CenEA Working Paper Series WP08/09

Count your hours: returns to education in Poland

Michał Myck
Anna Nicińska
Leszek Morawski

Abstract:
Combining information from two Polish surveys from 2005 and taking advantage of the Polish microsimulation model (SIMPL) we demonstrate how different the estimates of returns to education can be depending on whether we use net or gross, and monthly or hourly wages, and we examine the role of correcting for employment selection. Annual rates of return to university education for men vary from 6.7% to 9.7% and for women from 8.0% to 13.4%. We show that simple linear estimation performs relatively well for men, while family demographics seem to be the “second best" exclusion restriction in the case of the estimation for women.

Centre for Economic Analysis
CenEA
www.cenea.org.pl

Count your hours. Returns to education in Poland.*

Michał Myck†, Anna Nicińska‡, and Leszek Morawski§

July 29, 2009

Abstract

We show how significant may be the difference in the estimated returns to education in Poland conditional on the measure of wages used and the estimation approach applied. Combining information from two different Polish surveys from 2005 and taking advantage of the Polish microsimulation model (SIMPL) we demonstrate how different the results can be depending on whether we use net or gross, and monthly or hourly wages, and show how important selection correction is for the conclusion. While there are several papers examining the wage equation in Poland, so far none of them has provided a comprehensive analysis of the effects of using different methods and the issue of selection-correction in the estimation of the wage equation in Poland has not been examined in detail. Annual rates of return to university education for men vary from 6.7% to 9.7% and for women from 8.0% to 13.4% when we compare results using net monthly wages without correcting for labor market selection to those from a selection corrected specification using gross hourly wages. We also demonstrate that simple linear estimation performs relatively well for men in comparison to our preferred selection corrected estimation, while using family demographics as exclusion restrictions seems to be the "second best" in the case of the wage equation estimation for women.

Keywords: returns to education, wage equation, selection models, instrumental variables,
JEL Classification: J31, J21.

*Data from the Polish Household Budgets' Survey and the Polish Labor Force Survey have been provided by the Polish Central Statistical Office. The SIMPL micro-simulation model for Poland used in this paper has been developed with support of the Polish Ministry of Labor and Social Policy (for more details see: www.cenea.org.pl). This support is gratefully acknowledged. We are grateful to Costas Meghir for extremely valuable comments. The usual disclaimer applies.

†Michał Myck, corresponding author, (DIW-Berlin, Centre for Economic Analysis - CenEA, and Institute for Fiscal Studies), DIW-Berlin, 58 MohrenStr. 10117 Berlin, Germany; e-mail: mmyck@diw.de

‡Anna Nicińska (Department of Economics of the University of Warsaw), ul. Długa 44/50, 00-241 Warszawa, Poland; e-mail: anicinska@wne.uw.edu.pl

§Leszek Morawski (Department of Economics of the University of Warsaw and Centre for Economic Analysis - CenEA), ul. Długa 44/50, 00-241 Warszawa, Poland; e-mail: lmorawski@wne.uw.edu.pl

1 Introduction

The most common approach to the estimation of the determinants of wages uses the gross hourly wage as the dependent variable and relates it to a number of exogenous characteristics (e.g. Arellano and Meghir (1992), Harmon and Walker (1995), Heckman and Vytlacil (2001) and numerous others). At least since Hurd (1971) it has been recognized that non-random selection into employment will be likely to bias the estimated coefficients, and Heckman's (1979) seminal contribution, providing a parametric solution to account for this, has established the standard and common way of addressing this issue and has found ample application.[1] While due primarily to data availability, selection corrected estimations have most often used demographic characteristics as instruments, recently more structural approaches have been applied with instruments related more directly to labor supply decisions.[2]

In this study we provide a detailed analysis of wage determinants in Poland with a particular attention given to the estimation of returns to education and with the aim to address several important issues which in our view have not been dealt with sufficient care in the existing literature on Poland.[3] The novelty of our analysis consists of two elements. First of all we take advantage of the recently developed Polish micro-simulation model SIMPL. This on the one hand allows us to generate gross earnings from net values reported in the data, and on the other facilitates the use of the structural approach to selection correction as in Blundell, Reed, and Stoker (2003). Secondly, we combine information from two most important Polish sources of micro-data, the Polish Household Budgets' Survey (PHBS) and the Polish Labor Force Survey (PLFS) to make use of their respective advantages and as a result produce estimates of returns to *gross hourly wages*.

In addition to this we focus in detail on the consequences of "deviations" from our reference specification and demonstrate the role played by the assumptions concerning

[1] See e.g. Callam and Harmon (1999), Brunello and Miniaci (1999), Blundell, Dearden, and Sianesi (2005), and Boockmann and Steiner (2006). For a review of different methods of treating models with sample selection see Vella (1998).

[2] Examples of using demographics as instruments are presence and age of children, (Dustmann and Schmidt (2000), Hoynes (1996)), parent's education and whether mother ever worked (Neumark and Korenman (1994)). Gregg, Johnson, and Reed (1999), Blundell, Duncan, McCrae, and Meghir (2000), and Blundell, Reed, and Stoker (2003) used instruments more directly interpretable in a structural way.

[3] The principal studies estimating the wage equation in Poland in the literature are Duffy and Walsh (2001), Bejdechi, Hartog, and van Opheim (2004) Keane and Prasad (2006), and Newell and Socha (2007).

the exclusion restriction in the selection-corrected models. While many studies stress the importance of the instrument in estimating the Heckman-type models, in non-numerical studies one rarely finds comparisons of estimates derived under different assumptions. These assumptions are very often related to availability of data and we use comparisons with our preferred specification as a reference point to demonstrate their consequences for the resulting degree of collinearity and for estimation of the parameters of the wage equation.[4] We show that different exclusion restrictions may lead to significantly different results, and suggest a high degree of caution with regard to the choice of the instrument for the selection equation.

Due to data limitations we do not address two further potential issues which may bias the estimated education coefficients due to on the one hand omitted ability (upward bias, e.g. Grilliches (1977)) and on the other due to measurement error (downward bias, e.g. Ashenfelter and Krueger (1994) or Blackburn and Neumark (1995)). Our results will thus have to be considered with these two caveats in mind, and the solution to these sources of bias will have to wait for richer data sources. We are also far from claiming that the approach taken in this paper provides a "definitive" estimation of returns to education in Poland, and our preferred specification may still be questioned. We demonstrate in detail, however, how strongly the existing studies may be misreporting the role of education in determining the level of the wage.

The results show important differences in the estimates of returns to education depending on the use of a specific earnings measure as the dependent variable and the method used. Correcting for employment selection is particularly important for women but the results for both men and women can vary importantly depending on the exclusion restriction applied. We show that for estimates of returns to a year of higher education moving from net monthly earnings without selection correction to gross hourly wages correcting for selection means a change from 6.7% to 9.7% for men and from 8.0% to 13.4% for women. One of the most important results of the study is the difference in the estimated returns to education between monthly earnings and hourly wages. For men this difference is more important that accounting for sample selection. In fact, using the empirical Mean Squared Error we show that a simple OLS wage equation gives very close and in the case of vocational training and secondary

[4]From this point of view the article is similar in nature to e.g. Mroz (1987) who examines the consequences of different economic and statistical assumptions with regard to the estimates of the models of female working hours.

education even superior results to our reference specification. In the case of women the choice of the exclusion restriction is much more important. Instrumenting selection using the number and age of children, however, seems to be a good "second best" approach for estimating the parameters of the wage equation.

The paper is organized as follows. In Section 2 we discuss our approach with reference to the existing literature and available data, and discuss the role of the microsimulation model (SIMPL) used in the analysis. The data we use are presented in Section 3 and this is followed by a description of results in Section 4. In Section 4.1 we present results for different definitions of the dependent variable and demonstrate the difference in the estimated coefficients for the linear OLS specification and our reference specification which accounts for selection. Section 4.2 discusses the consequences of adopting different exclusion restriction assumptions in relation to the reference specification, while in Section 4.3 we summarize the results and outline the differences in the resulting annual rates of return. Conclusions follow in Section 5.

2 Model and estimation

2.1 Polish labor market data and existing estimates of the wage equation in Poland

There are relatively few studies focused on the determinants of wages in Poland and the existing literature reflects the shortcomings of available data, the very shortcomings which we attempt to overcome in this paper. Until recently labor market and incomes analysis in Poland could have been conducted using three sources of data: the Polish Labor Force Survey (PLFS), the Polish Households' Budget Survey (PHBS) and the Autumn Earnings Survey (AES).[5] The AES is a company based study, and as such while it contains information on gross earnings and the number of hours worked, there is no information on the non-working population, household structure, etc.

