Title: Family background variables as instruments for education in income regressions: A Bayesian analysis

Authors: Lennart Hoogerheide, Joern H. Block, Roy Thurik

PII:	S0272-7757(12)00026-X
DOI:	doi:10.1016/j.econedurev.2012.03.001
Reference:	ECOEDU 1290
To appear in:	*Economics of Education Review*
Received date:	16-1-2011
Revised date:	20-2-2012
Accepted date:	7-3-2012

Please cite this article as: Hoogerheide, L., Block, J. H., & Thurik, R., Family background variables as instruments for education in income regressions: a Bayesian analysis, *Economics of Education Review* (2010), doi:10.1016/j.econedurev.2012.03.001

Family background variables as instruments for education in income regressions: a Bayesian analysis

Lennart Hoogerheide [a], Joern H. Block [b], Roy Thurik [c]

[a] Department of Econometrics, Vrije Universiteit Amsterdam, De Boelelaan 1105, NL-1081 HV Amsterdam, the Netherlands; Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands. l.f.hoogerheide@vu.nl

[b] Centre for Advanced Small Business Economics, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands; Technische Universität München, München, Germany. block@wi.tum.de

[c] Centre for Advanced Small Business Economics, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands; EIM Business and Policy Research, P.O. Box 7001, 2701 AA Zoetermeer, the Netherlands and Max Planck Institute of Economics, Jena, Germany. thurik@ese.eur.nl.

Abstract: The validity of family background variables instrumenting education in income regressions has been much criticized. In this paper, we use data from the 2004 German Socio-Economic Panel and Bayesian analysis to analyze to what degree violations of the strict validity assumption affect the estimation results. We show that, in case of moderate direct effects of the instrument on the dependent variable, the results do not deviate much from the benchmark case of no such effect (perfect validity of the instrument's exclusion restriction). In many cases, the size of the bias is smaller than the width of the 95% posterior interval for the effect of education on income. Thus, a violation of the strict validity assumption does not necessarily lead to results which are strongly different from those of the strict validity case. This finding provides confidence in the use of family background variables as instruments in income regressions.

Keywords: education; family background variables; earnings; income; instrumental variables; Bayesian analysis; exclusion restriction

JEL codes: C11, C13, C15, J30

Corresponding author: Lennart Hoogerheide

The paper analyzes to what degree violations of the perfect validity of the exclusion restriction for family background variables in income regression affect the estimation results.

In case of moderate direct effects of the instrument on the dependent variable, the results do not deviate much from the benchmark case of no such effect (perfect validity of the instrument's exclusion restriction).

The finding provides confidence in the use of family background variables as instruments in income regressions.

Family background variables as instruments for education in income regressions: a Bayesian analysis

Abstract: The validity of family background variables instrumenting education in income regressions has been much criticized. In this paper, we use data from the 2004 German Socio-Economic Panel and Bayesian analysis to analyze to what degree violations of the strict validity assumption affect the estimation results. We show that, in case of moderate direct effects of the instrument on the dependent variable, the results do not deviate much from the benchmark case of no such effect (perfect validity of the instrument's exclusion restriction). In many cases, the size of the bias is smaller than the width of the 95% posterior interval for the effect of education on income. Thus, a violation of the strict validity assumption does not necessarily lead to results which are strongly different from those of the strict validity case. This finding provides confidence in the use of family background variables as instruments in income regressions.

JEL codes: C11, C13, C15, J30

Keywords: education; family background variables; earnings; income; instrumental variables; Bayesian analysis

1. Introduction

Education is a well-known determinant of income. However, the measurement of its influence suffers from endogeneity suspicion (Dickson and Harmon 2011; Griliches and Mason 1972, Blackburn and Neumark 1993, Webbink 2005). Instrumental variables (IV) regression is believed to yield an appropriate estimator in the presence of endogeneity (Angrist and Krueger 1991, Angrist et al. 1996, Card 2001). The difficulty arises in regard to finding an instrument that is strongly correlated with the endogenous variable *and* that satisfies the exclusion restriction (i.e., having no direct effect on income). In many studies, family background variables have been used as instruments for education (Blackburn and Neumark 1993, 1995, Parker and van Praag 2006). Compared with other instruments, family background variables have an advantage in that they are available in many datasets and that they are usually strongly correlated with the endogenous variable. Thus, the use of family background variables allows scholars to avoid a weak instruments problem (Bound et al. 1995). Recently, however, the use of family background variables, such as parents' or spouse's education levels, has been criticized (Trostel et al. 2002, Psacharopoulos and Patrinos 2004) because these variables do not meet the strict validity assumption that is required for IV regressions. Because family background variables are believed to have a *direct* effect on the respondent's income level, they cannot be used as an instrument for education. For example, it can be argued that family background variables are correlated with family wealth, which then may have a direct influence on the respondent's individual income. It may also be argued that family background variables are correlated with the preference for finding a job in a particular firm or industry. This preference may have a direct influence on the respondent's income.

This paper investigates the use of family background variables as instruments for education in detail. Using data from the 2004 German Socio-Economic Panel and Bayesian analysis, we analyze to what degree violations of the validity of family background variables as instruments have an effect on the estimation results of the IV model. Our research strategy is to begin with a tight prior around zero for the instrument's direct effect on the dependent variable, and subsequently to consider priors that allow for an increasing direct effect. As expected, our results demonstrate that if the instrument is assumed to have a sizeable direct effect of the instrument on the dependent variable, the coefficient of the IV model changes compared with the benchmark case where no direct effect of the instrument exists. For example, if the *direct* effect of the instrument (the father's education) on income, which works in addition to the instrument's *indirect* effect via own education (taking into account the effect of control variables), is 50% of the effect of a respondent's own education on income, then the estimated effect of an individual's own education on income decreases from $\beta=0.079$ to $\beta=0.044$. Indeed, the use of family background variables can lead to biased estimates. However, and more importantly, in many cases, the bias from using family background variables as instruments is lower than the width of the 95% posterior interval of the coefficient of the instrumented variable. Therefore, depending on the precision required of the estimated return to education – in terms of sign or level – and the strength of the assumed indirect effect, using family background variables is a viable option. In any case, the bias from using family background variables should be compared with the problems generated by alternative instrumentation strategies such as educational reforms (e.g., Oosterbeck and Webbink, 2007), which are rarely available. Across-the-board criticism of family background variables as instruments does not appear to be justified.

