LEMMA Working Paper

n° 2024-04

Schooling Endogeneity and the Rate of Return to Education : A Copula Approach

Peter Dolton

University of Sussex

Aki Skalli

Université Paris-Panthéon-Assas, LEMMA

LEMMA EA 4442 - 4 rue Blaise Desgoffe 75006 Paris

Schooling Endogeneity and the Rate of Return to Education: A Copula Approach.

Peter Dolton* and Ali Skalli**

Abstract

The rate of return to education (RoRtE) is one of the most important policy relevant parameters in economics. Variously, IV identification strategies have been used to overcome the problem of the endogeneity bias associated with years of education measurement. We explore a method to account for this endogeneity through instrument variable (IV) free estimation using Copula theory. We estimate the RoRtE based on the same earnings equation specification using data from five canonical data sets. Our estimate of the rate of return to education is consistently around 6% using these datasets, that have in the past, given rise to estimates for this parameter as high as 13-15%. Exacting robustness tests do not change our results.

JEL Codes: H52, I21, I26, I28.

Keywords: Copulas, Endogeneity, Return to Schooling

Address for Correspondence:
Professor Peter Dolton
Department of Economics
University of Sussex
Brighton, Sussex. BN1 9SL.

Acknowledgements: We wish to thank David Card, Josh Angrist, and Orley Ashenfelter for making their data available to other researchers on their websites and especially to Lennart Hoogerheide, Joern Block, and Roy Thurik for kindly giving us access to their data. We also wish to thank: Guido Imbens, Richard Blundell, Magne Mogstad, Lorraine Deardon, Richard Tol, Tim Barmby and seminar participants at Newcastle, Cardiff, and Sussex Universities and the Royal Economic Society conference at Warwick 2018 and an anonymous referee for insightful comments on an earlier version of this paper.

* University of Sussex and NIESR
** LEMMA, Université Panthéon-Assas (Paris 2)

1. Introduction.

Arguably the rate of return to education (RoRtE) is one of the most important policy relevant parameters in economics. Its estimation has been discussed in seminal papers by Griliches (1977) and Card (2004). Griliches was concerned with overcoming ability bias and Card compared different Instrumental variable (IV) estimation strategies. The RoRtE parameter has variously been estimated, to be as low as zero (e.g., Pischke and von Wachter, 2008) or as high as 15% (e.g., Harmon and Walker, 1995, and Buscha and Dickson, 2012). If the RoRtE is zero then a policy of high level incentives to acquire more education is difficult to justify on financial grounds. In contrast, if the RoRtE is 15% per annum or greater, then education authorities should provide strong incentives for young people to stay at school for longer.

Many other papers have estimated this parameter (e.g., Angrist and Krueger, 1991; Card, 1995; Harmon and Walker, 1995; Acemoglu and Angrist, 2001; Kling, 2001; Oreopoulos, 2006; Devereux and Hart, 2010; Carneiro, Heckman and Vytlacil, 2011) with different data, using alternative estimation methods, in various countries and at distinct points in time. Not surprisingly, there is considerable heterogeneity in the estimates retrieved for this single parameter. While the large number of estimates reflects the attention devoted to this parameter, the heterogeneity of these results is perplexing and makes it difficult to inform education policy.

The essential difficulty with the estimation of the RoRtE is that the acquisition of schooling is potentially endogenous to the determination of earnings. This is because the unobservable component in an earnings equation contains the influence of a person's: drive, energy, ambition, ability, determination, personality and many other variables. These influences will relate to how much schooling a person decides to acquire. Logically the decision to acquire more education – to stay on at school past the compulsory school leaving age, or go to higher education, will partly be influenced by what a person's expectation of their earnings prospects is, with and without, the extra educational investment. The standard way of attempting to overcome the endogeneity bias is to use IV estimation or a Regression Discontinuity Design (RDD). In the last 30 years a considerable emphasis has been placed in this literature on the use of IVs for schooling, for example, by using laws relating to the raising of the school leaving age (RoSLA).

At least four reasons may contribute to these widely differing estimates of the RoRtE. First, returns to education are likely to vary across different types of individuals (see, e.g., Kling 2001, Koop and Tobias 2004, and Carneiro et al. 2011) and may well vary according to whether lifecycle earnings are being modelled from (lifetime) panel data rather than diverse cross

sections of people of different ages (see Bhuller et al 2017). Second, heterogeneity in previous estimates may simply result from the fact that the RoRtE may differ across countries, different labour markets, and at different points in time. For example, Pischke and von Wachter (2008) document a zero return to education in Germany; Grenet (2013) finds a return to education close to zero in France in 1967; Devereux and Hart (2010) report a 6% return to education in Britain using an education reform that took place in 1947 in Britain. Third, the RoRtE may also differ at different margins of education: the return to one more year of school at, for example, age 14-15, may not be the same as the return to one more year of university.

A final reason for the large variation in the size of the estimates is that the econometric identification assumption and the estimation method employed may itself have a bearing on the size of the estimated coefficient. Generally, what is less well-known is the extent to which different estimation techniques applied to different equation specifications may change the size of the estimated coefficients that we find. This issue has been investigated using a single dataset from a single country in a previous paper (Dolton and Sandi, 2017) which examined the sensitivity of OLS, 2SLS estimates of the RoRtE to equation specification and to the definition of the instrumental variable (IV) used.

Since the estimation of the earnings returns to education may involve the well-known problem of endogeneity, much of the RoRtE literature has relied on IV or RDD estimation to retrieve unbiased and consistent treatment effects (e.g., Angrist and Krueger, 1991; Card, 1995; Harmon and Walker, 1995; Acemoglu and Angrist, 2001; Kling, 2001; Carneiro et al., 2011, Oreopoulos, 2006).

Recently there has been a growing scrutiny of IV and RDD methods of estimation. (See Abadie 2018). This suggests that an alternative identification strategy may avoid relying on a particular IV with its attendant interpretation problems (See Carneiro et al, 2011). This is the approach we adopt in this paper as we propose estimating the RoRtE while accounting for the endogeneity bias with an instrument-free strategy, that is based on copula theory.[1][2] More specifically, we use this copula-based approach to estimate the RoRtE in the canonical earnings

[1] Copulas have been used in a variety of contexts. Zimmer and Trivedi (2015) analyse selection and treatment effects in a demand for health care model where copulas help modelling the joint distribution of the error term of the demand for health care equation and the error terms in selection equations. Bonhomme and Robin (2006) as well as Dearden et al (2006) used this approach to model lifetime earnings. However, to our knowledge this is the first attempt to apply the technique to overcome the endogeneity problem.

[2] Other instrument free estimation strategies have been used to tackle this issue, see Park and Gupta (2012), p568. We do not examine these here.

equation specification from five prominent papers, while using their own data. The motivation for this is fourfold.

First, the use of the datasets examined by other authors in prominent papers provide us with the reference point of their results on the RoRtE as a logical comparator for our new suggested estimator. In the USA we examine the US Census data used by Angrist and Krueger (1991), the NLSYM data used by Card (1995) as well as the sample of identical twins used by Ashenfelter and Rouse (1998). In the UK we examine the Family Expenditure Survey (FES) in the Harmon and Walker (1995) paper. In Germany, we use the German Socio-Economic Panel Study (SOEP) used by Hoogerheide, Block, and Roy (2012) (henceforth HBT).

Most specifically, we avoid assuming that there is a satisfactory IV which is highly correlated with years of schooling but uncorrelated with the unobserved heterogeneity, which partially determines earnings. Rather, we examine an alternative assumption that there is a regular joint distribution over any sample of individuals between schooling and the unobserved heterogeneity. Arguably this identification assumption is weaker than that involved in finding valid IVs (which are not weak). Accordingly, we are interested in what difference the use of this estimation procedure may reveal about the RoRtE.

Second, by taking datasets from different countries, the UK, USA, and Germany with their different education systems and labour markets we explore whether our results have external validity and are not just the artifact of one particular system or the peculiarity of a specific set of data. In addition the estimated coefficients may have distinct interpretations when different IVs are used, for example it may be a LATE in the case of using the ROSLA, or an ATE in the case of using father's education as the IV.

Thirdly, these papers use different candidate IV variables in an attempt to find a variable which correlates with years of schooling but not with the assumed stochastic error term in the earnings equation. Card (1995) uses proximity of college, Ashenfelter and Rouse use twin's education, Angrist and Krueger (1991) use quarter of birth, Harmon and Walker (1995) use the Raising of the School Leaving Age, and Hoogerheide, Block, and Roy (2012) (henceforth HBT) argue that father's education can be considered to be a satisfactory IV. The advantage of our study is that we examine these different IVs and find that, for the most part, the type or form of the IV used does not change our copula results. A further advantage of copula methods is that we are not

debating whether the effect of education should be interpreted as a local average treatment effect or not (Imbens and Angrist, 1994).

The final advantage of our approach is that, conditional on our identification assumptions, we retrieve a direct estimate of the correlation between the years of schooling and the unobserved heterogeneity in the earnings equation. This is valuable as it potentially provides some evidence on the association of schooling with the error term.

In the next section, we review the standard OLS specification of the human capital earnings equation and the standard IV identification strategy. We then explain how the copula framework may be used to overcome the assumption that schooling and the unobserved heterogeneity in earnings determination are uncorrelated. In section 3 we briefly describe the replication of the canonical papers and their data, presenting the copula results in section 4 on both the RoRtE and the underlying correlation coefficient between schooling and earnings heterogeneity. In section 5 we then explore different distributional assumptions to examine the robustness of our conclusions. The surprising conclusion of our investigation is, irrespective of the copula distributional assumptions used, that the RoRtE is surprisingly stable at around 6%.

2. The Copula Model of Schooling and Earnings.

Applying and extending the logic of Park ad Gupta (2012) to our setting we can consider the log-linear earnings-schooling relationship:

$$\log(w_i) = X_i'\beta + \alpha S_i + \varepsilon_i, \qquad i = 1, \cdots, N \tag{1}$$

where, for $i = 1, \cdots, N$, $\log(w_i)$ denotes the log-wage of individual i, X_i is a $(k \times 1)$ vector of exogenous regressors, S_i is a (1×1) regressor, measuring individual i's schooling, ε_i is a structural error term and α and β are model parameters.

Despite the apparent simplicity of model (1), identification of parameter α, the average wage returns to schooling, is complex because of the endogeneity of the schooling measure, S_i, due to the latter possibly being correlated with ε_i. Specifically, the most common concern is that the unobserved heterogeneity, ε_i, includes the influence of individual unobservables and is therefore likely to be correlated with the realised observed years of full time education. So the individual may choose how much post-compulsory full time schooling to acquire based on their unobserved ability (i.e. $E(\varepsilon_i|S_i) \neq 0$) – possibly in the expectation that their future earnings are likely to be related to how much education they acquire.