PHBS and PLFS surveys collect information on the entire population but include only *net monthly earnings*. Moreover while PLFS contains information on the number

[5]The Polish names of the surveys are respectively, PLFS: Badanie Aktywności Ekonomicznej Ludności (BAEL), PHBS: Badanie Budżetów Gospodarstw Domowych (BBGD) and AES: Sprawozdanie o Strukturze Wynagrodzeń Według Zawodów (Z12). All these surveys are conducted by the Polish Central Statistical Office, GUS.

of hours of work it has no information on non-labor income, and the opposite is true for the PHBS. In addition to that the degree of non-response for wage information in the PLFS is close to 30%. The recently available SILC data in some respects addresses these shortcomings but relative to the PHBS and PLFS is much smaller and in addition to that has not yet been incorporated into a micro-simulation model which is central to our analysis.[6] As a result of data availability the existing estimates of wage equations in Poland have been conducted either on monthly net earnings (e.g. Bejdechi, Hartog, and van Opheim (2004) and Keane and Prasad (2006) using the PHBS or Duffy and Walsh (2001) using PLFS) or on hourly net wages using PLFS (Newell and Socha (2007)), and neither of these studies uses the structural approach to selection applied here. As we explain in Section 2.2 below our approach deals with these shortcomings by on the one hand imputing the hours information from PLFS to the PHBS data, and on the other by taking advantage of the possibility to simulate gross earnings and out-of-work income in the PHBS using the Polish micro-simulation model SIMPL. In Sections 2.3 and 2.4 we provide further details on our approach. Following the comparison of estimations using different definitions of the dependent variable, the analysis focuses on the consequences of assuming different specifications of the wage equation with gross hourly wage on the left-hand side of the equation.

2.2 From net monthly earnings to gross hourly wages

The final specifications of the wage equation we estimate in this paper use *gross hourly wages* and correct for non-random sample selection using the Heckman selection model. These specifications therefore take the well-known form of:

$$ln\left(\frac{w_i}{h_i}\right) = \beta' X_i + \beta_\lambda \hat{\lambda}_i + \epsilon_i \tag{1}$$

where w_i is the grossed monthly wage and h_i are the monthly hours of work of individual i; X_i is a vector of explanatory variables (such as age, education, region, etc.), $\hat{\lambda}_i$ is the inverse Mill's ratio from the participation equation, and ϵ_i is the individual iid residual. An important thing to note concerning regressions which use monthly

[6]For example O'Dorchai (2008) uses the SILC data to analyze gender and motherhood pay gap in Europe. Samples of working individuals in the Polish SILC data used in this paper are about a third of the size of the PHBS data used here.

or annual earnings and ignore intensity of work in the wage equation is the resulting omitted variable problem. If we were to regress monthly wages w_i on Xs and $\hat{\lambda}_i$ alone the regression would then take the form of:

$$ln(w_i) = \beta' X_i + \beta_\lambda \hat{\lambda}_i + \upsilon_i \tag{2}$$

where: $\upsilon_i = \epsilon_i + ln(h_i)$.

The expected consequences of such a formulation would be a negative bias of coefficients on variables negatively correlated with hours of work like advanced education or age (Abowd and Card (1989)), and a positive bias of coefficients on variables positively correlated with work intensity such as marital status or certain regional variables. To have an estimation which is unbiased due to unobserved hours we estimate returns to education using the following formulation:

$$ln(w_i) - ln(h_i) = \beta' X_i + \beta_\lambda \hat{\lambda}_i + \epsilon_i \tag{3}$$

and propose a procedure to substitute the actual hours of work $ln(h_i)$ with it's expected value so that the estimated returns are based on the following expression:[7]

$$E[ln(w_i)|h_i > 0, X_i] - E[ln(h_i)|h_i > 0, X_i] = \beta' X_i + \beta_\lambda \hat{\lambda}_i + E[\epsilon_i|h_i > 0, X_i, \hat{\lambda}_i] \tag{4}$$

These final estimations are conducted on individual earnings information in the Polish Household Budgets' Survey in 2005, but since hourly gross wages are not available in the data two intermediate steps are necessary prior to the final estimation. First of all we use the Polish microsimulation model, SIMPL, to calculate gross from net earnings reported in the data.[8] Secondly, because hours of work are not reported in the PHBS data we impute them using the information provided in the Polish Labor Force Survey.[9] Additionally we follow Blundell et al. (2003) and use simulated out-of-work income, also computed using the SIMPL model, as an instrument for the selection equation. In fact in our reference specification we use three instrumental variables for selection:

[7] We are grateful to Costas Meghir for his comments which helped us formulate this final expression.

[8] This requires an imputation of employees' social security contributions, the value of personal income tax and of universal health insurance. The net to gross conversion is conducted by backward inversion. As Bargain et al. (2007) showed this leads to a very good approximation of the administrative gross wage distribution.

[9] See Arellano and Meghir (1992) for an example of using complementary information from different data sets.

- (simulated) family disposable income in the scenario when the individual is out of work (and its interaction with the married dummy),
- a dummy variable conditional on whether a household is single- or multi-family,[10]
- for multi-family households additionally (simulated) equivalized income of household members who do not belong to the tax unit of the individual.

All three are assumed to affect the participation decision but not the wage level. Family disposable income in the non-employment scenario is computed as a sum of other family members' earnings and other incomes including all benefits and social assistance that a family is entitled to according to its financial and demographic situation. For the individuals observed as working this is simulated under the assumption of their zero earnings level. The out-of-work equivalized income of other household members captures other families' earnings and all benefits that the whole household is entitled to, such as for instance the housing benefit or family benefits.

To account for the fact that we use estimated and not observed hours of work the standard errors in the wage equation need to be corrected (see e.g. Arellano and Meghir (1992)). This is done using a triple-bootstrap procedure as follows. We first use a nonparametric bootstrap for the log (monthly) hours equation which is estimated on k bootstrapped PLFS samples. For each of those k samples we draw m sets of hours equation parameters given the estimated mean values and the estimated variance-covariance matrix. On the basis of these m sets of parameters (for each draw k) we compute the expected log monthly hours of work. Equation 3 thus becomes:

$$ln(w_i) - \widehat{ln(h_{i,k})} = \beta' X_i + \beta_\lambda \hat{\lambda}_i + \epsilon_i \tag{5}$$

where $\widehat{ln(h_{i,k})} = (\sum_m E[ln(h_{i,m,k})|Z_i, h_i > 0])/m$ and Z_i are variables used in the log hours equation estimated on the PLFS sample and for prediction on the PHBS sample.[11] The difference: $[ln(w_i) - \widehat{ln(h_{i,k})}]$ is our dependent variable in specifications which account for hours of work.

Each specification which accounts for the variation of hours worked is then estimated on n bootstrapped PHBS samples for each set of expected hours $\widehat{ln(h_{i,k})}$. This

[10] Family is defined similar to a tax unit, i.e. as a single adult or a couple with or without dependent children.

[11] These variables include industry dummies, regional dummies, part-time dummy, female dummy, age group dummies, and interactions of these variables. Details are available from the authors on request.

gives us $k * n$ vectors of the wage equation parameters, the distribution of which gives us the bootstrapped measures of confidence intervals.[12]

2.3 Collinearity tests of selection instruments

Two main issues concerning the performance of the selection-corrected estimators are the assumption of joint normality of error terms of the selection and wage equations and the potential for collinearity of the wage equation regressors. As noted by Puhani (2000) the literature gives little guidance in terms of the consequences of the violation of the first assumption, and there seems to be a consensus that collinearity is the more important problem of the two. Collinearity is a well documented problem of the Heckman-style selection correction approach. Even if estimations include appropriate exclusion restrictions, as will be the case in our reference specification, a high degree of collinearity can lead to unrobust results. While collinearity tests ought to be a crucial element of any selection analysis they are often neglected and/or unreported. We use the analysis in this study to show that the degree of collinearity changes significantly depending on the chosen instrument and show that in some cases collinearity may lead to a superior performance of a simple OLS estimation versus a selection corrected specification, even in relatively large samples that are at our disposal.

We use two measures of collinearity, the variance inflation factor (VIF) and the condition number. The first measure derives from the R^2_λ, a multiple correlation measure in a regression of $\hat{\lambda}_i$ on X_i, and is computed as:

$$VIF = \frac{1}{1 - R^2_\lambda}. \tag{6}$$

The condition number provides a more complete test of collinearity and is considered to be a better measure its degree (Leung and Yu (1996)).[13] The condition number of a matrix $\mathbf{A}$ is computed as a square root of the ratio of the largest to the smallest eigenvalue of the matrix:

$$\kappa(\mathbf{A}) = \left[\frac{\mu_{max}}{\mu_{min}}\right]^{1/2} \tag{7}$$

[12] In the estimation for each specification we estimate the wage equation 625 times taking $n = 25$ and $k = 25$; we also use $m = 25$ for the number of vectors of parameters on which the expected hours $\widehat{ln(h_{i,k})}$ are computed.

