Working papers series

WP ECON 14.06

Estimating Human Capital Externalities: The Case of the Spanish Provinces, 1995-2010

Manuel Hidalgo-Pérez (U. Pablo de Olavide)

Walter García-Fontes (Universitat Pompeu Fabra and Barcelona GSE)

Keywords: externalities, human capital, Spanish provinces.

JEL Classification: I21, J31, O47.

Department of Economics

Estimating Human Capital Externalities:

The Case of the Spanish Provinces, 1995-2010

Walter García-Fontes*
Manuel Hidalgo†

May 2013 ‡

Abstract

We estimate the strength of schooling externalities for Spanish provinces over the 1995-2010 period. Our empirical work employs both main approaches available in the literature, the Constant Composition Approach and the Mincerian Approach. Using data from the Continuous Sample of Working Records and change in province human capital stock we find that both methodologies yield significant externalities.

JEL codes: I21, J31, O47

1 Introduction

Estimating human capital externalities, the difference between the social and the private marginal returns to human capital, is important for various reasons. First, the strength of such externalities determines the optimal subsidies to education and to immigration of highly qualified workers. Second, human capital externalities have been emphasized as a key for understanding the process of economic growth (e.g. Lucas, 1988). It is therefore not too surprising that there are a variety of estimation approaches and estimates in

*Universitat Pompeu Fabra and Barcelona Graduate School of Economics
†Universidad Pablo de Olavide
‡Walter García-Fontes thanks partial support of grants ECO2011-25272 and ECO2011-30323-C03-02 (Gobierno de España) and Manuel Hidalgo thanks grants PAI SEJ-246 (Gobierno de España) and P07-SEJ-02479 (Proyecto de Excelencia de la Junta de Andalucía). We thank the comments from Antonio Ciccone and two anonymous referees. Corresponding author: Manuel Hidalgo. Universidad Pablo de Olavide, Ctra de Utrera, Km1, Sevilla. CP 41013, Spain. Email: mhidper@upo.es.

the literature (e.g. Rauch, 1993; Black and Henderson, 1999; Rudd, 2000; Acemoglu and Angrist, 2001; Moretti, 2004a; Moretti, 2004b).

For Spain there is much less work, however. The available estimates of the return to human capital almost all reflect private returns (e.g. Alba and San Segundo, 1995; Barceinas et. al., 2000; Raymond, 2002; De la Fuente, 2003; De La Fuente et al., 2003; Arrazola et al, 2003 ; Arrazola and Hevia, 2008). As far as we know, there are only three attempts to estimate social returns to education or externalities. De la Fuente and Domenech (2006) estimate social marginal returns to education in the '90s, while Alcalá and Hernández (2005) estimate human capital externalities at the firm and industry level. García-Fontes and Hidalgo (2008) find externalities using aggregate regional data for the period 1980-2000. All these papers find evidence of positive and significant externalities for Spain.

Compared to the previous papers for the Spanish case, the contributions of this paper are two. First, it uses new available information on earnings. Second, it uses the methodology of Ciccone and Peri (2006) at the provincial level. The provincial level estimation, as opposed to autonomous community level estimation, may allow for a more accurate estimation of human capital externalities.

Methodologically, there are currently two approaches in the literature to estimate human capital externalities. The first approach augments standard Mincerian wage equations with variables that measure the level of human capital at some geographical level (e.g. Rudd, 2000; Acemoglu and Angrist, 2001; Moretti, 2004a). This methodology estimates the strength of human capital externalities by looking at the effect of states or regional human capital levels on individual wages. The basic idea is that human capital externalities should show up in individual wages once all relevant individual characteristics are controlled for. A key assumption of this (Mincerian) approach is that workers with different levels of human capital are perfect substitutes in production. If different human capital levels are imperfect substitutes, the Mincerian approach yields a positive effect of aggregate human capital on individual wages even if the social return to human capital equals the private return (e.g. Ciccone and Peri, 2006). The intuition is that with imperfect substitution, an increase in the number of skilled workers implies an increase of the wages of unskilled workers that more than offsets the decrease in the wage of skilled workers. The empirical evidence for the United States (e.g. Katz and Murphy, 1992; Ciccone and Peri, 2006), as well as other countries (e.g. Angrist, 1995) including Spain (Hidalgo, 2010), indicates that different levels of human capital are imperfect substitutes. Therefore the Mincerian approach must be complemented with the

so-called Constant Composition approach (CC hereafter), which yields consistent estimates of the wedge between the social and the private return to human capital even if skilled and unskilled workers are imperfect substitutes (Ciccone and Peri, 2006). This approach estimates the strength of human capital externalities as the marginal effect of aggregate human capital levels on average wages holding the labor force composition constant. Ciccone and Peri show that this bias is directly related to the wage difference between skilled and unskilled workers, and inversely related to the elasticity of substitution. The smaller the wage premium of skilled workers the smaller the bias introduced by the Mincerian approach.

The estimation of externalities with the CC and the Mincerian approach consists of two steps. First, we need aggregate average wages at the provincial level for the Mincerian approach, and provincial and educational levels for the CC approach. To obtain these average wages we further need individual wage records. Using traditional wage equations, with different covariates such as education, experience or other individual observables, we obtain "cleaned" individual wages, i.e. wages without the variance these covariates could explain. With these residuals or "cleaned wages" we construct provincial aggregate wages as the average for each provincial and education level. The estimation of these aggregate average wages will be called first-step estimation (FSE hereafter). Secondly, externalities are obtained by the estimation of the marginal effect of a defined province aggregate human capital index on provincial cleaned-average wages in a second-step estimation (SSE hereafter). In this particular exercise the province aggregate human capital level is approximated by its average years of schooling.

Instead of García-Fontes and Hidalgo (2008), where three cross-sections of Spanish Household Budget Survey (1980, 1990 and 2000) are used in the FSE, here we use data at the province level from five waves, from 2006 to 2010, of the Continuous Sample of Working Records (CSWR 2006-2010, Muestra Continua de Vidas Laborales 2006-2010). These data offer employment records of workers between 2006 and 2010. The benefits of using the CSRW are the detailed information contained and the large size of the sample. However, the use of the CSWR data set has three important problems (García-Pérez, 2010) that we have to take into account.

First, the information on earnings is given by Social Security contributions and not directly by wages. Social Security contributions are censored both from above and from below, because of the legal maximum and minimum worker contributions. To obtain earnings from the CSWR data we use the methodology proposed by Boldrin et al. (2004) and Felgueroso et al. (2010). We present detailed descriptive statistics that compare the earnings

information obtained from the CSWR with other data sources that use wages directly. We also compare our results with estimations based on the median, since estimations using median wages would not be affected by censoring. In all cases we obtain that the CSWR is an appropriate data source for our goals.

A second problem with the CSWR is that education is obtained by matching of the Social Security records with municipal information (called "padrón" in Spanish cities). Since the municipal data is not updated often, it is possible that education attainment for some workers in the data is lower than their actual attainment. To deal with this problem, we adjust the education variable by using the occupation information found in the CSWR, and we construct an education variable that we call adjusted level of education (ALE).

A third problem comes from the nature of the CSWR, since it provides information corresponding to working records of workers with a contract during 2006 and 2010. For this reason, as we go back in time the data become less reliable since it will not reflect the actual structure of the labor force. For instance, average age of workers will decrease the farther away from the sampling period. To deal with this problem we restrict our estimations to the 1995-2010 period, and to obtain aggregate average wages we only use records for workers in the 25-65 and 25-55 age groups.

A key issue when estimating human capital externalities at the second step at the provincial level is that changes in aggregate human capital levels are endogenous as provinces with higher productivity and wages may attract more skilled workers. This makes it desirable to implement an instrumental-variables approach. Furthermore, the use of schooling proxies as a proxy for human capital levels introduces measurement errors. This is why we have to look for appropriate instruments to get a consistent estimate of education externalities. As an instrument, we use the age structure of the province with a five year lag. We think it is an appropriate instrument due to the correlation between the size of population cohorts with the total population acquiring information, and its power to predict future workers with a college degree. But the size of younger and older population must not be correlated with the change in the average wage in the long run. We include geographical dummies to control for similar productive structures that may affect the wage level and its change. Assuming that these instruments are exogenous, we can use them in our exclusion restrictions. Finally, heteroskedasticity has to be taken into account because of the provincial structure of the data. The estimation technique to be used is therefore Generalised Method of Moments (GMM).

Both approaches yield evidence of significant human capital externalities for each of the education-age estimations done. Also, we observe evidence that the Mincerian approach yields larger externalities than the constant composition approach. The difference in the point estimate of the externalities of human capital between the two approaches is tested. The results show a significant difference in favor of the Mincerian approach of around 8-15%, which is larger than what the theory predicts. It can be shown, however, that using previous estimates of the elasticity of substitution between skill levels in Spain, the bias of the Mincerian approach can be as high as 25%. Our results seem nevertheless consistent with this value, as it falls within the confidence interval around our estimate. It can be therefore considered that the Mincerian approach provides an upper bound for the estimation of human capital externalities, while the Constant Composition approach provides a lower bound (Ciccone and Peri, 2006).

