Block, Jörn H.; Hoogerheide, Lennart F.; Thurik, A. Roy

Working Paper

Are Education and Entrepreneurial Income Endogenous and Do Family Background Variables Make Sense as Instruments? A Bayesian Analysis

SOEPpapers on Multidisciplinary Panel Data Research, No. 329

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Block, Jörn H.; Hoogerheide, Lennart F.; Thurik, A. Roy (2010) : Are Education and Entrepreneurial Income Endogenous and Do Family Background Variables Make Sense as Instruments? A Bayesian Analysis, SOEPpapers on Multidisciplinary Panel Data Research, No. 329, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/150874

Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

www.diw.de

SOEP The German Socio-Economic Panel Study

SOEPpapers

on Multidisciplinary Panel Data Research

329

Jörn H. Block • Lennart F. Hoogerheide • A. Roy Thurik

Are Education and Entrepreneurial Income Endogenous and Do Family Background Variables Make Sense as Instruments? A Bayesian Analysis

Berlin, November 2010

SOEPpapers on Multidisciplinary Panel Data Research
at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | urahmann@diw.de

Are education and entrepreneurial income endogenous and do family background variables make sense as instruments? A Bayesian analysis

Jörn H. Block [a], Lennart F. Hoogerheide [b], A. Roy Thurik [c]

[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.

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

[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: Education is a well-known driver of (entrepreneurial) income. The measurement of its influence, however, suffers from endogeneity suspicion. For instance, ability and occupational choice are mentioned as driving both the level of (entrepreneurial) income and of education. Using instrumental variables can provide a way out. However, three questions remain: whether endogeneity is really present, whether it matters and whether the selected instruments make sense. Using Bayesian methods, we find that the relationship between education and entrepreneurial income is indeed endogenous and that the impact of endogeneity on the estimated relationship between education and income is sizeable. We do so using family background variables and show that relaxing the strict validity assumption of these instruments does not lead to strongly different results. This is an important finding because family background variables are generally strongly correlated with education and are available in most datasets. Our approach is applicable beyond the field of returns to education for income. It applies wherever endogeneity suspicion arises and the three questions become relevant.

First version: February 2010
Current version: February 2010

File name: endogeneity of education Block Hoogerheide Thurik_V31.doc

Save date: 2/26/2010 10:50:00 AM

JEL codes: C11, L26, M13, J24

Keywords: Education, income, entrepreneurship, self-employment, endogeneity, instrumental variables, Bayesian analysis, family background variables

Acknowledgements: Comments by Herman van Dijk are gratefully acknowledged. This paper has been written in cooperation with the research program SCALES, carried out by EIM and financed by the Dutch Ministry of Economic Affairs. It benefitted from a short stay of Roy Thurik at Université de Caen Basse-Normandie.

Corresponding author: Lennart Hoogerheide, Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, lhoogerheide@ese.eur.nl.

1. Introduction

The advent of a knowledge-intensive economy, together with the recognition that such an economy requires a vital entrepreneurial sector (Audretsch and Thurik 2001, Audretsch 2007), has renewed attention to the effect of education on entrepreneurial choice and financial performance (Bosma et al. 2004). Moreover, human capital (including age, experience and formal education) is shown to be a major determinant of entrepreneurial performance when compared to other explanatory variables (Parker 2009: chapter 13).[1] Finally, of the many factors known to influence entrepreneurial choice and performance (Van der Sluis et al. 2008, Grilo and Thurik 2008, Parker 2009), education is popular among politicians because it can be influenced by education policy (European Commission 2003, OECD, 2009).

The measurement of the influence of education, however, suffers from endogeneity suspicion because there may be a correlation between the education variable and the error term. Neglecting this endogeneity can lead to unreliable estimation results even in large samples because estimators of the model parameters are inconsistent.[2] There are several possible causes for this endogeneity. First, omitted factors may exist that impact both education and entrepreneurial performance. For instance, ability and occupational choice are mentioned as phenomena driving both the level of entrepreneurial income and education (Griliches and Mason 1972, Blackburn and Neumark 1993).[3] Second, measurement errors associated with education can also lead to endogeneity (Griliches 1977, Angrist and Krueger 1991). These measurement errors push the estimated return to education toward zero because they lead to variation in the education variable that has no effect on income.[4] Other causes include reverse causality, autoregression with autocorrelated errors and non-random samples (Kennedy 2008).