[13] Note also that unlike the VIF, the condition number will also signal collinearity between the X_i variables. In our early analysis it turned out for example that important collinearity issues emerged as a result of including an age polynomial. The degree of collinearity was reduced when age included using age group dummies.

where μ_{max} and μ_{min} are the maximum and minimum eigenvalues of the **A** matrix, and **A** = **X'X**, where **X** is the $p * n$ matrix of explanatory variables (including the ι vector of ones).[14]

There is no clear threshold determining when the level of collinearity should be considered as "high". Leung and Yu (1996) and Greene (2003) suggest that estimations generating condition numbers already above 20 should be treated with caution, while Belsley, Kuh, and Welsch (1980) (p.105) and Belsley (1991) (p.56) classify values between 30 and 100 as "moderate to strong relations", though at the same time point out that condition numbers in excess of 100 "are not uncommon in nonexperimental data matrices" (Belsley (1991), p.77).[15] In the analysis we will use the collinearity tests to assess the performance of the chosen specifications and exclusion restrictions and relate them to a broader measure of quality of estimation, namely the empirical mean square errors.

2.4 Instruments, collinearity and empirical mean square errors

As noted earlier our preferred specification of the wage equation will be a selection corrected equation using *log hourly gross* wages as the dependent variable, and instrumenting selection with simulated out of work income variables. As we shall see in Section 4 estimates of returns to education in Poland in this specification differ significantly from those using different measures of wages and/or neglecting the selection bias. An important additional issue we shall examine in detail is the question of the extent of estimation error resulting from different exclusion restriction assumptions. This is particularly important in various types of selection models as it is often the case that due to data availability one is restricted in the use of a specific instrument and can't use the approach taken in the case of our reference specification.

[14]The columns of **A** must be additionally scaled to have a length of 1, because eigenvalues are affected by column scale. Our analysis is conducted using STATA8 and we use the code provided by Blasnik (1998) to compute the condition numbers.

[15]There is no direct relationship between the VIF and the condition number and the correspondence between them relies on the nature of the data, though higher VIF generally imply higher condition indexes. In numerical experiments Belsley (1991) (p.107) detects the correspondence of condition numbers to correlations as roughly: 10≈0.5($VIF = 2$), 30≈0.9($VIF = 10$), 100≈0.99($VIF = 100$), 300≈0.999($VIF = 1000$). As we shall see in Section 4 our empirical results are very close to this pairing.

In our analysis we follow the approach used for example in Leung and Yu (1996) and Madden (2008) and apply the criterion of the mean square error (MSE) of the parameters of interest computed as the sum of the variance of a parameter plus the square of its bias. Since we do not know the true value of the parameters we assume the "true" vector of parameters to be those from our reference specification, and then use the so-called empirical MSE test (Toro-Vizcarrondo and Wallace (1968)) to judge the various different specifications against each other. By assumption the MSE of the reference model has no bias and the computed MSE for the parameters estimated in our reference specification will only be their variance. The MSEs of parameters from estimations compared to this reference specification will most likely include both the variance and some bias. For each specification the results reported in Section 4.2 will be given in terms of the empirical MSE ratio computed as:

$$\varphi_{mse} = \frac{MSE^{Sk}_{\beta_j}}{MSE^{S*}_{\beta_j}} \tag{8}$$

where $MSE^{S*}_{\beta_j}$ is the MSE of the β_j parameter from our reference specification $S*$ (i.e. by definition their variance), while $MSE^{Sk}_{\beta_j}$ is the MSE of the β_j parameter in some other specification Sk. Any value lower than 1 will imply superiority of the alternative specification Sk, while values greater than 1 will support the reference specification.

For each specification we shall also compute the two measures of collinearity in the wage regression data which, as we shall see, vary significantly depending on the exclusion restrictions we make. This will allow us to examine the importance of exclusion restrictions in the estimation of returns to education both for the resulting degree of collinearity *per se* and for its implied influence on the MSE ratio.

3 Data

The data we use for the estimation of the wage equation come from the Polish Household Budgets' Survey (PHBS) 2005. The sample is composed of individuals aged 18-54 (women) and 18-59 (men) whose employment status is known and indicates their capacity to work. From the sample we exclude students and dependent children, individuals on maternity leave, the self-employed and those helping in family enterprize. Disabled individuals with significant and medium level of disability are also excluded. We also select out individuals with wages higher than the 99.5 percentile and lower

than 0.5 percentile of the distribution. Some individuals who declare employee status and do not report their wages are also excluded.[16] The final PHBS sample we use contains 19,999 men and 20,743 women. Of these 14,015 men and 11,958 women belong to the employed sample with wage observations.

[TABLE 1 ABOUT HERE]

Table 1 presents descriptive statistics for a set of characteristics in the PHBS sample which we use for the estimation. Individuals with the higher education degree make up 15.8% of the total sample and only 6.9% of the non-working sub-sample. The distribution of individuals with secondary education is roughly symmetric over the working and not-working sub-samples (31.5% and 32.6% respectively). What's notable is the difference in educational achievement between men and women, which is especially strong in the case of the working sample. Working women are almost twice as likely to have a higher education degree compared to working men (27.6% vs. 14.9%), and are four times more likely to have professional post-secondary qualifications (6.0% vs 1.6%). 44.2% of men and 27.1% of women completed vocational education.

The individuals in the sample are about 39 years old on average, with the non-working sample slightly older on average than the working population. Most of the estimation sample (almost 60 per cent) have at least one child living with them in the household. Non-working women in the sample are more likely to be married compared to non-working men (67.1% vs. 55.8%). In the non-working sample men are much less likely to have children compared to non-working women, but the probability of having a child is almost the same in the working sample (61.3% vs 60.0%). Among the working individuals about 8% work part time, and 39% work in the public sector (31% of men and 47% of women).

The information on hours of work comes from the Polish Labor Force Survey (PLFS) 2005. The PLFS is a quarterly rolling panel and we use information on individuals who are observed for the first time in the panel in each of the four quarters in 2005. Similar selection criteria are applied to the PLFS as to the PHBS data which gives us a sample of 13,097 employees with hours observations. Some descriptive

[16]This is the case only for 638 individuals, which gives a non-response level of only 2.5%, and confirms the relatively high quality of the earnings data in the PHBS. The PHBS non-response is far lower compared to earnings non-reporting in the PLFS data, which grew from 8 per cent in 1998 to 27 per cent in 2002 (Newell and Socha (2007)).

statistics including a breakdown of hours of work by education is given in Table 8 in the Appendix.

4 Returns to education - results

4.1 Returns to education: the dependent variable and the role of selection correction

The first set of results combines estimates of returns to education generated using different specifications which have either been used in the existing literature on Poland or which serve as reference to judge the performance of our reference specification and the effect of various approaches on the estimated parameters. Given the focus of the paper we present only the key parameters on education variables. These are shown for the full sample (in which case we include also the coefficient on the female dummy variable), and for the separate estimations for men and women. Results for five different specifications are given in Table 3 and include:

- Specification 1: OLS estimates using *(log) net monthly earnings*,
- Specification 2: OLS estimates using *(log) gross monthly earnings*,
- Specification 3: OLS estimates using *(log) gross hourly wages*,
- Specification 4: selection corrected estimates (Limited Information Maximum Likelihood, LIML) using *(log) gross monthly earnings*,
- Specification 5: selection corrected estimates (LIML) using *(log) gross hourly wages*,

Education variables beyond secondary education are included in a "sequential" fashion, i.e. all those with reported higher and post-secondary education are also assigned secondary education, so that the reported coefficients can be more readily interpreted as returns to education. Apart from education variables all of these specifications use the same vector of control variables X_i. These include age group dummies, family composition variables, controls for disability, 15 regional dummies and controls for town size. Every specification estimated on the full sample includes also a female

dummy to control for gender. Details concerning the variables included are given in Table 2. In the case of Specifications 3 and 5 where we use gross hourly wages as the dependent variable, the standard errors are computed using the triple-bootstrap procedure outlined in Section 2.2. Details of the estimations for Specification 5 are presented in the Appendix in Tables 9 and 10.[17]

[TABLE 2 ABOUT HERE]

Several general conclusion can be drawn on the basis of results shown in Table 3. First of all, while there are noticeable differences between OLS results using net and gross monthly earnings (Specifications 1 and 2), the differences are relatively small and often not statistically significant. This is perhaps not so surprising given the limited degree of non-linearity in the Polish labor tax system in 2005 (Morawski and Myck (2009)) and the fact that we use the log of earnings as the dependent variable. However small the nonlinearity is though, it does imply greater estimates of returns to education in cases of all estimates and also a higher earnings penalty for women. The largest estimated difference in percentage points is on the secondary education coefficients which differ by 3.8 percentage points (pp) for the full sample, by 3.7pp for men, and by 4.1pp for the female sample. Coefficients on higher education are about 2.6pp higher for men and 3.4pp higher for women in Specification 2.[18]

The next two specifications (Specifications 3 and 4) demonstrate the difference in the estimated education coefficients resulting (separately) from using monthly earnings rather than hourly wages and using monthly values but omitting the non-random selection into employment. Relative to Specification 2 using *gross hourly wages* leads in some cases to economically large and statistically significant differences in the values of the estimated education coefficients. In cases of all studied samples, and consistent with differences in the hours distributions by education level (see Table 8 in the Appendix), estimates of coefficients on higher and post-secondary education are greater when we use hourly wages and those on secondary and vocational education are lower. The coefficient on higher education is about 9.2pp higher for men and 14.7pp higher for women when we use hourly rather than monthly earnings. There are also large

[17]For other specifications full details are available from the authors.

