

Hoogerheide, Lennart; Block, Joern H.; Thurik, Roy

Working Paper

Family Background Variables as Instruments for Education in Income Regressions: A Bayesian Analysis

Tinbergen Institute Discussion Paper, No. 10-075/3

Provided in Cooperation with:

Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Hoogerheide, Lennart; Block, Joern H.; Thurik, Roy (2010) : Family Background Variables as Instruments for Education in Income Regressions: A Bayesian Analysis, Tinbergen Institute Discussion Paper, No. 10-075/3, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:

<https://hdl.handle.net/10419/87059>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



TI 2010-075/3

Tinbergen Institute Discussion Paper

Family Background Variables as Instruments for Education in Income Regressions: A Bayesian Analysis

Lennart Hoogerheide^{1,2,4}

Joern H. Block^{1,3,5}

Roy Thurik^{1,2,3,6,7}

¹ Erasmus University Rotterdam, ² Tinbergen Institute, ³ Centre for Advanced Small Business Economics, ⁴ Econometric Institute, the Netherlands;

⁵ Technische Universität München,; Germany;

⁶ EIM Business and Policy Research, Zoetermeer, the Netherlands;

⁷ Max Planck Institute of Economics, Jena, Germany.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam

Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam

Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at
<http://www.tinbergen.nl>.

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

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

^a Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands, lhoogerheide@ese.eur.nl.

^a Centre for Advanced Small Business Economics, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands, block@ese.eur.nl; Technische Universität München, München, Germany.

^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 of the 2004 German Socio-Economic Panel and Bayesian analysis in order to analyze to what degree violations of the strong 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). The size of the bias is in many cases smaller than the standard error of education's estimated coefficient. Thus, the violation of the strict validity assumption does not necessarily lead to strongly different results when compared to the strict validity case. This provides confidence in the use of family background variables as instruments in income regressions.

First version: May 2010

Current version: July 2010

File name: Family background variables_V13

Save date: 27-7-2010 14:27:00

JEL codes: C11, C13, C15, J24, J30

Corresponding author: Joern Block

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

1. Introduction

Education is a well-known driver of income. The measurement of its influence, however, suffers from endogeneity suspicion (Griliches and Mason 1972, Blackburn and Neumark 1993, Web-bink 2005). Instrumental variables (IV) regression is considered to yield an appropriate estimator in the presence of endogeneity (Angrist and Krueger 1991, Angrist et al. 1996, Card 2001). The difficulty arises when it comes to finding an instrument that is strongly correlated with the endogenous variable *and* valid. In many studies, family background variables have been used as instruments for education (Blackburn and Neumark 1993, 1995, Parker and van Praag 2006). Compared to other instruments, family background variables have the advantage that they are available in many datasets and that they are usually strongly correlated with the endogenous variable, thus avoiding a weak instruments problem (Bound et al. 1995). Recently, however, the use of family background variables, such as parents' or spouse's education, has been criticized (Trostel et al. 2002, Psacharopoulos and Patrinos 2004): these variables would not meet the strict validity assumption that is required for IV regressions. Family background variables are believed to have a *direct* effect on the respondent's income level and therefore 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 income of an individual. It may also be argued that family background variables are correlated with the preference to find a job in a particular firm or industry, which then may have a direct influence on the respondent's income.

This paper investigates this problem, i.e., the possible invalidity of family background instruments, in detail. Using data of 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 show that when assuming a sizeable direct effect of the instrument on the dependent variable, the coefficient of the IV model changes compared to the benchmark case where no direct effect of the instrument exists. For example, if the *direct* effect of the instrument (father's education) on income, which works in addition to the instrument's *indirect* effect via own education (and 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 standard error of the coefficient of the instrumented variable. So, depending on the required precision 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 against alternative instrumentation strategies which are often hardly available. The across-the-board criticism of family background variables as instruments does not seem 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; 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 years.¹ Bayesian methods rely on Bayes' theorem of probability theory (Bayes 1763). This theorem is given by

$$\Pr(\theta | y) = \frac{\Pr(y | \theta) \Pr(\theta)}{\Pr(y)}, \quad (3)$$

where θ represents the set of unknown parameters, and y represents the data. $\Pr(\theta)$ is the prior density of the parameter θ that may be derived from theoretical or other a priori knowledge. $\Pr(y | \theta)$ 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 . $\Pr(y)$ is the marginal likelihood, the marginal density of the data y , and finally, $\Pr(\theta | y)$ 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 find out about a relationship between two variables, Bayesian analysis proceeds in the following steps: first, a priori beliefs about the relationship of interest are formulated (the prior distribution, $\Pr(\theta)$). Next, a probability of occurrence of the data given parameter values is assumed (the likelihood function, $\Pr(y | \theta)$). In a second step, data are used to update these beliefs. The result is the posterior density, $\Pr(\theta | y)$. 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 by a Gibbs sampling approach.

