	0.582

Note: The results using HT<sup>1</sup> is based on the estimates reported in Tables 3 and 4. In HT<sup>2</sup> we have included indicators for minimum school leaving ages as extra identifying variables.

Table 7. Discrimination dominance checks.

Cumulative sample share	Generalized Lorenz ordinates for $ \hat{r} - \hat{y} $		Inverse Generalized Lorenz ordinates for $ \hat{r} - \hat{y} $	
	OLS	HT	OLS	HT
0.10	0.439	0.257	0.883	0.731
0.20	0.921	0.570	1.614	1.217
0.30	1.438	0.918	2.441	1.804
0.40	2.042	1.347	3.188	2.389
0.50	2.637	1.746	3.828	2.860
0.60	3.282	2.221	4.428	3.263
0.70	4.030	2.806	5.033	3.692
0.80	4.856	3.392	5.550	4.040
0.90	5.588	3.879	6.031	4.353
1.00	6.471	4.610	6.471	4.610

Table 8. Decomposition of distributional index  
by sample subgroup (using HT)

Variable	Sample proportion (%)	$Df$	$J_\alpha$ $\alpha = 0.25$
<i>All</i>	<i>100</i>	<i>9.711</i>	<i>0.025</i>
Primary education	46	8.253	0.024
High school	35	9.552	0.024
College/University	19	13.697	0.031
Parents immigrants	9	0.662	0.018
Mother medium or high schooling	24	14.916	0.033
Father medium or high schooling	26	8.915	0.025
Reside outside urban area	72	8.704	0.024
Married or cohabiting	90	8.454	0.023
Age less than 35	18	5.354	0.017
Age between 35 and 55	68	11.062	0.027
Age over 55	14	9.270	0.027
Experience < 15	44	4.402	0.017
Experience between 15 and 25	38	11.337	0.026
Experience over 25	18	20.388	0.040
Year = 1983	26	5.510	0.018
Year = 1985	27	6.439	0.020
Year = 1992	26	12.140	0.027
Year = 1995	21	16.635	0.034

## 8 APPENDIX

Table A1. Estimated coefficients of wage equations  
using different instruments (Males).  
(T-statistics in Parenthesis)

Variable	IV-set 1	IV-set 2	IV-set 3
Intercept	2.483 (5.61)	2.682 (1.83)	3.052 (20.83)
Father medium schooling	-0.071 (1.80)	-0.059 (0.65)	-0.037 (1.29)
Father high schooling	-0.140 (1.92)	-0.113 (0.57)	-0.064 (1.44)
Mother medium schooling	0.007 (0.21)	0.011 (0.26)	0.018 (0.61)
Mother high schooling	0.059 (0.93)	0.073 (0.61)	0.101 (1.97)
Father was farmer	-0.037 (1.10)	-0.043 (0.83)	-0.053 (1.82)
Father was blue collar	-0.014 (0.38)	-0.023 (0.31)	-0.040 (1.46)
Mother never worked	-0.012 (0.49)	-0.012 (0.53)	-0.014 (0.64)
Both parents immigrants	-0.057 (1.10)	-0.071 (0.66)	-0.096 (2.33)
Education	0.087 (2.86)	0.073 (0.72)	0.047 (5.13)
Experience	0.020 (4.07)	0.019 (1.79)	0.016 (4.25)
Experience <sup>2</sup> /100	-0.010 (1.35)	-0.011 (1.32)	-0.012 (1.65)
Resides in medium-sized cities	-0.020 (0.82)	-0.025 (0.60)	-0.034 (1.64)
Resides in countryside	-0.024 (0.64)	-0.036 (-0.40)	-0.058 (2.12)
Married or cohabiting	0.005 (0.20)	0.008 (0.24)	0.014 (0.68)
Year 1986	0.007 (0.63)	0.009 (0.51)	0.013 (1.29)
Year 1993	0.075 (2.84)	0.085 (1.15)	0.103 (6.60)
Year 1996	0.056 (1.60)	0.070 (0.68)	0.095 (5.01)
Age less than 35	0.073 (1.57)	0.058 (0.50)	0.030 (0.92)
Age between 35 and 55	0.077 (2.78)	0.071 (1.50)	0.061 (2.58)
$\hat{\sigma}_\varepsilon$	0.139	0.135	0.131
$\hat{\sigma}_\mu$	0.234	0.232	0.232
Mean of $\theta_i$	0.646	0.643	0.643
$\overline{R}^2$	0.625	0.648	0.671

Note: In IV-set 1, the time dummy variables and age dummy variables are excluded from the instruments. In IV-set 2, the only time-varying instruments are the time dummy variables and the age dummy variables. In IV-set 3 we have added indicators for minimum school leaving age.