[18]It should be noted, though, that due to a significant increase in the non-linearity of the Polish labor tax system in the recent years (as documented in Morawski and Myck, 2009) this bias will be higher for estimates based on data from 2008 onwards.

differences in the coefficients on post-secondary education (about 4pp for men and women), though these are not statistically significant. The coefficient on secondary education is 8.4pp lower for women when we account for hours and this difference is statistically significant at 5%. It is also notable that the female dummy is reduced by 10.6pp once differences in hours of work are taken into account.

Looking at the difference in the estimates resulting from neglecting selection (Specification 2 vs Specification 4) confirms the well-known role of selection in biasing results, and as one could expect the bias resulting from neglecting selection is greater among women. The bias is always positive with the exception of post-secondary education among men, but is often not statistically significant. In the case of higher education the bias for men is about 2.2pp while for women is 7.0pp (and statistically significant). The highest bias resulting from neglecting non-random selection is found on secondary education coefficient for women (9.2pp). Interestingly the differences in secondary and vocational education coefficients both for men and for women between Specification 2 and 3 and between 2 and 4 are of similar magnitude but of opposite sign. When the two corrections are taken jointly in Specification 5 (i.e. when we use gross hourly wages and correct for selection) in the case of these two levels of qualification the two biases cancel each other out. The same applies for the coefficient on the female dummy variable which is almost exactly the same in Specifications 2 and 5.

However, as we can see in the final set of results presented in Table 3, i.e. for our *reference* Specification 5, the biases on higher and post-secondary education resulting from using monthly earnings and neglecting employment selection reinforce each other. This results in very large and statistically significant differences particularly in the level of coefficients on higher education. While when we use gross monthly earnings and neglect selection (Specification 2) the coefficient on higher education for men is 0.362, in Specification 5 it is 0.483, which gives a 12.1pp difference in the overall return to having higher education. For women this difference is even higher at 23.8pp, and the difference on post-secondary education for women is as high as 7.6pp though this is only significant at 10%. The specific bias resulting from using hourly wages rather than monthly earnings in the selection-corrected models can be examined by comparing Specification 5 and 4, while the bias in the estimated coefficients which results from correcting for selection by comparing Specification 5 and 3. The overall pattern of results confirms the differences we discussed with respect to Specifications 2, 3 and

4. Selection correction is far less important for men than for women, while neglecting differences in hours distributions has significant consequences for the estimates of coefficients both for men and for women. The results confirm that for education levels which tend to be negatively correlated with the number of hours worked, the estimates using monthly earnings as the dependent variable will be significantly biased downward. As we shall see in Section 4.3 the bias translates into economically large differences in the calculated rates of return.

The estimated returns to education are much larger compared to the rates found in other studies on Poland. For example the study of Keane and Prasad (2006), in which net monthly (or quarterly) earnings are used as the dependent variable and in which no corrections for sample selection are made, suggests that a higher education degree carried a premium from about 16pp in 1990 to about 34pp in 1996[19] . The result for their latest year of analysis is close to that in Specification 1, but as we showed above, it is significantly underestimated relative to our reference Specification 5, which suggests a premium of 63pp. A more recent study of Newell and Socha (2007) using the PLFS data and net hourly wages suggests a premium to higher education of 22pp in 2004 using an OLS specification and 27pp using a selection corrected model. Both of these are far lower compared to our specifications as well as to results of other studies on Poland. The estimates could perhaps be explained by the fact of a different focus of the study and relate to including a very large set of controls which are usually considered endogenous and have been left out in our analysis (e.g. occupation, industry, sector, hours of work, firm size, type of contract, etc.). In addition to this, the type of exclusion restriction made by the authors is also unclear.[20] The return to higher education for men estimated in our paper is similar to the estimates of Bejdechi, Hartog, and van Opheim (2004) for 1995. Estimates in this paper (based on the Luxembourg Income Study (derived from PHBS), however, are based on a sample of full-time working individuals including the self-employed, which may explain why it avoids the downward bias. Moreover their estimated return to secondary education is about 15pp (47percent) lower compared to our estimates.

[TABLE 3 ABOUT HERE]

[19]Estimates in Keane and Prasad (2006) are given only for the combined sample of men and women.

[20]For example Newell and Socha (2007) include occupation variables in the selection equation and they seem to treat the nonemployed in the same way as those not reporting wages.

4.2 Selection equation, collinearity and the role of the instrument

In this section we examine the role of the instrument in determining the estimated values of education parameters. All specifications analyzed here use the gross hourly wage as the dependent variable and are related to the results of our reference Specification 5 and to the linear equation, i.e. Specification 3. The specifications we consider have been chosen to allow us to examine the performance of different types of exclusion restrictions including their effect on the bias of the estimated coefficients and on the degree of collinearity they induce. Collinearity, with the consequent reduction in the precision of the estimates, together with the bias jointly contribute to the measure of precision of estimation, namely the ratio of the mean square errors, φ_{mse}, as defined in Section 2.4.

The specifications we estimate are outlined in Table 4. In Specifications 7, 10, and 12 we exclude respectively the marital status information, the information on the age of the youngest child and all information on children from the wage equation and use these as instruments in the selection equation (instead of the instruments used in Specification 5). These exclusion restriction are to approximate many approaches taken to the estimation of wage equations in the absence of detailed incomes information in the data and/or inability to simulate out-of-work incomes. In Specification 9 we examine the performance of extended household structure as an instrument for selection, again in the case of absence of detailed incomes data but in the situation where information on household structure is available to the researcher. Further four specifications (6, 8, 11 and 13) identify the models with functional form. These are estimated to examine the specific effects of making such identifying assumptions in various data availability scenarios, and to analyze the effects of the functional form identification on the bias and the level of collinearity. Since in all cases the hourly wages are used as the dependent variable we compute standard errors using the same procedure as for Specification 3 and 5. Results of the estimations are given in Tables 5, 6 and 7. The first of these tables gives the log likelihood and the Akaike Information Criteria (AIC) values, in Table 6 we present collinearity statistics for Specifications 5-13, and finally in Table 7 we present details of the bias, the empirical means square error and the MSE ratio for Specifications 3 and 6-13 relative to Specification 5. For clarity of

presentation the details are only given for the estimated education parameters and in the case of the full sample for the gender dummy variable.[21]

Looking at the overall performance of the different specifications the values of the log-likelihood and the AIC are favorable to our reference specification (Table 5). This is not surprising given the nature of the exclusion restrictions in this case and in particular the continuous nature of two out of three instruments used in Specification 5. It should be noted though, that the specification which includes only the complex household indicator (Specification 9) performs relatively well as far as the fit of the model is concerned. However as we shall see in Section 4.2 this specification induces a high degree of bias. Also unsurprisingly most of the functional form specifications perform worse with respect to their respective estimations which include an instrument (i.e. Specifications 6 vs. 5 and 9, Specification 8 vs. 7, Specification 11 vs. 10, and Specification 13 vs. 12). The only exceptions are Specifications 7 and 8 estimated on the male sample, which may suggest that - at least for Poland - marital status (once we control for other variables) may be a poor instrument for labor market selection for men. As we shall see below Specification 7 also induces a high degree of bias.

[TABLE 4 ABOUT HERE]

[TABLE 5 ABOUT HERE]

[TABLE 6 ABOUT HERE]

Collinearity statistics presented in Table 6 demonstrate that the degree of collinearity is relatively high for all of the specifications we estimate, with the VIF in the range of 7.5 - 61.0 and the condition numbers in the range of 35.7 - 107.0. These values, according to Belsley's classification, fall into the category of "moderate to strong relations". Collinearity is of course particularly high for specifications with functional form identification (6, 8, 11, 13), though in the case of Specifications 10 and 11 estimated on the sample of men it is actually marginally lower for the functional form identification. The values suggest a weak role of the instrument in Specification 10 estimated on the sample of men and Specification 7 estimated for women. In the latter case the estimation is instrumented with the marital status dummy, which in the case of men - as far as collinearity measures are concerned - performs relatively well and in fact induces less collinearity than the instruments chosen for Specification 5.

[21] Full details are available from the authors on request.