2.2 A Bayesian analysis of instruments

Bayesian analysis can be used to find out whether an instrument makes sense. An instrument makes sense if it is valid and strongly correlated with the endogenous explanatory variable.

Validity of the instrument: In principle, an instrument should not be correlated with the error term, i.e., it should *not* have a direct effect on the dependent variable—its only effect on the dependent variable should be via the endogenous explanatory variable.² Bayesian analysis can be used to analyze what happens when 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* equal to 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, it 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

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

² In the classical approach, one can perform the Sargan test on the validity of instruments (Kennedy 2008, pp. 154-156), if we have more instruments than endogenous explanatory variables. But this has no power (i.e., power equal to size) against cases where the instruments' direct effect on the dependent variable is proportional to their effect on the endogenous explanatory variable, a situation that is often plausible. 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.

meaningful conclusions. Bayesian analysis can be used to find out whether a weak instruments problem exists; it helps to identify weak instruments and the problems they cause regarding the accuracy of the estimates (Hoogerheide et al. 2007a, 2007b). Weak instruments are defined as those instruments that are only weakly correlated with the endogenous variable. When the dataset is large enough (and a statistically significant but weak correlation between the instrument and the endogenous variable can be found), classical IV regression using a weak instrument would result in a highly significant estimate for the endogenous variable. However, the estimate is likely to be strongly biased. In other words, one may obtain a seemingly precise but incorrect estimate (see the discussion of problems with weak instruments in Bound et al. 1995, who comment on Angrist and Krueger 1991). Bayesian analysis does not change the strength of an instrument (i.e., its correlation with the endogenous variable), but it allows for a precise statement of how the strength of the instrument influences the preciseness of the estimates. This is the case because the result of Bayesian analysis is not a point estimate (which is then either significant or not) but a probability distribution of the model coefficients, which is not only correct for hypothetical infinite data sets but also for real finite data samples. Hence, a Bayesian analysis provides a warning in case of weak instruments: using a weak instrument would lead to a ‘wide’ posterior distribution, and therefore, the danger of computing a seemingly precise but incorrect estimate is not present.

3. Econometric model

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

$$income = \alpha_1 + \beta education + \sum_{i=1}^m \beta_i x_i + u_1, \quad (1)$$

where *income* is the dependent variable, *education* is our explanatory variable of interest, x_i are exogenous variables, α_1 is a constant, and u_1 is an error term with $E(u_1)=0$. The variable *education*, however, 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 leads to the following equation:

$$education = \alpha_2 + \delta z + \sum_{i=1}^m \delta_i x_i + u_2, \quad (2)$$

where *education* is the endogenous variable, z refers to the instrument used (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. Yet, 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; the second condition refers to the *strength* of the instrument.

To estimate the bias when using family background variables as instruments, we suppose that there is a (small) *direct* effect γ of the instrument on income, which works in addition to the instru-

ment’s *indirect* effect via own education (and taking into account the effect of control variables). Then equation (1) reads as follows:

$$income = \alpha_1 + \beta education + \gamma z + \sum_{i=1}^m \beta_i x_i + u_1 \quad (3)$$

Define $\tilde{\gamma} = \gamma / \beta$ as the ratio of the effects of the instrument and the respondent’s education on income. We consider posterior results for different values of $\tilde{\gamma}$, iteratively simulating from the conditional posterior distributions by the Gibbs sampling method of Conley et al. (2008). The reason for considering $\tilde{\gamma}$ rather than γ is that it is easier to specify prior ideas about the relative effect of father’s education vis-à-vis own education than about the absolute effect of father’s education.

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 a longitudinal household survey conducted annually that provides amongst others detailed information about the participant’s occupational status (e.g., employee or self-employed). To construct our estimation sample, we make use of the year 2004 and select those persons who are either self-employed or employed. After excluding observations with missing values, we obtained a data set with 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 €) with the number of annual 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 father’s secondary education. As control variables, we included the respondent’s *labor market experience* (in its linear and squared term), *gender*, *wealth* (as proxied by the respondent’s income from assets), *status of marriage*, *nationality*, *duration of unemployment before employment*, whether the respondent lives in former *West-Germany*, whether the respondent is self-employed, as well as industry dummies. For more details regarding the construction of the variables, see Table A1 of our Appendix.

5. Results

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. Table A2 shows the detailed estimation results for all 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, refer to Wagner et al. (1993, 2007).