Table A2. Estimated coefficients of wage equations  
using different instruments (Females).  
(T-statistics in Parenthesis)

Variable	IV-set 1	IV-set 2	IV-set 3
Intercept	3.335 (4.26)	3.537 (14.40)	3.488 (16.56)
Father medium schooling	0.079 (0.78)	0.104 (2.14)	0.099 (2.12)
Father high schooling	0.011 (0.09)	0.038 (0.64)	0.032 (0.55)
Mother medium schooling	0.026 (0.23)	0.054 (1.09)	0.047 (1.01)
Mother high schooling	0.102 (0.64)	0.139 (1.83)	0.130 (1.80)
Father was farmer	-0.037 (0.76)	-0.045 (1.12)	-0.043 (1.08)
Father was blue collar	-0.031 (0.66)	-0.037 (0.96)	-0.036 (0.92)
Mother never worked	-0.007 (0.23)	-0.010 (0.34)	-0.009 (0.32)
Both parents immigrants	-0.011 (0.20)	-0.019 (0.42)	-0.017 (0.37)
Education	0.015 (0.21)	-0.003 (0.15)	0.001 (0.07)
Experience	0.009 (1.83)	0.008 (1.97)	0.008 (2.02)
Experience <sup>2</sup> /100	-0.010 (1.11)	-0.010 (1.10)	-0.010 (1.10)
Resides in medium-sized cities	0.007 (0.22)	0.011 (0.42)	0.010 (0.39)
Resides in countryside	-0.020 (0.55)	-0.026 (-0.79)	-0.024 (0.75)
Married or cohabiting	0.060 (1.09)	0.048 (1.38)	0.051 (1.51)
Year 1986	0.018 (1.12)	0.020 (1.40)	0.019 (1.38)
Year 1993	0.105 (2.32)	0.116 (5.15)	0.113 (5.31)
Year 1996	0.083 (1.41)	0.097 (3.54)	0.094 (3.63)
Age less than 35	0.018 (0.30)	0.030 (0.78)	0.028 (0.73)
Age between 35 and 55	0.047 (0.96)	0.057 (1.96)	0.055 (1.94)
$\hat{\sigma}_\varepsilon$	0.161	0.165	0.164
$\hat{\sigma}_\mu$	0.293	0.290	0.293
Mean of $\theta_i$	0.627	0.624	0.627
$\overline{R}^2$	0.523	0.509	0.513

Note: In IV-set 1, the time dummy variables and the age dummies are excluded from the instruments. In IV-set 2, the only time-varying instruments are the time dummy variables and the age dummies. In IV-set 3 we have added indicators for minimum school leaving age.

Table A3. Estimated coefficients of wage equations (Males).  
(T-statistics in Parenthesis)

Variable	Attrited Sample	Full Sample
Intercept	3.130 (29.99)	3.211 (41.49)
Father medium schooling	0.044 (1.46)	-0.015 (0.90)
Father high schooling	0.055 (1.38)	-0.081 (3.17)
Mother medium schooling	-0.005 (0.18)	0.021 (1.64)
Mother high schooling	0.012 (0.27)	0.135 (4.19)
Father was farmer	0.012 (0.42)	-0.071 (3.91)
Father was blue collar	0.027 (0.91)	-0.061 (3.60)
Mother never worked	0.024 (1.05)	-0.009 (0.65)
Both parents immigrants	-0.093 (2.84)	0.138 (5.05)
Education	0.034 (9.79)	0.035 (14.77)
Experience	0.019 (3.95)	0.024 (6.91)
Experience <sup>2</sup> /100	-0.025 (2.76)	-0.030 (-4.47)
Resides in medium-sized cities	0.004 (0.16)	-0.055 (3.55)
Resides in countryside	-0.045 (1.53)	-0.104 (5.53)
Married or cohabiting	0.079 (2.70)	0.001 (0.03)
Year 1986	-0.017 (0.72)	0.012 (0.72)
Year 1993	0.106 (4.26)	0.095 (5.10)
Year 1996	-	0.072 (3.79)
Age less than 35	-0.080 (1.32)	-0.007 (0.15)
Age between 35 and 55	-0.010 (0.22)	0.016 (0.43)
$\hat{\sigma}_\varepsilon$	0.257	0.233
Selection correction term	-	0.019 (0.13)



Table A4. Estimated coefficients of wage equations (Females)  
(T-statistics in Parenthesis)

Variable	Attrited Sample	Full Sample
Intercept	3.466 (47.90)	3.254 (40.99)
Father medium schooling	-0.030 (1.10)	-0.050 (2.38)
Father high schooling	0.026 (0.67)	-0.022 (0.79)
Mother medium schooling	-0.003 (0.10)	-0.002 (0.10)
Mother high schooling	0.075 (1.64)	0.081 (2.30)
Father was farmer	-0.061 (2.17)	-0.054 (2.56)
Father was blue collar	-0.058 (2.07)	-0.047 (2.26)
Mother never worked	-0.002 (0.08)	-0.006 (0.41)
Both parents immigrants	-0.042 (1.30)	-0.015 (0.60)
Education	0.022(6.22)	0.027 (10.05)
Experience	0.003 (0.74)	0.011 (2.83)
Experience <sup>2</sup> /100	0.004 (0.41)	-0.013 (1.72)
Resides in medium-sized cities	-0.028 (1.26)	-0.008 (0.49)
Resides in countryside	-0.052 (1.76)	-0.044 (2.09)
Married or cohabiting	0.048 (1.76)	0.058 (2.63)
Year 1986	-0.012(0.52)	0.014 (0.62)
Year 1993	0.066(2.76)	0.078 (3.95)
Year 1996	-	0.040 (1.54)
Age less than 35	-0.146 (3.55)	-0.052 (1.61)
Age between 35 and 55	-0.037 (1.11)	0.024 (1.00)
$\hat{\sigma}_\varepsilon$	0.259	0.248
Selection correction term	-	-0.033 (0.29)

**Figure 1. Generalized Lorenz Curves and Generalized Concentration Curves for OLS and HT**