The remainder of the paper is organized as follows: Section 2 describes our Bayesian approach. Section 3 shows our econometric model. Section 4 introduces our dataset and variables. Section 5 shows our results, and Section 6 concludes.

2. Method

2.1 The Bayesian approach

We use Bayesian methods to estimate the IV model. Bayesian analysis of IV models has become increasingly popular over the last several years.[1] Bayesian methods rely on Bayes' theorem of probability theory (Bayes 1763). This theorem is given by

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}, \tag{1}$$

where θ represents the set of unknown parameters, and y represents the data. $p(\theta)$ is the prior density of the parameter, which may be derived from theoretical or other a priori knowledge. $p(y \mid \theta)$ is the likelihood function, which is the density (or probability in the case of discrete events) of the data *y* given the unknown parameter θ. $p(y)$ is the marginal likelihood, the marginal density of the data *y*, and finally, $p(\theta \mid y)$ represents the posterior density which is the density of the parameter θ given the data *y*. In Bayesian analysis, inference comes from the posterior distribution which states the likelihood of a particular parameter value. To determine the relationship between two variables, Bayesian analysis proceeds as follows. First, a priori beliefs about the relationship of interest are formulated (the prior distribution, $p(\theta)$). Next, a probability of occurrence for the data given certain parameter values is assumed (the likelihood function, $p(y \mid \theta)$). Third, data are used

[1] See Kleibergen and Zivot (2003) and Lancaster (2005) for an overview of Bayesian analysis of IV models and a comparison with classical IV regression.

to update these beliefs. The result is the posterior density, $p(\theta \mid y)$. It allows for statements in terms of likely and unlikely parameter values. We compute and analyze the means, standard deviations, and percentiles of the respective parameter distributions. These posterior properties are computed as the sample statistics of a large set of draws from the posterior distribution, which are obtained using Gibbs sampling.

2.2 Bayesian analysis in the instrumental variables model

An instrument makes sense if it satisfies the exclusion restriction and is strongly correlated with the endogenous explanatory variable. Bayesian analysis can be used to get reliable estimation results, when a researcher doubts whether the instrument satisfies these requirements.

Exclusion restriction of the instrument: In principle, an instrument should not be correlated with the error term. That is, the instrument should *not* have a direct effect on the dependent variable; it should only affect the dependent variable via the endogenous explanatory variable.[2] Bayesian analysis can be used to analyze the outcome if this crucial assumption is violated. Through Bayesian analysis, it is possible to incorporate a prior distribution for the instrument's direct effect on the dependent variable. In many situations, researchers believe that there is a direct effect that is *approximately* zero rather than one that is *exactly* zero. By beginning with a tight prior around zero and subsequently considering priors that allow for an increasing direct effect, one can analyze the robustness of the results with respect to the *validity assumption*.

Strength of the instrument: an instrument should be correlated with the endogenous explanatory variable. Preferably, the instrument should have a strong effect on the endogenous explanatory variable. Otherwise, one is faced with the issue of *weak instruments*, which may make it difficult to draw meaningful conclusions. Prior research has used Bayesian methods to generate reliable and accurate estimation results when weak instruments were used (Hoogerheide et al. 2007a, 2007b). However, as expected, our family background variable (having a correlation of 0.38 with education, p<0.001) certainly does not constitute a weak instrument. Therefore, our only concerns regarding the instrument concern the validity of the strict exclusion restriction.

3. Econometric model

We estimate the effect of education on income, expressed in the following equation:

[2] In the classical approach, one can perform the Sargan test on the validity of instruments (Kennedy 2008, pp. 154-156), if one has more instruments than endogenous explanatory variables. But this has no power (i.e., power is equal to size) against cases in which the instruments' direct effects on the dependent variable are proportional to their effects on the endogenous explanatory

$$income = \alpha_1 + \beta\ education + \sum_{i=1}^{m} \delta_{1i} w_i + u_1 \tag{2}$$

where *income* is the dependent variable, *education* is our explanatory variable of interest, w_i are the exogenous variables, α_1 is a constant, and u_1 is an error term with E(u_1)=0. However, the variable *education* is assumed to be endogenous, i.e. the variable is correlated with the error term u_1. IV regression is considered to be an appropriate estimator in the presence of endogeneity (Angrist et al. 1996; Card 2001). The basic idea is to find an instrument that is uncorrelated with the errors u_1 in the model but that is correlated with the endogenous variable *education*. In our case, this idea leads to the following equation:

$$education = \alpha_2 + \delta z + \sum_{i=1}^{m} \delta_{2i} w_i + u_2 \tag{3}$$

where *education* is the endogenous variable, *z* refers to the instrument used (the father's education), δ measures the strength of the relationship between the instrument and the endogenous variable, α_2 is a constant, and u_2 is an error term. The idea of the IV approach is to estimate both equations simultaneously. However, for this approach to work and to produce meaningful estimates, two conditions need to be satisfied: (1) cov(z, u_1) = 0 (i.e., the instrument should not be correlated with the error term of the performance equation), and (2) $\delta \neq 0$ (i.e., there should be a non-zero relationship between the instrument and the endogenous explanatory variable). The first condition refers to the *validity* of the instrument, whereas the second condition refers to the *strength* of the instrument.