In the related literature that addresses occupational performance in general (including wage earners), many methods have been proposed to deal with this problem (for summaries, see Ashenfelter et al. 1999, Card 2001, Webbink 2005). Instrumental variables regression (hereafter, IV regression) is considered to be an appropriate estimator in the presence of endogeneity (Card 2001), independent of what its possible cause may be. The idea behind IV regression is to find one variable (or more) that is strongly correlated with the endogenous explanatory variable but has no direct effect on the dependent variable beyond its indirect effect via the endogenous regressor and to use this variable as an instrument to estimate the effect of the endogenous variable. A wide range of variables have been proposed as instruments for education, e.g., differences in schooling laws across regions, regional college proximity, father's education, or number of siblings (Angrist and Krueger 1991, Card 2001, Webbink 2005, Parker and Van Praag 2006).[5] To our knowledge, Parker and Van Praag (2006) and Van Praag et al. (2009) are the only studies that employ IV techniques to estimate the effect of education (formal schooling) on entrepreneurial income. Parker and Van Praag (2006) find that its effect on entrepreneurial income increases from 7.2% in the ordinary least squares (OLS) model to a 13.7% reward of additional income per extra year of education in the IV model (an increase of about 90%).[6]

---

1 Parker's overview of the literature on entrepreneurial choice points to age and experience as playing prominent and positive roles, while education plays a lesser role (Parker 2009: chapter 4).

2 In the linear regression model, neglecting the endogeneity of explanatory variables is a more severe error than ignoring heteroskedasticity (a non-constant variance of the error term) or non-zero correlation between the errors. The latter leads to unreliable standard errors of the parameter estimates but leaves the least-squares estimator consistent.

3 If an explanatory variable is positively correlated with an omitted variable that has a positive effect, then its positive effect is overestimated. For instance, individuals with higher ability typically obtain higher education levels but also earn higher income given a certain education level. If this were the only reason for the endogeneity of education, ignoring the issue would lead to an overestimated return to schooling.

4 In analyzing the effect of education on income, the variable for years of education is arguably a poor proxy for education level, even more so for accumulated human capital. Further, on-the-job-training is often ignored, and the length of education is sometimes simply misreported.

5 Webbink (2005) discriminates between instruments based upon 'controlled experiments,' 'natural experiments,' 'institutional rules' and 'natural variation.'

6 The influences of specific entrepreneurship education programs are discussed in Oosterbeek et al. (2009).

Our study contributes to this literature on the effect of education on entrepreneurial income using data on self-employed German individuals. Three important questions are at the heart of our analysis: whether endogeneity of the education variable is really present, whether it matters and whether the selected family background instruments make sense. The second contribution this paper makes is dealing with these questions using Bayesian techniques, which allows us to be very precise about the degree of endogeneity and its effect and to tolerate (small) deviations from the strict validity assumption of the instruments. The third contribution we make is to show that there is a positive effect of education on entrepreneurial income using a longitudinal household survey that was conducted in Germany.

Concerning the first two questions, the results obtained using Bayesian techniques are straightforward: the relationship between education and entrepreneurial success is indeed endogenous. In addition, the impact of endogeneity on the estimated relationship between education and income is sizeable. The bias-corrected mean estimate is 75% higher than the uncorrected estimate. We conclude that the endogeneity problem needs to be addressed when analyzing the relationship between education and entrepreneurial income.