The rest of the paper is organized as follows. Section 2 summarizes the relevant literature, while Section 3 reviews the two main empirical methodologies used and how the bias might be calculated. Section 4 presents the Mincerian approach, while section 5 discusses some possible econometric problems. Section 6 presents the data sources used. The main results are in Section 7. Finally, section 8 concludes.

2 Related Literature

The strength of human capital externalities is defined as the difference between the social and private marginal return to an additional unit of human capital. Most empirical work focuses on the return to an additional year of (formal) schooling. There is a very large literature on the private return to an additional year of schooling, which has found this return to lie between 5 and 12% depending on the country and time period considered (e.g. Card, 1999). There is less work estimating the strength of schooling externalities. Rauch (1993) estimates schooling externalities in the US in 1980 using a Mincerian wage equation augmented for state-level schooling measures. The idea behind the Mincerian approach is that if there are externalities, individual wages should be increasing in aggregate schooling levels controlling for individual characteristics, such as education, experience, gender, etc. Rauch finds schooling externalities between 3 and 5%. Later contributions refine the Mincerian approach by using panel data to control for province fixed effects and by employing instruments for the change in aggregate schooling levels, see Acemoglu and Angrist (2001), Rudd (2000), Conley, Flier, and Tsiang (2003), Moretti (2004a), and Moretti (2004b). The results vary

with the time period, the level of spatial aggregation, the country and the specification. For example, while Acemoglu and Angrist (2001) do not find state-level average schooling externalities in the US over the 1960-1980 period, Moretti (2004b) finds externalities from the share of college workers in the US to be significant at the city level for 1981-1991.

The Mincerian approach to human capital externalities assumes that workers with different human capital levels are perfect substitutes in production. Perfect substitutability simplifies identification because it implies that changes in the relative supply of human capital do not affect the relative wages of the different human capital groups, holding total factor productivity constant. Consequently, all the effects that human capital supply changes have on workers with a given level of human capital have to come through total factor productivity and can be interpreted as externalities. Ciccone and Peri (2006) show that when workers with different human capital are imperfect substitutes, the Mincerian approach overestimates the strength of schooling externalities. They propose an alternative methodology that estimates externalities as the marginal effect of human capital on log average wages holding labor-force composition constant.

De la Fuente and Domenech (2006) relate province productivity growth to human capital and other variables. They estimate a 16% return of human capital, which is larger than the elasticity estimated in previous works which is around 8%. They attribute the difference to the social return to education. Using an internal rate of return approach, they estimate the social rate of return of education to be between 10% and 12%. Taking the difference between the private and the social return, externalities are estimated to be between 4% and 5%. Alcalá and Hernández (2005) estimate the externalities of human capital at the firm level. Their estimates show a private return of 8% and externalities equal to 4.7%. García-Fontes and Hidalgo (2008), using the General and Continuous Expenditure Surveys and regional data, estimate externalities to be in the range 4 to 5%.

This paper extends the previous literature for the Spanish case and applies a methodology that provides a consistent estimate of human capital externalities.

3 Methodology and Econometric Specification

The main contribution of this paper is to estimate externalities using more recent information and using both methodologies available, Mincerian and Constant Composition. In this section we will provide the conceptual framework behind the novel Constant Composition approach following Ciccone

and Peri (2006), and how to go from the theory to the empirical implementation.

3.1 The Constant Composition Approach

In this section we follow Ciccone and Peri (2006). Let us assume that provincial output Y at province p in year t depends from total low skill workers L and total high skill workers H, as well from the stock of physical capital K and technological level A.

$$Y_{pt} = A_{pt} K_{pt}^{\gamma} F(L_{pt}, H_{pt}). \tag{1}$$

Let us assume that the aggregate production function is twice continuously differentiable and subject to constant returns to scale. Let us also assume that the labor market in each province is perfectly competitive and therefore, equilibrium real wages are equal to marginal productivity, holding aggregate technology constant. We can write:

$$w_{pt}^{L} = A_{pt} K_{pt}^{\gamma} F_1(L_{pt}, H_{pt}) = A_{pt} K_{pt}^{\gamma} F_1(l_{pt}, (1 - l_{pt})) \tag{2}$$

$$w_{pt}^{H} = A_{pt} K_{pt}^{\gamma} F_2(L_{pt}, H_{pt}) = A_{pt} K_{pt}^{\gamma} F_2(l_{pt}, (1 - l_{pt})). \tag{3}$$

where $l = L/N$ and N is the total number of workers, with $N = L + H$. Equations (2) y (3) assume that each province is a labor market, in other words, identical workers should obtain the same wages. There is no assumption on (the absence of) migration. Specifically, firms can hire workers at the national level.

We introduce externalities allowing aggregate technology A in (1) to be incremented by some measure x of work force qualification in the province. The functional form used is

$$A_{pt} = B_{pt} x_{pt}^{\theta} N_{pt}^{\delta}, \tag{4}$$

where x is an aggregate measure of high skill intensity of workers in the province. B captures all the other determinants of total factor productivity. See Sveikauskas (1975), Segal (1976), Moomaw (1981), Henderson (1986) and Rauch (1993) for the effects of scale in US cities.

We write now the fraction of low skill workers and high skill workers as a function of the aggregate measure of high skill intensity:

$$\begin{aligned} l_{pt} &= g^{L}(x_{pt}), \\ h_{pt} &= 1 - l_{pt} = g^{H}(x_{pt}). \end{aligned} \tag{5}$$

For instance if $x = h/l$ then $l = l + (1 + x)$ and $h = x/(1 + x)$. We will use this result to write average wage as a function of the high skill intensity measure x.

Average wage w in province p at time t can be defined as the wage of the two groups weighted by their relative size:

$$w_{pt}(x_{pt}, l_{pt}) = w^L_{pt}(x_{pt})l_{pt} + w^H_{pt}(x_{pt})(1 - l_{pt}), \tag{6}$$

since (2)-(5) imply that high skill and low skill wages can be written as a function of the aggregate measure of high skill intensity.

Equation (6) shows that an increase in observed average wage in a province may be due to changes in the composition of the labor force in favor of high skill workers or an increase in the wage of workers of all skills, but with a fixed skill composition. The Ciccone and Peri methodology consists in using this second effect to estimate the effects of human capital intensity on wages.

To see how this methodology works, it is useful to think on what would happen if there were no externalities associated with the intensity of human capital. In this case if there is a small increment in x in a province, average wages would not change if the labor force skill composition is held constant. Intuitively this is so because high skill workers are paid their marginal product in each province and therefore there are no externalities. Consequently, an increase in total wages associated with a small increment in the intensity of high skill workers in the province would go all to the new high skill workers, who are responsible of the increase in the high skill intensity. For this reason total wages of the existing workers remains unchanged despite the increase in the high skill intensity. In sum, aggregate average wage would not change if we hold the labor force composition constant.

This can summarized in the following proposition:

Proposition 1:

The elasticity of average wages with respect to an aggregate measure of labor force quality (x) when we hold labor force composition constant is equal to the externality of human capital θ.

$$\frac{\partial \ln(w^L_{pt}(x_{pt})l_{pt} + w^H_{pt}(x_{pt})(1 - l_{pt}))}{\partial x_{pt}} x_{pt} = \theta. \tag{7}$$

Appendix A include the proof of this proposition.

Notice that the elasticity of average wages with respect to measure x of labor force quality allows to identify human capital externalities only when labor force composition is held constant. If labor for composition is not constant, then the elasticity will show the total effect of the change in labor

force composition on average wages, according to the two effects mentioned in equation (6).

In sum, holding labor force composition constant, an percentage increment in average wages as a response to a percentage in a labor force quality measure can be interpreted as the external effect of human capital. This methodology implies also a more general view on human capital externalities, since it suggests that any observation that high skill workers are being paid below their marginal product at the provincial level should be evidence in favor of human capital externalities.

3.2 Empirical approximation to the constant composition approach

The theoretical basis presented in the previous subsection for the constant composition approach proposed by Ciccone and Peri (2006) is summarized as follows: under general conditions, the value of the externalities of human capital is equal to the average weighted effect that human capital has on wages, which in turn is equal to the marginal effect of human capital over average wages holding constant the composition of the labor force.

Therefore to apply this estimation methodology we need to perform a two-stage procedure. In the first stage estimation (FSE) it is necessary to obtain a measure of weighted average wages holding the composition constant for each schooling level for the different periods. For this purpose average wages are computed for each of the education groups defined and they are used to obtain weighted average wages using the weights for one of the years included in the period. To implement this we estimate average wages by schooling levels once we eliminate wage differences unrelated to education. Call w_{ispt} the wage of individual i with schooling level s in province p at period t, and z_{ispt} the characteristics of this individual that we want to clean out from our measure of average wages. We construct a measure of adjusted average wages of workers with schooling level s in province p and period t as the estimated constant in the following regression:

$$\log w_{ispt} = \sum_{q=0}^{P} \alpha_{qst} D(p=q) + \sum_{j=0}^{J} \beta_{st}^{j} z_{ispt}^{j} + u_{ispt} \tag{8}$$

where P is the total number of provinces, J are the different individual characteristics that we want to control, $D(p=q)$ is a dummy variable which is equal to 1 if $p=q$ and 0 otherwise, and u_{ispt} is a residual. This equation provides estimations for $\hat{\alpha}_{qst}$, the average wage of workers with schooling level s in province p at moment t, adjusted by characteristics z, that in

this case we use experience, type of contract (full-partial time), gender, firm tenure and (four) sectors.