To estimate the bias when using family background variables as instruments, we assume that there is a (small) *direct* effect γ of the instrument on income, which works in addition to the instrument's *indirect* effect via own education (and taking into account the effects of the control variables). Then equation (2) is rewritten as follows:

$$income = \alpha_1 + \beta\ education + \gamma\, z + \sum_{i=1}^{m} \delta_{1i} w_i + u_1 \tag{4}$$

Define $\tilde{\gamma} = \gamma / \beta$ as the ratio of the effects of the instrument and the respondent's education on income. We consider the posterior results for various values of $\tilde{\gamma}$, iteratively simulating from the

variable (a common situation). The data simply contain no information as to whether this particular violation is present or not, so *a priori* assumptions about this aspect are crucial for estimation results.

conditional posterior distributions by the Gibbs sampling method of Conley et al. (2012). We consider $\tilde{\gamma}$ rather than γ because it is easier to specify prior ideas about the relative effect of father's education vis-à-vis the effect of own education than to specify ideas about the absolute effect of the father's education. Appendix 1 summarizes our approach in a technical way.

With respect to the validity of the exclusion restriction, it is useful to consider two extreme cases. First, the model (2) can be considered an 'extreme' case of model (3) with $\gamma = 0$, where the father's education has not direct effect on income. A second extreme case is the situation, where $\gamma = \hat{\gamma}_{OLS}$, the OLS estimator in the model that results if we delete the explanatory variable education from (3). For $\gamma = \hat{\gamma}_{OLS}$, the posterior of β will be centered on 0, as the whole effect of one's father's education on one's income will be considered a direct effect such that the indirect effect of the father's education via one's own education will be 0. In other words, if γ approaches 0, the posterior mode of β will be close to $\hat{\beta}_{TSLS}$, the two-stage least squares estimator of β. However, if γ approaches $\hat{\gamma}_{OLS}$, the posterior mode of β approaches 0.

We assume that γ has a 'moderate' value ($\gamma < \hat{\gamma}_{OLS}$) for three reasons. First, one would expect one's own education to be more important than one's father's education. A direct effect of one's father's education on one's income may stem from various factors such as access to networks and connections, family wealth, and work values. However, if one's father's education affects these circumstances and attitudes, then it is implausible that one's own education would have no (or a smaller) effect. In other words, if education has a causal effect on earnings, then it is implausible that an additional year of education will benefit one's son or daughter, but not (or to a lesser extent) oneself. Second, using data from Chile, Patrinos and Sakellariou (2011) also consider a different instrument based on the introduction of a nationwide school choice system in 1981. These researchers estimate the direct effect γ of the father's education on income as merely 0.010, and $\tilde{\gamma} = \gamma / \beta$ as approximately 0.18. Third, evidence for a significantly positive β is found in multiple studies involving different instruments than family background variables. In the seminal paper by Angrist and Krueger (1991), instrumental variables based on quarter-of-birth dummies are used for multiple cohorts and for several model specifications (including a diverse number of control variables). For each cohort and model, a significantly positive estimate of $\hat{\beta}_{TSLS}$ is found.

4. Data and Variables

4.1 Data

Our estimations are based on a data set that is made available by the German Socio-Economic Panel Study (SOEP) at the German Institute for Economic Research (DIW), Berlin.[3] The SOEP is an annually conducted longitudinal household survey that provides amongst others detailed information about, for example, the participant's occupational status (e.g., employee or self-employed). To construct our estimation sample, we selected the year 2004 and those persons who are either self-employed or employed. After excluding observations with missing values, we obtained a data set containing 8,244 observations.

4.2 Variables

Income is measured as the natural logarithm of hourly wage, which is determined by dividing the annual gross income (in €) by the annual number of hours worked. The endogenous explanatory variable *education* is measured as the number of years of schooling. The instrument used in the education equation is the number of years of the father's secondary education. As the control variables, we included the respondent's *labor market experience* (in its linear and squared terms), *gender*, *wealth* (as proxied by the respondent's income from assets), *marriage status*, *nationality*, *duration of unemployment before employment*, whether the respondent lives in the former *West-Germany*, whether the respondent is self-employed, and industry dummies. For more details regarding the construction of the variables, see Table A1 of Appendix 2.

5. Results and Discussion

If we assume a perfectly valid instrument, that satisfies the exclusion restriction (i.e. $\tilde{\gamma}$ =0), then the posterior density of β is given by Figure 1. The posterior mean is 0.079; the 2.5% and 97.5% posterior percentiles are 0.066 and 0.092, respectively. Table A2 of Appendix 2 lists the detailed estimation results for all of the variables included in the instrumental variables regression. That is, an extra year of education leads on average to a 7.9% increase of the hourly wage.

[3] For more information about the SOEP, we refer to Wagner et al. (1993, 2007).

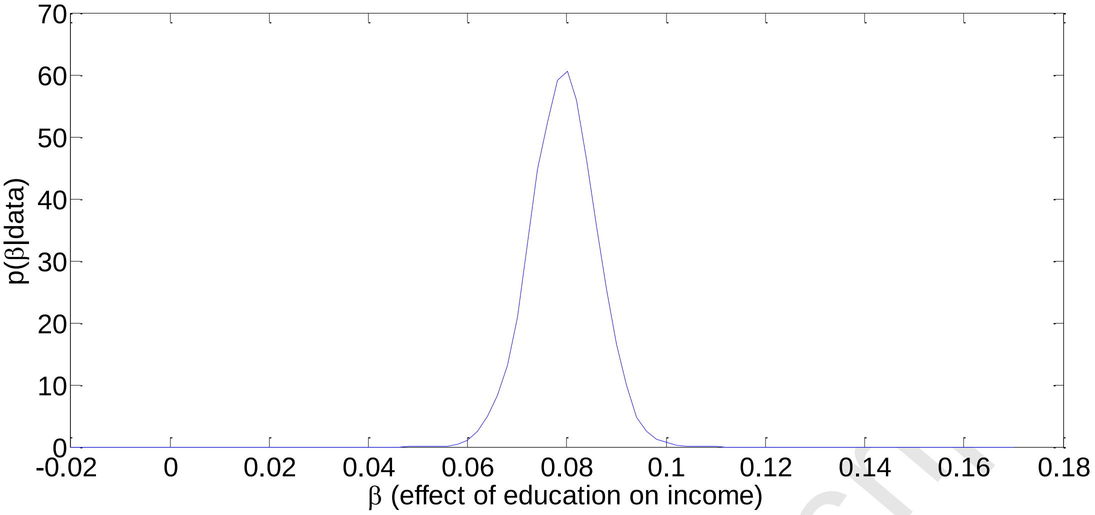

Figure 1: The posterior density $p(\beta \mid data)$ of β, the effect of (years of) education on the logarithm of income, when perfect validity of the instrument is assumed

Figure 2 illustrates the effect of choosing various values of $\tilde{\gamma}$ on the estimated posterior distribution of β. The vertical line at $\tilde{\gamma}$ =0 corresponds to the results illustrated in Figure 1. Table A3 in Appendix 2 gives a full account of the estimated posterior distribution of β.

