

The Effects of Schooling on Lifetime Earnings

Giorgio Brunello* Guglielmo Weber† Christoph T. Weiss‡

September 30, 2011

PRELIMINARY AND INCOMPLETE

Abstract

The empirical literature typically estimates returns to education by using current rather than lifetime income, because of data limitations. We use the detailed retrospective information provided by the third wave of the Survey of Health, Ageing and Retirement in Europe (SHARE) to estimate a measure of lifetime income at age ten, which includes both labour and pension income. Using a multi-country setup, we estimate the marginal effect of education on lifetime income using an instrumental variables strategy. There are two main results: first, estimated returns are in the order of 10 percent. Second, the size of these effects is substantially lower for those groups with a poor socio-economic background at age 10.

1 Introduction

The empirical literature typically estimates returns to education by using current rather than lifetime income. This practice has been recently challenged on the grounds that a better measure of economic success is life-time earnings or life-time income (Haider and Solon, 2006; Heckman, Lochner and Todd, 2006). Indeed, some recent papers have shown that the returns of education based on current earnings are significantly biased compared to those

*University of Padua, CESifo and IZA

†University of Padua, CEPR and IFS

‡University of Padua

based on life-time earnings, particularly if the sample includes many older workers (Bhuller, Mogstad and Salvanes, 2011). The evidence on this point is based on administrative data from just one country (Norway): on the one hand, the earnings data are highly accurate, on the other hand the data set does not contain information on a number of important variables that researchers typically control for in their estimation.

We are able to investigate this issue in a number of European countries using a rich data set which contains detailed retrospective information on earnings, pensions and many variables of potential interest (including childhood characteristics). The data we use is drawn from the third wave of the Survey of Health, Ageing and Retirement in Europe (SHARE) and allows us to estimate a measure of the net present value of lifetime income at age ten, which includes both labour and pension income. We estimate the marginal effect of education on lifetime income of men in a multi-country setting by adopting an instrumental variables strategy.

Our estimates agree with the Norwegian evidence presented in Bhuller, Mogstad and Salvanes (2011) that an additional year of education on average increases earnings by almost 10 percent. This is reassuring, given that the Norwegian earnings data are unlikely to be affected by substantial measurement error.

However, we are also able to show that returns vary markedly with with socio-economic background early in life: in particular, returns to education are much lower for those in poor conditions at age ten. This result contributes to the growing literature on the importance of early life interventions, that shows, for instance, lower returns to college for individuals who grew up in disadvantaged households (Heckman, 2000; Cunha and Heckman, 2007).

Another contribution of this paper is that we include in our measure of life-time resources also (actual and/or expected) pension income until death using cohort and country specific mortality tables. This measure seems appropriate as pension benefits are typically affected by either earnings or contributions, and can be seen as part of the return to the investment in education.

2 Review of the Literature

[to be written]

3 Theoretical framework

Economic theory suggests that rational individuals select their optimal education by equalising (expected) marginal returns to the marginal costs of schooling, which include both direct and opportunity costs. In an inter-temporal planning horizon, marginal returns refer to the entire stream of earnings and benefits. Starting with Mincer, however, empirical practice has estimated the returns to schooling by regressing (log) current earnings on schooling conditional on other covariates, such as age or experience.

This practice has long been justified by the lack of adequate data on lifetime earnings. As discussed by Haider and Solon (2006) and Heckman, Lochner and Todd (2006), the practice correctly identifies the internal rate of returns to education only if some very tight assumptions are verified. For instance, we require that earnings profiles are parallel in age or experience, an assumption usually rejected by the data¹. When earnings profiles are not parallel, estimating the returns to education using current rather than lifetime earnings generates a life-cycle bias.

To illustrate, define W_t as earnings at time since labour market entry t and I_0 as lifetime earnings at labour market entry. Let schooling be S . The life cycle bias associated with estimating returns using current rather than lifetime earnings is

$$LCB_t = \left| \frac{\partial \ln W_t}{\partial S} - \frac{\partial \ln I_0}{\partial S} \right| \quad (1)$$

Further suppose that the relationship between current earnings and schooling is described by the following Mincerian equation

$$\ln W_t = a + bS + ft + gSt + kSt^2 + \varepsilon_t \quad (2)$$

The marginal return to schooling in this case is

$$\frac{\partial \ln W_t}{\partial S} = b + gt + kt^2 \quad (3)$$

Next define lifetime earnings in the absence of uncertainty and unemployment spells as

$$I_0 = W_1 + \frac{W_2}{1+r} + \frac{W_3}{(1+r)^2} + \dots + \frac{W_T}{(1+r)^T} \quad (4)$$

¹See for instance the evidence for Europe in Brunello and Comi (2004).