Next, we use these adjusted wages by schooling level, province and time to construct an average adjusted wage holding composition constant:

$$\log \hat{w}^F_{pt} = \log \sum_{s=1}^{S} l_{spT} \hat{\alpha}_{spt}$$

Notice that the proportions l_{spT} correspond to the base year T. The average wages computed allow us to evaluate the increase in adjusted average wages holding constant the composition of the labor force, $\log(\hat{w}^F_{pt}) - \log(\hat{w}^F_{p,t-1})$. Finally, in a second stage (SSE), we estimate the intensity of externalities by an empirical formulation of a discrete version of (7):

$$\log(\hat{w}^F_{p,t}) - \log(\hat{w}^F_{p,t-1}) = controls_{pt} + \theta(\log(h_{pt}) - \log(h_{p,t-1})) + u_{pt}, \quad (9)$$

where h_{pt} is our province human capital index for year t. Controls include variables that are present in (8),i.e. the log in total employment in the provinces to take into account scale effects, as well as log of physical capital over GDP, since the level of this factor may also increase average wages.[1] Since we are working with growth rates, permanent changes in wages at the province level do not affect our results. For instance, if firms in the service sector of Madrid or Barcelona pay wages which are 30% larger than the firms in Seville due to higher living costs, then these differences in wages will not affect our results as long as they are constant over the period considered. Generally, shocks that increase average wages in all provinces equally (such as the national inflation rate) do not affect the results since they simply get absorbed by the constant of our regressions. However, differences in the inflation rate may have an effect, since the increase in the level of prices is included as a control. The error term u_{pt} is allowed to have a province fixed term, representing differences in weighted wages growth changes due to the province non-observed heterogeneity.

We reduce the sample periods to four periods, which allows to use three differences in average wages as well as the covariates, including the human capital intensity measure, namely 1995-2000, 2000-2005 and 2005-2010. In sum the estimations will use 3 differences and 50 provinces.

An important issue in this method is that the estimation of externalities depends on the weights chosen to compute the constant composition average

[1]It can be observed that any proportional transformation of the variables in (8) that does not affect the provincial schooling level does not modify the result in (9). Therefore we can use the ratio instead of the human capital stock in levels since it helps in the interpretation of the coefficients.

wages. Ciconne and Peri (2006) show that if the production function is concave with respect to high skilled workers, that is, if it has marginal returns to scale with respect to high skilled workers net of externalities, then the values of the estimated externalities that we obtain choosing the weights of the initial and final period constitute a lower and an upper bound of the true externalities (not necessarily respectively). The true value of externalities lies in an intermediate unknown point between the two bounds. In this work we are going to present one estimation using a base year equal to the year 2000[2].

Finally, the estimation of externalities in (9) may be affected by endogenity of the schooling levels, since migration across provinces implies that schooling levels are endogenous. Higher productivity and wages in a province may lead to it attracting high skilled workers from elsewhere. Another factor that may work in the same direction is that high income provinces may have amenities that are especially attractive for high skilled workers. Such concerns should be much attenuated by our panel data approach, which eliminates all permanent differences across provinces. But residual endogeneity of province schooling levels could lead to inconsistent least-squares estimates. In section 5 we describe the instruments and methodology we will use to deal with this problem.

4 Other Empirical Approaches to Human Capital Externalities: the Mincerian Approach

As mentioned earlier, since the work by Rauch (1993) there have been many attempts to estimate social returns to education using a typical Mincerian equation, as in Mincer (1974). The main idea is to introduce an additional variable proxying the average endowment of human capital at the city, province or country level. Let's look at a simple case including only two types of workers (qualified and non-qualified). In this case, a simple version of the wage equation can be written as:

$$\log(w_{pt}) = \theta h_{pt} + \alpha_{pt} + bD_{it},$$

where w_{ipt} is the wage of worker i in province p and year t, h_{pt} represents the value of human capital in the province and D_{it} is a dummy which is equal to 1 for qualified workers and 0 otherwise. In this case, θ can be interpreted as the social marginal return of the province human capital on individual wages, showing therefore the value of externalities.

[2]Results using other weights are not presented to save space, but do not change the results and are available upon request from the authors.

In order to estimate θ a two-stage procedure can be used again. In the FSE, a classical Mincerian regression is estimated on the log of individual wages against individual characteristics which are supposed to affect wage determination. The goal of this first stage is again to estimate average province wages clean of individual characteristics, including private returns to education:

$$\log(w_{ipt}) = \alpha_{pt} + \gamma s_{ipt} + \sum_{j=0}^{J} \beta_t^j z_{ipt}^j + u_{ipt} \tag{10}$$

where w_{ipt} is the wage of individual i at province p during year t, s_{ipt} is average years of schooling, z_{ipt} are the same J individual characteristics as in the constant composition approach that we want to control, and u_{ipt} captures the effect of heterogeneous non-observable variables and estimating errors. The constant terms α_{pt} correspond to average wages for each province and year.

Since there may be unobservable variables included in the error term, which may be correlated with the schooling variable, the model may present endogeneity problems. These omitted unobservable variables, for instance innate ability, can cause estimation biases and have to be dealt with by using appropriate instruments. For the Spanish case it is not clear which instruments to use, since individual data are not available [3].

As we have shown earlier, the estimation of the coefficients related to education in (10) will be biased. A possible solution would be to use instruments, but we do not follow this route because of two reasons. First, it is hard to find appropriate instruments using Spanish data, and in our case the ones suggested in the literature would have reduced explanatory power. For instance Arrazola et al (2003) y Arrazola y Hevia (2008) have used instruments related to the opportunities to study as instruments, since the different reforms in the educational system and events as Spanish Civil War. In our case this instrument does not have sufficient explanatory power since most workers in our sample started working after the reforms and therefore there is little heterogeneity with respect to this variable. A second reason is that if we assume that we have measurement error in our dependent variable which is average wages, this will reduce the coefficient precision but will not imply a biased estimation. Furthermore if the measurement error is constant over time, since we use the differences in average wages, this error will be

[3] There are several examples of the estimation of private returns to education without controlling for possible endogeneity, for instance Alba and Segundo (1995), Oliver, Raymond, Roig, and Barceinas (1999) and Vila and Mora (1998). Nevertheless there are some examples of the use of instruments, for instance Arrazola et. al. (2001).

mitigated [4]. For these reasons we have chosen to estimate the Mincerian equation without using instruments that would not correct appropriately for the endogeneity problem.

In the SSE, in order to obtain the marginal effect of the human capital of province p over average wages, we compute the first difference in constant terms for each province ($\Delta\hat{\alpha}_{pt} = \hat{\alpha}_{pt} - \hat{\alpha}_{p,t-1}$) and we regress these differences in average wages on the change in our human capital indicator ($\Delta h_{pt} = h_{pt} - h_{p,t-1}$):

$$\Delta\hat{\alpha}_{pt} = controls + \theta\Delta h_{pt} + v_{pt}. \tag{11}$$

The value of human capital externalities according to the Mincerian approach is represented by θ. Similar controls used in the CC approach are also introduced corresponding to variables which affect changes in average wages which are not related with changes in the endowment of average human capital in the province. Finally, as in the CC approach, estimation of (11) has to be performed using instrumental variables due to the possible endogeneity of average wages and human capital stocks.

5 Endogeneity, Measurement Error and Estimation

Migration across provinces implies that schooling levels are endogenous in the SSEs (9) and (11). We therefore have to implement an instrumental estimation procedure, and we propose to use the beginning-of-period population structure as an instrument for the change in province schooling over the following year(s). The underlying assumption is that a higher share of younger (older) people implies a greater increase in average schooling levels in the province. Also, following Ciccone and Peri (2006) we use as instruments the geography of the provinces, trying to capture possible effects on human capital accumulation due to the similarity across provinces. Lastly, we use the total population as an instrument, under the assumption that a larger population implies more educational services such as the existence of local universities. Despite the fact that students may cross provincial borders and return to their original province after attaining some education level, we assume that the higher the chances to get education near their home, the higher must be the average schooling level of local workers (Card, 1993).

Assuming that endogeneity bias leads to an overestimation of externalities, the use of proxies for human capital in the estimation of externalities

[4] If the estimation of γ is biased not only the return to education will be affected but also the estimation of the *alpha*'s. In this case a positive correlation between education and the error term will underestimate the value of adjusted average wages obtained in (10).

may introduce a bias because of measurement errors of opposite sign to the endogeneity bias. Despite the fact that both biases can be corrected using instrumental variable estimation, it is not possible a priori to predict the sign of the bias as it is not known which of these two biases will prevail.

Suppose that $y_{pt} = \log(\hat{w}^F_{r,t}) - \log(\hat{w}^F_{r,t-1})$ and $x_{pt} = (controls'_{pt}, \triangle h_{pt})$ are our k explanatory variables in (9) and $y_{pt} = \Delta\hat{\alpha}_{pt}$ with the same x_{pt} in (11) for each province p. Then, the two models we estimate are rewritten as:

$$y_{pt} = x'_{pt}\gamma + v_{pt} \tag{12}$$

where $\gamma' = (\mu', \theta)$, with θ being the strengh of externalities. Suppose that

$$v_{pt} = \phi_p + \varepsilon_{pt}$$

where ϕ_p is a province fixed effect and ε_{pt} is the random part of the error term.