Naturally, the consequence of excluding a variable, or a set of variables, from the wage equation and using these as instruments may result in biasing the estimates of the coefficients we are interested in due to an omitted variables problem. Thus a gain in the precision of estimation resulting from lower degree of collinearity comes at the cost of the bias. In Table 7 we produce results which combine the two and judge the performance of the specifications using the MSE ratio, φ_{mse}. As noted in Section 2.4 any values of the φ_{mse} greater than 1 imply a better performance of our reference Specification 5 relative to the compared specification, and those below 1 signify a superiority of the latter.

As we can see in Table 7 there is a general pattern for the bias on the reported coefficients for the full sample, with education coefficients being underestimated relative to Specification 5, and the female dummy parameter being overestimated. In the latter case the bias is especially high for Specifications 7, 9 and 11 when it is of similar range to the bias induced by the OLS estimation. It is notable that the bias on higher education coefficients is similar for Specifications 3, 7 and 10, which shows that using a poor instrument might do very little to reduce the bias. As the values of the MSE ratio suggest in such cases it may be far better to use the functional form specification even if it means inducing a higher degree of collinearity. This is the case also for the separate subsamples of men and women in the case of Specification 9 (vs. Specification 6) and for men in Specificaiton 7 (vs. Specification 8). An explanation may rest in the fact that the bias in the specifications with exclusion restrictions (i.e. 7 and 9) may result from endogeneity of the instrument with respect to the wage level, which in the case of both marital status and complex household structure seems plausible. Excluding an endogenous instrument in these cases and relying on functional form identification significantly reduces the bias.

[TABLE 7 ABOUT HERE]

As we noted earlier, the bias resulting from omitting the selection correction is relatively small for men. This results from a smaller proportion of men who are censored compared to the female sample (see Table 9 in the Appendix), but also suggests that the degree of non-randomness in the employment selection process may be lower in the case of men. This could be the case if non-employment was less of a choice in the case of men, which is likely to be the case.

What is notable in the case of our specifications is the fact that not only is the bias small in the sample of men relative to our reference specification, but the performance of the OLS estimator as measured by the empirical MSE is very close to or even superior to that of Specification 5. The reason for that is on the one hand the low degree of bias and on the other the higher level of precision of estimation of the OLS estimates, which reflects the collinearity induced by controlling for selection. This is despite the relatively large samples we use for the estimation. The MSE ratio in the case of OLS (Specification 3) for education variables is as low as 0.6 for vocational education and 0.9 for secondary education. This suggests that in the case of men using the OLS may actually be a better approach compared to applying selection correction. For the sample of men, with the exception of Specification 7 and 9 the MSE ratio is relatively close to 1 suggesting on the one hand that the chosen instruments perform relatively well, but on the other stressing that performing selection correction is not as important in the case of the male sample.

Looking at the full sample results and at the sample of women, the estimations suggest that none of the examined specifications outperforms Specification 5. The ratios show how important in many cases the choice of explanatory variables may be for the bias and the resulting empirical MSE. For example in Specification 7 (instrumented by marital status) the MSE ratio on the female dummy is as large as 125.7. In the female sample Specification 12 is closest to our reference set of exclusion restrictions. For this specification in the case of most education parameters the bias is in the range of 1%, and the MSE ratio is never higher than 2. What is notable are the high values of the MSE ratio for Specification 3 (OLS) and those identifying selection using functional form identification (especially Specifications 11 and 13). In the latter case omitting variables related to the family structure from both the wage and the selection equation is responsible for the high degree of the bias.

4.3 The dependent variable, selection and the rate of return

To compute the annual rate of return to the analyzed levels of education one has to account for the duration of the specific level of schooling. In the case of the Polish education system it is safe to assume that higher education takes five years, post-secondary education two years, secondary education four years and vocational schooling three

years. There are naturally many exceptions from these general rules and individually it may take more or less time to reach a specific qualification.

Assuming these general durations, we can compute the differences in the estimated rates of return resulting from several of the estimated Specifications. Using the estimates from Table 3 our results suggest that the return to higher education for men grows from 6.7% to 9.7% when we go from returns estimated using linear OLS and use net monthly earnings as the dependent variable to our reference specification using gross hourly wages and correcting for selection. Given the canceling out of the two biases the return to secondary and vocational education for men grow only slightly, respectively from 8.2% to 9.0% and from 4.8% to 5.2%, and in these cases the changes are not statistically significant. For women, the returns to higher education for these two specifications grow from 8.0% to 13.4%, and the returns to secondary education from 8.1% to 9.7%.

The values of the bias relative to Specification 5 given in Table 7 show that a misspecification of the model may result in very significant biases of estimated returns. Our examples of misspecification included estimates of up to -1.4 percentage points on annual returns in the case of higher education and 3.8pp in the case of secondary education for men (Specification 7), and up to 6.3pp and 11.2pp respectively for women (Specification 13). The results suggest a significant and economically important difference in the estimated returns to education conditional on the specific dependent variable used, and confirm a very important role of employment selection, especially in the estimates for women. They also point to a high degree of caution with respect to the choice of the exclusion restriction used for the identification of the selection process.

5 Conclusion

The existing literature on determinants of wages in Poland has been based almost exclusively on net monthly earnings and there are no studies which would comprehensively treat the issue of labor market selection. In this analysis we showed that moving from net monthly earnings to gross hourly wages implies a substantial and statistically significant difference in the estimated returns to education for both men

and women in Poland. Hours of work are on average lower among the better educated and omitting this relationship leads to a downward bias on returns to higher and post-secondary education and to an upward bias on returns to secondary and vocational education. In the case of men this bias is much more important than that induced by lack of controlling for selection into employment. The annual rate of return to higher education grows from 6.7% to 9.7% for men and from 8.0% to 13.4% for women when we use *gross hourly wages* and correct for employment selection rather than run the OLS estimation on *net monthly earnings* as the dependent variable. These results are significantly higher compared to other estimates of wage equations using Polish data.

The analysis also showed the importance of the choice of exclusion restrictions in the selection corrected estimates for the implied level of collinearity, the induced bias and the consequent level of the empirical MSE. Since selection generally seems of less relevance in the estimation of the wage equation for men, the choice of the exclusion restriction is also of less importance although some exclusion restrictions may still lead to a significant bias. The estimations for women are much more sensitive to the choice of the exclusion restriction and perform particularly poorly in cases of functional form identification. Judged by level of the MSE our analysis suggests that OLS estimates are good approximations to our reference specification for the sample of Polish men, and in fact are superior in the case of secondary and vocational education. On the other hand a complex set of demographic variables (including the number of children and the age of youngest child) seems to be the "second best" in the case of estimating returns to education in Poland for women.

References

Abowd, J., and D. Card (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57(2), 411–445.

Arellano, M., and C. Meghir (1992): "Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets," *Review of Economic Studies*, 59(3), 537–559.

Ashenfelter, O., and A. Krueger (1994): "Estimates of the Economic Return to Schooling from a New Sample of Twins.," *American Economic Review*, 84(5), 1157–73.

Bejdechi, S., J. Hartog, and H. van Opheim (2004): "Investment in Education in Nine Nations-Return and Risk," Discussion Paper, Univeristy of Amsterdam and Tinbergen Institute.

Belsley, D. (1991): *Conditioning diagnostics: Collinearity and Weak Data in Regression*. John Wiley and Sons, New York.

Belsley, D., E., E. Kuh, and R. Welsch (1980): *Regression diagnostics: Identifying influentioal data and sources of collinearity*. Wiley, New York.

Blackburn, M. L., and D. Neumark (1995): "Are OLS Estimates of the REturn to Schooling Biased Downward? Another Look.," *Review ofEconomics and Statistics*, 77(2), 217–29.

Blasnik, M. (1998): "CNDNMB3: Stata module to calculate condition number of regressor matrix," Statistical Software Components, Boston College Department of Economics.

Blundell, R., L. Dearden, and B. Sianesi (2005): "Evaluating the Impact of Education on Earnings in the UK: Models, Methods and Results from the NCDS," *Journal of the Royal Statistical Society, Series A*, 168(3), 473–512.

Blundell, R., A. Duncan, J. McCrae, and C. Meghir (2000): "The Labour Market impact of the Working Families Tax Credit," *Fiscal Studies*, 21(1), 75–104.

Blundell, R., H. Reed, and T. Stoker (2003): "Interpreting Aggregate Wage Growth: The Role of Labor Market Participation," *American Economic Review*, 94(3), 1114–1131.

Boockmann, B., and V. Steiner (2006): "Cohort effects and the returns to education in West Germany," *Applied Economics*, 38, 1135–1152.

Brunello, G., and R. Miniaci (1999): "The Economic Returns to Schooling for Italian Men. An Evaluation Based on Instrumental Variables," *Labour Economics*, 6, 509–519.

Callam, T., and C. Harmon (1999): "The Economic Returns to Schooling in Ireland," *Labour Economics*, 6, 543–550.

Duffy, F., and P. Walsh (2001): "Indicidual Pay and Outside Options: Evidence from the Polish Labor Survey," Working Paper No. 364, Trinity College.