Using eq.(2), the relationship between earnings at time t and at time 1 is given by

$$W_t = W_1 \exp \{ (t-1)f + (t-1)gS + (t^2-1)kS + \varepsilon_t - \varepsilon_1 \} \quad (5)$$

Useful linear approximations are

$$\begin{aligned} \ln(1+x) &\simeq x \\ e^x &\simeq 1+x \end{aligned}$$

Using eq.(4) into eq.(2) and applying the approximations above yields

$$\begin{aligned} \ln I_0 = &\lambda + \left\{ b + g \left[1 + \frac{1}{1+r} + \frac{2}{(1+r)^2} + \dots + \frac{T-1}{(1+r)^{(T-1)}} \right] \right. \\ &\left. + k \left[1 + \frac{3}{1+r} + \frac{8}{(1+r)^2} + \dots + \frac{T^2-1}{(1+r)^{(T-1)}} \right] \right\} S + v \end{aligned}$$

where v is a combination of error terms. Defining $A = \left[1 + \frac{1}{1+r} + \frac{2}{(1+r)^2} + \dots + \frac{T-1}{(1+r)^{(T-1)}} \right]$ and $B = \left[1 + \frac{3}{1+r} + \frac{8}{(1+r)^2} + \dots + \frac{T^2-1}{(1+r)^{(T-1)}} \right]$, the life-cycle loss is given by

$$LCB_t = g(t - A) + k(t^2 - B) \quad (6)$$

As expected, this loss is equal to zero if current earnings profiles are parallel ($g = 0$; $k = 0$). The value of t in the current earnings function which minimizes this earnings loss function is

$$t^* = \frac{-\frac{2g}{3k} \pm \sqrt{\left(\frac{2g}{3k}\right)^2 + \frac{4}{3}}}{2} \quad (7)$$

if $k \neq 0$ and $t = 0$ if $k = 0$.

3.1 Introducing unemployment spells

Unemployment spells affect both lifetime earnings and the associated returns to education for two reasons: first, a period of unemployment is a period when human capital is not put to use in the labour market. Second, unemployment can have a scarring effect on future earnings. To illustrate, consider a simple two-period model and let U be the incidence or the duration of unemployment in the first period. There is no unemployment in the second and final period. One way to introduce the scarring effect is to write second period wages

for those who have experienced unemployment in the first period as

$$\ln W_2 = a + bS + 2f + 2gS + 4kS - \lambda U_1 + \varepsilon_2 \quad (8)$$

Assume that the unemployment yields zero income. Then the lifetime earnings of those who experienced unemployment in the first period are given by

$$I_1 = \frac{W_2}{1+r} \quad (9)$$

Using eq.(5) the second period wages can be written as

$$W_2 = W_1 \exp \{f + gS + 3kS - \lambda U_1 + \varepsilon_2 - \varepsilon_1\} \quad (10)$$

so that lifetime earnings become

$$\ln I_0 = \delta_0 + \left[\frac{g}{1+r} + \frac{3k}{1+r} \right] S - \lambda U_1 + v_0 \quad (11)$$

which compares with the following lifetime earnings for those without unemployment

$$\ln I_0 = \delta_1 + \left\{ b + g \left[1 + \frac{1}{1+r} \right] + k \left[1 + \frac{3}{1+r} \right] \right\} S + v_1 \quad (12)$$

In a sample which includes both types of individuals, the relationship between lifetime earnings and schooling should be specified as follows

$$\ln I_1 = \lambda_0 + \lambda_1 S + \lambda_2 U + \lambda_3 US + e \quad (13)$$

4 Data and empirical model

We use the Survey of Health, Ageing and Retirement in Europe (SHARE), a multidisciplinary and cross-national panel database of micro data containing rich information on health and socio-economic status of more than 25,000 individuals aged 50 or over. In particular, we exploit the survey's third wave of data collection, SHARELIFE, which collects detailed retrospective life-histories. We focus on males because of the potential problems related to female labor force participation and exclude the self-employed.

We define permanent income at age 10 as the Net Present Value at age 10 of all wages and

pension benefits earned over the life cycle from age 10 using a discount rate of 2% ($r = 0.02$). This sum is multiplied by r because we recover the annuitized stream of income received by each individual. To make the wages and pension benefits comparable across time and country, we transform them using PPP-exchange rates and CPI measures into 2006 Euro DM. More details on how we construct the measure of permanent income can be found in Appendix A.