Instead of relying on a presumed valid instrument, we adopt an instrument-free approach, based on copula theory, to identify the returns to schooling, α, in the possible presence of endogeneity[3]. The idea is that, even without instruments, one can obtain consistent estimates if the joint distribution of the endogenous regressor and the structural error term is known.

Let us assume one knows the bivariate distribution, $f(S, \varepsilon)$, that S_i and ε_i are generated from. Then, one can obtain consistent estimates of the model parameters by maximizing the log-likelihood function derived from $f(S, \varepsilon)$. Identification in this case stems from knowledge of the information contained in the joint distribution of the endogenous regressor and the structural error term, not from instrumental variables.

Conventionally marginal distrubutions are derived by integration from an assumed (limited choice of) parametric joint distribution. But this limits the choice of marginals for the endogenous regressor, S_i, and the structural error term, ε_i, thereby imposing restrictions on the stochastic nature of the observed data. This potentially causes bias in estimated parameters of interest. The copula approach overcomes this shortcoming of the standard model as it allows one to start from the marginals. Using information contained in the observed data, one can separately select appropriate marginal distributions for the endogenous regressor and for the structural error term. The joint distribution can then be inferred from these marginals with an assumption about their structure of dependence, described by a given copula.

We start by assuming that the marginal distribution of the structural error term is normal. This is a reasonable default which has been nearly universally used for inference in this literature (e.g. Angrist and Krueger, 1991, Harmon and Walker, 1995, Card, 1999). Of course, other assumptions are possible. For example, if one believes that the actual error distribution may have thicker tails than the Normal, then the Student's *t*-distribution can be used. In section 5, we assess the sensitivity of the proposed model to possible misspecification by showing that the model estimates are robust to violations of the distributional assumption of ε_i.

To avoid problems of distributional misspecification, we use the empirical marginal distribution of the endogenous regressor, S_i. Unlike the structural error term, which is unobserved, sample data from the true distribution of S_i is observed and allows us to let the data 'speak for themselves'.

[3] Such an approach has previously been suggested by Park and Gupta (2012) in the general context of a linear regression model.

Using copula theory we can now construct the joint distribution function from the marginal distributions of the endogenous regressor and the structural error term, denoted by $H(s)$ and $G(\varepsilon)$, respectively. Letting $F(s,\varepsilon)$ be a joint distribution function with marginals $H(s)$ and $G(\varepsilon)$, Sklar's theorem (Sklar, 1959) states that there exists a copula function, C, such that for all S and ε,

$$F(s,\varepsilon) = C\big(H(s), G(\varepsilon)\big) = C(U_S, U_\varepsilon) \tag{3}$$

where $U_S = H(s)$ and $U_\varepsilon = G(\varepsilon)$. This theorem is central to the theory of copulas and clarifies the role that copulas play in the relationship between multivariate distribution functions and their univariate marginals. Since $H(\cdot)$ and $G(\cdot)$ are marginal distribution functions, U_S and U_ε are probability integral transformations, and thus U_S and U_ε are uniform$(0,1)$ random variables and the copula can be viewed as a function $[0,1]^2 \rightarrow [0,1]$. It becomes clear that the copula provides a way to study and model the scale-free dependence structure between variables after filtering out the influence of marginal distributions. Let $f(s,\varepsilon)$ denote the joint density function. Then, from (3), we have

$$f(s,\varepsilon) = c\big(H(s), G(\varepsilon)\big)h(s)g(\varepsilon), \tag{4}$$

where $c\big(H(s), G(\varepsilon)\big) = \partial^2 C/\partial s \partial \varepsilon$ is the copula density and where $h(s)$ and $g(\varepsilon)$ are the marginal density functions of the endogenous regressor, S and the structural error, respectively.

In the statistics literature, alternative copula functions have been studied (see Nelsen, 2006), each of these functions being more appropriate to model a specific structure of dependence between the marginals. We start by following Park and Gupta (2012) who use the Gaussian copula to model the dependence between the endogenous regressor and the structural error term:

$$C(U_S, U_\varepsilon) \;= \Psi_\rho\big(\Phi^{-1}(U_S), \Phi^{-1}(U_\varepsilon)\big) \tag{5}$$

$$= \frac{1}{2\pi(1-\rho^2)^{1/2}} \cdot \int_{-\infty}^{\Phi^{-1}(U_S)} \int_{-\infty}^{\Phi^{-1}(U_\varepsilon)} \exp\left[\frac{-(m^2 - 2\rho \cdot m \cdot n + n^2)}{2(1-\rho^2)}\right] dmdn,$$

where Φ denotes the univariate standard normal distribution function and Ψ_ρ denotes the bivariate standard normal distribution function with the correlation coefficient, ρ. If we let S^* and ε^* denote $\Phi^{-1}(U_S)$ and $\Phi^{-1}(U_\varepsilon)$, respectively, the Gaussian copula models $[S^*, \varepsilon^*]'$ as the standard bivariate normal distribution with correlation coefficient, ρ.

From (4) and (5), we can write the joint density function of S_i and ε_i :

$$f(S_i, \varepsilon_i) = \frac{1}{(1-\rho^2)^{1/2}} \exp\left[-\frac{\rho^2\left(\Phi^{-1}\left(U_{S,i}\right)^2 + \Phi^{-1}\left(U_{\varepsilon,i}\right)^2\right)}{2(1-\rho^2)} + \frac{\rho\Phi^{-1}\left(U_{S,i}\right)\Phi^{-1}\left(U_{\varepsilon,i}\right)}{(1-\rho^2)} \right] h(S_i)g(\varepsilon_i), \tag{6}$$

We can obtain consistent estimates by maximizing the following log-likelihood function, derived from (6), with respect to model parameters $\Theta = \{\alpha, \beta, \sigma_\varepsilon, \rho\}$:

$$\begin{aligned} \ln l(\{S_i, \varepsilon_i\}|\Theta) = & -\frac{N}{2}\ln(1-\rho^2) \\ & -\sum_{i=1}^{N}\left[\frac{\rho^2\left(\Phi^{-1}\left(U_{S,i}\right)^2 + \Phi^{-1}\left(U_{\varepsilon,i}\right)^2\right)}{2(1-\rho^2)} - \frac{\rho\Phi^{-1}\left(U_{S,i}\right)\Phi^{-1}\left(U_{\varepsilon,i}\right)}{(1-\rho^2)}\right] + \ln\phi_{(0,\sigma_\varepsilon^2)}(\varepsilon_i), \end{aligned} \tag{7}$$

where $\phi_{(0,\sigma_\varepsilon^2)}(\cdot)$ is the normal density with mean 0 and variance σ_ε^2. Note that the nonparametric density $h(S)$ does not include any parameters and thus disappears from the log-likelihood function. However, it influences the log-likelihood function via $U_{s,i}$. To obtain $U_{S,i}$, we simply exploit the fact that S_i is coded as a discrete variable in all the datasets that we use and compute it as the cumulative frequency corresponding to S_i. $U_{\varepsilon,i}$ can be easily calculated from the normal distribution function. Recall that $U_{\varepsilon,i} = G(\varepsilon_i) = \Phi_{(0,\sigma_\varepsilon^2)}(\varepsilon_i)$, where $\Phi_{(0,\sigma_\varepsilon^2)}(\cdot)$ is the distribution funtion of the normal with mean 0 and variance σ_ε^2. This implies that:

$$\Phi^{-1}\left(U_{\varepsilon,i}\right) = \varepsilon_i^* = \Phi^{-1}\left(\Phi_{(0,\sigma_\varepsilon^2)}(\varepsilon_i)\right) = \varepsilon_i/\sigma_\varepsilon. \tag{8}$$

As shown by Park and Gupta (2012), the Gaussian copula model implies that $[S^*, \varepsilon^*]'$ follows the standard bivariate normal distribution[4] with correlation coefficient, ρ which allows one to rewrite (1) as follows:

[4] This can be written as follows:

$$\begin{pmatrix} S_i^* \\ \varepsilon_i^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}\begin{pmatrix} \widetilde{\omega}_{1,i} \\ \widetilde{\omega}_{2,i} \end{pmatrix}, \qquad \begin{pmatrix} \widetilde{\omega}_{1,i} \\ \widetilde{\omega}_{2,i} \end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right),$$

$$\log(w_i) = X_i'\beta + \alpha S_i + \sigma_\varepsilon \cdot \rho \cdot S_i^* + \sigma_\varepsilon \cdot \sqrt{1-\rho^2} \cdot \widetilde{\omega}_{2,i}, \qquad i = 1, \cdots, N \tag{9}$$

This representation provides another way to estimate the model. Specifically we can consistently estimate the model in (9) using the least squares estimator, where S_i^* is included as an additional regressor and $\widetilde{\omega}_{2,i}$ is not correlated with any other terms on the right-hand side of this equation. The additional regressor, S_i^*, corrects for the endogeneity bias and in this respect, is similar to the control function approach of others (Heckman, 1978, Hausman, 1978). Note however that if S were normally distributed, the parameter α would not be identifiable, as in this case, we would have $S_i^* = \Phi^{-1}(U_{S,i}) = \Phi^{-1}(\Phi(S_i)) = S_i$. We explore the empirical justification of these assumptions in Appendix D where we explicitly investigate the non-normality of S and compare its distribution to that of S^*.

Our estimation strategy occurs in two stages. In the first stage, $U_{S,i}$ or S_i^* are inferred from the observed distribution of S. In the second stage, computed $U_{S,i}$ or S_i^* are used to construct the likelihood function, and the parameters are estimated. However, the usual standard errors of parameters based on the information matrix would treat the computed $U_{S,i}$ and S_i^* as (measurement) error-free variables so that the standard errors may be incorrect. We use bootstrap standard errors instead.

3. Data and Empirical Replication Strategy.

In our empirical estimation we use a variety of different data sources: both cross section, cohort, censuses and survey data, as well as data collected on specific sub-populations. We do this because we wish to 'stress test' our estimation procedure to establish its robustness to different kinds of data. We also use survey from one-off special sub-populations – i.e. e.g. twins in the case of Ashenfelter and Rouse (1998) or black Americans, as used by Angrist and Krueger (1991) in the case of one of their tables.

The ideal dataset to be used to compare the results of different estimation techniques is one which: incorporates the possibility of using different common IV techniques; has a large sample

---

or $\varepsilon_i^* = \rho \cdot \widetilde{\omega}_{1,i} + \sqrt{1-\rho^2} \cdot \widetilde{\omega}_{2,i} = \rho \cdot S_i^* + \sqrt{1-\rho^2} \cdot \widetilde{\omega}_{2,i}$. Combined with (8), this yields:

$$\varepsilon_i = \sigma_\varepsilon \cdot \varepsilon_i^* = \sigma_\varepsilon \cdot \rho \cdot S_i^* + \sigma_\varepsilon \cdot \sqrt{1-\rho^2} \cdot \widetilde{\omega}_{2,i}.$$

consisting of many people of different ages who have experienced different labour market conditions; has been used to publish articles in highly ranked journals so that a meaningful comparison may take place with generally accepted and well cited findings.