For our third question about family background instruments, we investigate the degree to which the accurate measurement of the causal effect and its interpretation depend upon the instrument.[7] The use of one or more instrumental variables is a general solution in the sense that it works regardless of the reason for the endogeneity. An instrumental variable must satisfy certain criteria. *First*, it should be a *valid* instrument that is not correlated with the error term, which amounts to the *exclusion* restriction that the instrument should not have a direct effect on the dependent variable; its only effect on the dependent variable is via its effect on the endogenous explanatory variable. *Second*, the instrument should be *statistically relevant* in that it is correlated with the endogenous explanatory variable. Preferably, the instrument has a strong effect on the endogenous explanatory variable. With weak instruments, it may be difficult to draw meaningful conclusions (Bound et al. 1995): classical confidence intervals or Bayesian posterior intervals may be too wide to base any useful statement on the estimation results.[8] *Third*, the instrument should not only be correlated with the endogenous explanatory variable, but it should affect it *in a relevant way*. As an example, the quarter of birth dummies may only affect education for those who desire to leave school as early as the law allows it. Hence, the resulting estimate may only be relevant for those who are interested in this particular effect, i.e., the effect of (secondary) education laws. The estimate may provide much less insight into the effect of a Master's degree (versus a Bachelor's degree).

To analyze the impact of family background variables as an instrument on our results, we build on a recent contribution of Conley et al. (2008), who developed Bayesian methods of performing inference while relaxing the exclusion restriction, thereby providing tools for applied researchers who want to proceed with less than perfectly valid instruments. Their Bayesian approach has several advantages over its classical counterparts. It allows in a natural way for the inclusion of prior beliefs in the extent of the instruments' invalidity. Moreover, the Bayesian inference method easily allows for dependence of the instruments' invalidity on other model parameters. Thus, we extend their Bayesian approach to panel data and show that relaxing the strict validity assumption on the family background instruments does *not* lead to strongly different results. The results remain qualitatively the same when the validity of the instrument is substantially violated compared to the benchmark case where the instrument is assumed to be strictly exogenous. We conclude that the wholesale critique of the use of family background variables is unjustified (Trostel et al. 2002, Psacharopoulos and Patrinos 2004), which is an important result for applied researchers because family background variables are available in many household surveys.

---

[7] Two recent papers (Deaton 2009; Heckman and Urzua 2009) are so critical of the well-established method of IV (and the type of instruments used in most studies) that they even caught the attention of *The Economist* (2009).

[8] In the education-income literature, famously weak instruments include Angrist and Krueger's (1991) quarter of birth dummies. Hoogerheide et al. (2007b) show that only in a few southern U.S. states do these instruments have a strong effect on education. For other regions, wide confidence intervals or posterior intervals are found regarding the return to schooling.

The remainder of the paper is organized as follows. Section 2 summarizes the theoretical arguments for why education and entrepreneurial income are presumed to be endogenous. Section 3 explains our Bayesian test for endogeneity and the quality of the instrument. Section 4 describes our data and variables. Section 5 gives the results of our empirical analysis, which are then discussed in Section 6. The conclusion is given in Section 7. All technicalities are discussed in the four attached appendices.

2. Theory: why might education and entrepreneurial income be endogenous?

A variable is called endogenous if it appears as a causal variable in an econometric equation system while it is correlated with the errors in the model. A possible reason is that there is a variable that has an impact on the causal and the dependent variables at the same time. If the correlation between the omitted variable and the causal variable is positive, the effect of the causal variable is overestimated; if the correlation is negative, the effect is underestimated. Previous studies from the labor economics literature have argued that the relationship between education and labor income is endogenous (Blackburn and Neumark 1993; for a summary, see Ashenfelter et al. 1999). We argue that this endogeneity problem from omitted variables could be even stronger for entrepreneurs. There are at least three types of omitted variables that are difficult to capture in an entrepreneurial income equation.

The first group of omitted variables concerns the relationship between education and (entrepreneurial) ability. Certain factors, such as intelligence or stamina, lead to both higher education levels *and* higher levels of entrepreneurial income. If these ability measures are missing in the income equation, the estimate of education is biased (Griliches and Mason 1972, Blackburn and Neumark 1993).