Table 1: Summary statistics

Variable	Mean	Std. Dev.	Min	Max
Permanent income	10051.54	6124.33	328.732	39597.88
Years of education	11.91	4.03	1	25
Year of birth	1942.13	8.73	1920	1956
Years of compulsory education	7.55	1.57	4	10
Number of jobs	3.12	2.05	1	18
Few books at age 10	0.39			
Rural area or village during childhood	0.43			
Ever unemployed	0.08			
Austria	0.04			
Belgium	0.11			
Czech Republic	0.11			
Denmark	0.13			
France	0.12			
Germany	0.12			
Italy	0.13			
Netherlands	0.13			
Sweden	0.11			
Sample size		6,809		

The sample used in the regressions consists of all males born between 1920 and 1956 with complete information on the characteristics. The European countries used in our study include Austria, Belgium, the Czech Republic, Denmark, France, Germany, Italy, the Netherlands and Sweden. Individuals from Spain, Switzerland and Greece are excluded because there is no reform of compulsory schooling that we can exploit for our cohorts of interest in these countries.² Some summary statistics on our sample are provided in Table 1.

²In Belgium, we only keep people who went to school in Flanders because the school reform of 1953

We regress permanent income on education and a series of control variables. Consider the linear model

$$Y = \beta_1 S + X^\top \beta_2 + U \quad (14)$$

where Y is a continuous random variable denoting the log of permanent income, S a continuous random variable denoting years of education, and $X \equiv \{X_k\}_{k=1}^K$ is a vector of covariates. Assume that the unobservables U are independently and identically distributed with $U \sim F_u$ and that $U \perp\!\!\!\perp (S, X)$.

We are interested in the parameter β_1 , which is the first partial derivatives of wages with respect to years of education S given values of X . The assumptions of linearity in the parameters and separability simplify the analysis and the interpretation and are a convenient point of departure.

Let us be a bit more specific about the variables in the list of covariates X . This list includes cohort dummies, country dummies, whether the individual lived in a rural area or a village during his childhood, the number of books at age ten in the place where the individual was living (excluding magazines, newspapers or school books) and country-specific age trends - the interactions between age (and the square of age) with each country.

Suppose that education is a potentially endogenous variable in this model. Consider the linear model

$$Y = S\beta_1 + X^\top \beta_2 + U + V \quad (15a)$$

$$S = Z^\top \gamma_1 + X^\top \gamma_2 + V \quad (15b)$$

where S , X and U are defined as above and $Z \equiv \{Z_l\}_{l=1}^L$ is a vector of instruments. Let $V \sim F_v$ and assume that the unobserved errors U and V are stochastically independent and identically distributed.

To be considered valid and relevant, the instruments should affect post-compulsory years of schooling but should not be correlated with the disturbances in the main equation. The list of instruments Z include compulsory years of education and an interaction term between schooling and whether the individual lived in a rural area or a village during his childhood.³

took place in this region and not in the rest of the country. We do not consider individuals from Poland because of unreliable income data. Trevisan, Pasini and Rainato (2011) argue that among others Poles seem to get confused between new and old Zloty around the devaluation in 1995 and misreport amounts during the hyper inflation of the 80s and 90s.

³Note that the compulsory years of education are computed using the country where the individual was living when the reform potentially affected him and not with the country where he is residing now. This is an important issue for Germany where we use information at the State level because some people have

Table 2: Estimated coefficients from log permanent income regressions

Variable	OLS	IV	IV - first stage
Years of education	0.0269*** [0.003]	0.0938*** [0.033]	
Compulsory years of education			-0.1131 [0.083]
Rural \times Compulsory years of education			0.3116*** [0.059]
Rural	-0.0712*** [0.016]	-0.0116 [0.035]	-3.2381*** [0.472]
Few books	-0.0344* [0.019]	0.1261 [0.085]	-2.3947*** [0.112]
Observations	6,809	6,809	6,809
R-squared	0.169	0.060	0.290
F(2, 331)			13.78

Note: All models include controls for birth cohort, country and country-specific age trends (interaction of age and its square with each country). Cluster standard errors in square brackets below the estimated coefficients. Standard errors are clustered by country and birth cohort. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The models in eqs. (14) and (15) are estimated by OLS and two-stage least squares, respectively. The first two columns in Table 2 report the estimated coefficients of years of education. The third column reports the estimated coefficients of the instruments in the first stage regression. The joint F-test statistic is above 10: the threshold value commonly used to assess the relevance of instruments.