The choice of data set and paper was also guided by the fact that we wish to explore the effect of different IV variables – in our selected papers we have five very different IVs which genuinely reflect the diversity of instruments which have been variously used in the literature. Card (1999) uses the proximity of college, HBT (2012) use father's education, Ashenfelter & Rouse (1998) use the other twin's reported years of education, Harmon and Walker (1995) use the RoSLA, and Angrist & Krueger (1991) use the quarter of birth. Each of these IVs have their merits and constitute genuine attempts to find a variable highly correlated with years of schooling, but uncorrelated with unobserved heterogeneity in the earnings equation. In each paper the suitability of the use of the IV estimation technique depends essentially on the strong (but untestable) assumption of the degree to which the chosen IV was not correlated with the unobserved heterogeneity in the earnings equation. It is likely that this may vary between, for example, using a characteristic of the family of the respondent – e.g HBT (2012) using father's education, rather than using, for example a natural experiment like a change in the law – as per the RoSLA.

We also wished our data to be taken from different countries – in our case: the USA, UK and Germany and to have an array of different sample sizes and other possible available controls for the earnings equation. The virtue of data from different countries is that it means that the education experiences of the respondents is different with different institutional structures and school examination systems. Our sample sizes vary from 658 in the case of Ashenfelter & Rouse (1998) to 329,509 in the case of Angrist & Krueger (1991).

The basic characteristics of the datasets used are summarized in Appendix A and specifically in Table A1. Further details of these data are available in the original papers cited in the references. Three features of our comparative datasets are worth mentioning. Firstly, two of the datasets contain men and women, namely those studied by Ashenfelter & Rouse (1998) and HBT (2012) whilst the others use data from men only. Secondly, one dataset is a census dataset which contains limited personal conditioning regressors, namely that by Angrist & Krueger (1991) whilst the others are based on survey data with many richer individual co-variates. Thirdly, it should be pointed out that our different data come from very different time periods – the earliest having been sampled between 1966 and 1970 and the latest sampled only in 2004.

The paper by Card (1999) uses US cohort survey data from the National Survey of Young Men (NSLYM) and consists of an estimation sample of only 3010 men. This dataset has been used by many researchers over the years. American Census data is used by Angrist & Krueger (1991) over the years 1970, and 1980 which consists of over 329,509 observations. They also have a sample of 26,913 black men which we exploit in one of our estimation replications. Ashenfelter & Rouse (1998) use a dataset of their own collection which relates to a special survey of identical twins collected over the years 1991-3. Unfortunately, this dataset is limited in size with only 684 twins but does afford us the opportunity to examine a specialist more highly educated subset of the population for whom we have an interesting IV.

The paper by HBT (2012) uses the German Socio-Economic Panel (SEOP) from 2004. This is a yearly household panel survey and is recognized as the best available data of households in Germany. It consists of 8,244 men and women and is characterized by a rich set of possible conditioning regressors. Their paper has a different focus than the other papers in that it examines the possible underlying person specific heterogeneity in the RoRtE by using Bayesian methods. Their suggested IV of using father's education, if it had been used in a more conventional setting may have been more controversial. It affords us the opportunity to possibly examine a candidate IV which arguably is more likely to be endogenous.

Finally, the paper by Harmon and Walker (1995) uses the UK Family Expenditure Survey Data which is a cross section dataset that has been collected for over 65 years since 1961. Arguably this data is appropriate for the examination of the rate of return to education as it encompasses detailed information on individuals of different ages over many years which include those which saw the changing of the compulsory schooling reforms.

A further consideration in our choice of sample was to try to include less representative datasets to see how far our findings could be generalizable to other, less typical, data. This aim is realised by the use of the Angrist & Krueger (1991) subsample of only black men (who typically have less years of completed education) in their Table VIII. Here we see that the RoRtE is substantially lower at around 0.045. The second atypical sample we use is the data from Ashenfelter & Rouse (1998) which uses only identical twins. A major reason they are atypical is that they have around a year's worth more of education than the population at large. Here we find the RoRtE is much higher at around 0.12. This is potentially due to the non-representative nature of the twins sample. Another crucial aspect of our study is the heterogeneity of schooling years in different datasets. In particular, the empirical distribution of years of schooling may

not be normally distributed. This is important for the identification of the returns to schooling in this approach as it, indeed, requires that the schooling variable be non-Normal. This can easily be seen from (9). If S_i were normally distributed, S_i^* would be a linear transformation of it since $S_i^* = \Phi^{-1}\big(H(S_i)\big)$. Hence, we would not be able to separately identify α and $\sigma_\varepsilon \cdot \rho$ in (9).

As mentioned above, we have chosen to let the data speak for themselves by simply letting $H(S_i)$ be the cumulative frequency up to S_i. Of course, one could alternatively consider kernel density estimation. We have also done this using Harmon and Walker's (1995) data but it turned out that this makes little difference if any to the estimated coefficients.[5]

Our estimation and identification strategy involves first replicating what estimates may be retrieved by the use of OLS estimation of the earnings equation in each paper 'as if' there was no concern about the possibility of endogeneity. This also establishes that we have indeed got the right data and can retrieve the same estimates as the original authors. We then use their published details to replicate the authors own IV estimation of the RoRtE coefficient. Finally, we then use our suggested copula estimation to directly compare our results with the OLS and IV estimates of the same parameter. We experiment with simple changes to the standard human equation specification to see how the RoRtE parameter behaves. We find that the estimated coefficient varies between .05 to .065 depending on what other regressors are used.

4. Empirical Results from Copula Estimation.

In this section, we report the main results of the paper where we compare the OLS, IV and copula estimates of the RoRtE for the 5 different papers. We have replicated the authors OLS and IV estimation strategies and then re-estimated the authors preferred specifications via the use of the Gaussian copula through maximization of the log-likelihood function (7) for the five papers under consideration. The results are reported in Table 1. In Panel A, we report the results for the preferred specifications in the papers by Card (1993), HBT (2012), Harmon & Walker (1995) and Ashenfelter & Rouse (1998) papers. In Panel B, we report the results for three

[5] School leaving age being a discrete variable with bounded support, we actually used local polynomial kernel density methods with Gaussian kernel, jittering (Nagler, 2018) and direct plug-in methodology for bandwidth selection (Sheeter and Jones, 1991).

different specifications in the Angrist & Krueger (1991) paper[6]. For each paper, we report the author's and our own OLS estimates and the Gaussian copula estimates. In nearly all cases we replicate the authors OLS and 2SLS IV results to within 3 decimal places.

The most eye-catching result is that whatever specification one considers, the estimated IV RoRtE is systematically larger than its OLS and Copula counterparts in Table 1. The copula approach we used also suggests that OLS estimates can be higher or lower that the copula estimates. However, the size of this difference is far from being large. Overall, the copula-based method suggests the unbiased estimate of the RoRtE is not much larger than its OLS counterpart: with 6.21%, it is not even the half of the 15.4% Harmon and Walker suggested it is, based on the same data and the same earnings equation specification.

That fact that the estimate suggested by the copula approach is not very different from its OLS counterpart is probably due to potential endogeneity not being so severe. At least this is what the estimated linear correlation between school leaving age and the structural error term of the estimated earnings equation would logically predict.

There are basically four other substantive conclusions from the results in Table 1. Firstly, it would seem that the copula estimates of the RoRtE is approximately around 6% whichever data we use. (The only exception is the Ashenfelter and Rouse (1998) data on twins which is a special case). The copula RoRtE estimate always seems to be much lower than IV and closer to the OLS estimate (although mostly slightly higher than OLS too).

[6] In Table 1 we report our results using the data from Angrist & Krueger (1991) Table VI relating to all men born 1920-29 from the 1970 US Census and their Table VIII relating to black men born 1930-39 from the 1980 US census.

<u>Table 1:</u> OLS, IV and Copula-based Maximum Likelihood Estimates of the Return to Education Estimate by Paper.

PANEL A									
		Card (1993)		HBT (2012)		Harmon & Walker (1995)		Ashenfelter & Rouse (1998)	
Regressor		Authors'	Ours	Authors'	Ours	Authors'	Ours	Authors'	Ours
OLS RoRtE		0.073*** (0.004)	0.0725** (0.0037)	0.0658*** (NA)	0.0658*** (0.0021)	0.0613*** (0.001)	0.0597*** (0.0010)	0.1100*** (0.0096)	0.1100*** (0.0096)
IV RoRtE		0.132*** (0.049)	0.1304*** (0.0520)	0.0792*** (NA)	0.0792*** (0.0069)	0.1525*** (0.015)	0.1548*** (0.0143)	0.1160*** (0.0104)	0.1160*** (0.0104)
Gaussian copula	RoRtE		0.0638*** (0.0051)		0.0569*** (0.0044)		0.0619*** (0.0023)		0.1193*** (0.0189)
	ρ		0.0376** (0.0187)		0.0217** (0.0109)		-0.0104 (0.0098)		-0.0225 (0.0395)

PANEL B							
		Angrist & Krueger (1991)					
		Table IV, Specification 3/4		Table IV, Specification 7/8		Table VIII, Specification 7/8	
Regressor		Authors'	Ours	Authors'	Ours	Authors'	Ours
OLS RoRtE		0.0802*** (0.0004)	0.0802*** (0.0004)	0.0701*** (0.0004)	0.0701*** (0.0004)	0.0576*** (0.0013)	0.0576*** (0.0014)
IV RoRtE		0.1310*** (0.0334)	0.1310*** (0.0352)	0.1007*** (0.0334)	0.1007*** (0.0354)	0.0391*** (0.0199)	0.0391 (0.0205)
Gaussian copula	RoRtE		0.0832*** (0.0007)		0.0692*** (0.0007)		0.0460*** (0.0010)
	ρ		-0.0122*** (0.0023)		0.0035 (0.0025)		0.0476*** (0.0038)

Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The second is that the IV estimates are larger than the OLS estimates. Comparing the RoRtE obtained by OLS to the IV estimate it seems that the estimate is around half of the IV estimate. Again, a special case relates to black men in the 1980 US census data from the Angrist and Krueger (1991) data, where the RoRtE is lower at around 5%. This is in their Table VIII, Specification 7/8 where it is possible that this specification exhibits endogeneity as shown by the copula estimate of $\rho = 0.0476$.

A third substantive finding from Table 1 is that it does not make much difference what IV is used – these estimations variously use: ROSLA, Quarter of Birth, Nearness of college, Twin data, and father's education. The OLS and copula results are similar in all cases – it is only the IV results which are markedly different – up to 200% larger.