Dustmann, C., and C. Schmidt (2000): "The Wage Performance of Immigrant Women: Full-Time Jobs, Part-Time Jobs, and the Role of Selection," Discussion Paper No. 2702, Centre for Economic Policy Research.

Greene, W. H. (2003): *Econometric Analysis*. New York University, New York.

Gregg, P., P. Johnson, and H. Reed (1999): "Entering work and the tax and benefit system," Report, Institute for Fiscal Studies.

Grilliches, Z. (1977): "Estimating the Returns to Schooling: some Econometric Problems.," *Econometrica*, 45(1), 1–22.

Harmon, C., and I. Walker (1995): "Estimates of the Economic Return to Schooling for the United Kingdom," *American Economic Review*, 85(3), 1278–1286.

Heckman, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47 (1), 153–161.

Heckman, J., and E. Vytlacil (2001): "Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return to Schooling," *The Review of Economics and Statistics*, 83(1), 1–12.

Hoynes, H. (1996): "Welfare Transfers in Two-Parent Families: Labor Supply and Welfare Participation Under AFDC-UP," *Econometrica*, 64(2), 295–332.

Hurd, M. (1971): "Changes in Wage Rates Between 1959 and 1967," *Review of Economics and Statistics*, 53(2), 189–199.

Keane, M., and E. Prasad (2006): "Changes in the Structure of Earnings During the Polish Transition," *Journal of Development Economics*, 80(2), 389–427.

Leung, S., F., and S. Yu (1996): "On the choice between sample selection and two-part models," *Journal of Econometrics*, 72(1), 197–229.

Madden, D. (2008): "Sample selection versus two-part models revisited: The case of female smoking and drinking," *Journal of Health Economics*, 27, 300–307.

Morawski, L., and M. Myck (2009): "'Klin'-ing up - effects of Polish labour tax reforms on those in and on those out," *Labour Economics*, 00(0), 00–00.

Mroz, T. (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economis and Statistical Assumptions.," *Econometrica*, 55(4), 765–799.

Neumark, D., and S. Korenman (1994): "Sources of Bias in Women's Wage Equations: Results Using Sibling Data," *The Journal of Human Resources*, 29(2), 379–405.

Newell, A., and M. Socha (2007): "The Polish Wage Inequality Explosion," *Economics of Transition*, 15(4), 733–758.

O'Dorchai, S. (2008): "Pay Inequality in 25 European Countries," Working Paper No. 08-06.rs, DUBLEA.

Puhani, P., A. (2000): "The Heckman Correction for Sample Selection and its Critique," *Journal of Economic Surveys*, 14(1), 53–68.

Toro-Vizcarrondo, C., and T. Wallace (1968): "A test of the mean square error criterion for restrictions in linear regression," *Journal of the American Statistical Association*, 63(322), 558–572.

Vella, F. (1998): "Estimating Models with Sample Selection Bias: A survey," *Journal of Human Resources*, 33(1), 127–169.

Tables

Table 1: Stylized facts about the individuals in the selected populations in BBGD 2005 samples.

	All			Working			Not working		
	All	Men	Women	All	Men	Women	All	Men	Women
Age	38.71	39.48	37.98	38.43	38.60	38.25	39.21	41.51	37.62
Number of children (column percentages)									
- 0 children	0.433	0.479	0.391	0.393	0.387	0.400	0.506	0.690	0.378
- 1 child	0.261	0.236	0.286	0.279	0.269	0.290	0.230	0.158	0.280
- 2 children	0.217	0.202	0.231	0.240	0.242	0.237	0.176	0.109	0.222
- >2 children	0.088	0.083	0.092	0.088	0.101	0.073	0.088	0.043	0.119
Married	0.682	0.701	0.664	0.714	0.763	0.658	0.625	0.558	0.671
Education (column percentages)									
- Higher	0.158	0.125	0.190	0.209	0.149	0.276	0.069	0.067	0.069
- Post-sec. prof.	0.033	0.016	0.049	0.037	0.016	0.060	0.026	0.016	0.033
- Secondary	0.343	0.304	0.352	0.315	0.395	0.396	0.326	0.278	0.359
- Vocational	0.354	0.442	0.271	0.331	0.437	0.210	0.396	0.454	0.356
- Primary or none	0.112	0.113	0.138	0.108	0.003	0.058	0.183	0.185	0.183
Part-time work	-	-	-	0.079	0.064	0.095	-	-	-
Public sector	-	-	-	0.386	0.314	0.468	-	-	-
Observations	40742	19999	20743	25973	14015	11958	14769	5984	8785

Source: Authors' calculations using PHBS, 2005.

Table 2: Explanatory variables for the wage equation.

Education$^{\diamondsuit}$	Family composition*	Residence$^{\sharp}$	Other
higher	married	town2: $200k$ up to $500k$	age-group dummies
post-secondary	one child	town3: $100k$ up to $200k$	seasonal dummies
secondary	two children	town4: $20k$ up to $100k$	disability dummy
vocational	three children	town5: town up to $20k$	
	four children	town6: village	
	five or more children		
	child aged <7 in family	regional dummies	

Notes: Specifications for the full sample (men and women together) include a gender dummy indicator. Reference categories: $^{\diamondsuit}$ - primary or no education; * - no children; ** - no significant disability; $^{\sharp}$ - town size $>= 500k$.

Table 3: Returns to education levels under alternative specifications of the dependent variable.

Estimation	Specification 1 OLS monthly net		Specification 2 OLS monthly gross		Specification 3 OLS hourly gross		Specification 4 LIML monthly gross		Specification 5 LIML hourly gross	
	Coeff.	s.e.	Coeff.	s.e.	Coeff.	s.e.	Coeff.	s.e.	Coeff.	s.e.
All										
Education:										
- Higher	0.3724	(0.008)	0.4034	(0.008)	0.5284	(0.009)	0.4947	(0.009)	0.6308	(0.016)
- Post-sec.	0.0696	(0.015)	0.0763	(0.016)	0.1148	(0.016)	0.1170	(0.018)	0.1537	(0.022)
- Secondary	0.3224	(0.011)	0.3607	(0.012)	0.3066	(0.012)	0.4893	(0.013)	0.4331	(0.025)
- Vocational	0.1263	(0.011)	0.1432	(0.012)	0.1172	(0.012)	0.2230	(0.013)	0.1925	(0.020)
Female dummy	-0.2292	(0.006)	-0.2482	(0.006)	-0.1421	(0.006)	-0.3556	(0.007)	-0.2426	(0.014)
Men										
Education:										
- Higher	0.3355	(0.012)	0.3619	(0.013)	0.4540	(0.013)	0.3838	(0.014)	0.4831	(0.021)
- Post-sec.	0.0025	(0.031)	0.0047	(0.033)	0.0467	(0.033)	0.0012	(0.033)	0.0475	(0.034)
- Secondary	0.3290	(0.015)	0.3658	(0.016)	0.3255	(0.016)	0.4127	(0.017)	0.3580	(0.039)
- Vocational	0.1434	(0.015)	0.1609	(0.016)	0.1364	(0.015)	0.1952	(0.016)	0.1564	(0.032)
Women										
Education:										
- Higher	0.3990	(0.010)	0.4332	(0.010)	0.5798	(0.011)	0.5030	(0.012)	0.6712	(0.019)
- Post-secondary	0.0915	(0.016)	0.1001	(0.018)	0.1420	(0.019)	0.1289	(0.019)	0.1762	(0.023)
- Secondary	0.3243	(0.016)	0.3656	(0.017)	0.2820	(0.019)	0.4571	(0.020)	0.3871	(0.029)
- Vocational	0.0946	(0.017)	0.1094	(0.018)	0.0736	(0.019)	0.1509	(0.019)	0.1213	(0.025)

Notes: LIML - Limited Information Maximum Likelihood.
Source: Authors' calculations using PHBS, 2005.

Table 4: Alternative specifications of wage equation models.

	Identifying selection with:	Label
Specification 5	equivalised household income if out of work simulated family income if out of work multifamily household indicator simulated family income if out of work if married	PREF
Specification 6	functional form to Specification 5	ff-PREF
Specification 7	married indicator	MARST
Specification 8	functional form to Specification 6	ff-MARST
Specification 9	multifamily household indicator	MFH
Specification 10	child aged <7 in the family	YCHLD
Specification 11	functional form to Specification 10	ff-YCHLD
Specification 12	1 child, 2 children, 3 children, 4 children, 5 children child aged <7	KIDS
Specification 13	functional form to Specification 12	ff-KIDS

Notes: For Specifications 7, 10, and 12 variables indicated as instruments are excluded from the wage equation relative to Specification 5. This carries through to the corresponding specifications where identification is through functional form (8, 11, 12). For example in Specification 8 the married dummy (which is used as instrument for selection in Specification 7) is excluded from the wage equation. Note also that functional form to Specification 9 is identical with the functional form to Specification 5.

Table 5: Alternative specifications of wage equation models.