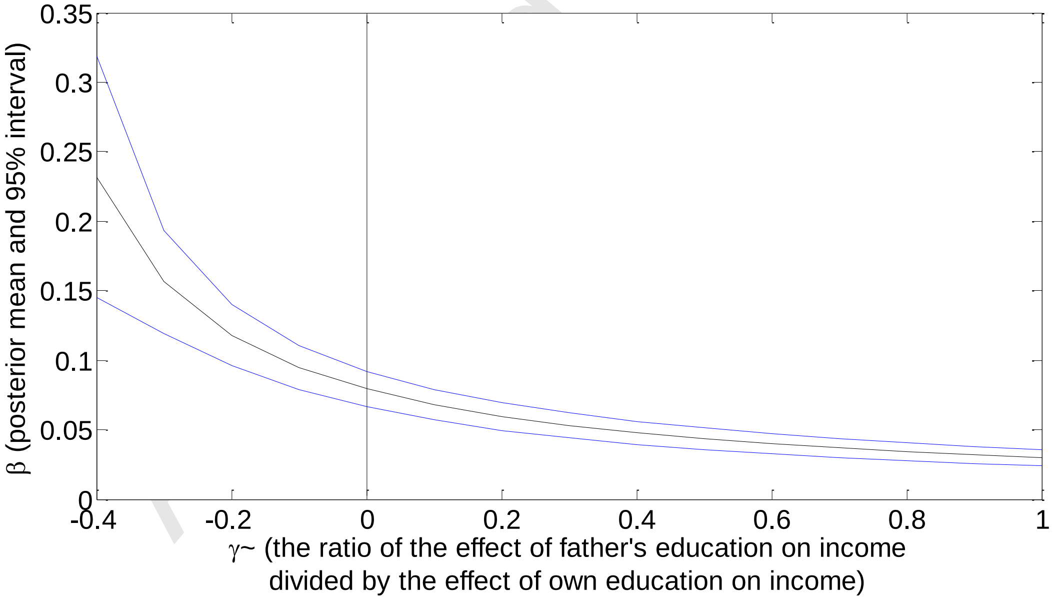

Figure 2: The mean and the 2.5% and 97.5% percentiles of the posterior distribution of β, the effect of (years of) education on the logarithm of income, for different values of $\tilde{\gamma}$, the ratio of the effect of the father's education on the logarithm of income to the effect of own education on the logarithm of income.

Notice that the posterior results do not change substantially if we choose plausible, small positive values of $\tilde{\gamma}$. For example, consider $\tilde{\gamma}$=0.35, which assumes that the effect of an extra year of the father's (secondary) education is 35% of the effect of an extra year of an individual's own education on income. For $\tilde{\gamma}$=0.35, the 2.5% posterior percentile of β is 0.042, which is 0.024 lower than the 2.5% percentile for $\tilde{\gamma}$=0. This difference of 0.024 is smaller than the 0.026 width of the 95% interval for $\tilde{\gamma}$=0. In other words, incorporating the uncertainty regarding the validity of the instrument leads to an increase in the posterior uncertainty of β that is no larger than the uncertainty that we face in the case of a perfectly valid instrument.

For increasingly positive values of $\tilde{\gamma}$, the posterior of β moves to 0; an increasingly large part of the *total* effect of the father's education on income is considered as a *direct* effect on income, rather than as an *indirect* effect via own education. This relationship is illustrated in Figure 3.

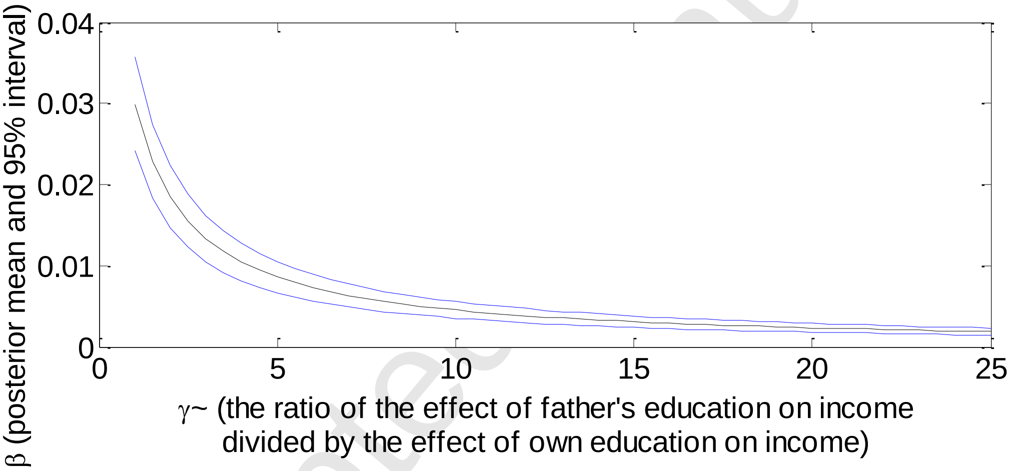

Figure 3: The mean and the 2.5% and 97.5% percentiles of the posterior distribution of β, the effect of (years of) education on the logarithm of income, for different values of $\tilde{\gamma}$, the ratio of the effect of the father's education on the logarithm of income to the effect of own education on the logarithm of income.