The fourth area of our substantive conclusions relates to the indicator for the possible presence of endogeneity. In our case studies the copula estimation suggests that the estimation of ρ is

usually very low and often insignificant. This may mean that for the papers we examine the schooling years variable is possibily not endogenous. This can be seen from last row in Panel A and Panel B of Table 1. This suggests that it is possible that endogeneity of schooling may not actually be a problem in the, Ashenfelter and Rouse (1998) and Harmon and Walker (1995) papers. Although it should be noted that ρ is statistically significantly different from zero in the Card (1993) paper, the HBT (2012) study, and two of the Angrist and Krueger (1991) specifications.

5. Robustness.

There are three areas of potential concern regarding the identification and estimation of the results in Table 1. Specifically, it is possible that the results we have obtained are sensitive to the exact specification of the earnings equation (and/or the first stage IV equation). We examine this issue in sub-section 5.1. Secondly, we may question whether the assumption about the distribution of the unobserved heterogeneity term in the earnings equation is really Normally distributed and how much it matters if it deviates from Normality. We discuss this aspect in sub-section 5.2. where the Student distribution is considered as an alternative to normality. Thirdly, the results in Table 1 are based on the assumption that, school leaving age being endogenous, identification of the RoRtE is possible if one knows the joint distribution of school-leaving age and individuals' unobserved characteristics. Copula theory is very helpful in this respect as it allows one to model that joint distribution in a quite flexible way. According to Sklar's theorem, knowledge of the joint distribution is guaranteed if one knows the marginal distributions and their structure of dependence. So far, we have assumed the error term has a normal distribution and considered that the structure of dependence between the latter and our schooling variable can be described by a Gaussian copula. In sub-section 5.3., we assess the sensitivity of our results based on the Gaussian copula to alternative dependence structures.

5.1. Sensitivity to Equation Specification.

One important question which may have a bearing on the exact nature of the RoRtE estimation results is the exact form of the specification of the earnings function. In this section, we report the estimation of the RoRtE when different controls are used in the earnings equation. Using the Harmon and Walker (1995) data we compare how the RoRtE coefficient varies as the controls vary for both the OLS, IV and copula estimation procedures in Table 2. The first thing

to notice is that the copula RoRtE estimates only vary between 0.0538 and 0.0672 between no controls and a full set of controls and does not get monotonically smaller in size as more controls are added. The second feature of the results is that the copula estimates are around 0.004 larger than the OLS results on a consistent basis irrespective of the specification.

What is clear about the results in Table 2 Panel B, is the extent to which the exact form of the specification of the first and second stage IV changes the estimated RoRtE coefficient. Literally from negatively significant, to insignificantly different from zero, to 0.15 and statistically significant. The literature has chosen to emphasize the latter as the most appropriate estimate and this has had huge attendant policy implications. It is appropriate for us to examine the extent to which this coefficient might change if we change out basic underlying identification assumption.

However, as can be seen from the alternative specification, the results critically depend on the age controls used, and on the inclusion of controls for the birth cohort. This provides support to Card's (1999) criticism of Harmon and Walker (1995) regarding adequate controls for systematic inter-cohort changes in educational attainment and earnings.

A final observation on these results is that the ρ parameter estimate is uniformly small irrespective of specification. Although it is statistically significantly different from zero in specification columns (1) and (4). This suggests that it is possible that endogeneity of schooling may not actually be a problem and that it could be dependent on the specification of the first or second stage equations.

We also find that the estimates can be sensitive to the specification of the equations used in the first and second stage of ordinary IV in the Angrist and Kreuger (1991) paper in all their tables. In addition, holding the specification constant but changing the data years – for example see Angrist and Kreuger (1991) – we saw in Table 1 that the results on the size of the estimated RoRtE and ρ can change markedly. For ρ it goes from negatively significant, to insignificantly different from zero, to positively significant. But the differences are never large.

Table 2: Different Specification Estimates of Returns to Education and Log Hourly Earnings, Harmon & Walker (1995): Male Sample (FES 1978-86 Data)

PANEL A: OLS

VARIABLES	(1) lnW_hourly	(2) lnW_hourly	(3) lnW_hourly	(4) lnW_hourly	(5) lnW_hourly
RoRtE	0.0538***	0.0647***	0.0507***	0.0521***	0.0597***
	(0.00105)	(0.00101)	(0.00105)	(0.00105)	(0.00101)
Observations	34,335	34,335	34,335	34,335	34,335
R-squared	0.071	0.244	0.089	0.082	0.269

PANEL B: IV

VARIABLES	(1) lnW_hourly	(2) lnW_hourly	(3) lnW_hourly	(4) lnW_hourly	(5) lnW_hourly
RoRtE	-0.0068**	0.1210***	-0.0044	-0.0162***	0.1548***
	(0.0033)	(0.0136)	(0.0032)	(0.0034)	(0.0143)
Observations	34,335	34,335	34,335	34,335	34,335
R-squared	0.036	0.155	0.027	0.017	0.197

PANEL C: Gaussian Copula

VARIABLES	(1) lnW_hourly	(2) lnW_hourly	(3) lnW_hourly	(4) lnW_hourly	(5) lnW_hourly
RoRtE	0.0580***	0.0672***	0.0538***	0.0569***	0.0619***
	(0.00213)	(0.00217)	(0.00236)	(0.00231)	(0.00181)
σ_ε	0.4297***	0.3876***	0.4254***	0.4270***	0.3809***
	(0.00246)	(0.00235)	(0.00261)	(0.00258)	(0.00247)
ρ	-0.0179***	-0.0119	-0.0131	-0.0205***	-0.0104
	(0.00823)	(0.00874)	(0.00878)	(0.00875)	(0.00949)
Observations	34,335	34,335	34,335	34,335	34,335
Log-likelihood	-81571.16	-78035.57	-81236.86	-81362.51	-77444.41

CONTROLS

Constant	Yes	Yes	Yes	Yes	Yes
age and age squared	No	Yes	No	No	Yes
regional dummies	No	No	Yes	No	Yes
year dummies	No	No	No	Yes	Yes

Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes on Panel C:

1. These estimation results are based on the observed (empirical) distribution of school leaving age.

2. These results assume the structural error term is normal and the structure of dependence between the latter and school leaving age is describable by a Gaussian copula.

3. Bootstrap standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

5.2. Non-Normal Error Term in the Earnings Equation

In the copula results so far discussed we have assumed the error term for the earnings equation is normally distributed. This is the assumption typically made in the OLS general linear model for inference purposes. But it is also the assumption which is generally used as a reasonable approximation to reality. In Appendix C we show the conditions under which this assumption may be justified formally, if the earnings distribution is log normal. We also invoke the logic of why the assumption of Normality of the stochastic error term in the earnings equation may reasonably be assumed to be approximated by the Normal invoking Central Limit arguments. If the logic in Appendix C is accepted then we can have some confidence in the results in Table 1.

Notwithstanding the above logic and the fact that the assumption of Normality is usually adopted in the literature, we now assess the robustness of our estimates of the RoRtE to possible misspecification in this respect. We do so by considering a distribution of the error term that involves thicker tails than the Normal, namely the Student's *t*-distribution.

More specifically, the density of the error term is given by:

$$g(\varepsilon_i) = \frac{1}{\sqrt{\pi\delta}} \frac{\Gamma\left(\frac{\delta+1}{2}\right)}{\Gamma\left(\frac{\delta}{2}\right)} \left(1 + \frac{{\varepsilon_i}^2}{\delta}\right)^{-\frac{\delta+1}{2}},$$

where δ denotes the number of degrees of freedom and Γ is the gamma function.[7]

Assuming the Gaussian copula (5) describes the structure of dependence between school leaving age and the error term, the log-likelihood function to be maximized is given by:

$$\ln l(\{S_i, \varepsilon_i\}|\Theta) = N\ln\left(\frac{1}{\sqrt{\pi\delta}} \frac{\Gamma\left(\frac{\delta+1}{2}\right)}{\Gamma\left(\frac{\delta}{2}\right)}\right) - \frac{N}{2}\ln(1-\rho^2) - \left(\frac{\delta+1}{2}\right)\ln\left(1 + \frac{{\varepsilon_i}^2}{\delta}\right) - \sum_{i=1}^{N}\left[\frac{\rho^2\left(\Phi^{-1}\left(U_{S,i}\right)^2 + \Phi^{-1}\left(U_{\varepsilon,i}\right)^2\right)}{2(1-\rho^2)} - \frac{\rho\Phi^{-1}\left(U_{S,i}\right)\Phi^{-1}\left(U_{\varepsilon,i}\right)}{(1-\rho^2)}\right].$$

where $U_{\varepsilon,i} = T_\delta(\varepsilon_i)$, $T_\delta(\cdot)$ being the CDF of the Student distribution with δ degrees of freedom.

[7] $\Gamma(z) = \int_0^{+\infty} x^{z-1}e^{-x}dx$.

The results are reported in Table 3 where the same set of 5 specifications from Table 2 for the Harmon and Walker (1995) paper are again considered. Although they only can be compared with caution, the optimum values of the log-likelihood are much smaller than those in the bottom of Table 2, hence suggesting the fit to the Student's *t*-distribution is less appropriate than the fit of the Normal distribution. This intuition is confirmed by the very large estimated degrees of freedom. A number of degrees of freedom as large as 341 suggests the distribution of the error term does not have thick tails and can therefore be very reasonably approximated by a Normal distribution.

Table 3: Copula-based Maximum Likelihood Estimates of Returns to Education with Student Error Term in the Log Hourly Earnings equation for the Harmon and Walker (1995): Male Sample (FES 1978-86 Data)

VARIABLES	(1) lnW_hourly	(2) lnW_hourly	(3) lnW_hourly	(4) lnW_hourly	(5) lnW_hourly
RoRtE	0.0545***	0.0747***	0.0509***	0.0523***	0.0680***
	(0.00260)	(0.00228)	(0.00258)	(0.00257)	(0.00226)
Constant	Yes	Yes	Yes	Yes	Yes
age and age squared	No	Yes	No	No	Yes
regional dummies	No	No	Yes	No	Yes
year dummies	No	No	No	Yes	Yes
ρ	-0.0048	-0.0288***	-0.0025	-0.0039	-0.0239***
	(0.00428)	(0.00352)	(0.00423)	(0.00422)	(0.00343)
δ	341.24***	341.24***	341.24***	341.24***	341.24***
	(4.60^{E}-5)	(4.81^{E}-5)	(3.29^{E} -4)	(5.92^{E}-5)	(2.30^{E}-6)
Observations	34,335	34,335	34,335	34,335	34,335
Log-likelihood	-96633.62	-96040.97	-96572.60	-96595.50	-95954.24

Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes :

2. These results assume the structural error term follows a distribution of Student (with δ degrees of freedom) and the structure of dependence between the latter and school leaving age is describable by a Gaussian copula.