The second group of omitted variables concerns the occupational choice itself (Block et al. 2009, Parker 2009). Riley (1979, 2002) argues that if employers demand a high degree of education from their employees as an otherwise unproductive screening device or if potential employees use a high degree of education to signal their ability to potential employers, then individuals who want to become entrepreneurs and do not face this requirement should have a lower degree of education. If this argument holds true, then education and the willingness to become an entrepreneur are correlated negatively, and the estimate of education is biased. A related argument can be constructed from the jack-of-all-trades theory (Wagner 2003; Lazear 2004, Lazear 2005), according to which entrepreneurs are generalists who do not excel at a particular skill but are competent at many skills. Accordingly, individuals who want to become entrepreneurs do not engage in a lengthy specialized education but instead choose a rather short generalist education. Again, the length of education and willingness to become an entrepreneur are negatively correlated, and if this is not controlled for, the estimate of education will be biased.

The third group of omitted variables refers to the types of motivations for becoming an entrepreneur. Since 2001, the Global Entrepreneurship Monitor (GEM) has discussed necessity and opportunity entrepreneurship (Reynolds et al. 2002, Block and Sandner 2008; Block and Wagner 2010). Opportunity entrepreneurs are those who start their businesses to pursue an entrepreneurial opportunity, whereas necessity entrepreneurs start their businesses due to a lack of alternative employment options. For example, they might have experienced a long period of unemployment before starting their business (Meager 1992, Pfeiffer and Reize 2000). Almost by definition, necessity entrepreneurs have few alternatives for earning their living other than through entrepreneurship. The reason for this situation may be related to having a low level of education. In this case, necessity entrepreneurship and the level of education are correlated negatively, which leads to biased estimates of the impact of education on entrepreneurial income.

3. Method

3.1. Instrumental variables regression

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

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

where *entrepreneurial income* is the dependent variable, *education* is our variable of interest, x_i are exogenous variables, α is a constant, and u_1 is an error term with E(u_1)=0. For the theoretical reasons discussed above, 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:

$$education = \gamma + \delta z + u_2, \quad (2)$$

where *education* is the endogenous variable, *z* refers to the instrument used, δ measures the strength of the relationship between the instrument and the endogenous variable, γ 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 variable). The first condition refers to the *validity* of the instrument; the second condition refers to the *strength* of the instrument.

Our IV model is actually somewhat more involved than (1)-(2). It involves multiple instruments and describes panel data with (random) individual effects. See appendix 3.

3.2. The Bayesian approach

We use Bayesian methods to estimate the IV regression and a simple linear model used as a benchmark case. Bayesian analysis of IV models has become increasingly popular over the last years (for an overview and a comparison to classical IV regression, see Kleibergen and Zivot 2003, Lancaster 2005). Bayesian methods rely on Bayes' theorem of probability theory (Bayes 1763). This theorem is given by

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

where θ represents the set of unknown parameters, and y represents the data. $\Pr(\theta)$ is the prior distribution of the parameter θ that may be derived from theoretical or other a priori knowledge. $\Pr(y \mid \theta)$ 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. $\Pr(y)$ is the marginal likelihood, the marginal density of the data y, and finally, $\Pr(\theta \mid y)$ 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 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 these a priori beliefs is assumed (the likelihood function, $\Pr(y \mid \theta)$). In a second step, data are used to update these beliefs. The result is the posterior distribution, $\Pr(\theta \mid y)$.

This posterior distribution gives a probability density function of the relationship between these two variables. That is, it allows for statements in terms of likely and unlikely parameter values. We report 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 the Gibbs sampling approach described in appendix 3.

3.3. A Bayesian analysis of endogeneity

We estimate the IV regression described above with Bayesian methods and use the results to analyze: (1) whether endogeneity is present, (2) whether it matters, and (3) whether the use of family background variables as instruments makes sense.