However, large values of $\tilde{\gamma}$ can be considered implausible. It is plausible that an individual's own education is more important than the father's education.

In addition, negative values of $\tilde{\gamma}$ are implausible because one may expect the effects of the father's education and one's own education to have the same (typically positive) sign. For increasingly negative values of $\tilde{\gamma}$, the posterior of β moves away from 0. In these (implausible) cases, one assumes that the effect of own education is particularly large because it 'compensates' for the negative effect of the father's education. The *total* effect of the father's education is then split into a *negative direct* effect and a more positive *indirect* effect via own education than in the case of a strictly valid instrument. Therefore, this assumption of $\tilde{\gamma}$ <0 would only make the estimated effect of own education on income stronger.

To assess the effect of incorrectly specifying the relation between family background and education, we have considered the following simulation experiment. We simulated 1,000 data sets for which expected education is a non-linear (convex or concave) function of family background, and estimated the linear IV model. As a result, the 95% posterior interval for β became somewhat wider, but this 95% posterior interval for β still contains the true value of β for 95% of the simulated data sets. Moreover, the posterior mean remained an (at least approximately) unbiased estimator of β. Therefore, this type of misspecification is *'not dangerous'* in the sense that it does not lead to a wrongly located, biased posterior of β. It may only be *more efficient,* in the sense that it leads to a more precise estimator of β (i.e. a smaller posterior standard deviation) by considering a different functional form for the effect of father's education on own education. This issue has been left as a topic for further research. Here, we mention only that adding the square of the father's education to the education equation (3) causes only minor changes, as compared with the linear IV model.

We now discuss the possible effects that some other types of specification errors may have. First, omitted variables in the income equation may cause endogeneity. The use of the IV model implies that this possibility has already been taken into account. Second, non-normality and heteroskedasticity can seriously affect the estimation results for weak instruments; see Hoogerheide, Opschoor and Van Dijk (2011). However, our instrument is very strong; father's education has a substantial and significant effect on own education. Third, we did not find significant evidence for the presence of a non-linear (quadratic) effect of education on (the logarithm of) income. The analysis of different functional forms for the income equation has been left as a topic for further research.

The framework of this paper used to evaluate the use of family background variables as instruments in income regressions can be extended easily to the case of a heterogeneous return to education. In this case, one can choose between two alternatives. One can specify a constant ratio $\tilde{\gamma}$ of the direct effect of father's education to the effect of one's own education, while these effects both vary across individuals. Alternatively, one can specify different ratios $\tilde{\gamma}$ for different groups of individuals.

6. Conclusions

Our results imply that the across-the-board criticism of family background variables as instruments is unjustified. Most researchers are very critical about the use of family background variables as instruments because these variables may have a direct effect on the respondent's income level, violating the exclusion restriction. Our Bayesian analysis investigates the severity of this problem. We find that relaxing the strict exclusion restriction on the family background instruments does lead to different results. However, the size of the bias is often smaller than the width of the 95% posterior interval of the education coefficient in the IV model. The results remain qualitatively similar even when the validity of the instrument would be substantially violated compared with the benchmark case where the instrument is assumed to be strictly exogenous. In conclusion, depending on the precision required of the estimated return on education, using the father's education as an instrument in an income regression is a viable option for solving the endogeneity problem with regard to education.

It is unclear how generalizable our findings are for other family background variables, such as the mother's or the spouse's education and the parent's social class or profession (Block et al. in press). As a general guideline to judge the suitability of a family background variable as an instrument, we propose the following two steps. First, one should perform a Bayesian analysis under the assumption that the exclusion restriction is exactly satisfied. If this analysis results in a 95% posterior interval for the coefficient of interest that is too wide for any practical purposes, then the instruments are apparently too weak for any useful inference to be drawn. If not, then in the second step, one should consider the priors that allow for an increasing direct effect of the instrument, to assess whether the potential bias is so large that the estimation results become unusable for one's particular research problem. Future research in this area could analyze the suitability of other family background variables as instruments. Other than using other family background variables as instruments, it would also be fruitful to learn more about the use of family background variables in other areas of education or labor market research, such as occupational choice decisions (Block et al. in press; Evans and Jovanovic 1989) or the determinants of job satisfaction (Fabra and Camisón 2009).

Our findings have practical implications for the empirical research in labor and education economics. Unlike other instruments, such as quarter of birth in combination with differences among schooling laws (Angrist and Krueger 1991; Deaton 2009; see Webbink 2005 for a survey), family background variables are available in many household surveys, including the German Socio-Economic Panel (SOEP), the British Household Panel Survey (BHPS), and the US panel study of income dynamics (PSID). Household surveys such as the European Community Household Panel (ECHP) enable cross-country IV regressions. Cross-country research about the economic effects of educational attainment (Ashenfelter et al. 1999; Behrman 1978; Brunello and Comi 2004; Flabbi et al. 2008*;* García-Mainar and Montuenga-Gómez 2005) is facilitated. Furthermore, family background variables are usually highly correlated with the respondent's level of education. Hence, the issue of having a (statistically) weak (Bound et al. 1995; Dickson and Harmon 2011) or (economically) irrelevant instrument (Deaton 2009) can be avoided.

References

Angrist, J.D., Krueger, A.B. 1991. Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics 106(4): 979-1014.

Angrist, J.D., Imbens, G.W., Rubin, D.B. 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91(434): 444-455.

Ashenfelter, O., Harmon, C., Oosterbeek, H. 1999. A review of estimates of the schooling/earnings relationship, with tests for publication bias. Labour Economics 6(4): 453-470.

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 53: 370-418.

Behrman, J.R. 1987. Schooling in developing countries: Which countries are the Over- and underachievers and what is the schooling impact? Economics of Education Review 6(2): 111-127.