3.We were not able to obtain bootstrap standard deviations as log-likelihood maximization failed to converge for every sample replication. We present approximate standard errors instead, computed as the square roots of the approximate variance-covariance matrix, $H^{-1}g^{T}gH^{-1}$ where $H = \nabla^{2}f(x)$, is the Hessian of the log-likelihood function, $f(x) = \sum_{i=1}^{N} f_i(x)$ and $g = (\nabla f_1, \cdots, \nabla f_N) = \left(\partial f_i / \partial x_j\right)$, its gradient. See Cramer (1986) or Gallant (1987).

Also, remarkable in Table 3 is the relative instability of the estimates of RoRtE as well as of the coefficient of correlation between school leaving age and the error term. In specifications (2) and (5), inclusion of age and its square as regressors results in both a significantly higher negative correlation and a significantly larger RoRtE. But even in these two specifications,

where the estimated RoRtE is the largest under the Student *t*-distribution hypothesis, these coefficients remain relatively close to their OLS counterparts.

5.3. Non-Gaussian Copulas

Another question of interest is that of the potential effect on estimated RoRtE of considering alternative copula assumptions; that is, of considering alternative descriptions of the structure of dependence between school leaving age, the endogenous regressor, and the error term of the earnings function. In Table 4, we compare the results from estimating the preferred specification in each article under different copula hypotheses. In the different rows of the table, we report results from three alternative copula assumptions: the Student copula, the Gumbel copula and the Clayton copula, respectively. In Appendix B, we give a formal presentation of the model under consideration in each of the three scenarios and we specify the corresponding log-likelihood function that we maximize to estimate the model parameters.

<u>Table 4:</u> Alternative Copula Maximum Likelihood Estimates of Returns to Education and Log Hourly Earnings by Paper.

PANEL A					
		Card (1993)	HBT (2012)	Harmon & Walker (1995)	Ashenfelter & Rouse (1998)
Gaussian copula	RoRtE	0.0638*** (0.0051)	0.0569*** (0.0044)	0.0619*** (0.0023)	0.1193*** (0.0189)
	ρ	0.0376 (0.0187)	0.0217** (0.0109)	-0.0104 (0.0098)	-0.0225 (0.0395)
Student copula	RoRtE	0.0607*** (0.0056)	0.0508*** (0.0059)	0.0524*** (0.0026)	0.1190*** (0.0207)
	Ndf	26.84 (0.0011)	52.70 (0.0038)	40.63 (0.0029)	23.20 (0.0026)
	ρ	-0.0612** (0.0195)	-0.0438** (0.0145)	-0.0570** (0.0128)	0.0277 (0.0467)
Clayton copula	RoRtE	0.0722*** (0.0025)	0.0667** (0.0021)	0.0587** (0.0014)	0.1243** (0.0192)
	θ	0.0034 (0.0763)	-0.0086 (0.0103)	0.0206 (0.0194)	-0.1128 (0.1197)
Gumbel copula	RoRtE	0.0725*** (0.0042)	0.0628*** (0.0023)	0.0594** (0.0013)	0.1059** (0.0112)
	γ	1.0000*** (0.0001)	1.0008*** (00004)	1.0002 (0.0002)	1.0010 (0.0013)

PANEL B				
		Angrist & Krueger (1991)		
		Table IV, Specification 3/4	Table IV, Specification 7/8	Table VIII, Specification 7/8
Gaussian copula	RoRtE	0.0832*** (0.0007)	0.0692*** (0.0007)	0.0460*** (0.0010)
	ρ	-0.0122*** (0.0023)	0.0035 (0.0025)	0.0476*** (0.0038)
Student copula	RoRtE	0.0823*** (0.0014)	0.0655*** (0.0016)	*-0.1749*[8] *(0.0105)*
	Ndf	28.71 (0.63^{E}-5)	26.23 (0.0005)	*5.06* *(0.3980)*
	ρ	0.0102* (0.0056)	-0.0193*** (0.0066)	*-0.7647**** *(0.0149)*
Clayton copula	RoRtE	0.0792*** (0.0003)	0.0697*** (0.0003)	0.0554*** (0.0004)
	θ	0.0071*** (0.0010)	0.0035*** (0.0010)	0.0145*** (0.0010)
Gumbel copula	RoRtE	0.0802*** (0.0004)	0.0701*** (0.0004)	0.0575*** (0.0004)
	γ	1.0000*** (7.22^{E}-6)	1.0000 (7.7^{E}-6)	1.0000 (0.0000)

Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

[8] Estimation of this specification under the Student copula assumption yields strange results. We believe this is due to the specific sample that is used to estimate this specification. This, in turn, resulted in the log-likelihood maximization algorithms we have considered being unable to reach a global maximum.

Like the Gaussian copula, the Student copula is elliptical so that the correlation coefficient is a flexible measure of the structure of dependence. In both cases, if the correlation coefficient is, say, negative, then the tails of the joint distribution with young school leaving ages and positive error terms and with old school leaving ages and negative error terms would be fatter than the tails on the other corners of the joint distribution. However, even for the same value of the correlation coefficient, the Student copula with a relatively small number of degrees of freedom will always show fatter tails than the Gaussian. This to say that if the data suggest that pairs of extreme values of the two variables under consideration are more frequent (with greater tail probability mass) than suggested by the Gaussian copula, then the Student copula is a reasonable alternative to consider.

The results in Table 4 show that the estimated number of degrees of freedom is quite large. With the exception of specification 7/8 from Angrist and Krueger's (1991) Table 8, it is systematically above 20. Such a large number suggests that the tails of the joint distribution of school leaving age and the error term are not that fat and therefore that the Gaussian copula is an equally acceptable description of the structure of dependence in the data. The results also confirm that despite the estimated correlation coefficient between school leaving age and the error term is now almost six times larger than suggested by the Gaussian copula hypothesis, it remains negative and very modest in absolute value (around -0.06), again suggesting endogeneity of school leaving age is not severe. It is interesting to note that the estimated RoRtE under the Student copula hypothesis is even slightly smaller (5.24%) than its OLS counterpart (5.97%).

We have also considered two Archimedean copulas, namely the Gumbel copula and the Clayton copula. The corresponding results are reported in the lower rows of Table 4, Panel A and Panel B.[9]

The bivariate Gumbel copula with parameter, $\gamma \in [1, +\infty[$, is given by :

$$C_{Gu,\gamma}(U_S, U_\varepsilon) = \exp[-((-\ln U_S)^\gamma + (-\ln U_\varepsilon)^\gamma)^{\frac{1}{\gamma}}]$$

[9] An Archimedean copula is one that is generated using a function, φ, called a generator, that meets the conditions of the following theorem. Given a function $\varphi: [0,1] \to [0, +\infty]$, that is continuous and strictly decreasing over $[0,1]$ and such that $\varphi(1) = 0$, the function :

$$C(u_1, u_2) = \begin{cases} \varphi^{-1}(\varphi(u_1) + \varphi(u_2)) \; if \; 0 \leq \varphi(u_1) + \varphi(u_2) \leq \varphi(0) \\ 0, \qquad\qquad\qquad if \; \varphi(0) < \varphi(u_1) + \varphi(u_2) < \infty \end{cases}$$

is a copula iff φ is a convex function.

It is useful for the modelling of dependence structures yielding more probability mass or larger dependence on the right of the joint distribution. More specifically, it simplifies to independence copula for $\gamma = 1$ but as γ gets larger, it tends towards the comonotonicity copula, yielding tail dependence to the right of the joint distribution. It therefore allows a potentially more satisfactory modelling of those who have higher ability and extra-long participation in further education. Another property of the Gumbel copula that is worth noting is the relationship that exists between the copula parameter, γ, and the Kendall's rank correlation coefficient, τ, between the two variables the structure of dependence of which is described by the copula:

$$\gamma = \frac{1}{1-\tau}.$$

The results reported in the last row of Table 4, Panel A and Panel B, suggest an estimate of the parameter γ of 1, meaning Kendall's rank correlation coefficient, τ, between school leaving age and the error term is not significantly different from 0. Like the Gaussian and then the Student copula hypotheses, the Gumbel copula specification also indicates independence between schooling, our endogenous regressor, and unobservable characteristics. Again, this is perhaps the reason why the resulting estimate of the RoRtE is similar to its OLS counterpart.

The bivariate Clayton copula with parameter, $\theta \in [-1, +\infty[- \{0\}$, is given by :

$$\boldsymbol{C}_{Cl,\theta}(U_S, U_\varepsilon) = \left(\max\left\{(U_S)^{-\theta} + (U_\varepsilon)^{-\theta} - 1, 0\right\}\right)^{-\frac{1}{\theta}}$$

It is useful for the modelling of dependence structures yielding more probability mass or larger dependence on the left of the joint distribution. A value of $\theta = -1$ indicates countermonotonicity. As θ tends to 0, the structure of dependence approaches independence and as θ tends to infinity, the structure of dependence becomes more and more comonotonic. In contrast to the Gumbel copula, the Clayton copula allows a potentially more satisfactory modelling of those on the left tail of the error term (e.g. who have lower ability) and who are not willing or able to invest in education beyond compulsory levels. Interestingly, the parameter of the Clayton copula, θ, can also be linked to Kendall's rank correlation coefficient, τ, through the following relationship:

$$\theta = \frac{2\tau}{1-\tau}.$$

The results reported in the middle row of Table 4, Panel A and B, suggest an estimate of the parameter θ of 0, again meaning that Kendall's rank correlation coefficient, τ, between school leaving age and the error term is not significantly different from 0. Like the Gaussian, the

Student and the Gumbel copula hypotheses, the Clayton copula specification also provides some support for independence between schooling and unobservable characteristics. Accordingly, the resulting estimate of the RoRtE is again not so different from its OLS counterpart.

The major finding of this section which is of importance is that the distributional form of the copula used does not seem to make much difference – irrespective of whether one puts weight in the tails or asymmetrically in one tail – the results do not differ much from the Gaussian copula. This is a substantive finding which speaks to the robustness of our results to the charge that they may depend on the precise form assumed on the copula.

6. Implications and Conclusions.

While the rate of return to education is one of the most important policy relevant parameters, its determination is a complex task due to the acquisition of schooling possibly being endogenous to the determination of earnings. This explains the large body of literature which proposed instrumenting strategies to overcome the endogeneity problem. Among these, IV and RDD methods have been widely used. In particular, in the last 30 years, an emphasis has been placed in this literature on the use of instrumental variables for schooling – notably laws relating to the raising of the school leaving age (RoSLA). Recently, however, there has been increased scrutiny of instrumental variable methods of estimation. In particular, even in the presence of a valid instrument, it is not necessarily clear whether one is identifying the average treatment effect or some specific local average treatment effect. In addition, the recent literature shows that IV estimates may be sensitive, not only to the chosen instrument but to the estimated model specification as well.