3.3.1. Question 1: Is endogeneity present?

To answer the question of whether endogeneity is present, we calculate the posterior distribution of the correlation between education and the error term of the income equation, that is, u_1. A correlation (much) different from zero would indicate a (strong) degree of endogeneity. Furthermore, a positive correlation suggests that the endogeneity derives from factors that have similar influences on both income and education (e.g., the influence of the omitted variable is positive in both cases); in turn, a negative correlation shows that the sources of endogeneity are factors that have differing influences on education and income.[9]

3.3.2. Question 2: Does endogeneity matter?

Even if endogeneity is present, it may only have a weak effect on the coefficients. To find out whether this is the case, we estimate the effect of education on entrepreneurial income in both an IV and a non-IV model and compare the results of these two models. We conclude that endogeneity matters when the results of the IV model deviate strongly from the results of the non-IV model, e.g., when the respective posterior distribution functions have very different properties.

3.3.3. Question 3: Does a particular instrument make sense?

Bayesian analysis can be used to find out whether an instrument makes sense. An instrument makes sense if it is (1) valid, (2) strongly correlated with the endogenous variable, and (3) effects the endogenous variable in a relevant way.

Validity: 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.[10] 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 increasingly large direct effect, one can analyze the robustness of the results with respect to the validity assumption.

Statistically relevant: 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 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

[9] Lancaster (2005, p. 332-335) describes this approach of measuring the level of endogeneity in more detail.

[10] In the classical approach, one can perform the Sargan test on the validity of instruments (Kennedy 2008, pp. 154-156), 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 necessarily crucial for estimation results.

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 very 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. Using a weak instrument would lead to a non-normal and wide distribution, and therefore, the danger of computing a precise but incorrect estimate is not present.

The endogenous variable should be affected in a relevant way: An instrument should not only be correlated with the endogenous explanatory variable; it should affect it in a relevant way. That is, if an instrument only affects the endogenous variable for a particular subsample of the population, then we should interpret the estimation results as referring to that particular subpopulation, which may imply that estimates are of limited utility. Examples of such subsamples are geographical regions, social groups or subsets of observations that have the endogenous variable's value falling in a certain interval. In these cases, the *local* character of the estimation results should be stressed. In this paper, we will consider simple descriptive graphs of the data to assure that the instruments' effect is not restricted to a particular subsample of the population.

4. Data and variables

4.1. Data

Our estimations are based on an unbalanced panel data set that is made available by the German Socio-Economic Panel Study (SOEP) at the German Institute for Economic Research (DIW), Berlin.[11] 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 years from 1984-2004 and select those persons who are self-employed. After excluding observations with missing values, we obtained a panel data set with 8,288 observations from 2,280 individuals.

4.2. Variables

Entrepreneurial 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 two instruments used in the education equation are based on the *respondent's father's education* (measured by his secondary school certificate). The respondent's father's education falls in one of three categories: category 1 refers to "Hauptschule" (corresponding to approximately 9 years of primary and secondary education); category 2 refers to either "Realschule" (approximately 10 years) or "Fachhochschulreife" (approximately 12 years); category 3 refers to "Abitur" (approximately 13 years).[12] The three categories provide two category indicators (with category 1 as the reference category). 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*

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

[12] The three categories have 5,490, 1,245, and 1,553 observations, respectively. Within the second category, there are only 82 observations with father's education "Fachhochschulreife", which is the reason for not including this as a separate category. The effect of "Fachhochschulreife" appeared close to "Realschule", so that these are included in the second category.

marriage, *nationality*, *duration of unemployment before self-employment*, whether the respondent lives in former *West-Germany*, time dummies as well as industry dummies. For more details regarding the construction of the variables, see Table A1 in appendix 1.

5. Results

Table 1 shows the results of the Bayesian random effects IV regression, while Table 2 shows the results of a non-IV Bayesian random effects model used as a benchmark case.[13] In both models, we used a non-informative prior for β, a normally distributed prior with mean zero and standard deviation of one.[14] The use of alternative priors and their effects on the results is discussed in the robustness section. For each coefficient, we report the mean and standard deviation of the posterior distribution as well as the 50% and 95% density intervals.