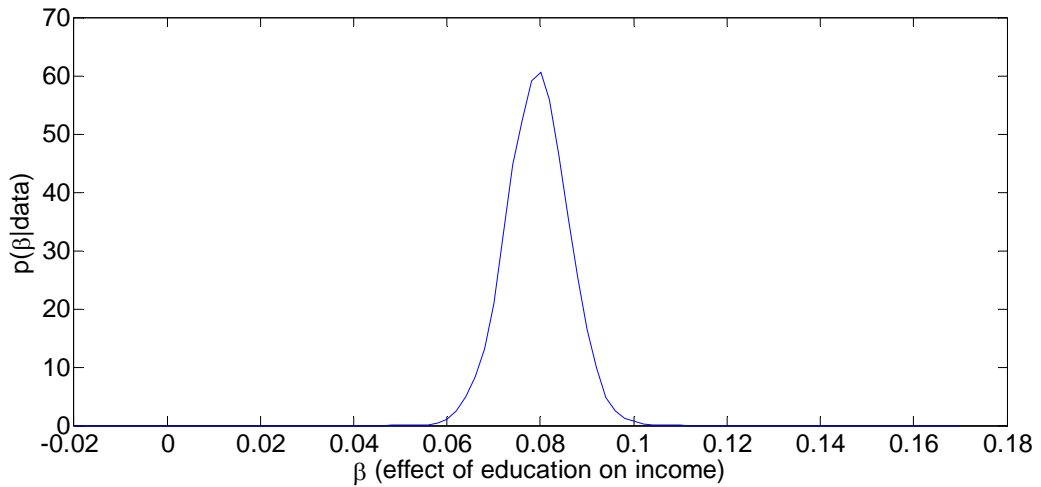


Figure 1: posterior density $p(\beta | data)$ of β , the effect of (years of) education on the logarithm of income, when perfect validity of the instrument is assumed

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

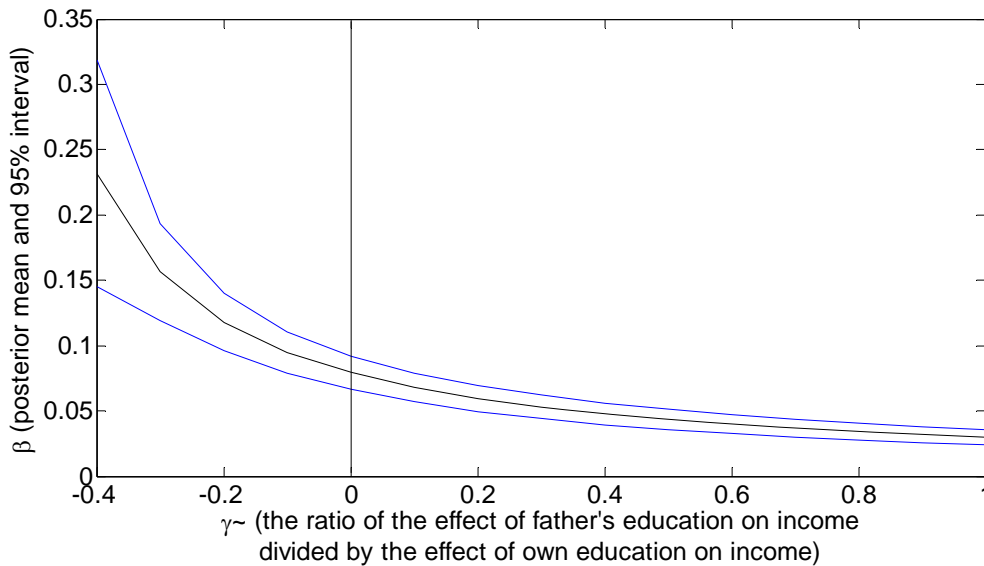


Figure 2: mean and 2.5% and 97.5% percentiles of posterior distribution of β , the effect of (years of) education on the logarithm of income, for different values of $\tilde{\gamma}$, the ratio of the effect of father's education on logarithm of income divided by the effect of own education on logarithm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

Notice that posterior results do not change substantially if we choose plausible, small positive values of $\tilde{\gamma}$. For example, consider $\tilde{\gamma}=0.35$, which amounts to the assumption that the effect of an extra year of 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 on the validity of the instrument leads to an increase of the posterior uncertainty on β that is no larger than the uncertainty that we face in the case of a perfectly valid instrument.