This paper has contributed to this literature by proposing an alternative identification strategy that does not rely on an instrument, but rather, involes identifying the return to education under the assumption that full information is available to the researcher about the joint distribution of schooling and the unobserved determinants of earnings. Indeed, copula theory implies that such a joint distribution can be inferred from two basic ingredients: the marginal distributions of schooling and the stochastic error and secondly the structure of dependence. The method is flexible since it facilitates model estimation conditional on reasonable assumptions regarding the distribution of the error term and of dependence structure. It also allows one to check the robustness of the resulting estimate of the return to education to a number of variants of these

hypotheses. Most importantly the method, conditional on our identification assumptions, allows us to retrieve a direct estimate of the extent of the correlation between the years of schooling and the unobserved heterogeneity in the earnings equation. This may in turn be suggestive of whether there may be some bias in the OLS estimates.

It is also worth noting that the data sets were chosen for their diversity which allows us to 'stress test' the estimation method. These datasets have been rigorously studied in the literature and resulted in influential papers published in prominent journals. Identification in each study relies on exploitation of different, plausibly exogenous, IVs so we are testing that our results are not the artefact of scrutinising only one specific IV.

The main conclusion of using copula estimation is that the RoRtE based on the different datasets in different countries is around 6% for males in large representative samples (although it may be lower for black men and higher for twins). We consistently find that the IV estimate is around twice the size of the OLS estimate but that the copula estimate is most closely similar to the OLS estimate. This is an important result because of its implications in terms of education funding policies and the provision of incentives to invest in education. Interestingly, this estimate of 6% is not dependent on the choice of the copula. First, as our results show, the estimated RoRtE is robust to specification changes. Our approach involves no direct first stage estimation of a schooling equation but rather lets the 'data speak for itself' and directly exploits the observed distribution of schooling. On the other hand, even modifying the distributional assumptions does not result in a significant variation in the estimated RoRtE. Second, the method is robust and applicable to qualitiatively different types of IVs whether they be the quarter of birth, ROSLA, father's education or whatever.

This paper is not alone in questioning whether the RoRtE may actually be as lower than the 8 percent that was the received wisdom in the 1990s when Card (1999) wrote his definitive survey. Pischke and Watcher (2008) have found similar results for Germany – although not using copula estimation methods. Our results also support the case for more detailed examination of the results obtained by the use of IV and RDD estimation procedures.

In conclusion, it is worth briefly reflecting on the policy context of why our result may be important. It is widely understood that the optimism of the 1990s regarding the returns to education contributed to the political context in which governments sought a rapid expansion of higher education. These priorities were justified at the time because the leading economists of the day were suggesting that the return to an extra year of education in the USA could be

13% (Card, 1995 and Angrist and Kreuger, 1991), or in the UK could be as high as 15% (Harmon and Walker, 1995) or that the return to a degree could be as high as 27% (Blundell and Dearden, 1999). In policy terms such an option is a 'win-win'. If young people could be encouraged to go on to higher education they would reap the rewards personally in terms of the financial investment – but so would the public exchequer in terms of future tax revenue.

The potential implications of this investigation are twofold. Firstly, that we need to be cautious in our claims for single policy parameter estimates when they are based on specific IV/RDD estimation with strong identification assumptions. Secondly, we may be reassured as economists by the 6% RoRtE we get as it vindicates much of our simpler earlier analysis from OLS estimation.

References

Abadie, A. (2018) 'Statistical Non-significance in Empirical Economics', MIT mimeo.

Acemoglu, D., and Angrist. J. (2001). "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws," in *Ben S. Bernanke and Kenneth, eds., Rogoff, NBER Macroeconomics Annual 2000*. Cambridge, MA: MIT Press, pp. 9–59.

Albouy, V, and Lequien, L. (2009). "Does Compulsory Education Lower Mortality?" *Journal of Health Economics* 28 (1): 155–68.

Angrist, Joshua D., and Alan B. Krueger. (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106(4), pp. 979–1014.

Ashenfelter, O. and Rouse, C. (1998) "Income Schooling and Ability: Evidence from a New Sample of Identical Twins", *Quarterly Journal of Economics*, 113(2), pp. 253–284..

Bhuller, M. , Mogstad, M, and K.G.Salvannes (2017) Life-Cycle Earnings, Education Premiums, and Internal Rates of Return, *Journal of Labor Economics*, 35(4), 993-1030.

Black, S., Devereux, P., and Salvanes. K. (2008). "Staying in the Classroom and Out of the Maternity Ward? The Effect of Compulsory Schooling Laws on Teenage Births*."* *Economic Journal* 118 (530): 1025–54.

Bonhomme, S. and Robin, J-M. (2006), Modeling Individual Earnings Trajectories Using Copulas: France, 1990–2002, in Henning Bunzel, Bent Christensen, George R. Neumann, Jean-Marc Robin (ed.) *Structural Models of Wage and Employment Dynamics (Contributions to Economic Analysis, Volume 275)* Emerald Group Publishing Limited, pp.441 – 478.

Buscha, F. and Dickson. M. (2012). "The raising of the school leaving age: returns in later life." *Economics Letters*, vol. 117 (2), pp. 389-393.

Card, D. (1995). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky, eds., Aspects of labour market behaviour: Essays in honour of John Vanderkamp*. Toronto: University of Toronto Press, pp. 201–22.

Card, D. (1999). "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, Vol. 3A, edited by Orley Ashenfelter and David Card, 1801–63. Amsterdam: Elsevier.

Card, D. (2001) "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica*, *69*(5), pp. 1127–60.

Carneiro, P., Heckman, J., and Vytlacil. E. (2011). "Estimating Marginal Returns to Education." *American Economic Review*, 101 (October 2011): 2754–2781.

Cramer, J. S. (1986). *Econometric Applications of Maximum Likelihood Methods*. Cambridge: Cambridge University Press.

Del Bono, E., and Galindo-Rueda, F. (2007). "The Long-Term Impacts of Compulsory Schooling: Evidence from a Natural Experiment in School Leaving Dates." *CEE Discussion Papers 0074*, Centre for the Economics of Education, LSE.

Dearden, L., Fitzsimons, E., Goodman, A. and Kaplan, G. (2006) 'Estimating Lifetime Earnings Distributions Using Copulas' IFS mimeo.

Devereux, P., and Hart, R. (2010). "Forced to Be Rich? Returns to Compulsory Schooling in Britain." *Economic Journal*, 120 (549): 1345–64.

Dickson, M,. and Smith, S. (2011). "What determines the return to education: an extra year or a hurdle cleared?" *Economics of Education Review*, vol. 30 (6), pp. 1167-1176.

Dolton, P. and Sandi, M. (2017) 'Returning to returns: Revisiting the British education evidence', *Labour Economics,* 48, 87-104.

Feller, W. (1957) *An Introduction to Probability Theory and its Applications*, 2nd ediction, John Wiley & Sons, New York.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. New York: John Wiley & Sons.

Grenet, J. (2013). "Is Extending Compulsory Schooling Alone Enough to Raise Earnings? Evidence from French and British Compulsory Schooling Laws." *The Scandinavian Journal of Economics*, 115 (1): 176–210.

Griliches, Z (1977), "Estimating the returns to schooling: some econometric problems", *Econometrica,* 45:1 22.

Harmon, C, and Walker, I. (1995). "Estimates of the Economic Return to Schooling for the United Kingdom." *American Economic Review*, 85(5), pp. 1278–86.

Harrison, A. (1981) Earnings by Size: A Tale of Two Distributions, *The Review of Economic Studies*, 48, 621-631.

Hausman, J. A., (1978), Specification Tests in Econometrics, *Econometrica*, 46(6), 1251-1272.

Heckman, J. J., (1978), Dummy Endogenous Variables in a Simultaneous Equation System, *Econometrica*, 46(4), 931-959.

Hoogerheide, L., Block, J.H., and Roy T (2012), "Family background variables as instruments for education in income regressions: A Bayesian analysis", *Economics of Education Review*, 31, 515– 523

Imbens, G, W. and Angrist, Joshua D. (1994) "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, (2), 467-475

Johnston, J. (1972) *Econometric Methods*, 2nd Ed, McGrawHill, New York.

Kenny, L. Lung-Fei Lee, G. S. Maddala and R. P. Trost (1979) Returns to College Education: An Investigation of Self-Selection Bias Based on the Project Talent Data', *International Economic Review*, 20(3), 775-789

Kling, Jeffrey R. (2001). "Interpreting Instrumental Variables Estimates of the Returns to Schooling." *Journal of Business and Economic Statistics*, 19(3): 358–64.

Koop, G., and J. L. Tobias. (2004). "Learning about heterogeneity in returns to schooling." *Journal of Applied Econometrics*, vol.19: pp. 827–849.

Lleras-Muney, A. (2005). "The Relationships Between Education and Adult Mortality in the United States." *Review of Economic Studies*, 72 (250): 189–221.

Lydall, H. (1968) *The Structure of Earnings*, Clarendon Press, Oxford.

Nagler, T. (2018) 'Asymptotic analysis of the jittering kernel density estimator', Methods of Statistics, Forthcoming, arXiv:1705.0543

Nelsen, R.B. (2006) *An Introduction to Copulas*, (2nd ed) *Lecture Notes in Statistics*, Springer.

Oreopoulos, P. (2006). "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter." *American Economic Review*, 96 (1): 152–75.

Park, S. and Gupta, S. (2012) 'Handling Endogenous Regressors by Joint Estimation Using Copulas'. *Marketing Science,* 31(4):567-586.

Pischke, J-S., and von Wachter, T. (2008). "Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation", *Review of Economics and Statistics*, 90 (2008), 592-598.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, Series B, 53, 683–690.

Shih, J. and Louis, T. (1995) 'Inferences on the association parameter in copula models for bivariate survival data', *Biometrika*, 51, 1384-1399.

Sklar, A. (1959) 'Fontions de répartition a n dimensions et leurs marges', *Publ Inst Statist Univ, Paris*, 8, 229-231.

Theil, H. (1971) *Principles of Econometrics*. John Wiley and Sons, London.

Willis, R., and Rosen, S. (1979) 'Education and Self-Selection', *Journal of Political Economy*, 87, (5)(2), S7-S36

White, H. (1984) *Asymptotic Theory for Econometricians*, Academic Press, New York.

Zimmer, D. M. and Trivedi, P. K. (2006), 'Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand', *Journal of Business & Economic Statistics*, 24(1) (Jan., 2006), pp. 63-76

ON-LINE APPENDICES

APPENDIX A: Data in the Papers Examined

APPENDIX B: Alternative Copulas.

APPENDIX C: The Distribution of Errors in the Earnings Equation.

APPENDIX D: Comparing the distribution of S and S*

APPENDIX A. Data in the Papers Examined

For the analysis of this paper we rely on data from other prominent published papers. In guiding our choice of paper, we needed the paper to be published in a highly reputable top ranked general or field journal and the data to be available on the author's own website in three cases, namely: Card (1995), Angrist & Krueger (1991) and Ashenfelter & Rouse (1998). In one case, HBT (2012) we obtained the data directly from the authors. In the case of the Harmon and Walker (1995) paper we replicated the data sample extraction from the raw survey data. so that we could be sure to be able to replicate their original estimation.