It is worth discussing our choice of instruments. We follow standard practice in the literature on returns to education and use reforms to compulsory years of school as an instrument for actual years of education - the evidence provided by Brunello, Fort and Weber (2009) suggests this instrument has an impact not only on those directly affected by the reform, but also on others whose educational attainment was somewhat higher. In this paper we are able to interact this instrument with the place of residence at age of ten - we conjecture that the reform was likely to be most effective for children living in rural areas, whose direct and indirect costs of attending school was much higher (child labour was of course quite common in rural areas in Europe for those cohorts, and travelling to the nearest school was much expensive for children living in remote villages).

The estimates presented in Table 2 are very much in line with the Norwegian evidence of changed State and moved inside the country during their life.

Bhuller, Mogstad and Salvanes (2011) - see for instance their Table 2 on page 29, where OLS produces an estimate of 2.5% (our estimate is instead 2.69%) and IV produces an estimate of 9.9%, that compares to our estimate of 9.38%. This we take as evidence that the measurement error that is likely to be present in our data more than in administrative data does not have a major impact on the estimates once a life-time measure is constructed.

Table 3: Estimated coefficients from a log permanent income regression (with first stages)

Variable	IV	Edu.	Few books \times Edu.
Years of education	0.1212*** [0.028]		
Few books \times Years of education	-0.0913*** [0.034]		
Compulsory years of education		-0.2181** [0.090]	-0.0765 [0.065]
Rural \times Compulsory edu.		0.3005*** [0.071]	-0.0024 [0.021]
Few books \times Compulsory edu.		0.2857*** [0.099]	0.1817* [0.109]
Rural \times Few books		0.8516 [0.898]	-2.5235*** [0.745]
Rural \times Few books \times Comp. Edu.		-0.0293 [0.113]	0.2726*** [0.095]
Rural	-0.0114 [0.029]	-3.4077*** [0.565]	-0.0258 [0.179]
Few books	1.1536*** [0.399]	-4.8647*** [0.800]	9.3593*** [0.887]
Sample size	6,809	6,809	6,809
R-squared	0.019	0.294	0.843
F(5, 331)		10.60	6.75

Note: The model includes controls for birth cohort, country and country-specific age trends (interaction of age and its square with each country). Cluster standard errors in square brackets below the estimated coefficients. Standard errors are clustered by country and birth cohort. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

We also consider a model where schooling and the interaction between schooling and the number of books at age ten are endogenous regressors. S in equation (15) now includes schooling and an interaction term between schooling and the number of books at age ten. That is, there are two endogenous variables. We add to the previous list of covariates X an

interaction term between the number of books at age ten and rural area during childhood. The list of instruments Z is augmented with an interaction term between number of books at age ten and rural area during childhood and an interaction term between rural area, number of books and years of compulsory education.

The first column in Table 3 reports the estimated coefficients from a regression of log permanent income where years of education and the interaction between the number of books at age 10 and years of education are the endogenous variables. The second and third columns report the estimated coefficients in the two first stage regressions.

In Tables 3 and 4 we show that returns to education vary substantially according to socio-economic conditions at age ten. In Table 3 we focus on what has been found to be a very good indicator of the cultural background of the family, that is the number of non-school books present in the household at age ten. We construct an indicator taking value 1 if the household had less than ten books (“a shelf”). This indicator is added to the specification and is also interacted with years of education, on the one hand, and with all the instruments on the other. The key result is that the overall return of education is higher, at 12.1%, but it is substantially lower for those with few books.

5 Conclusion

In this paper we have investigated how life-time income relates to education and socio-economic background during childhood in a number of European countries using a rich data set containing detailed retrospective information on earnings, pensions and many variables of potential interest (including childhood characteristics).

Our estimates suggest that an additional year of education on average increases earnings by almost 10 percent. However, returns vary markedly with socio-economic background early in life: in particular, returns to education are much lower for those in poor conditions at age ten.