		Log likelihood			Akaike Information Criterion		
	Label	All	Men	Women	All	Men	Women
Specification 5	PREF	-41030.3	-19536.1	-19032.0	82150.6	39160.2	38152.0
Specification 6	ff-PREF	-41198.2	-20037.3	-20742.7	82486.4	40162.6	41573.4
Specification 7	MARST	-41278.4	-20374.4	-20459.2	82644.8	40834.8	41004.4
Specification 8	ff-MARST	-41366.6	-20330.8	-20678.8	82821.2	40747.6	41443.6
Specification 9	MFH	-41174.9	-20012.6	-20164.0	82439.8	40113.2	40416.0
Specification 10	YCHLD	-41214.5	-20037.4	-20160.7	82517.0	40160.8	40407.4
Specification 11	ff-YCHLD	-41264.8	-20039.3	-22144.2	82617.6	40164.6	44374.4
Specification 12	KIDS	-41243.5	-20043.0	-20157.4	82565.0	40162.0	40390.8
Specification 13	ff-KIDS	-41343.8	-20124.5	-24094.8	82765.6	40325.0	48265.6

Source Authors' calculations using PHBS, 2005.

Table 6: Collinearity diagnostics for selection corrected specifications.

		All		Men		Women	
	Label	VIF	Cond. num.	VIF	Cond. num.	VIF	Cond. num.
Specification 5	PREF	8.6	39.4	17.7	51.5	7.5	39.2
Specification 6	ff-PREF	30.4	70.6	35.0	71.6	47.4	93.3
Specification 7	MARST	17.7	52.4	8.4	35.7	46.2	87.9
Specification 8	ff-MARST	32.9	71.2	38.8	73.5	47.3	89.0
Specification 9	MFH	26.5	65.9	30.8	67.1	41.3	87.0
Specification 10	YCHLD	19.9	57.4	40.0	70.2	12.0	49.5
Specification 11	ff-YCHLD	30.7	70.7	34.4	69.6	52.7	102.0
Specification 12	KIDS	17.1	51.4	21.8	54.5	7.5	38.7
Specification 13	ff-KIDS	31.3	68.6	33.8	67.4	61.0	107.0

Source: Authors' own calculations using PHBS, 2005.

Notes: VIF - Variance inflation factor; Cond. num. - Condition number.

Table 7: Bias, MSE and φ_{mse} for all Specifications 3 and 6-13 relative to Specification 5.

		All					Men				Women			
	Label	Higher	Post.sec.	Second.	Vocat.	Female	Higher	Post.sec.	Second.	Vocat.	Higher	Post.sec.	Second.	Vocat.
Bias														
Specification 3	OLS	-0.1024	-0.0389	-0.1265	-0.0753	0.1005	-0.0291	-0.0008	-0.0325	-0.0199	-0.0914	-0.0341	-0.1052	-0.0477
Specification 6	ff-PREF	-0.0230	-0.0100	-0.0323	-0.0189	0.0244	0.0129	-0.0022	0.0317	0.0215	0.1609	0.0721	0.2246	0.1099
Specification 7	MARST	-0.1394	-0.0641	-0.1968	-0.1202	0.1590	-0.0726	0.0101	-0.1544	-0.1133	0.1339	0.0574	0.1921	0.0949
Specification 8	ff-MARST	-0.0243	-0.0147	-0.0285	-0.0146	0.0180	0.0178	-0.0124	0.0417	0.0316	0.1541	0.0689	0.2191	0.1094
Specification 9	MFH	-0.1075	-0.0474	-0.1531	-0.0938	0.1277	-0.0289	0.0036	-0.0610	-0.0466	0.0233	0.0104	0.0317	0.0160
Specification 10	YCHLD	-0.0090	-0.0047	-0.0136	-0.0081	0.0092	0.0131	-0.0028	0.0306	0.0214	0.0460	0.0213	0.0648	0.0331
Specification 11	ff-YCHLD	-0.0245	-0.0109	-0.0344	-0.0209	0.1385	0.0135	-0.0023	0.0317	0.0226	0.2219	0.1023	0.3178	0.1594
Specification 12	KIDS	-0.0347	-0.0143	-0.0423	-0.0252	0.0383	-0.0241	0.0018	-0.0395	-0.0295	0.0107	0.0056	0.0172	0.0089
Specification 13	ff-KIDS	-0.0257	-0.0110	-0.0269	-0.0168	0.0262	0.0156	-0.0054	0.0480	0.0353	0.3129	0.1413	0.4487	0.2238
MSE:														
Specification 5	PREF	0.0002	0.0005	0.0006	0.0004	0.0002	0.0004	0.0011	0.0015	0.0010	0.0003	0.0005	0.0008	0.0006
Specification 3	OLS	0.0106	0.0018	0.0161	0.0058	0.0101	0.0010	0.0011	0.0013	0.006	0.0085	0.0015	0.0114	0.0026
Specification 6	ff-PREF	0.0010	0.0005	0.0020	0.0009	0.0010	0.0006	0.0011	0.0022	0.0013	0.0276	0.0062	0.0540	0.0136
Specification 7	MARST	0.0197	0.0045	0.0395	0.0149	0.0256	0.0056	0.0018	0.0244	0.0134	0.0200	0.0045	0.0411	0.0107
Specification 8	ff-MARST	0.0011	0.0006	0.0019	0.0008	0.0008	0.0008	0.0014	0.0033	0.0019	0.0255	0.0057	0.0516	0.0135
Specification 9	MFH	0.0120	0.0027	0.0246	0.0094	0.0169	0.0012	0.0014	0.0052	0.0031	0.0035	0.0011	0.0064	0.0019
Specification 10	YCHLD	0.0004	0.0005	0.0008	0.0005	0.0002	0.0006	0.0011	0.0022	0.0013	0.0027	0.0012	0.0056	0.0019
Specification 11	ff-YCHLD	0.0010	0.0005	0.0022	0.0010	0.0194	0.0006	0.0011	0.0022	0.0014	0.0516	0.0118	0.1058	0.0273
Specification 12	KIDS	0.0015	0.0006	0.0024	0.0010	0.0017	0.0009	0.0013	0.0025	0.0015	0.0005	0.0006	0.0013	0.0008
Specification 13	ff-KIDS	0.0011	0.0006	0.0017	0.0008	0.0011	0.0007	0.0012	0.0035	0.0021	0.1009	0.0216	0.2080	0.0529
MSE ratio,φ_{mse}:														
Specification 3	OLS	42.8	3.8	25.1	14.5	49.7	2.5	1.0	0.9	0.6	28.2	3.0	14.3	4.4
Specification 6	ff-PREF	4.0	1.1	3.1	2.3	5.0	1.4	0.9	1.5	1.3	79.9	11.8	65.8	21.1
Specification 7	MARST	79.8	9.4	61.4	37.2	125.7	13.0	1.5	16.5	12.9	58.0	8.5	50.0	16.6
Specification 8	ff-MARST	4.3	1.4	3.0	2.0	3.9	1.9	1.2	2.2	1.9	73.8	10.8	62.9	20.9
Specification 9	MFH	48.7	5.7	38.3	23.5	82.9	2.9	1.2	3.5	3.0	10.1	2.0	7.8	2.9
Specification 10	YCHLD	1.5	1.0	1.3	1.2	1.1	1.4	1.0	1.5	1.3	7.9	2.2	6.8	2.9
Specification 11	ff-YCHLD	4.2	1.2	3.5	2.5	95.1	1.5	1.0	1.5	1.3	149.1	22.4	128.8	42.3
Specification 12	KIDS	6.0	1.3	3.7	2.6	8.4	2.2	1.1	1.7	1.5	1.4	1.1	1.5	1.2
Specification 13	ff-KIDS	4.5	1.2	2.7	2.0	5.3	1.6	1.1	2.4	2.1	291.7	41.0	253.2	81.9

Source: Authors' calculations using PHBS, 2005.

Notes: MSE ratio, φ_{mse}, computed with reference to Specification 5.

Appendix

Table 8: Selected features of employment by education status in the PLFS 2005 data.