The choice of data set and paper was also guided by the fact that we which to explore the effect of different IV variables – in our selected papers we have five very different IVs which genuinely reflect the diversity of instruments which have been variously used in the literature. It was thought that the potential for the more suitable use of the IV estimation technique may depend on how likely it was that the IV chosen was not correlated with the unobserved heterogeneity in the earnings equation.

Finally, we wished our data to be taken from different countries – in our case: the USA, UK and Germany and to have an array of different sample sizes and other possible available controls for the earnings equation. Our sample sizes vary from 658 in the case of Ashenfelter & Rouse (1998) to 329,509 in the case of Angrist & Krueger (1991).

The basic characteristics of the datasets used are summarized in the Table A1 below. Further details of these data are available in the original papers cited in the references. Three features of our comparative datasets are worth mentioning. Firstly, two of the datasets contain men and women, namely those studied by Ashenfelter & Rouse (1998) and HBT (2012) whilst the others use data from men only. Secondly, one dataset is a census dataset which contains limited personal conditioning regressors, namely that by Angrist & Krueger (1991) whilst the others are based on survey data with many richer individual co-variates. Thirdly, it should be pointed out that our different data come from very different time periods – the earliest having been sampled between 1966 and 1970 and the latest sampled only in 2004.

A further consideration in our choice of sample was to try to include less representative datasets to see how far our findings could be generalizable to other less typical data. This aim is realised by the use of the Angrist & Krueger (1991) subsample of only black men in their Table VIII. Here we see that the RoRtE is substantially lower at around 0.045. The second atypical sample we use is the data from Ashenfelter & Rouse (1998) which uses only identical twins. A major reason they are atypical is that they have around a year's worth more of education than the population at large. Here we find the RoRtE is much higher at around 0.12. This is undoubtedly due to the non-representative nature of the twins sample.

Table A1. Data Characteristics from Datasets Used.

	Survey Source	Date of Survey	Original Sample Size/Subset Criteria	Observations	IV
Author(s)					
Card (1995)	US – National Survey of Young Men (NSLYM)	1966	5525 men aged 14-24. Valid Wage and Education Data	3010	Proximity to College
Angrist & Krueger (1991)	US Census data, 1970, and 1980	1970, 1980,	Men born For Table IV: 1920-1929 For Table VIII; Blacks, 1930-39	247,199 26,913	Quarter of Birth
Ashenfelter & Rouse (1998)	US Survey of Identical Twins	1991-93	Identical Twins	658	Other Twin's Years of Education
Harmon & Walker (1995)	UK – Family Expenditure Survey	1978-1986	Employed males 18-64.	34,335[10]	Raising of the School Leaving Age Reform.
Hoogerheide, Block, and Roy. (2012)	German Socio-Economic Panel (SEOP)	2004	Male and Female employed.	8,244	Father's Education.

The logic for investigating: very different sample sizes; data from different countries; data collected in different ways; data from non-representative as well as a representative sample; data which had different sets of regressors available to it; and data which exploited different IVs, is that we wished to establish that our conclusions about the use of the Copula estimator are robust to different circumstances and hence establish some reasonable degree of external validity for our findings.

[10] In the original Harmon and Walker (1995) paper they had 34,336 observations. Since we replicated their data extraction from scratch but could not match their observation sample count exactly. However, with only one observation missing we were able to replication all their estimation results to within 3 decimal places. See Dolton and Sandi (2017) for details.

APPENDIX B: Alternative Copulas.

The Student copula

Let ρ denote Spearman's coefficient of correlation between our endogenous regressor, S_i and the structural error term, ε_i. The bivariate Student copula with δ degrees of freedom could be written as :

$$C^{t}_{\delta,\rho}(U_S, U_\varepsilon) = \mathbf{T}_{\delta,\rho}\left(T^{-1}_{\delta}(U_S), T^{-1}_{\delta}(U_\varepsilon)\right)$$

$\mathbf{T}_{\delta,\rho}$ is the bivariate Student CDF with δ degrees of freedom while T_δ is the univariate Student CDF with δ degrees of freedom.

The joint density function of S_i and ε_i (the analogue to equation (6)) is then given by:

$$f(S_i, \varepsilon_i)$$
$$= \frac{\Gamma\left(\frac{\delta+2}{2}\right)\Gamma\left(\frac{\delta}{2}\right)}{\sqrt{(1-\rho^2)}\Gamma^2\left(\frac{\delta+1}{2}\right)} \frac{\left(1+\frac{1}{\delta(1-\rho^2)}\left(\left(U_{S,i}\right)^2 - 2\rho U_{S,i}U_{\varepsilon,i} + \left(U_{\varepsilon,i}\right)^2\right)\right)^{-\frac{\delta+2}{2}}}{\left(\left(1+\frac{\left(U_{S,i}\right)^2}{\delta}\right)\left(1+\frac{\left(U_{\varepsilon,i}\right)^2}{\delta}\right)\right)^{-\frac{\delta+1}{2}}} h(S_i)g(\varepsilon_i),$$

and the corresponding log-likelihood function (the analogue to equation (7)) is

$$\begin{aligned}\ln l(\{S_i, \varepsilon_i\}|\Theta) = N\ln&\left(\frac{\Gamma\left(\frac{\delta+2}{2}\right)\Gamma\left(\frac{\delta}{2}\right)}{\sqrt{(1-\rho^2)}\Gamma^2\left(\frac{\delta+1}{2}\right)}\right)\\ &+\frac{\delta+1}{2}\sum_{i=1}^{N}\ln\left(\left(1+\frac{\left(U_{S,i}\right)^2}{\delta}\right)\left(1+\frac{\left(U_{\varepsilon,i}\right)^2}{\delta}\right)\right)\\ &-\frac{\delta+2}{2}\sum_{i=1}^{N}\ln\left(1+\frac{1}{\delta(1-\rho^2)}\left(\left(U_{S,i}\right)^2 - 2\rho U_{S,i}U_{\varepsilon,i} + \left(U_{\varepsilon,i}\right)^2\right)\right)\\ &+\ln\phi_{(0,\sigma_\varepsilon^2)}(\varepsilon_i),\end{aligned}$$

The Gumbel copula

The generator of the Gumbel copula is (see footnote 7):

$$\varphi_{Gu,\gamma}(u) = (-\ln u)^\gamma, \ \gamma \in [1, +\infty[$$

so that the bivariate Gumbel copula with parameter, γ is given by :

$$C_{Gu,\gamma}(U_S, U_\varepsilon) = \exp[-((-\ln U_S)^\gamma + (-\ln U_\varepsilon)^\gamma)^{\frac{1}{\gamma}}.$$

$$\gamma = \frac{1}{1-\tau}$$

where τ denotes Kendall's correlation coefficient of S_i and ε_i.

This yields the following expression of the joint density function of S_i and ε_i (the analogue to equation (6)):

$$\begin{aligned} f(S_i, \varepsilon_i) = \exp[-\big(\big(-\ln U_{S,i}\big)^\gamma + \big(-\ln U_{\varepsilon,i}\big)^\gamma\big)^{\frac{1}{\gamma}}]\big(U_{S,i}U_{\varepsilon,i}\big)^{-1}\big(\big(-\ln U_{S,i}\big)^\gamma \\ + \big(-\ln U_{\varepsilon,i}\big)^\gamma\big)^{-2+\frac{2}{\gamma}}\Big(\big(\ln U_{S,i}\big)\big(\ln U_{\varepsilon,i}\big)\Big)^{\gamma-1} \\ \times\left\{1 + (\gamma-1)\big(\big(-\ln U_{S,i}\big)^\gamma + \big(-\ln U_{\varepsilon,i}\big)^\gamma\big)^{-\frac{1}{\gamma}}\right\} h(S_i)g(\varepsilon_i), \end{aligned}$$

and therefore to the following log-likelihood function (the analogue to equation (7))

$$\begin{aligned} \ln l(\{S_i, \varepsilon_i\}|\Theta) = -\sum_{i=1}^{N}\big(\big(-\ln U_{S,i}\big)^\gamma + \big(-\ln U_{\varepsilon,i}\big)^\gamma\big)^{\frac{1}{\gamma}} \\ -\frac{2(\gamma-1)}{\gamma}\sum_{i=1}^{N}\ln\big(\big(-\ln U_{S,i}\big)^\gamma + \big(-\ln U_{\varepsilon,i}\big)^\gamma\big) \\ +\sum_{i=1}^{N}\ln\left(1 + (\gamma-1)\big(\big(-\ln U_{S,i}\big)^\gamma + \big(-\ln U_{\varepsilon,i}\big)^\gamma\big)^{-\frac{1}{\gamma}}\right) - \sum_{i=1}^{N}\ln\big(U_{S,i}U_{\varepsilon,i}\big) \\ +(\gamma-1)\sum_{i=1}^{N}\ln\Big(\big(\ln U_{S,i}\big)\big(\ln U_{\varepsilon,i}\big)\Big) + \ln\phi_{(0,\sigma_\varepsilon^2)}(\varepsilon_i), \end{aligned}$$

The Clayton copula

The generator of Clayton copula is (see footnote 7):

$$\varphi_{Cl,\theta}(u) = \frac{1}{\theta}\big(u^{-\theta} - 1\big),\ \ \theta \in [-1, +\infty[\ -\{0\}$$

The bivariate Clayton copula with parameter, θ is then:

$$C_{Cl,\theta}(U_S, U_\varepsilon) = \big(\max\big\{(U_S)^{-\theta} + (U_\varepsilon)^{-\theta} - 1, 0\big\}\big)^{-\frac{1}{\theta}}$$

$$\theta = \frac{2\tau}{1-\tau}$$

where τ denotes Kendall's correlation coefficient of S_i and ε_i.

The joint density function of S_i and ε_i (the analogue to equation (6)) is then given by

$$f(S_i, \varepsilon_i) = (\theta + 1)\left(U_{S,i}U_{\varepsilon,i}\right)^{-(\theta+1)}\left({U_{S,i}}^{-\theta} + {U_{\varepsilon,i}}^{-\theta} - 1\right)^{-\frac{2\theta+1}{\theta}} h(S_i)g(\varepsilon_i),$$

whereas the corresponding log-likelihood function (the analogue to equation (7)) is :

$$\ln l(\{S_i, \varepsilon_i\}|\Theta) = N\ln(\theta + 1) - (\theta + 1)\sum_{i=1}^{N} \ln\left(U_{S,i}U_{\varepsilon,i}\right)$$
$$- \frac{2\theta + 1}{\theta}\sum_{i=1}^{N} \ln\left({U_{S,i}}^{-\theta} + {U_{\varepsilon,i}}^{-\theta} - 1\right) + \ln\phi_{(0,\sigma_\varepsilon^2)}(\varepsilon_i),$$

.

APPENDIX C: The Distribution of Errors in the Earnings Equation.