References

- Bhuller, Manudeep, Magne Mogstad and Kjell G. Salvanes (2011). "Life-Cycle Bias and the Returns to Schooling in Current and Lifetime Earnings." IZA Discussion Paper No. 5788.
- Brunello, Giorgio and Simona Comi (2004). "Education and Earnings Growth: Evidence from 11 European Countries." *Economics of Education Review*, 23(1), 75-83.
- Brunello, Giorgio, Margherita Fort and Guglielmo Weber (2009). "Changes in Compulsory Schooling, Education and the Distribution of Wages in Europe." *Economic Journal*, 119, 516-539.
- Christelis, Dimitris (2011). "Imputation of Missing Data in Waves 1 and 2 of SHARE." CSEF Working Paper No. 278, 2011.
- Cunha, Flavio and James J. Heckman (2007). "The Technology of Skill Formation." *American Economic Review*, 97(2), 31-47.
- Haider, Steven and Gary Solon (2006). "Life-Cycle Variation in the Association between Current and Lifetime Earnings." *American Economic Review*, 96(4), 1308-1320.
- Heckman, James J. (2000). "Policies to Foster Human Capital." *Research in Economics*, 54(1), 3-56.
- Heckman, James J., Lance J. Lochner and Petra E. Todd (2006). "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond." In *Handbook of the Economics of Education*, Vol.1. Hanushek, E. and F. Welch (eds.), North-Holland, Amsterdam, 307-458.
- Heitjan, Daniel F. and Roderick J.A. Little (1991). "Multiple Imputation for the Fatal Accident Reporting System." *Journal of the Royal Statistical Society C*, 40(1), 13-39.
- Horton, Nicholas J. and Ken P. Kleinman (2007). "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *American Statistician*, 61(1), 79-90.
- Schenker, Nathaniel and Jeremy M.G. Taylor (1996). "Partially Parametric Techniques for Multiple Imputation." *Computational Statistics & Data Analysis*, 22(4), 425-446.
- Trevisan, Elisabetta, Giacomo Pasini and Roberta Rainato (2011). "Cross-Country Comparison of Monetary Values from SHARELIFE." SHARE Working Paper, 02/2011.

Appendix A. Constructing a measure of permanent income

We define permanent income at age 10 as the Net Present Value at age 10 of all wage and pension benefits earned over the life cycle from age 10 using a discount rate of 2% ($r = 0.02$).

Wages and pension benefits in 2006 EURO DM

We mainly use data on work history from SHARELIFE, Release 1 but also some data from SHARE waves 1 and 2, release 2.5.0. We modify the raw data to create a measure of permanent income for each individual using the wages earned during the lifecycle and pension benefits.

Wages and pension benefits are transformed using PPP-exchange rates and CPI measures into 2006 Euro DM. PPP-adjusted exchange rates and CPI measures are taken from the OECD and national sources.⁴

Length of an employment spell

The first step is to compute the length of each employment spell. Note that when the years at the beginning and at the end of the spell are identical, we assume that the individual spent an entire year in the job, i.e. working from Jan 1 to Dec 31. When the years are different, we assume that they started and stopped working in the same month, e.g. working from March 1954 to March 1966. This implies that someone who reports to have started working in an employment spell in say 1954 and stopped in 1954 will be treated equally to someone who started in 1954 but stopped in 1955.

Missing current income

Whenever the current income from SHARELIFE wave 3 is missing but an income measure was reported at the beginning of the current employment spell, we use the income measure from the imputation module in wave 2 (if the current employment spell started before the interview year of wave 2) or from wave 1 (if the current employment spell started before the interview year of wave 1). The imputation modules in waves 1 and 2 contain measure of annual net income from employment (or self-employment) in the previous year.⁵

⁴More details can be found in Trevisan, Pasini and Rainato (2011).

⁵ For more details on the imputed variables in SHARE, see Christelis (2011).

Implausible wages or pension benefits

We set annual wages that are above the 99th percentile or below the 1st percentile of the wage distribution to missing. We proceed in a similar fashion for pension benefits. This ‘censoring’ should not create too much harm to the data. One could (or should) of course consider other floors or ceilings.

Imputing missing wage values

We impute missing wage values using predictive mean matching. Predictive mean matching method is an imputation method used for continuous variables. It is similar to a regression method except that for each missing value, it imputes a value randomly from a set of observed values whose predicted values are closest to the predicted value for the missing value from the simulated regression model.⁶ We first create a panel using all wage measures across employment spells. Annual wage is then regressed on the following list of variables: ISCED education level (3 different levels), birth cohort (3 cohorts), decade of the start of the employment spell (4 different decades), whether the worker is a white collar during the spell, whether he worked part-time during the spell, and country. Approximately 25% of the wage values are missing. Perhaps unsurprisingly, there are more missing values for jobs that started in earlier decades and less missing values for jobs that started in later decades.