To take into account a provincial fixed effect in the estimation of (12) we demean all variables (we take the difference between the variable and its mean). We keep the same notation used so far but it has to be taken into account that now we refer to demeaned variables.

Let Z be an array $(r \times t)$ that contains the r instruments we have defined above. Given that we suposed that all these instruments are uncorrelated with respect to v_{rt}, we can impose the following r moment equations:

$$E^r[z^r_{pt}(y_{pt} - x'_{pt}\gamma)] = 0 \tag{13}$$

Just because in this exercise $r > k$, the number of instruments exceeds the number of endogenous variables, the moments equations represented in (13) implies a system with more equations than unknowns in γ. In such a case to maintain these restrictions we need that the matrix $E^r[z^r_{pt}(y_{pt} - x'_i\gamma)]$ has reduced rank k.

However, even in that case, the sample counterpart of (13)

$$\frac{1}{N}\sum_{i=1}^{N} z_i(y_i - x'_i\gamma) = 0 \tag{14}$$

will not have reduced rank because of sampling error. Therefore there will not be an unique solution for γ.

The objective is then to find an estimation of our coefficients of interest minimizing the quadratic distance of (14). To do that we estimate this coefficients using the Generalized Method of Moments (GMM). Then we

need to impose a $r \times r$ non-negative definite weight matrix A_N. Then, $\hat{\gamma}$ is a GMM estimator of γ if

$$\hat{\gamma} = argmin_{c\in\Gamma} \left(\frac{1}{N} \sum_{i=1}^{N} z_i(y_i - x_i'c) \right)' A_N \left(\frac{1}{N} \sum_{i=1}^{N} z_i(y_i - x_i'c) \right) \quad (15)$$

Given $y = (y_1, y_2, ..., y_N)$,$X = (x_1, x_2, ..., x_N)$, and $Z = (z_1, z_2, ..., z_N)$ where $N = p \times t$, the matrix sample-counterpart of (15) is the GMM objective function

$$N^{-2}(y - Xc)'ZA_NZ'(y - Xc)$$

The estimator of γ that minimizes this objective function is

$$\hat{\gamma}_{GMM} = N^{-2} \left[X'ZA_NZ'X \right]^{-1} X'ZA_NZ'y$$

Under conditional homoskedasticity $E(\varepsilon_{pt}|z_{pt}) = \sigma^2$ it is known that the optimal weighting matrix is $A_N = (Z'Z)^{-1}/N$. In that case, the GMM estimator for γ is:

$$\hat{\gamma}_{GMM} = \left[X'Z(Z'Z)^{-1}Z'X \right]^{-1} X'Z(Z'Z)^{-1}Z'y$$

which is the 2SLS estimator

However, we consider that the errors ε_{pt} are not homoskedastic. We assume that each province has a difference error variance. So, in this case, $E(\varepsilon\varepsilon'|z_{pt}) = \sigma^2\Omega$. Then, the GMM estimator for γ is

$$\hat{\gamma}_{GMM} = \left[X'(\hat{\Omega}^{-1} \otimes Z(Z'Z)^{-1}Z')X \right]^{-1} X'(\hat{\Omega}^{-1} \otimes Z(Z'Z)^{-1}Z')y$$

or

$$\hat{\gamma}_{GMM} = \left[\hat{X}'\hat{\Omega}^{-1}\hat{X} \right]^{-1} \hat{X}'\hat{\Omega}^{-1}y$$

where

$$\hat{X} = \hat{\Pi}'Z$$

with $\Pi = (Z'\hat{Z})^{-1}Z'X$. In that case, $\gamma_{G\hat{M}M}$ is the FGLS estimator of $\hat{\gamma}$.

To verify the validity of the instruments we show the results for the test of overidentifying restrictions and from underidentification and weak identification. In the first case the Hansen's J statistic is computed, which under the null hypothesis is distributed as chi-squared in the number of overidentifying restrictions. The null hypothesis says that the instruments are valid instruments, i.e., uncorrelated with the error term, and that the

excluded instruments are correctly excluded from the estimated equation, The under and weak identification test means that at the null hypothesis of the test (the matrix of reduced form coefficients has rank=k-1 where) the equation is underidentified. Under the null of underidentification, the statistic is distributed as chi-squared with degrees of freedom=$(L-k+1)$ where L=number of instruments (included+excluded). A rejection of the null hypothesis indicates that the model is identified.

5.1 Mincerian estimation bias

According to Ciccone and Pieri (2006), the bias of the Mincerian approach when skilled and unskilled workers are substitutable, can be estimated as:

$$Bias = \frac{1}{\sigma}(\frac{w_H - w_L}{w}),$$

where σ is the elasticity of substitution between more (skilled, H) and less (unskilled, L) educated workers and the term in brackets is the wage premium of skilled workers. If we use Hidalgo, O'Kean and Rodriguez (2008) average estimate for Spain of the elasticity of substitution between college-educated workers and the rest of workers of 1.6 and assume a wage premium between 1988 and 2007 of approximately 40% as estimated by Hidalgo (2010), the bias of the Mincerian estimation is approximately 25% (0.4/1.6). We will test if our results are significantly different from this prediction.

6 Data and Instruments

For the first step estimation we need individual information on wages and workers characteristics We also need province information for schooling levels and some additional controls for the second step estimation.

6.1 Individual Data

Our main individual data source for FSE is the Continuous Sample of Working Records from waves 2006 to 2010 (CSWR, Muestra Continua de Vidas Laborales), a yearly sample of working histories and benefits from the Spanish Social Security records. We restrict the sample period to 1995-2010 due to the fact that we are using working histories so the data from previous periods is less accurate and we prefer to use the last year for which we have data from the five waves.

We will not use yearly differences, as externalities are less likely to be detected in the short term. We will use instead years 1995, 2000, 2005 and

2010, taking into account the change in periods of five years. Nevertheless for some schooling levels such as college educated workers there are sample size problems for some provinces, which can be overcome by pooling two years in each of the periods considered, and so finally we use the data in 1995-1996, 2000-2001, 2005-2006 and 2009-2010 to estimate average provincial wages in the first stage.[5]

The CSWR data set is composed by three basic files: affiliation, contribution and benefit files. Each record in the affiliation file contains detailed information from each of the different relationships between the individual and the Social Security. Each of these relationships includes information about the starting and ending dates of the affiliation spell and a number of characteristics of the job, including some firm and personal characteristics (education, age, gender, type of contract, province, sector of activity and much more). Thus, for each person, we have as many records as changes he/she has had with the Social Security from his/her initial register.

Obviously, only those relationships originating a salary are interesting for our purposes, given that this means that the worker is paying the corresponding contributions and, thus, we know the corresponding monthly contribution narrowly related with the wages. Therefore, the unemployed and pensioners in all forms are not included in the sample. Of the remaining workers, we exclude self-employed, since the information about contributions is unrelated with earnings for them, and all those who do not belong to the General Regime of the Social Security (domestic helpers and other special regimes are excluded from the analysis). From those contributing to the General Regime we restrict the analysis to those in groups of contribution 1 to 10. With regard to the employment relationship, some special cases, as those contributors who have some peculiarities that make them not being registered or those with a learning contract, have been eliminated. Workers hired through temporary employment agencies were also eliminated. Finally, those workers with missing information that might be relevant for the subsequent analysis have also been eliminated.

The information contained in the working records provided by the CSWR is very detailed, providing monthly information and all changes in labor contracts that could happen within a year. Since we do not need such level of detail for our estimation purposes, we just take into account the record of October of each year. This allows us to compare the earnings information extracted from the CSWR with other data sources and therefore to check

[5]Notice that because of data availability the last period change will be computed over a one year shorter period.

its validity. After applying this filter, we end up with 6,519,158 observations for the eight years that we use.

To correct for the double censoring problem mentioned above we use the procedure proposed by Boldrin et al (2004) and predict wages for the records that are right censored (see appendix B for a description). As them, we consider that censoring from below is too noisy (because of part time jobs and other incidences) and decide not to treat them at all.

The second problem with the CSWR is how to identify correctly education for the first stage estimation. The CSWR has an education variable that comes from a match of individuals with municipal information (the Spanish *Padrón*). For each worker this information is matched only once, and therefore it is possible that the actual education attainment for some workers is larger than the one recorded in the CSWR. This is the more likely the higher the education level of the workers. Given that, we generate three subsamples in terms of education level (primary, secondary and college).

In table 1 we show the percentage of workers with primary, secondary and college education comparing the CSWR and the Active Population Survey (EPA *Encuesta de Población Activa*) for 2000. It can be seen that the CSWR underreports workers with college education and overreports workers with primary education. If the sample selection of workers according to their education is not random, the estimation in the FSE both of the CC and Mincerian approach will be biased[6].