Blackburn, M., Neumark, D. 1993. Omitted-ability bias and the increase in the return to schooling. Journal of Labor Economics 11(3): 521-544.

Blackburn, M., Neumark, D. 1995. Are OLS estimates of the return to schooling biased downward? Another look. The Review of Economics and Statistics 77(2): 217-230.

Block, J., Hoogerheide, L., Thurik, R. in press. Education and entrepreneurial choice: an instrumental variables analysis. International Small Business Journal: doi:10.1177/0266242611400470.

Brunello,G. 2002. Absolute risk aversion and the returns to education. Economics of Education Review 21(6): 635-640.

Brunello, G., Comi, S. 2004. Education and earnings growth: evidence from 11 European countries. Economics of Education Review 23(1): 75-83.

Bound, J., Jaeger, D.A., Baker, R.M. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association 90(430): 443-450.

Card, D. 2001. Estimating the returns to schooling: progress on some persistent econometric problems. Econometrica 69(5): 1127-1160.

Conley T.G., Hansen C.B., Rossi P.E. 2012. Plausibly exogenous. The Review of Economics and Statistics, 94(1): 260-272.

Deaton, A.S. 2009. Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development. NBER working paper 14690.

Dickson, D., Harmon, C. 2011. Economic returns to education: What we know, what we don't know, and where we are going—some brief pointers. Economics of Education Review 30(6): 1118-1122.

Evans, D.S. and Jovanovic, B. 1989. An estimated model of entrepreneurial choice under liquidity constraints. Journal of Political Economy 97(4): 808-827.

Fabra, M.E., Camisón, C. 2009. Direct and indirect effects of education on job satisfaction: A structural equation model for the Spanish case, Economics of Education Review 28(5): 600-610.

Flabbi, L., Paternostro, S., Tiongson. E.R. 2008. Returns to education in the economic transition: A systematic assessment using comparable data. Economics of Education Review 27(6): 724-740.

Geman, S., Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6(6): 721-741.

García-Mainar,I., Montuenga-Gómez,V.M. 2005. Education returns of wage earners and self-employed workers: Portugal vs. Spain. Economics of Education Review 24(2): 161-170

Griliches, Z., Mason, W.M. 1972. Education, income, and ability. Journal of Political Economy 80(3): S74-S103.

Hoogerheide, L.F., Kaashoek, J.F., Van Dijk, H.K. 2007a. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks. Journal of Econometrics 139(1): 154-180.

Hoogerheide, L.F., Kleibergen, F., Van Dijk, H.K. 2007b. Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. Journal of Econometrics 138(1): 63-103.

Hoogerheide, L.F., Opschoor, A., Van Dijk, H.K. 2011. A Class of Adaptive EM-based Importance Sampling Algorithms for Efficient and Robust Posterior and Predictive Simulation. Tinbergen Institute Discusion Paper 11-004/4.

Kennedy, P. 2008. A Guide to Econometrics. 6th edition. Blackwell Publishing: Oxford.

Kleibergen, F., Zivot, E. 2003. Bayesian and classical approaches to instrumental variable regression. Journal of Econometrics 114(1): 29-72.

Lancaster, T. 2005. An introduction to modern Bayesian econometrics. Blackwell Publishing: Oxford.

Oosterbeek, H., Webbink, D. 2007. Wage effects of an extra year of basic vocational education. Economics of Education Review 26(4): 408-419

Parker, S.C., Van Praag, C.M. 2006. Schooling, capital constraints, and entrepreneurial performance: the endogenous triangle. Journal of Business & Economic Statistics 24(4): 416-431.

Patrinos, H.A., Sakellariou, C. 2011, Quality of schooling, returns to schooling and the 1981 vouchers reform in Chile. World Development, doi:10.1016/j.worlddev.2011.04.018.

Psacharopoulos, G., Patrinos, A. 2004. Returns to investment in education: a further update. Education Economics 12(2): 111-134.

Trostel, P., Walker, I., Wooley, P. 2002. Estimates of the economic return to schooling for 28 countries. Labour Economics 9(1): 1-16.

Wagner, G.G., Burkhauser, R.V., Behringer, F. 1993. The English language public use file of the German Socio-Economic Panel Study. The Journal of Human Resources 28(2): 429-433.

Wagner, G.G., Frick, J.R., Schupp, J. 2007. The German Socio-Economic Panel Study (SOEP) – scope, evolution and enhancements. Schmollers Jahrbuch 127(1): 139-169.

Webbink, D. 2005. Causal effects in education. Journal of Economic Surveys, 19(4): 535-560.

Appendix 1: Bayesian analysis of an instrumental variables model

We write our model as:

$$y_j = x_j\beta + z_j\gamma + w_j{}'\delta_1 + u_{1j} \qquad \text{(j=1,…,n)} \qquad \text{(A1)}$$

$$x_j = \tilde{z}_j{}'\delta_2 + u_{2j} \qquad \text{(j=1,…,n)} \qquad \text{(A2)}$$

where $y_j =$ log(income) of individual j;

$x_j =$ education of individual j ;

$w_j =$control variables for individual j (including a constant term);

u_{1j}, u_{2j} = error term for individual j ;

$\tilde{z}_j =$ instruments (including education z_j of individual j's father and control variables w_j).