In the estimates in section 4 of this paper we have assumed that the error in the earnings (*w*) equation (of regressors X_n) is Normally distributed for our benchmark Copula estimation. The pertinent questions here are: is this likely to be a valid assumption and what happens if the assumption is not valid.

By Bernstein's Theorem, if $X_1, X_2, \ldots\ldots X_n$ are independent Normally distributed random variables, then

$$\sum_{k=1}^{n} b_k X_k \sim \mathbb{N}$$

where b_k is a set of fixed constants. Furthermore, if earnings *w* are lognormally distributed then $Log(w)$ will be Normally distributed. The best available empirical evidence on the distribution of earnings concludes that it is Lognormally distributed with a Pareto tail (Harrison, 1981). The presence of this possible fat tail could advance the claim for us to model the resulting distribution of ε as having fat tails. This we do with a Student's t distribution in Section 5.2 in the paper.

In this case, since the difference between two Normally distributed variables is also Normal we may suggest that:

$$Log(w) - X\beta = \varepsilon \sim \mathbb{N}$$

In the event that $X_1, X_2, \ldots\ldots X_n$ are not independent Normally distributed random variables then we may appeal to the Lindberg-Levy Central Limit Theorem[11] to suggest that in large samples,

$$\sum_{k=1}^{n} b_k X_k \xrightarrow{asy} \mathbb{N},$$

[11] See White (1984), p64-65.

i.e. that the sum of the predicted regressors multiplied by their OLS estimated coefficients is asymptotically Normal.[12]

Or as Johnston explains[13], in the event that the assumption that ε cannot be assumed to be Normally distributed then it is approximately true in large samples *"by making no explicit assumption about the form of the distribution and appealing to the Central Limit Theorem to justify our regarding the tests as approximately correct."*

Another logic regarding the distribution of w is to reflect on the result that the product of Normally distributed random variables is LogNormal. Lydall (1968) provides the logic and rationale to suggest that if the conditioning variables in the determination of earnings act not in a linear additively separable way but in the form of multiplicative interactions then the resulting product – in this case earnings – may be distributed as Lognormal. Such results generate an interest in modelling 'fat tail' distributions as a logical alternative as we do in our robustness checks.

[12] See Feller (1957) pp 229, 238-241 or Theil (1971) p370 for details.
[13] See Johnston (1972) p135-6.

APPENDIX D: Comparing the distribution of S and S*

Our strategy to identify the parameter of interest, α, relies on copulas as a means of modelling the structure of dependence between the structural error term, ε_i, and the endogenous variable, S_i. As shown on page 9, maximization of the log-likelihood given in equation (7) is equivalent to estimating, through least squared methods, the augmented specification (9) where the term below is included instead of ε_i

$$\sigma_\varepsilon \cdot \rho \cdot S_i^* + \sigma_\varepsilon \cdot \sqrt{1-\rho^2} \cdot \widetilde{\omega}_{2,i}$$

where $S^* = \Phi^{-1}(U_S)$ and $\widetilde{\omega}_{2,i} = N(0,1)$.

Importantly, the additional regressor, S_i^*, is, by construction, drawn from the normal distribution where $U_{S,i}$ is nothing but $F(S_i)$, $F(\cdot)$ denoting the empirical CDF of S. Therefore, if S were normally distributed, the parameter α would not be identifiable, as in this case, we would have $S_i^* = \Phi^{-1}\left(U_{S,i}\right) = \Phi^{-1}\left(\Phi(S_i)\right) = S_i$.

It is therefore important to show that S_i significantly departs from the normal distribution and, consequently, from S_i^* which is again drawn from the normal distribution. We first conduct several tests the null of which is the assumption that S is normally distributed; namely the Kolmogorov-Smirnov test, the Cramer-von Mises test and the Anderson-Darling test. The results are reported in Table D1 below which also reports basic descriptive statistics as well as the results from the Shapiro-Wilk test when using the sample from the Ashenfelter and Rouse (1998) study where the sample size is 680 observations only. For the five data sets under consideration, these tests unanimously suggest the rejection of the normality hypothesis at the 1% level.

We also produce quantile-quantile plots as a means of highlighting the extent to which the schooling variable, departs from the normal distribution in each of the five data sets that we examine. These are reported on the left of Figure D1 below where each individual QQ-plot contrasts the percentiles of the distribution of the standardized (zero-mean and unit variance) version of S (on the vertical axis) to the percentiles from $N(0,1)$ (on the horizontal axis). The resulting figures highlight the discrete nature of S in each case, but also the fact that in each of our five datasets and for both plots we see a meaningful departure from the 45 degree line. This indicates that in each case the schooling variable is significantly different from Normal.. Moreover, in each case the nature of the departure is different. In the case of the Harmon and Walker, Card and Angrist and Kreuger data, the departure is mainly in the extreme left tail and after the zero point. In contrast in the Ashenfelter and Rouse and HBT data the departure is more specifically at various definite points in the support of S.

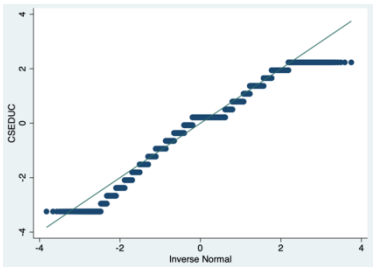
In addition, we investigate the relationship between S and S^*. One means of highlighting the differences between these two variables is by drawing QQ-plot-like diagrams where the two variables are contrasted

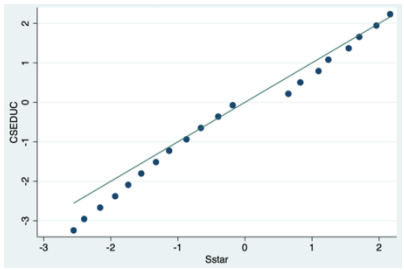
with each other. To be more specific, for each observed value S_i of S, we are able to empirically compute the corresponding cumulative frequency, $U_{S,i}$ as well as the corresponding value $S_i^* = \Phi^{-1}(U_{S,i})$; that is the percentile of the normal distribution that corresponds to $U_{S,i}$. Also, to make the comparison of S and S^* easier, we normalize the former. The resulting diagrams are reported to the right of Figure D1. Again, for any given observed percentile, departure from the 45 degree line reveals the difference between S and S^* and, in a sense, the extent of departure of S from normality.

Figure D1: Comparing the distributions of S and S^*.

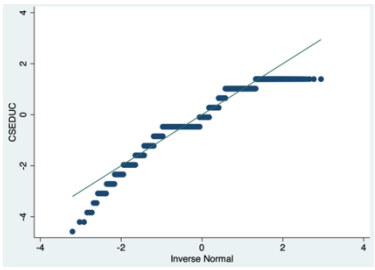
$S(0,1)$ vs. $N(0,1)$	$S(0,1)$ vs. S^*

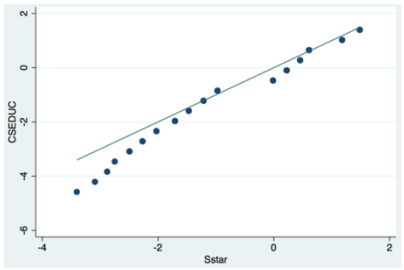
Angrist & Krueger (1991)



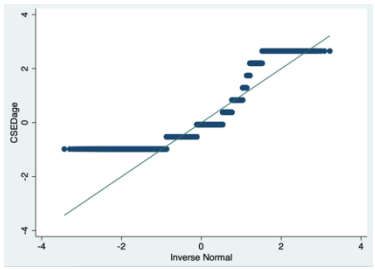


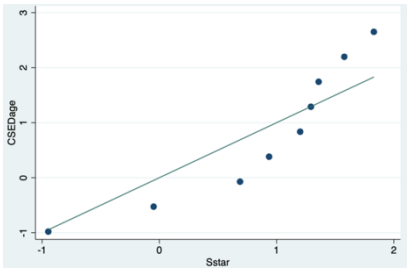
Card (1995)



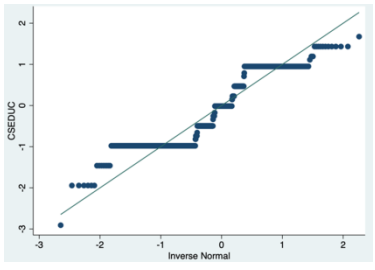


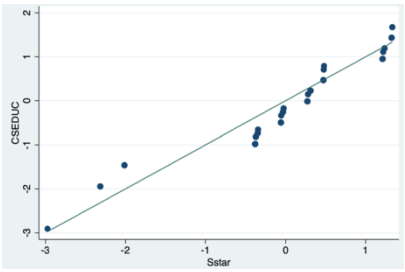
Harmon & Walker (1995)



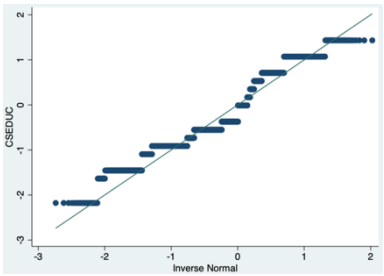


Ashenfelter & Rouse (1998)





Hoogerheide et al. (2012)



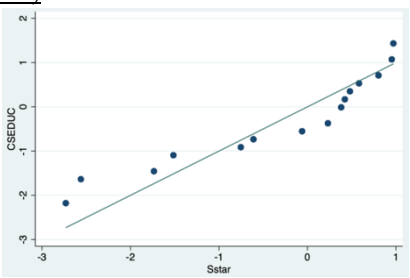


TABLE D1: Normality Test on Schooling by Dataset.

Authors	Descriptive Statistics						Normality Tests			
	Variable	N. Obs.	Mean	S.D.	Skewness	Kurtosis	Kolmogorov-Smirnov	Cramer-von-Mises	Anderson-Darling	Shapiro-Wilk
Angrist & Krueger (1991)	Yrs of Sch.	247,199	11.49	3.36	-0.26	0.36	0.1648*** (0.0100)	880.67*** (0.0050)	4158.99*** (0.0050)	-
Card (1995)	Yrs of Sch.	3,010	13.26	2.68	-0.23	0.28	0.1762*** (0.0100)	14.27*** (0.0050)	73.78*** (0.0050)	-
Harmon & Walker (1995)	Age left ed.	34,353	16.16	2.20	1.69	2.36	0.2831*** (0.0100)	465.26*** (0.0050)	2637.3*** (0.0050)	-
Ashenfelter & Rouse (1998)	Yrs of Sch.	680	14.03	2.07	0.42	-0.91	0.1876*** (0.0100)	4.60*** (0.0050)	29.64*** (0.0050)	0.8898*** (0.0001)
Hoogerheide et al. (2012)	Yrs of Sch.	8,244	13.03	2.77	0.69	-0.76	0.2365*** (0.0100)	72.05*** (0.0050)	434.81*** (0.0050)	-