[table 1 about here]

To correct for this possible bias we also use only two education levels: obligatory and non-obligatory. In the first group workers with primary and non-obligatory secondary education are included, while in the second group workers with obligatory secondary and college education are included. In the first group we are sure to capture all workers with primary education and non-obligatory secondary education, and the ones that report primary education but actually have completed non-obligatory secondary education. For the ones who actually have college education, we use occupation information in the CSWR, since occupation groups (*grupos de cotización*) 1 and 2 require college degrees. We therefore define workers with college level to those either being reported in the CSWR or having occupation levels 1 or 2 in the CSWR. In table 2 we compare now the EPA with this new variable and as it can be seen the differences are much smaller. We will call this new education variable the Adjusted Level of Education (ALE hereafter).

[table 2 about here]

[6]The weights used in the CC approach are not affected as they are based on the EPA.

Finally to correct for possible problems of attrition of working records for those workers with jobs between 2006 and 2010 we repeat the analysis for two different age cohorts, 25-55 and 25-65 years old.

To estimate the individual wage equations for the FSE's we define *general experience, firm tenure, gender, industrial sector and type of contract* as controls. The experience-tenure variables are constructed using only the effective time the workers have had a labor relationship. Thus, general experience is not potential experience. So general experience is the time a worker has had a job since his first real register, while firm tenure is defined as the duration of the current spell within the same firm[7]. As economic sector we use a broad classification in agriculture, manufacturing, construction and services. We also distinguish between fixed-term and long-term contracts.

6.2 Province Data

We use the following province data to implement the externalities estimation in our second step:

- *Province schooling level*: average years of schooling (Source: Human Capital Project Data).

  Controls: We use four different controls in the first step estimation:

  - *Total workers:* Used to measure the size of provinces (Source: Labor Active Survey (INE).).
  - *Physical capital stock over GDP:* Used as a robustness check (Source: IVIE Database Data).
  - *Province Consumer Index Price:* Used as robustness check for hypothetical differences on inflation that may influence in the externalities estimation (Source: Spanish National Institute of Statistic (INE)).

6.3 Instruments

As it was stated earlier, as instruments we will use provincial demographic structure with a 5-year lag, measured as the weight of young and old workers (defined as two age groups), as well as a variable showing the geographic position of the province. To construct this variable we use municipal (*Padrón*) information provided by the National Statistic Institute.

[7]Other variables included in the CSWR 2008 may be a priori interesting (as sector and firm size) but the number of missing values is so large that we do not include them.

6.4 Descriptive statistics

Table 3 shows descriptive statistics of our sample data. In figure (1) we can see the evolution in wages comparing our sample with official statistics in Spain, Quartely Survey of Labor Costs and EU-Klems. The patterns are very similar. Table 4 compares average wages for different years, and as it can be seen the level is smaller for our sample, since our data are based on Social Security contributions which are upper censored. But since we are using wage differences what matters for our purposes is the change in wages and their level, and according to figure 1 changes are well measured with our sample.

[figure 1 about here]

[table 3 about here]

[table 4 about here]

Table 5 show wage differences with respect to gender and education. Gender differences are smaller than the ones found in other studies[8]. This is another consequence of using Social Security contributions and not actual wages. The same can be said with respect to the education premium, which is lower using the CSWR than in other sources, but this could also be related to the fall in the wage premium during the last two decades[9].

[table 5 about here]

In table 6 average salaries are shown for the 10 provinces with largest average salaries and for the 10 provinces with the lowest average salaries. The three Basque provinces are in the first group while Castilla la Mancha, Canarias and Extremadura are in the second group. As it can be seen, the ranking is very stable for our period.

[table 6 about here]

7 Estimation and Results

In this section we present our main estimation results.

7.1 Exclusion restrictions

We regress the log change of our province schooling level (Δh_{pt}), change on province's average years of schooling, on the weights for young and old population cohort shares five years before. As the young age group we use the 10-19 age group, while as the old group we use the 50 to 59 age group. We also

[8] See among others Hernández and Méndez (2005), De la Rica and Ugidos (1995), Simón (2006), Gardeazabal and Ugidos (2005), and Amuedo-Dorantes and De la Rica (2005).

[9] As shown in Felgueroso, Hidalgo, and Jiménez-Martín (2010)

include geographical fixed effects, in order to capture similar socioeconomic structures that could imply similar education decisions[10]. The results in table 7 show that the instruments work well for our purposes. While the only significant coefficient is the one associated with the "young" age group, the F statistic shows that the joint distribution is far from being significant. Also, the F statistic associated with the excluded instruments shows that they are appropriate to explain the evolution of our endogenous variable. Finally, the R^2 reaches a value of around 36%. We will show endogeneity and overidentification tests in the Mincerian and Constant Composition approach estimations results.

[table 7 about here]

After getting reassurance about the validity of our instruments, we can now tackle the rest of the estimations.

7.2 The Constant Composition approach

First, we aggregate the province-year and schooling level fixed effects to obtain province-year average wages holding the labor-force composition constant. Specifically, we obtain $\log(\hat{w}^F_{pt}) = \log\left(\sum_{s=1}^{s} l_{spT}\alpha_{spt}\right)$ where l_{spt} is the average of the shares for the 2000 period coming from the IVIE Human Capital series. The results of the SSE of the Constant Composition approach are reported in tables 8 and 9. Columns (1)-(4) refer to the GLS estimations, while columns (2)(3)(5) and (6) refer to the GMM estimation with the instruments (GMM1 with demographic and geographic instruments while GGM2 only with demographic instruments). Columns (1) to (3) are estimations for the 25-65 years sample and columns (4) and (6) for the 25-55 years sample. Finally table 8 presents results with our first education variable, while in table 9 we use our ALE variable for the calculus of the province average wages.

[table 8 about here]

[table 9 about here]

Our results show significant schooling externalities, which are robust to the different education variables. For the GMM estimation, the strength of the estimated externalities are between 3.2% to 8.9% depending on the education variable, the age sample and instruments used. All the GMM coefficients are significant. At the bottom of the tables we present the tests for endogeneity and overidentifying restriction, finding that all tests show the validity of our instruments. Also, we identify an increase in coefficients

[10] We take into account six geographical areas, namely Eastern and Western Cantábrico, North-Centre, South-Center, South and Eastern.

once we instrument. This implies that generally the measurement error bias is more important than the bias generated by endogeneity. A possible explanation for this is that migration between Spanish provinces has not been important for most part of the period analyzed, and therefore the bias introduced may be negligible, but measuring human capital using education levels may introduce important measurement errors, more than compensating the bias introduced by the migration of skilled workers.

Comparing our results with previous results obtained for Spain, for instance by De La Fuente and Domenech (2005) and Alcalá and Hernandez (2005), our results are similar to what they found for the 25-65 sample. However, we find higher externalities with our 25-55 sample. The reason behind this results may be that the productivity improvement based on human capital externalities is not homogeneous across worker age cohorts, being stronger for those groups who potentially can benefit more from wage increases. We also find stronger externalities when we use our adjusted education variable. This may be due to a lower accuracy in the estimation of average provincial wages when we use only two education categories.

7.3 The Mincerian approach

Once we run the SSE, we use the province-year specific fixed effects in (10) to obtain the change in "cleaned" average province wages necessary to implement the SSE of the Mincerian approach. The results are in tables 9 and 11. Again columns (1)-(4) refer to the GLS estimations, while columns (2)(3)(5) and (6) refer to the GMM estimation with the instruments. Columns (1) to (3) are estimations for the 25-65 years sample and columns (4) to (6) for the 25-55 years sample.. Table 10 presents results with our first education variable, while table 11 for ALE variable.

[table 10 about here]

[table 11 about here]

The strength of the externalities using the Mincerian approach is clearly positive and higher than those found using the CC approach, with estimates statistically significant at the 1% and 10% levels. Again, for these estimations we test the endogeneity of our proxy variables and an over-identifying restriction. We find again that none of the usual hypothesis can be rejected at the usual significance levels. Hence, the Mincerian approach yields statistically significant externalities. Again, we identify an increase in coefficients once we instrument.

Nevertheless we find a larger difference between the Mincerian and CC approach than expected. Our expected bias, according to our previous dis-

cussion, should be around 25% but we find a higher difference of 57% especially for the estimations of the 25 to 55 age group using our original education variable.

7.4 Comparison of results: significance and bias

In this section we analyze if the difference in estimation between the two approaches is statistically significant. In order to perform a test of the significance or the bias of the Mincerian approach we use bootstrapping to estimate the covariance between the two estimations.

With a small sample of 100 registers we replicate our data thousand times using a structural model of the Mincerian and Constant Composition approach. We estimate the average, standard deviation and covariance of the estimates. We then proceed to perform a test on the Mincerian approach estimate being larger than the Constant Composition approach estimate. The results are shown in table 12. As it can be observed in the table we find a positive bias for the estimations using the MA approach as compared to the CC approach. The difference in estimation between the two approaches is only significant for the 25-65 age group. For the "adjusted education level" human capital variable, the difference is not significant for either group, but close to the 10% significance for 25-65 age group.

8 Conclusions

The strength of human capital externalities in Spain is important for growth accounting and from a public-policy perspective. We have applied two different approaches to quantify the wedge between the social and the private return to schooling at the province level for the 1995-2010 period. Our results yield evidence of significant schooling externalities. The two approaches used yield different point estimates, being the estimates using the Mincerian approach larger than using the constant composition approach although bootstrapping methods leaves little evidence about its significance. However, our estimate of the externality of human capital is significantly different from the theoretical prediction arising from previous work by Ciccone and Peri (2006).