The error terms (u_{1j}, u_{2j}) are independently and normally distributed $(u_{1j}, u_{2j})' \sim N(0, \Sigma_u)$ with

$$\Sigma_u = \begin{pmatrix} \sigma_{u_1}^2 & \sigma_{u_1,u_2} \\ \sigma_{u_1,u_2} & \sigma_{u_2}^2 \end{pmatrix}.$$

We have $\gamma = \tilde{\gamma}\,\beta$ such that (A1) is

$$y_j = (x_j + z_j\tilde{\gamma})\beta + w_j{}'\delta_1 + u_{1j}\,. \qquad \text{(A3)}$$

We specify a flat prior for δ_1, δ_2: $p(\delta_1, \delta_2) \propto 1$; for β an uninformative, proper normal prior β ~ $N(\mu_{\beta,prior}, \sigma^2_{\beta,prior})$; for $\Psi_u \equiv \Sigma_u^{-1}$ an uninformative limit case of the Wishart distribution: $p(\Psi_u) \propto |\Psi_u|^{-3/2}$. The joint prior for $\theta = \{\beta, \delta_1, \delta_2, \Psi_u\}$ is therefore:

$$p(\theta) = p(\beta, \delta_1, \delta_2, \Psi_u) \propto \exp\left(-\frac{1}{2}\frac{(\beta - \mu_{\beta,\,prior})^2}{\sigma^2_{\beta,\,prior}}\right) |\Psi_u|^{-3/2}.$$

The likelihood is:

$$p(y, x \mid z, w, \theta) = (2\pi)^{-n} |\Psi_u|^{n/2} \exp\left[-\frac{1}{2}\sum_{j=1}^{n}\begin{pmatrix} y_j - (x_j + z_j\tilde{\gamma})\beta - w_j{}'\delta_1 \\ x_j - \tilde{z}_j{}'\delta_2 \end{pmatrix}' \Psi_u \begin{pmatrix} y_j - (x_j + z_j\tilde{\gamma})\beta - w_j{}'\delta_1 \\ x_j - \tilde{z}_j{}'\delta_2 \end{pmatrix}\right]$$

The posterior density kernel is:

$p(\theta \mid y, x, z, w) \propto$

$$|\Psi_u|^{(n-3)/2} \exp\left(-\frac{1}{2}\frac{(\beta-\mu_{\beta,prior})^2}{\sigma^2_{\beta,prior}}\right)\exp\left[-\frac{1}{2}\sum_{j=1}^{n}\begin{pmatrix} y_j-(x_j+z_j\tilde{\gamma})\beta-w_j'\delta_1 \\ x_j-\tilde{z}_j'\delta_2 \end{pmatrix}' \Psi_u \begin{pmatrix} y_j-(x_j+z_j\tilde{\gamma})\beta-w_j'\delta_1 \\ x_j-\tilde{z}_j'\delta_2 \end{pmatrix}\right]$$

.

We will use the notation $\theta_{-\eta}$ to denote the set of all parameters in θ except for η. We apply the Gibbs sampler (Geman and Geman (1984)) to simulate the draws from the posterior distribution, iteratively sampling from the full conditional posteriors:

(i) $\Psi_u \mid y, x, z, w, \theta_{-\Psi_u} \sim$

$$Wishart\left(n, \left[\sum_{j=1}^{n}\begin{pmatrix} y_j-(x_j+z_j\tilde{\gamma})\beta-w_j'\delta_1 \\ x_j-\tilde{z}_j'\delta_2 \end{pmatrix}\begin{pmatrix} y_j-(x_j+z_j\tilde{\gamma})\beta-w_j'\delta_1 \\ x_j-\tilde{z}_j'\delta_2 \end{pmatrix}'\right]^{-1}\right).$$

(ii) $(\beta, \delta_1')' \mid y, x, z, w, \theta_{-(\beta,\delta_1)} \sim N\left(\mu_{\beta,\delta_1}, V_{\beta,\delta_1}\right)$ with

$$V_{\beta,\delta 1} = \left[\begin{pmatrix} \left(\sigma^2_{\beta,\,prior}\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix} + (\sigma^2_{u_1|u_2})^{-1}\sum_{i=1}^{n}\begin{pmatrix} (x_j+z_j\tilde{\gamma})^2 & (x_j+z_j\tilde{\gamma})w_j' \\ (x_j+z_j\tilde{\gamma})w_j & w_jw_j' \end{pmatrix}\right]^{-1}$$

$$\mu_{\beta,\delta 1} = V_{\beta,\delta 1}\left[\begin{pmatrix} \left(\sigma^2_{\beta,\,prior}\right)^{-1}\mu_{\beta,\,prior} \\ 0 \end{pmatrix} + (\sigma^2_{u_1|u_2})^{-1}\sum_{j=1}^{n}\begin{pmatrix} x_j+z_j\tilde{\gamma} \\ w_j \end{pmatrix}\left(y_j-\mu_{u_{1j}|u_{2j}}\right)\right]$$

where $\mu_{u_{1j}|u_{2j}} = (x_j-\tilde{z}_j'\delta_2)\sigma_{u_1,u_2}/\sigma^2_{u_2}$ and $\sigma^2_{u_1|u_2} = \sigma^2_{u_1} - \sigma^2_{u_1,u_2}/\sigma^2_{u_2}$.

(iii) $\delta_2 \mid y, x, z, w, \theta_{-\delta_2} \sim N\left(\mu_{\delta_2}, V_{\delta_2}\right)$ with $V_{\delta_2} = \left[\left(\sigma^2_{u_2|u_1}\right)^{-1}\sum_{j=1}^{n}\tilde{z}_j\tilde{z}_j'\right]^{-1}$

$$\mu_{\delta_2} = V_{\delta_2}\left[(\sigma^2_{u_2|u_1})^{-1}\sum_{j=1}^{n}\tilde{z}_j\left(x_j-\mu_{u_{2j}|u_{1j}}\right)\right]$$

where $\mu_{u_{2j}|u_{1j}} = (y_j-(x_j+\tilde{z}_j\tilde{\gamma})\beta-w_j'\delta_1)\,\sigma_{u_1,u_2}/\sigma^2_{u_1}$ and $\sigma^2_{u_2|u_1} = \sigma^2_{u_2} - \sigma^2_{u_1,u_2}/\sigma^2_{u_1}$.