Recovering the income at the end of the employment spell

In SHARELIFE, individuals are asked about the amount they were paid monthly after taxes when they started a job. They are not asked how much they were paid at the end of the spell except for the main spell in their career (if they have retired) or the current employment spell (if they are still working). That is, only the current and the main employment spells have wage measures both at the beginning and the end of the spell.

We predict the wage at the end of the spell using potential experience as the running variable. Potential experience is defined as $A_t - S - IS1$ where A_t denotes age in year t , S years of education and $IS1$ age at ISCED 1 entry. We regress (the log of) current wage on a series of characteristics: potential experience, potential experience squared, years of education, whether the job is full-time or part-time, industry⁷, country, 3 birth cohorts, a

⁶See, e.g., Heitjan and Little (1991), Schenker and Taylor (1996) or Horton and Kleinman (2007).

⁷The industries are: agriculture, manufacturing, services, public sector, community services.

concentration index in the household at age ten⁸, number of books at age ten in the place where the individual was living (excluding magazines, newspapers or school books), whether the individual was better (or much better) to others in mathematics at age ten (as opposed to about the same, worse or much worse), whether the individual was better (or much better) to others in the country’s language at age ten.

$$y_c = \beta_1 E_c + \beta_2 E_c^2 + \beta_3 E_c X_c + \beta_4 X_c + \beta_5 W + U$$

Using experience as the running variable implies that it is interacted with all the regressors X that are specific to an employment spell. W denotes regressors that are constant across employment spells (e.g. country and childhood variables). We estimate a linear-in-the-parameters model by OLS. We use the (logarithm of) wage at the beginning of employment spell j and the estimated coefficients from the regression on the current spell to predict the wage at the end of employment spell j

$$\hat{y}_{t1j} = y_{t0j} + b_1 (E_{t1j} - E_{t0j}) + b_2 (E_{t1j}^2 - E_{t0j}^2) + b_3 (E_{t1j} X_j - E_{t0j} X_j)$$

where \hat{y}_{t1j} is the predicted log wage at the end of the spell, y_{t0j} is the observed (or imputed) log wage at the beginning of spell, E_{t1j} and E_{t0j} denote potential experience at the end and the beginning of the spell respectively. Armed with the log incomes at the beginning and the end of the spell, we compute the annual growth of income during an employment spell as follows: $(\hat{y}_{t1j} - y_{t0j}) / len_j$ where len_j denotes the length of the employment spell.

To test the accuracy of our procedure, we can use the current and main employment spells where we have wage measures at both the beginning and the end of the employment spell. Table 4 shows that the predicted wage values are close to the ‘true’ values. Obviously, the standard deviation of the predicted wage measures are significantly higher. Moreover, the predicted current income is closer to the ‘truth’ because the estimated coefficients used for the prediction are taken from a regression on current income.⁹

⁸The concentration index is defined as the number of people living in the household at age ten divided by the number of rooms occupied by the household at age ten (including bedrooms but excluding kitchen, bathrooms, and hallways). If this ratio is above 10, it is set to 10.

⁹While the means of the current income and the predicted current income are not statistically different at 5% confidence level, this is not the case for the means of the main income and the predicted main income.

Table 4: Prediction error for current and main wage

Variable	Obs.	Mean	Std. Dev.
Log current income	2280	9.9434	0.4739
Predicted log current income	2280	9.9566	0.8281
Prediction error of log current income	2280	-0.0131	0.7752
Log main income	4630	9.8234	0.6850
Predicted log main income	4630	9.7706	1.0227
Prediction error of log main income	4630	0.0528	1.0816

Earnings over the life cycle up to the interview year

We first multiply all monthly income measures by 12. We annualize them because the time period for an employment spell is expressed in years (i.e. year started job, year ended job) while the income measure is in months (e.g first monthly wage). Of course, in some countries, some individuals are paid 13 or 14 months of salary per year. This is ignored here.

For each individual, the mathematical formula used to compute the discounted sum of all annual incomes is

$$A = r \cdot \sum_{j=1}^J Y_j \sum_{k=1}^K (1 + gr_j)^{(k-1)} / (1 + r)^{(S_j - (BY+11) + k)}$$

where j refers to job (or employment spell) j , J is the total number of jobs, Y_j is the annual income at the beginning of employment spell j , k refers to each year spent in the employment spell, K is the total length of each employment spell (in years), $1 + gr_j$ the annual growth rate of income during the employment spell j , S_j the year in which the employment spell j started, BY is the year of birth and r is the interest rate.