Future research could combine our estimates with the Spanish tax system and education subsidies to examine whether the incentives to human capital accumulation are consistent with the social returns implied.

References

Acemoglu, D., and J. Angrist (2001): "How Large are the Social Returns to Education: Evidence from Compulsory Schooling Laws," in *NBER Macroeconomic Annual 2000*, ed. by B. Bernanke, and K. Rogoff, pp. 9–59. Cambridge, MA: The MIT Press.

Alba, R., and M. S. Segundo (1995): "The Return to Education in Spain," *Economics of Education Review*, 14(2).

Alcalá, F., and P. J. Hernández (2005): "Las Externalidades del Capital Humano en las Empresas Españolas," Revista de Economía Aplicada (forthcoming).

Amuedo-Dorantes, C., and S. De la Rica (2005): "The impact of gender segregation on male-female wage differentials: Evidence from matched employer-employee data for Spain," Iza discussion papers 1742, Institute for the Study of Labor (IZA).

Angrist, J. (1995): "The Economic Returns to Schooling in the West Bank and Gaza Strip," *American Economic Review*, 85, 1065–1087.

Arrazola, M., and J. de Hevia (2008): "Three measures of returns to education: An illustration for the case of Spain," *Economics of Education Review*, 27(3), 266–275.

Arrazola, M., J. De Hevia, M. Risueno, and J. F. Sanz (2003): "Returns to education in Spain: Some evidence on the endogeneity of schooling," *Education Economics*, 11(3), 293–304.

Barceinas, F., J. Oliver, J. Raymond, and J. Roig (2000): "Los Rendimientos de la Educación en España," *Papeles de Economía Española*, 86.

Black, D., and V. Henderson (1999): "A Theory of Urban Growth," *Journal of Political Economy*, 107(2), 252–284.

Boldrin, M., S. Jiménez-Martín, and F. Peracchi (2004): *Micro-Modeling of Retirement Behavior in Spain*pp. 499–578. University of Chicago Press.

Card, D. (1999): "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1801–1864. Elsevier, Amsterdam.

——— (2003): "Estimating the return to schooling: progresss on some persistent econometric problems," *Econometrica*, 69, 1127–1160.

Ciccone, A., and G. Peri (2006): "Identifying Human-Capital Externalities: Theory with Applications," *Review of Economic Studies*, 73(2), 381–412.

Conley, T., F. Flier, and G. Tsiang (2003): "Spillovers from local market human capital and spatial distribution of productivity in Malysia," *Advances in Economic Analysis & Policy*, 3(1), Article 5. http://www.bepress.com/bejeap/advances/vol3/iss1/art5.

De la Fuente, A. (2003): "Human Capital in a Global Knowledge-based Economy, part II: assesment at the EU Country Level," European Commission Report, Employment and Social Affairs.

de la Fuente, A., and R. Domenech (2006): "Human capital in Growth Regressions: How much difference does data quality make?," *Journal of the European Economic Association*, 4(1), 1–36.

De la Fuente, A., R. Domenech, and J. F. Jimeno (2003): "Human Capital as a Factor of Growth and Employment at the Regional Level. The Case of Spain," UFAE and IAE Working Papers 610.04.

De la Rica, S., and A. Ugidos (1995): "¿Son las diferencias en capital humano determinantes de las diferencias salariales observadas entre hombrs y mujeres?," *Investigaciones Económicas*, 19, 395–414.

Felgueroso, F., M. Hidalgo, and S. Jiménez-Martín (2010): "Explaining the fall of the skill wage premium in Spain," *DOCUMENTO DE TRABAJO*, 2010, 19.

García-Fontes, W., and M. A. Hidalgo (2008): "¿Es Posible Estimar las Externalidades del Capital Humano? Evidencia para las Regiones Españolas," *Temas actuales de economía*, 2, 49–71, Instituto de Análisis Económico y Empresarial de Andalucía.

García-Pérez, J. (2008): "La muestra continua de vidas laborales: una guía de uso para el análisis de transiciones," *Revista de Economía Aplicada*, 16(1), 5–28.

Gardeazabal, J., and A. Ugidos (2005): "Gender wage discrimination at quantiles," *Journal of Popultion Economics*, 18, 165–179.

Hernández, P. J., and I. Méndez (2005): "La corrección del sesgo de selección en los análisis de corte transversal de discriminación salarial por sexo: estudio comparativo en los países de la Unión Europea," *Estadística Española*, 47, 179–214.

Hidalgo, M. A. (2010): "A Demand and Supply Analysis of the Spanish Education Wage Premium," *Revista de Economía Aplicada*, XVIII.

Katz, L. F., and K. M. Murphy (1992): "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *The Quarterly Journal of Economics*, 107(1), 35–78.

Lucas, R. J. (1988): "On the Mechanics of Economic Development," *Journal of Monetary Economics*, 22(1), 3–42.

Mincer, J. (1974): "Scholling, Experience and Earnings," Discussion paper, National Bureau of Research, New York.

Moretti, E. (2004a): "Estimating the Social Return to Higher Education: Evidence from Longitudinal and Repeated Cross-sectional Data," *Journal of Econometrics*, 121(1-2), 175–212.

——— (2004b): "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions," *American Economic Review*, 94(3), 656–690.

Oliver, J., J. Raymond, J. Roig, and F. Barceinas (1999): "Returns to Human Capital in Spain: A Survey of the Evidence," in *Returns to human capital in Europe: A literature review*, chap. 13, pp. 280–298. R. Asplund and P.Pereira, Helsinki.

Rauch, J. (1993): "Productivity Gains from Geographic Concentration of Human Capital: Evidence from the Cities," *Journal of Urban Economics*, 34(3), 380–400.

Raymond, J. (2002): "Convergencia Real de las Regiones Españolas y Capital Humano," *Papeles de Economía Española*, 93, 109–121.

Rudd, J. (2000): "Empirical Evidence on Human Capital Spillovers," Finance and Economics Discussion Series 2000-46, Board of Governors of the Federal Reserve System (U.S.).

Simón, H. (2006): "Diferencias salariales entre hombres y mujeres en España: Una comparación internacional con datos emparejados empresa-trabajador," *Investigaciones Económicas*, 30(1), 55–87.

Vila, L., and J. Mora (1998): "Changing Returns to Education in Spain During the 1980s," *Economics of Education Review*, 17(2), 173–178.

Appendix A. Proof of Constant Composition Approach validity to estimate externalities

See that $w_{pt}^L = B_{pt}x_{pt}^{\theta}N_{pt}^{\delta}K_{pt}^{\gamma}F_1(l_{pt}(x),(1-l_{pt}(x)))$ and consequently

$$\begin{aligned}\frac{\partial w_{pt}^L(x_{pt})}{\partial x_{pt}} &= \theta B_{pt}x_{pt}^{\theta-1}N_{pt}^{\delta}K_{pt}^{\gamma}F_1(l_{pt}(x),(1-l_{pt}(x)))+ \\ &+B_{pt}x_{pt}^{\theta}N_{pt}^{\delta}K_{pt}^{\gamma}(F_{11}(l_{pt}(x),(1-l_{pt}(x))- \\ &-F_{12}(l_{pt}(x),(1-l_{pt}(x)))l'(x_{pt}).\end{aligned} \tag{16}$$

Similarly:

$$\begin{aligned}\frac{\partial w_{pt}^H(x_{pt})}{\partial x_{pt}} &= \theta B_{pt}x_{pt}^{\theta-1}N_{pt}^{\delta}K_{pt}^{\gamma}F_2(l_{pt}(x),(1-l_{pt}(x)))+ \\ &+B_{pt}x_{pt}^{\theta}N_{pt}^{\delta}K_{pt}^{\gamma}(F_{21}(l_{pt}(x),(1-l_{pt}(x))- \\ &-F_{22}(l_{pt}(x),(1-l_{pt}(x)))l'(x_{pt}).\end{aligned} \tag{17}$$

Constant returns to scale of high and low skill workers in the aggregate production function for a given technological level imply that $F_1 l + F_2 h = F$. Furthermore $F_{11}l + F_{12}h = 0$ and $F_{21}l + F_{22}h = 0$. Combining these two last equations we obtain $(F_{11} - F_{12})l + (F_{12} - F_{22})h = 0$. Consequently the weighted average of (16) y (17) with weights equal to the fraction of workers of each type is

$$\begin{aligned}&\frac{\partial w_{pt}^L(x_{pt})}{\partial x_{pt}}l_{pt} + \frac{\partial w_{pt}^H(x_{pt})}{\partial x_{pt}}(1-l_{pt}) = \\ &= \theta B_{pt}x_{pt}^{\theta-1}N_{pt}^{\delta}K_{pt}^{\gamma}(F_1(l_{pt}(x),(1-l_{pt}(x))l(x_{pt})+ \\ &+F_2(l_{pt}(x),(1-l_{pt}(x))(1-l(x_{pt}))) = \\ &= \theta B_{pt}x_{pt}^{\theta-1}N_{pt}^{\delta}K_{pt}^{\gamma}F(l_{pt}(x),(1-l_{pt}(x)).\end{aligned} \tag{18}$$

Consequently:

$$\begin{aligned}&\frac{\partial \ln(w_{pt}^L(x_{pt})l_{pt} + w_{pt}^H(x_{pt})(1-l_{pt}))}{\partial x_{pt}}x_{pt} = \\ &= \frac{\theta B_{pt}x_{pt}^{\theta-1}N_{pt}^{\delta}K_{pt}^{\gamma}F(l_{pt}(x),(1-l_{pt}(x))}{B_{pt}x_{pt}^{\theta}N_{pt}^{\delta}K_{pt}^{\gamma}F(l_{pt}(x),(1-l_{pt}(x))}x_{pt} = \theta.\end{aligned} \tag{19}$$

Appendix B. Adjusted wages

The salary information is inferred from the monthly contributions to Social Security. The main problem is that the highest wages are censored because of the existence of a maximum for contributions. There is also a minimum for contributions, but this it not as problematic as there are also minimum legal wages.