Appendix 2

Table A1: Description of variables

Variable	Description
	Categorical variables
Male	Dummy for an individual who is male
Non-German	Dummy for an individual who is Non-German by nationality
Married	Dummy for an individual who is married
West Germany	Dummy for an individual who lives in West Germany
Industry dummies	Dummies for the following industries: agriculture (NACE 1,2, and 5); manufacturing (NACE 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 40, 41, 96, 97, and 100); retail (NACE 51 and 52); hotel and restaurant (NACE 55); financial services (NACE 65, 66, 67, and 70); firm services (NACE 50, 72, and 74); construction (NACE 45); health (NACE 85); transportation (NACE 60, 61, 62, and 63); culture, sports, and leisure (NACE 92); and other (NACE 10, 11, 12, 13, 14, 64, 71, 73, 75, 80, 90, 91, 93, 95, 98, and 99)
Self-employed	Dummy for an individual who is self-employed
	Continuous variables and ordinal variable
Income	Log (annual gross income [in €] divided by annual hours worked [in hrs.])
Education	Years of schooling (including time at university)
Education of respondent's father	Years of education required to reach the father's secondary school certificate: 9 years for "Hauptschule", 10 years for "Realschule", 12 years for "Fachhochschulreife", 13 years for "Abitur".
Experience	Current age minus age at first job
Unemployment duration	Number of months that an individual has been unemployed in his or her entire working life before entering self-employment
Wealth	Log (household income from assets)

Table A2: Posterior results of the instrumental variables model for a perfectly valid instrument

Dependent variable: *income* (=log hourly wage)

	Mean and standard dev. of posterior distribution		**Percentiles of posterior distribution**			
Variables	**Mean**	**Std. Dev.**	**2.5%**	**97.5%**	**25%**	**75%**
Education (instrumented) [1]	0.079	0.007	0.066	0.092	0.075	0.084
Experience	0.034	0.002	0.031	0.038	0.033	0.035
Experience²/10	-0.006	0.000	-0.007	-0.005	-0.006	-0.006
Unemployment duration	-0.049	0.006	-0.060	-0.038	-0.053	-0.045
Male	0.138	0.013	0.113	0.163	0.130	0.147
Married	0.051	0.012	0.028	0.076	0.043	0.059
Non-German	0.033	0.033	-0.032	0.097	0.011	0.055
Wealth	0.029	0.003	0.022	0.035	0.026	0.031
West Germany	0.328	0.013	0.302	0.354	0.319	0.337
Agriculture [2]	-0.393	0.049	-0.490	-0.299	-0.426	-0.360
Manufacturing [2]	0.076	0.019	0.039	0.113	0.063	0.088
Retail [2]	-0.104	0.025	-0.153	-0.056	-0.120	-0.087
Hotel and Restaurant [2]	-0.246	0.045	-0.332	-0.159	-0.276	-0.216
Financial Services [2]	0.164	0.025	0.117	0.213	0.148	0.181
Firm Services [2]	-0.014	0.021	-0.055	0.026	-0.028	0.000
Construction [2]	-0.051	0.030	-0.110	0.009	-0.072	-0.030
Health [2]	0.043	0.020	0.004	0.081	0.030	0.056
Transportation [2]	-0.012	0.034	-0.078	0.053	-0.035	0.011
Culture, Sports, and Leisure [2]	-0.068	0.044	-0.154	0.019	-0.098	-0.039
Self-employed	0.001	0.019	-0.036	0.038	-0.011	0.014

Notes: N = 8,244 observations; data source: GSOEP
The posterior moments and percentiles are estimated on the basis of 10,000 simulated draws, that are generated using the Gibbs sampling method (using the pseudo-random number generators in Matlab^{TM}) after a burn-in of 1000 discarded draws. A non-informative proper prior is specified for β, a standard normal distribution N(0,1). Non-informative improper priors are specified for the other parameters. The results are robust with respect to considerable deviations in the non-informative prior specification.

[1] Instrument used: *education of respondent's father*

[2] Reference category: industry category *other*.

Table A3: Posterior distribution of β, the effect of education (years) on the logarithm of income, for different values of $\tilde{\gamma} = \gamma / \beta$

	Mean and standard dev. of the posterior distribution of β		Percentiles of the posterior distribution of β	
$\tilde{\gamma}$	**Mean**	**Std. Dev.**	**2.5%**	**97.5%**
-0.40	0.232	0.044	0.145	0.319
-0.30	0.157	0.019	0.113	0.194
-0.20	0.118	0.011	0.096	0.140
-0.10	0.095	0.008	0.079	0.111
0	0.079	0.007	0.066	0.092
0.05	0.073	0.006	0.061	0.085
0.10	0.068	0.006	0.057	0.079
0.15	0.064	0.005	0.053	0.074
0.20	0.060	0.005	0.050	0.070
0.25	0.056	0.005	0.047	0.066
0.30	0.053	0.005	0.044	0.062
0.35	0.050	0.004	0.042	0.059
0.40	0.048	0.004	0.039	0.056
0.45	0.046	0.004	0.038	0.054
0.50	0.044	0.004	0.036	0.051
0.60	0.040	0.004	0.033	0.047
0.70	0.037	0.004	0.030	0.044
0.80	0.034	0.003	0.028	0.041
0.90	0.032	0.003	0.026	0.038
1	0.030	0.003	0.024	0.036
2	0.019	0.002	0.015	0.022
3	0.013	0.002	0.011	0.016
4	0.010	0.001	0.008	0.013
5	0.009	0.001	0.007	0.011
10	0.005	0.0005	0.004	0.006
25	0.002	0.0002	0.001	0.002
100	0.0005	0.0001	0.0004	0.0006
1000	0.00005	0.00001	0.00004	0.00006

Notes: $\tilde{\gamma}$ = the ratio of γ, the direct effect of the father's education on the logarithm of income, to β, the effect of one's own education on the logarithm of income.

The posterior moments and percentiles are estimated on the basis of 10,000 simulated draws that are generated using the Gibbs sampling method after a burn-in of 1000 discarded draws. A non-informative proper prior is specified for β, a standard normal distribution N(0,1). Non-informative improper priors are specified for the other parameters. The results are robust with respect to considerable deviations in the non-informative prior specification. The control variables are the explanatory variables given in Table A2, with the exception of education and the inclusion of a constant term.