	All	Education groups:				
		Higher	Post-sec.	Secondary	Vocational	Primary/none
All						
Observations	13097	3013	548	4546	4138	852
Hours worked (column percentages)						
- 1-14 hours worked	0.006	0.011	0.002	0.005	0.004	0.013
- 15-24 hours worked	0.060	0.139	0.050	0.031	0.030	0.067
- 25-34 hours worked	0.037	0.078	0.048	0.023	0.018	0.033
- 35-44 hours worked	0.695	0.631	0.777	0.758	0.671	0.651
- 45+ hours worked	0.202	0.140	0.123	0.183	0.277	0.236
Part-time	0.058	0.040	0.065	0.063	0.055	0.120
Public sector	0.400	0.616	0.529	0.383	0.248	0.285
Men						
Observations	7070	1182	142	2269	2935	542
Hours worked (column percentages)						
- 1-14 hours worked	0.004	0.005	0.000	0.004	0.002	0.010
- 15-24 hours worked	0.026	0.055	0.020	0.017	0.014	0.054
- 25-34 hours worked	0.020	0.039	0.048	0.020	0.008	0.026
- 35-44 hours worked	0.687	0.672	0.754	0.721	0.673	0.622
- 45+ hours worked	0.264	0.229	0.178	0.238	0.302	0.288
Part-time	0.037	0.022	0.050	0.045	0.028	0.091
Public sector	0.327	0.495	0.467	0.341	0.243	0.246
Women						
Observations	6027	1831	406	2277	1203	310
Hours worked (column percentages)						
- 1-14 hours worked	0.010	0.015	0.003	0.006	0.008	0.018
- 15-24 hours worked	0.102	0.197	0.061	0.045	0.068	0.090
- 25-34 hours worked	0.058	0.106	0.048	0.026	0.044	0.045
- 35-44 hours worked	0.705	0.603	0.786	0.797	0.666	0.702
- 45+ hours worked	0.126	0.078	0.102	0.126	0.213	0.145
Part-time	0.083	0.052	0.070	0.082	0.122	0.172
Public sector	0.488	0.701	0.552	0.426	0.261	0.355

Source: Authors' calculations using PLFS, Q1-Q4, 2005 (only first-time observations).

Table 9: Selection corrected results for the reference Specification 5, wage equation.

	All		Men		Women	
	Coeff.	St. error	Coeff.	St. error	Coeff.	St. error
Education						
- higher	0.6308**	(0.016)	0.4831**	(0.021)	0.6712**	(0.019)
- post-sec.	0.1537**	(0.022)	0.0475	(0.034)	0.1762**	(0.023)
- secondary	0.4331**	(0.025)	0.3580**	(0.039)	0.3871**	(0.029)
- vocational	0.1925**	(0.020)	0.1564**	(0.032)	0.1213**	(0.025)
Female	-0.2426**	(0.014)	-	-	-	-
Age group						
- age (24,29]	0.1010**	(0.017)	0.0300	(0.030)	0.1010**	(0.022)
- age (29,34]	0.3115**	(0.021)	0.1987**	(0.033)	0.3084**	(0.028)
- age (34,39]	0.3982**	(0.023)	0.2418**	(0.033)	0.4159**	(0.030)
- age (39,44]	0.4325**	(0.023)	0.2434**	(0.034)	0.4734**	(0.029)
- age (44,49]	0.4116**	(0.021)	0.2222**	(0.027)	0.4626**	(0.029)
- age (49,54]	0.3950**	(0.020)	0.2271**	(0.024)	0.4686**	(0.024)
- age >54	0.2737**	(0.027)	0.2571**	(0.050)	-	-
Town size:						
- town2	-0.0919**	(0.017)	-0.0520*	(0.023)	-0.1359**	(0.023)
- town3	-0.1505**	(0.015)	-0.1114**	(0.025)	-0.1804**	(0.023)
- town3	-0.1925**	(0.014)	-0.1339**	(0.018)	-0.2264**	(0.018)
- town5	-0.2094**	(0.014)	-0.1632**	(0.021)	-0.2288**	(0.018)
- town6	-0.2073**	(0.012)	-0.1786**	(0.017)	-0.2197**	(0.019)
Seasonal dummies						
- 2nd quarter	0.0176*	(0.009)	0.0098	(0.014)	0.0001	(0.013)
- 3rd quarter	0.0193*	(0.009)	0.0041	(0.012)	0.0113	(0.013)
- 4th quarter	0.0690**	(0.009)	0.0553**	(0.015)	0.0549**	(0.012)
Family characteristics						
- 1 child	0.0688**	(0.010)	0.0684**	(0.019)	0.0171	(0.012)
- 2 children	0.0816**	(0.011)	0.1110**	(0.020)	0.0113	(0.015)
- 3 children	0.0299	(0.016)	0.0713**	(0.025)	-0.0467*	(0.023)
- 4 children	0.0272	(0.027)	0.0756	(0.039)	-0.0468	(0.038)
- >4 children	-0.1166**	(0.041)	-0.0524	(0.052)	-0.1894**	(0.069)
- child aged <7	0.0083	(0.011)	0.0038	(0.012)	0.0074**	(0.017)
- married	0.1500**	(0.012)	0.2479**	(0.038)	0.0481**	(0.012)
Disability	-0.3712**	(0.048)	-0.2536**	(0.092)	-0.1646**	(0.050)
Regional dummies	included		included		included	
Constant	1.5156**	(0.043)	1.6706**	(0.125)	1.4969**	(0.060)
Number of observations:						
- censored:	14609		5849		8760	
- uncensored:	26133		14150		11983	
Log likelihood	-41030.3		-19536.1		-19032.0	

Note: ** - 1 per cent significance level , * - 5 per cent significance level.

Source: Authors' own calculations using PHBS, 2005.

Table 10: Selection corrected results for the Preferred Specification 5, selection equation.

	All		Men		Women	
	Coeff.	St. error	Coeff.	St. error	Coeff.	St. error
Education:						
- higher	0.6561**	(0.024)	0.3682**	(0.041)	0.7725**	(0.031)
- post-sec.	0.2290**	(0.040)	-0.0316	(0.082)	0.2813**	(0.046)
- secondary	0.6771**	(0.023)	0.5554**	(0.035)	0.7389**	(0.033)
- vocational	0.3722**	(0.023)	0.3807**	(0.032)	0.3324**	(0.033)
Female	-0.6306**	(0.015)	-	-	-	-
Age group						
- age (24,29]	0.3032**	(0.027)	0.3523**	(0.040)	0.3258**	(0.038)
- age (29,34]	0.5242**	(0.029)	0.4133**	(0.046)	0.6149**	(0.041)
- age (34,39]	0.6161**	(0.032)	0.3767**	(0.049)	0.7451**	(0.045)
- age (39,44]	0.5942**	(0.031)	0.2317**	(0.048)	0.7627**	(0.043)
- age (44,49]	0.4916**	(0.029)	0.0797	(0.046)	0.6607**	(0.040)
- age (49,54]	0.1640**	(0.029)	-0.0934*	(0.045)	0.2131**	(0.040)
- age >54	-0.5860**	(0.039)	-0.7806**	(0.049)	-	-
Town size:						
- town2	-0.0572	(0.034)	-0.0139	(0.053)	-0.0707	(0.046)
- town3	-0.1190**	(0.036)	-0.0548	(0.055)	-0.1524**	(0.048)
- town4	-0.2289**	(0.027)	-0.1826**	(0.043)	-0.2606**	(0.037)
- town5	-0.2585**	(0.030)	-0.2290**	(0.047)	-0.2761**	(0.041)
- town6	-0.1713**	(0.026)	-0.0690	(0.041)	-0.2402**	(0.036)
Seasonal dummies						
- 2nd quarter	0.1070**	(0.019)	0.1656**	(0.029)	0.0597*	(0.026)
- 3rd quarter	0.1232**	(0.019)	0.1618**	(0.029)	0.0920**	(0.026)
- 4th quarter	0.1384**	(0.019)	0.1901**	(0.029)	0.0931**	(0.027)
Family characteristics						
- 1 child	0.1359**	(0.021)	0.2023**	(0.033)	0.0122	(0.028)
- 2 children	0.0817**	(0.024)	0.1849**	(0.039)	-0.0301	(0.034)
- 3 children	0.0049	(0.034)	0.2352**	(0.056)	-0.1671**	(0.046)
- 4 children	-0.0331	(0.056)	0.2971**	(0.095)	-0.2676**	(0.075)
- >4 children	-0.1973*	(0.082)	0.2492	(0.141)	-0.5863**	(0.120)
- child aged <7	-0.2368**	(0.022)	0.0313**	(0.038)	-0.4217**	(0.030)
- married	0.2255**	(0.023)	0.5266**	(0.041)	-0.0575	(0.031)
Disability	-1.1313**	(0.036)	-1.2661**	(0.048)	-0.9375**	(0.056)
Selection instruments:						
FINC0	-0.3176**	(0.023)	-0.1769**	(0.033)	-0.5343**	(0.040)
HHINC0	-0.1544**	(0.020)	-0.2221**	(0.031)	-0.0607*	(0.030)
Multifamily household	-0.0614**	(0.018)	-0.0303	(0.029)	-0.0904**	(0.026)
Married*FINC0	0.2314**	(0.024)	0.1633**	(0.035)	0.4301**	(0.040)
Regional dummies	included		included		included	
Constant	0.1295**	(0.043)	-0.0791	(0.064)	-0.2313**	(0.060)
Number of observations:						
- censored:	14609		5849		8760	
- uncensored:	26133		14150		11983	
Log likelihood	-41030.3		-19536.1		-19032.0	

Notes: Significance: ** - 1 %, * - 5 %; FINC0/HHINC0 - family/household income if not working.

Source: Authors' own calculations using PHBS, 2005.