Although the percentage of records who have their contributions censored is not too high (10%), it may bias the estimates. We correct for this by transferring the distribution structure of wages near to the censoring point. The methodology is based on the estimation of a Tobit model, where the log of the wage of a worker belonging to contribution group g is expressed as:

$$\begin{aligned} w_{ig} &= l_g \qquad si \quad x_{ig}\beta_g + \varepsilon_{ig} \leq l_g \\ w_{ig} &= u_g \qquad si \quad x_{ig}\beta_g + \varepsilon_{ig} \geq u_g \\ w_{ig} &= x_{ig}\beta_g + \varepsilon_{ig} \leq l_g \\ & \quad contrary \quad to \quad the \quad previous \quad cases \end{aligned} \tag{20}$$

where l_g and u_g are lower and upper limits of the contribution base for contribution group g; x_{ig} is a group of characteristics associated with worker i, β_g is the return to each of the above characteristics and ε_{ig} is the error term.

The idea is to estimate model (20) using a double-censored Tobit. Once the model has been estimated we simulate the wage and contribution for the censored workers.

If s_g is the standard error of the original wage series w_g, and defining $\widehat{u}_g$ as the estimated standard error and adjusted such that:

$$\widehat{u}_g = \frac{u_g - \widehat{w}_g}{s_g}$$

so that $\widehat{u}_g \sim n(0,1)$, we re-estimate wages for the censored workers by the expression:

$$\widehat{w}_{ig} = \widehat{w}_g + s_g \frac{\phi(\widehat{u}_g)}{1-\phi(\widehat{u}_g)} + s_g \phi^{-1}\left(\phi(\widehat{u}_g) + \theta_i(1-\phi(\widehat{u}_g))\right)$$

where $\theta_i \sim U(0,1)$ and ϕ is the normal density function with mean equal to 0 and variance equal to 1 and ϕ^{-1} is its inverse. That is to say, given a probability value a, $\phi^{-1}(a)$ gives us a value in $\mathbb{R}$. The second term on the right $\widehat{w}_g$ corrects the estimate of the bias introduced by censorship. The

third term on the right introduces randomness to the individual i and that is a function of the distribution of estimated errors with the information available for non-censored individuals.

Using this procedure we correct the salaries of those workers whose contribution base is equal to the legal limits. This method ensures the maintenance of the wage structure for the majority of workers.

Table 1: Percentage of workers by education level, 25-65 year old

	Primary		Secondary		College	
	CSRW	APS	CSRW	APS	CSRW	APS
1995	43.2	38.5	47.8	45.1	9.1	16.4
2000	41.8	31.3	48.3	49.9	9.9	18.9
2005	43.1	20.0	46.9	55.2	10.1	24.8
2010	43.3	14.4	46.8	55.8	9.8	29.8

Notes: Data come from the Continuos Sample of Working Lives (CSWR) and the Active Population Survey (APS).

Table 2: Percentage workers by education level, 25-65 years old (ALE definition)

	Compulsory		Non-Compulsory	
	CSRW	APS	CSRW	APS
1995	77,1	83,6	22,9	16,4
2000	76,4	81,2	23,6	18,9
2005	76,3	75,2	23,7	24,8
2010	74,4	70,2	25,6	29,8

Notes: ALE stands for adjusted level of education. Compulsory includes below college degree (primary and secondary) and low vocational education. Data come from the Continuos Sample of Working Lives (CSRW) survey and the Active Population Survey (APS).

Table 3: Sample Data Main Descriptive Statistics

	age	experience	tenure	female	% college	% non-compulsory ALE def
25-65 years old sample						
1995	38.3	7.3	5.2	32.7	9.1	22.9
2000	39.7	8.7	4.7	36.4	9.9	23.6
2005	40.9	10.1	4.4	40.2	10.1	23.7
2010	41.8	13.1	6.0	44.5	9.8	25.6
25-55 years old sample						
1995	38.2	7.3	5.2	32.8	9.5	20.0
2000	38.9	8.5	4.6	37.0	10.5	21.8
2005	39.1	9.5	4.1	41.2	10.6	22.1
2010	39.6	12.3	5.5	45.5	10.2	23.5

Notes: Experience is potencial experience obtained as age minus years of schooling minus six. Tenure is obtained directly from the survey information. College def 1 is derived directly from the CSRW information. College def 2 is College def 1 plus those workers who has not college degree in the CSRW but is occupied in occupations where college education is required.

Table 4: Average Wages per Year and Surveys (euros)

	CSRW		SWS	ASLC	QSLC(*)
	25-65	25-55			
1995	1096.9	1096.8	1235.9	-	-
2000	1260.6	1257.9	-	-	1334.8
2002	-	-	1414.4	1433.4	1431.5
2005	1453.7	1450.9	-	1574.3	1571.8
2006	-	-	1400.6	1620.4	1646.8
2010	1621.0	1627.7	1627.9	1878.3	1874.8

Notes: Data come from Continuous Sample of Working Lives (CSRW) survey, Structural Wage Survey (SWS), ASLC stands for Annual Survey of Labor Costs (ASLC) and Quartely Survey of Labor Costs (QSLC, Total Wage Cost).

Table 5: Wages per Level of Education and Gender (euros) and Wage Premia

	1995	2000	2005	2010
Level of education				
Primary	884.6	1018.2	1191.5	1334.3
Secondary	1145.1	1316.9	1545.1	1756.0
College	1462.8	1648.3	1862.1	2136.7
College premia				
College-primary wage premia	0.65	0.62	0.56	0.60
College-secondary wage premia	0.28	0.25	0.21	0.22
ALE				
Compulsory	932.5	1084.2	1268.8	1414.7
Non-Compulsory	1572.2	1753.3	2000.6	2274.9
Non-comp. premia				
Non-comp. wage premia	0.69	0.62	0.58	0.61
Gender				
Male	1091.3	1281.5	1505.9	1711.7
Female	989.0	1116.3	1289.0	1461.9
Gender Gap	0.10	0.15	0.17	0.17

Notes: ALE stands for adjusted level of education.

Table 6: Provinces with higher and lower wages (euros) in 2000

rank	Province	Average wage
1	Alava	1725.9
2	Gipuzcoa	1724.1
3	Bizcaia	1669.3
4	Navarra	1615.6
5	Madrid	1584.2
46	SC Tenerife	1245.3
47	Cáceres	1238.4
48	Ourense	1209.8
49	Lugo	1196.9
50	Cuenca	1180.0

Table 7: First Stage Regressions

	Educ. definition 1		Adjusted level of Educ.	
Log of Employment	0.619	0.401	0.772	0.554
	(0.585)	(0.544)	(0.583)	(0.538)
CPI	-0.085	-0.091*	-0.081	-0.088*
	(0.054)	(0.052)	(0.054)	(0.051)
Capital/GDP	0.563	0.664**	0.535	0.640*
	(0.352)	(0.334)	(0.356)	(0.339)
young	17.369**	17.448**	18.494**	18.545**
	(6.377)	(6.392)	(6.067)	(6.046)
old	-8.262	-7.766	-5.643	-5.463
	(15.771)	(15.193)	(15.116)	(14.349)
North	-0.154		-0.155	
	(0.115)		(0.115)	
North-Center	-0.116		-0.112	
	(0.129)		(0.129)	
South	-0.163		-0.161	
	(0.111)		(0.112)	
South-Center	-0.110		-0.095	
	(0.117)		(0.117)	
East	-0.146		-0.148	
	(0.104)		(0.104)	
Constant	0.019	-0.093*	0.003	-0.108**
	(0.088)	(0.052)	(0.089)	(0.052)
N. of cases	150	150	150	150
R2	0.357	0.345	0.360	0.347
F	8.232	16.172	8.074	16.233
Prob>F	0.000	0.000	0.000	0.000

Notes: Controls include the Consumer Price Index (CPI) for each province, Capital/GDP ratio, percentage in the 10-19 age group (Young), percentage in the 55-64 age group (Old), dummies for North (Basque country and Cantabria), North-Center (Castilla-Leon), South (Andalusia, Murcia, and Canary Islands) South-Center (Extremadura and Castilla-La Mancha), and East (Catalonia, Valencia, and Balearic-Islands). *, ** and *** significant at the 1%, 5% and 10% significance level.

Table 8: Constant composition approach estimation.