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

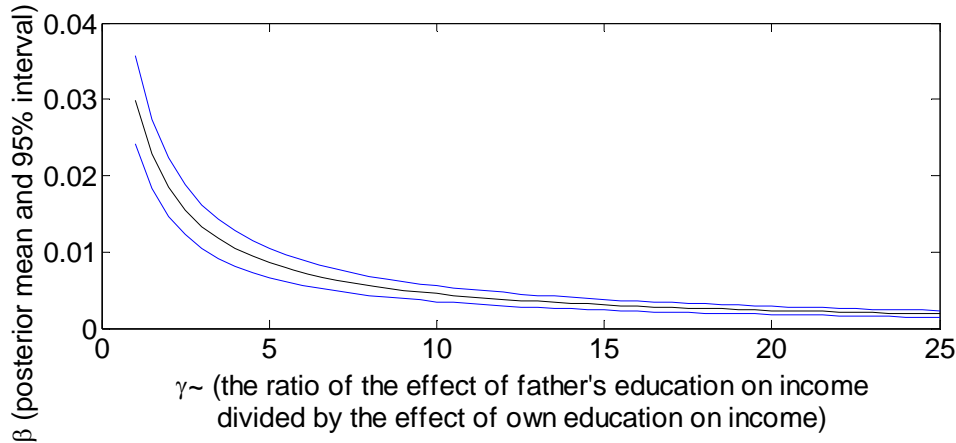


Figure 3: mean and 2.5% and 97.5% percentiles of posterior distribution of β , the effect of (years of) education on the logarithm of income, for different values of $\tilde{\gamma}$, the ratio of the effect of father’s education on logarithm of income divided by the effect of own education on logarithm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

Also negative values of $\tilde{\gamma}$ are implausible, since one may expect the effects of father’s and 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 such (implausible) cases, one assumes that the effect of own education is particularly large since it ‘compensates’ for the negative effect of father’s education. The *total* effect of 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. So, this assumption of $\tilde{\gamma} < 0$ would only make the estimated effect of own education on income higher.

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 since they may have a direct effect on the respondent’s income level. Our Bayesian analysis investigates the severity of this problem: relaxing the strict validity assumption on the family background instruments does indeed lead to different results. However, the size of the bias is in many cases smaller than the standard error of education’s estimated coefficient in the IV model. The results remain qualitatively similar even when the validity of the instrument would be

substantially violated compared to the benchmark case where the instrument is assumed to be strictly exogenous. In conclusion, depending on the required precision of the estimated return to education, using family background variables is a viable option to solve the endogeneity problem with regards to education. This result has practical implications for empirical research in labor economics. Unlike other instruments, such as quarter of birth in combination with differences between schooling laws (Angrist and Krueger 1991; 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). Furthermore, family background variables are usually highly correlated with the respondent's level of education. Hence the issue of having a weak instrument (Bound et al. 1995) 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.
- 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.
- 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. 2008. Plausibly exogenous. working paper. <http://ssrn.com/abstract=987057>.
- 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.
- 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.
- Griliches, Z., Mason, W.M. 1972. Education, income, and ability. *Journal of Political Economy* 80(3): S74-S103.
- Heckman, J.J., Urzua, S. 2009. Comparing IV with structural models: what simple IV can and cannot identify. NBER working paper 14706.
- 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.
- 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.
- 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.
- 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

Table A1: Description of variables

Variable	Description
Categorical variables	
Male	Dummy for individual who is male
Non-German	Dummy for individual who is Non-German by nationality
Married	Dummy for individual who is married
West Germany	Dummy for individual who lives in West Germany
Industry dummies	Dummies for the following industries: agriculture (NACE 1,2,5), 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, 100), retail (NACE 51, 52), hotel and restaurant (NACE 55), financial services (NACE 65, 66, 67, 70), firm services (NACE 50, 72, 74), construction (NACE 45), health (NACE 85), transportation (NACE 60, 61, 62, 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, 99)
Self-employed	Dummy for 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 (incl. time at university)
Respondent's father's education	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	Months that an individual has been unemployed in her entire working life before entering self-employment
Wealth	Log (household income from assets)

Table A2: Posterior results of instrumental variables model in case of a perfectly valid instrument

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

Variables	Mean and standard dev. of posterior distribution		Percentiles of posterior distribution			
	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
Experience ² /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 by the Gibbs sampling method (using the pseudo random number generators in MatlabTM) 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. Results are robust with respect to considerable deviations in the non-informative prior specification.

¹ Instrument used: *respondent's father's education*

² Reference category: industry category *other*.

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

$\tilde{\gamma}$	Mean and standard dev. of posterior distribution of β		Percentiles of posterior distribution of β	
	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 $\tilde{\gamma}$ = ratio of γ , the direct effect of father's education on logarithm of income, divided by β , the effect of own education on logarithm of income.

The posterior moments and percentiles are estimated on the basis of 10,000 simulated draws that are generated by 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. Results are robust with respect to considerable deviations in the non-informative prior specification. The control variables are the explanatory variables in Table A2, except for education, and including a constant term.