In other words, if someone is born in 1940 and starts working in 1950, the income in this job in 1950 is not discounted, in 1951 it is discounted with $1 + r$, in 1952 with $(1 + r)^2$ (denominator of the equation). The income in this job in 1950 is the annual income reported by the individual, the income in 1951 is the annual income time the annual growth rate of the employment spell, the income in 1952 is the annual income time the growth rate squared and so forth and so on (numerator of the equation).

Probability of death in the future

We then add to this discounted sum of all annual incomes up to the interview year all expected incomes and pension benefits up to age 110 (the time of death for all individuals). Each annual income and pension benefit received after the interview year of wave 3 is multiplied by the survival probability within elementary age interval, i.e. $1 -$ the probability of death between age_t and age_{t+1} . This is because we are considering income and pension benefits that will be received in the future. We hence need to take into account the possibility that individuals may die after the interview year of SHARELIFE wave 3. Of course, all individuals in our sample have survived up to the interview year. The probability of death qx_m varies across country, gender, year of birth and age. The data are based the Human Mortality Database (Department of Demography at the University of California, Berkeley and Max Planck Institute for Demographic Research). We make some modifications to the raw data. As Austria does not have data before 1947, all individuals born before 1947 are assigned the probability of death of individuals born in the year 1947. Data from Belgium replace Germany (German data start in 1991) and the Czech Republic (no data).

Pension benefits and expected pension benefits for the individuals who have retired

Some people are still working at the time of their interview, others have already retired. For those who have retired before the interview, the rest of their permanent income consists of pension benefits and expected pension benefits. For each individual who has retired, the mathematical formula for pension benefits up to the interview year is

$$B_{ret} = r \cdot pension \cdot \sum_{l=1}^L 1/(1+r)^{(RET-(BY+11)+l)}$$

where *pension* refers to the annual pension benefit currently received, l refers to each year spent in the current pension spell, L is the current length of the pension spell (i.e. the difference between the interview year and the retirement year), RET is the retirement year, BY is the year of birth and r is the interest rate. Note that we assume that pension benefits do not increase or decrease during retirement: the numerator of the expression in the sum is simply 1. Similarly, the mathematical formula for expected pension benefits (from interview

year up to age 110) is

$$D_{ret} = r \cdot pension \cdot \sum_{m=1}^M (1 - qx_m) / (1 + r)^{(INT - (BY + 11) + m)}$$

where *pension* refers to the annual pension benefit currently received (which is expected to be continuously received up to death), *m* refers to each year spent in the expected pension spell, *M* is the expected length of the pension spell (i.e. the difference between age 110 and the age at the time of the interview), qx_m is the probability of death within elementary age interval $[BY + m, BY + m + 1)$, *INT* is the interview year, *BY* is the year of birth and *r* is the interest rate. Note that we assume that expected pension benefits do not increase or decrease during the expected pension spell: the numerator of the expression in the sum is simply 1 minus the probability of death.

Measure of pension benefits

We use information from SHARELIFE wave 3 on monthly benefits after tax from social security or pensions, i.e. the sum of all pensions (public, occupational or private). We multiply the monthly benefits by 12 to obtain annual measures. When the sum of pension benefits is missing or below/above the censoring thresholds (1st and 99th percentiles), we use information from wave 1.

We compute the sum of all annual pension benefits reported in wave 2. We include public old age pension, public early or pre-retirement pension, public disability insurance, public unemployment benefit or insurance, public survivor pension from partner, war pension, private (occupational) old age pension, private (occupational) early retirement pension, private (occupational) disability insurance, private (occupational) survivor pension from partner's job, public old age supplementary pension, secondary public disability insurance pension, secondary public survivor pension from spouse/partner, occupational old age pension from a second job, occupational old age pension from a third job, and private (occupational) disability insurance. We censor this sum for values below/above the censoring thresholds (1st and 99th percentiles). We then use this measure to replace the missing values of pension benefits that could not be recovered using SHARELIFE.

When the sum of pension benefits is still missing, we use information from wave 1. We include public old age pension, public early or pre-retirement pension, public disability insurance, public unemployment benefit or insurance, public survivor pension from partner, public

invalidity or incapacity pension, war pension, private (occupational) old age pension, private (occupational) early retirement pension, private (occupational) disability insurance, and private (occupational) survivor pension from partner's job. We censor this sum for values below/above the censoring thresholds (1st and 99th percentiles). We then use this measure to replace the missing values of pension benefits that could not be recovered up to this point.