Dep. variable: growth of log province constant composition adjusted wages $\log(\hat{w}^F_{p,t}) - \log(\hat{w}^F_{p,t-1})$						
		25-65			25-55	
	OLS	GMM 1	GMM 2	OLS	GMM 1	GMM 2
H. Cap. measure	0.006	0.048***	0.048**	0.003	0.032**	0.034**
	(0.004)	(0.014)	(0.015)	(0.004)	(0.012)	(0.014)
Log of Employment	0.087***	0.128**	0.109**	0.094***	0.114***	0.101**
	(0.024)	(0.040)	(0.042)	(0.023)	(0.034)	(0.036)
CPI	0.001	0.003	0.003	0.000	0.002	0.002
	(0.003)	(0.004)	(0.005)	(0.003)	(0.004)	(0.004)
Capital over PIB	0.047**	0.075**	0.083***	0.057***	0.082***	0.089***
	(0.015)	(0.024)	(0.024)	(0.015)	(0.021)	(0.021)
Constant	-0.027***	-0.031***	-0.031***	-0.027***	-0.030***	-0.030***
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)
N. of cases	150	150	150	150	150	150
Hansen J St.		0.438	0.931		0.211	0.702
Kleibergen-Paap Under. test		0.003	0.000		0.003	0.000

Notes: Estimation methods include Ordinary Least Squares (OLS), General Method of Moments with population and geographic instruments (GMM 1), and General Method of Moments with population as instruments (GMM 2). The human capital is the growth rate of province average workers years of schooling. All estimates are using province fixed effects. Hansen's J statistic is computed for the test of overidentifying restrictions, which under the null hypothesis is distributed as chi-squared in the number of overidentifying restrictions. The Kleibergen-Paap test computes the under and weak identification test; under the null of underidentification, the statistic is distributed as chi-squared with degrees of freedom equal to $L - k + 1$, where L is the number of instruments (included+excluded). CPI is the Consumer Price Index for each province. Capital over GDP is the ratio between Physical Productive Capital and total GDP. Estimations for the 25-55 and 25-65 age groups. Standard errors in parentheses. *, ** and *** significant at the 1%, 5% and 10% significance level.

Table 9: Constant composition approach estimation. Adjusted Level of Education (ALE)

Dep. variable: growth of log province constant composition adjusted wages $\log(\hat{w}^F_{p,t}) - \log(\hat{w}^F_{p,t-1})$						
		25-65			25-55	
	OLS	GMM 1	GMM 2	OLS	GMM 1	GMM 2
H. Cap. measure	0.008*	0.064***	0.064***	0.024***	0.088***	0.089***
	(0.005)	(0.016)	(0.018)	(0.005)	(0.019)	(0.021)
Log of Employment	0.141***	0.196***	0.182***	0.175***	0.247***	0.231***
	(0.027)	(0.046)	(0.049)	(0.034)	(0.061)	(0.064)
CPI	0.004	0.004	0.004	0.007	0.005	0.005
	(0.004)	(0.005)	(0.005)	(0.004)	(0.006)	(0.006)
Capital over PIB	0.060***	0.077**	0.083**	0.070***	0.095**	0.100**
	(0.016)	(0.028)	(0.029)	(0.018)	(0.034)	(0.035)
Constant	-0.047***	-0.049***	-0.050***	-0.042***	-0.045***	-0.045***
	(0.003)	(0.004)	(0.005)	(0.004)	(0.005)	(0.006)
N. of cases	150	150	150	150	150	150
Hansen J St.		0.563	0.775		0.776	0.669
Kleibergen-Paap Under. test		0.003	0.000		0.003	0.000

Notes: Estimation methods include Ordinary Least Squares (OLS), General Method of Moments with population and geographic instruments (GMM 1), and General Method of Moments with population as instruments (GMM 2). The human capital is the growth rate of province average workers years of schooling. All estimates are using province fixed effects. Hansen's J statistic is computed for the test of overidentifying restrictions, which under the null hypothesis is distributed as chi-squared in the number of overidentifying restrictions. The Kleibergen-Paap test computes the under and weak identification test; under the null of underidentification, the statistic is distributed as chi-squared with degrees of freedom equal to $L - k + 1$, where L is the number of instruments (included+excluded). CPI is the Consumer Price Index for each province. Capital over GDP is the ratio between Physical Productive Capital and total GDP. Estimations for the 25-55 and 25-65 age groups. Standard errors in parentheses. *, ** and *** significant at the 1%, 5% and 10% significance level.

Table 10: Mincerian Approach estimation.

Dep. variable: growth of log province adjusted wages $\Delta\hat{\alpha}_{pt}$						
		25-65			25-55	
	OLS	GMM 1	GMM 2	OLS	GMM 1	GMM 2
H. Cap. measure	0.035**	0.130***	0.126**	0.013	0.056*	0.055*
	(0.016)	(0.036)	(0.038)	(0.013)	(0.032)	(0.033)
Log of Employment	-0.005	0.037	0.039	0.060	0.075	0.077
	(0.079)	(0.111)	(0.116)	(0.059)	(0.095)	(0.100)
CPI	0.028**	0.022*	0.017	0.020**	0.022*	0.018
	(0.011)	(0.013)	(0.014)	(0.009)	(0.012)	(0.012)
Capital over PIB	0.239***	0.295***	0.286***	0.260***	0.303***	0.276***
	(0.043)	(0.071)	(0.073)	(0.038)	(0.058)	(0.061)
Constant	-0.015	-0.013	-0.011	-0.034***	-0.037**	-0.032**
	(0.009)	(0.015)	(0.016)	(0.008)	(0.015)	(0.015)
N. of cases	150	150	150	150	150	150
Hansen J St.		0.810	0.994		0.625	0.831
Kleibergen-Paap Under. test		0.003	0.000		0.003	0.000

Notes: Estimation methods include Ordinary Least Squares (OLS), General Method of Moments with population and geographic instruments (GMM 1), and General Method of Moments with population as instruments (GMM 2). The human capital is the growth rate of province average workers years of schooling. All estimates are using province fixed effects. Hansen's J statistic is computed for the test of overidentifying restrictions, which under the null hypothesis is distributed as chi-squared in the number of overidentifying restrictions. The Kleibergen-Paap test computes the under and weak identification test; under the null of underidentification, the statistic is distributed as chi-squared with degrees of freedom equal to $L - k + 1$, where L is the number of instruments (included+excluded). CPI is the Consumer Price Index for each province. Capital over GDP is the ratio between Physical Productive Capital and total GDP. Estimations for the 25-55 and 25-65 age groups. Standard errors in parentheses. *, ** and *** significant at the 1%, 5% and 10% significance level.

Table 11: Mincerian Approach estimation. Adjusted Level of Education (ALE)

Dep. variable: growth of log province adjusted wages $\Delta\hat{\alpha}_{pt}$						
		25-65			25-55	
	OLS	GMM 1	GMM 2	OLS	GMM 1	GMM 2
H. Cap. measure	0.039**	0.130***	0.126**	0.022	0.120***	0.115**
	(0.016)	(0.036)	(0.038)	(0.016)	(0.036)	(0.038)
Log of Employment	0.016	0.037	0.039	-0.027	0.039	0.042
	(0.079)	(0.111)	(0.116)	(0.075)	(0.107)	(0.111)
CPI	0.029**	0.022*	0.017	0.033**	0.024*	0.019
	(0.011)	(0.013)	(0.014)	(0.011)	(0.014)	(0.014)
Capital over PIB	0.245***	0.295***	0.286***	0.265***	0.292***	0.282***
	(0.044)	(0.071)	(0.073)	(0.044)	(0.069)	(0.071)
Constant	-0.019*	-0.013	-0.011	-0.018*	-0.015	-0.014
	(0.010)	(0.015)	(0.016)	(0.009)	(0.016)	(0.016)
N. of cases	150	150	150	150	150	150
Hansen J St.		0.810	0.994		0.794	0.857
Kleibergen-Paap Under. test		0.003	0.000		0.003	0.000

Notes: Estimation methods include Ordinary Least Squares (OLS), General Method of Moments with population and geographic instruments (GMM 1), and General Method of Moments with population as instruments (GMM 2). The human capital is the growth rate of province average workers years of schooling. All estimates are using province fixed effects. Hansen's J statistic is computed for the test of overidentifying restrictions, which under the null hypothesis is distributed as chi-squared in the number of overidentifying restrictions. The Kleibergen-Paap test computes the under and weak identification test; under the null of underidentification, the statistic is distributed as chi-squared with degrees of freedom equal to $L - k + 1$, where L is the number of instruments (included+excluded). CPI is the Consumer Price Index for each province. Capital over GDP is the ratio between Physical Productive Capital and total GDP. Estimations for the 25-55 and 25-65 age groups. Standard errors in parentheses. *, ** and *** significant at the 1%, 5% and 10% significance level.

Table 12: Differences in estimation of externalities (boostrapping)

			Adjusted Level of Education	
Age Groups	Difference in coef. (CCA-MA)	Std. Dev	Difference in coef. (CCA-MA)	Std. Dev
25-65	-0.0802*	0.0498	-0.0704	0.0513
25-55	-0.0283	0.0499	-0.0389	0.0514

Notes: Boostrapping with one thousand iterations and with a random selection of size 125 over the 150 sample data. CCA-MA means the difference between coefficients estimated by the constant composition and Mincerian approaches.