--- Insert Tables 1 through 3 and Figures 1 through 6 here ---

5.1. Question 1: Is endogeneity present in the relationship between entrepreneurial income and education?

As stated above, we measure the level of endogeneity by calculating the correlation between education and the error term in the entrepreneurial income equation. The result is clear: the posterior distribution of the correlation has a mean of -0.122 and a 95% posterior interval that lies between -0.199 and -0.044 (Table 1). Figure 1 shows the marginal posterior density function of this correlation graphically. A value of zero for the correlation is clearly rejected. Based on these results, we conclude that education is an endogenous variable in the entrepreneurial income equation.

5.2. Question 2: Does endogeneity matter for the size of the estimated education coefficient?

In the IV model, the posterior distribution of the variable *education* has a mean value of 0.105; the 95% posterior interval lies between 0.079 and 0.130 (Table 1). In other words, with a probability of 95%, an additional year of education results in an increase of hourly gross earnings of between 7.9% and 13%. In the non-IV Bayesian random effects model, which we use as a benchmark case, the mean coefficient of the variable *education* is only 0.060, and the 95% density interval is between 0.053 and 0.066 (Table 2). Compared to the linear model, the IV model estimate of education is about 4.5 percentage points higher (an increase of about 75%). We conclude that using IV methods makes a great difference in the results.

5.3. Question 3: Do family background variables as instruments make sense?

As described above, the use of Bayesian methods allows us to discover whether an instrument makes sense. That is, we can comment on its validity, whether it is statistically relevant and whether it affects the endogenous variable in a relevant way.

Validity: The instruments should not have a direct effect on the dependent variable; their only effect on the dependent variable should be via their effect on the endogenous explanatory variable. However, there is a fundamental uncertainty regarding the validity of this assumption; see footnote

[13] To perform this analysis, we used the software package Matlab™. The exact Matlab™ code used to run the regressions can be requested from the corresponding author. It is also possible to estimate Bayesian models with the software package WinBUGS. The software is freely available online from the Medical Research Council at the University of Cambridge website. See http://www.mrc-bsu.cam.ac.uk/bugs (retrieved on February 7, 2009). The exact code needed can be found in Lancaster (2005, p. 321).

[14] For all other parameters, we specify non-informative improper priors. See Appendix 3.

10. Fortunately, a Bayesian analysis in which one specifies different prior distributions of the direct effect of the instruments on the dependent variable can provide a vital check of the robustness of estimation results.

We investigate the results for several priors in which the direct effect is *approximately* zero rather than *exactly* equal to zero. In each case, we specify a prior for the *relative* size of the direct effect of one extra year of the respondent's father's secondary education compared with one extra year of the respondent's own education. For this ratio, we consider a normal prior $N(0,\tau^2)$ with standard deviation $\tau = 0.05$ or $\tau = 0.10$. The latter corresponds to the assumption that one's father's secondary education has a *direct* effect on one's income (*in addition to* the effect that is captured in one's own education and controls) between approximately -20% and 20% of the effect of one's own education, which seems to be a rather conservative assumption. Appendix 4 describes the simulation method that we used for the computation of the posterior results. Table 3 and Figure 4 show that estimation results change little if we substitute the exact validity assumption for this conservative assumption of approximate validity. There is only a rather small change in the posterior distribution of β. The difference between the posterior distributions in the models with and without IV remains huge; the presence of endogeneity still matters. We also consider truncated normal priors $N(0,\tau^2)$, where the effect of father's education is restricted to that with the same sign as own education (typically positive). In this case, the specification of $\tau = 0.10$ reflects a 95% prior belief that the ratio of one's father's and one's own education's effects on income is between 0% and approximately 20%. Again, estimation results change little (see Table 3 and Figure 5).

We conclude that estimation results are robust with respect to the assumption of exact validity of the instruments.

Statistically relevant: Preferably, the instruments have a strong effect on the endogenous explanatory variable. Figure 3 shows that our instruments, the category indicators of father's secondary education, clearly have this property. The effect of the respondent's father's education on the respondent's education is clearly substantial, so the problem of *weak instruments* is not present here (Bound et al. 1995).