Expected incomes and expected pension benefits for the individuals who are still working

For those who are still working, we do not know the amount of their pension benefits. The rest of their permanent income consists of expected income up to expected pension age and expected pension benefits from expected pension age up to death. To compute a measure of permanent income that includes all working episodes over the life cycle for all individuals, we create a new artificial employment spell that should correspond to the last employment spell until retirement. Obviously, for those who have already retired, the length of this artificial employment spell is equal to zero. For those who are still working, the length of the employment spell is the difference between the age at which they expect to collect pension benefits and their current age. If these two ages are equal, we assume that they retire immediately and start collecting pension benefits.

The creation of this new artificial employment spell implies a number of assumptions. First, we assume that all individuals who are still working at the time of the interview in SHARE-LIFE wave 3 will work up to their expected retirement age. That is, they will not stop working before retirement age. Moreover, they will never be unemployed until retirement. This also implies that individuals who are still working but have passed the retirement age will immediately stop working and retire. We predict the wage at the end of this spell in a similar fashion to the way we predict the wage at the end of each employment spell. We then compute the growth of income from interview year up to retirement year. When individuals do not report at what age they will start collecting pension, we use information on statutory retirement age in their country. Sometimes the statutory retirement also varies across gender within a country but we in this study focus on males.

We compute the discounted sum of expected incomes up to expected pension age for each individual

$$C_{work} = r \cdot Y_{curr} \cdot \sum_{s=1}^S (1 + gr)(1 - qx_s) / (1 + r)^{(INT - (BY + 11) + s)}$$

where Y_{curr} refers to the current income, s refers to each year spent in the current employment

spell up to expected pension age, S is the expected length of this new artificial employment spell (i.e. the difference between the expected pension age and the interview year), $1 + gr$ the annual growth rate of income during the employment spell, qx_s is the probability of death within elementary age interval $[BY + s, BY + s + 1)$, INT is the interview year, BY is the year of birth and r is the interest rate. Similarly, the mathematical formula for expected pension benefits (from expected pension age up to age 110) is

$$D_{work} = r \cdot reparte \cdot Y_{curr} \cdot \sum_{t=1}^T (1 - qx_t) / (1 + r)^{(PY - (BY + 11) + t)}$$

where *reparte* refers to the replacement rate (or percentage of salary received as pension), Y_{curr} refers to the current income, t refers to each year spent in the expected pension spell, T is the expected length of the retirement spell (i.e. the difference between age 110 and the expected pension age), qx_t is the probability of death within elementary age interval $[BY + t, BY + t + 1)$, PY is the expected retirement year (the year in which the individual will start receiving pension benefits), BY is the year of birth and r is the interest rate. In the formula above, we use current wage and not the predicted wage at expected retirement age because individuals are asked what is the percentage of current wage that will be received as pension (and not the percentage of expected wage at retirement age). Note that we assume that expected pension benefits do not increase or decrease during the expected pension spell: the numerator of the expression in the sum is simply 1 minus the probability of death.

Measure of expected pension benefits

The expected percentage of salary received as pensions are reported by the individuals who are working in wave 2. We take the sum of all expected percentages. We set this sum of percentages to missing when the value is below 50% or above 100%.¹⁰ If the value from wave 2 is missing, we use data from wave 1. Because there are still some missing or implausible values at the individual level, we compute the median replacement rate within country and 3 birth cohorts (we hence have 3 possible values for each country). We then applies this median replacement rate to each individual who had a missing value. The replacement rate multiplied by the current income should be a good approximation to the pension benefits.

As an alternative measure for the replacement rate, we could compute the ratio of pension benefits over the main wage in the career. We obtain this information from the individuals who have already retired. We could then use the median of this ratio by country and apply

¹⁰This is an issue for some individuals in Sweden, Denmark and the Netherlands.

it to the individuals who are still working. We are currently using information provided by individuals who are working and report their own expected percentage of salary received as pensions. We believe that this is superior to using information reported by individuals who have retired at different points in time.

Total permanent income

For the individuals who have retired at the time of the interview, their permanent income is given by

$$NPV_{10} = A + B_{ret} + D_{ret}$$

where NPV_{10} refers to the net present value at age 10 of all wages, pension benefits and expected pension benefits earned up to death.

For the individuals who are still working at the time of the interview, their permanent income is given by

$$NPV_{10} = A + C_{work} + D_{work}$$

where NPV_{10} refers to the net present value at age 10 of all wages, expected income and expected pension benefits earned up to death.

We set the value of permanent income to missing if it is below the 1st percentile or above the 99th percentile of the permanent income distribution.