The endogenous variable should be affected in a relevant way: Figure 6 illustrates that the father's secondary school instruments have explanatory power for education at different levels. A respondent's father's high level of education (category 3) implies higher probabilities of a medium versus low level and a high versus medium level of the respondent's education. That is, the instruments do not merely affect the respondent's education level for a subgroup of individuals with low or high education or for years of schooling around a particular value. The instruments are relevant for estimating the effect of education on income for self-employed individuals in general, across the entire spectrum of education levels.

5.4. Robustness checks

5.4.1. Using alternative priors

To check whether our results are robust to prior specification, we estimated our Bayesian IV model also with alternative priors. The results did not change. For example, using a normally distributed prior with mean -1 (instead of mean 0) and variance 1 for the variables *respondent's father's education* and *education* results in a mean coefficient of 0.105 for the variable *education* and a mean coefficient of -0.122 for the correlation between the error terms (which exactly correspond to the results reported above). Also differences between the graphs of the posterior densities under these two priors are not visible. These results confirm that our standard normal prior on β is truly non-informative.

5.4.2. Comparison with the results of a classical IV model

We also compare the results of our Bayesian IV model to the results of a classical IV regression,[15] which is shown in Table A2 in Appendix 2. The coefficient of the education variable in the classical IV model is 0.113 (with $p<0.001$), which is only slightly different from the mean coefficient of 0.105 that results from the Bayesian model. The effect of education in the IV model was found to be 88% higher than in the OLS model (with the Bayesian approach: 75%). The Durbin-Wu-Hausman test (Hausman 1978) for endogeneity is significant at the 0.1% significance level.

5.5. Results with regard to control variables

The results concerning the control variables entered into the income equation are in line with our predictions and do not differ between the linear and the IV models. We report only the results of the IV model. Experience is found to have a positive effect in its linear term and a negative effect in its squared term. Male self-employed individuals earn more than female self-employed individuals (mean coefficient: 0.304). The same applies to West German versus East German self-employed individuals (mean coefficient: 0.450). A long period of unemployment before entering self-employment is found to have a negative impact on self-employment earnings (mean coefficient: -0.034). The level of wealth of the self-employed before entering self-employment is found to have a positive effect (mean coefficient: 0.037). The impacts of the variables *married* and *non-German* are unclear, which is illustrated by the high standard deviations of the posterior distributions and the change of signs in the 95% density interval.

6. Discussion

6.1. Are education and entrepreneurial income endogenous?

To establish whether education and entrepreneurial income are endogenous, a benchmark is needed. What is the minimum level of correlation between the causal variable and the errors in the model necessary to fulfill the criterion of endogeneity? In a theoretical study using simulated data, Hoogerheide et al. (2007a) introduce *strong endogeneity* when the correlation is ρ=0.99, *medium endogeneity* when ρ=0.5, and *no endogeneity* when ρ=0. Using this classification, the relationship between education and entrepreneurial success should be classified as weak. We estimate the mean correlation of the error terms to be 0.122 (95% density interval is between -0.199 and -0.044, Table 1). The result of the Durbin-Wu-Hausman test for endogeneity, which is estimated for the classical IV model, goes in the same direction. The test is highly significant ($p<0.001$, Table A2).

6.2. Is endogeneity a problem and, if relevant, how can it be solved?

We find that the endogeneity-corrected estimate of education is about 75% higher than the uncorrected estimate, which is similar to the 90% difference found in Parker and Van Praag (2006). Although the degree of endogeneity is rather low, the increase in the effect of the estimated coefficient is sizeable. We conclude that it is essential to address endogeneity and treat it as a problem, in particular when the size–and not only the direction–of the effect is analyzed.

Our findings have an important implication for dealing with the endogeneity of education in the entrepreneurial income equation. First, the results of the Bayesian and the classical IV models are similar: the estimated coefficient for education in the classical IV model (β=0.113, Table A1) is close to the posterior mean in the Bayesian IV model (β=0.105, Table 1). Thus, contingent on the low degree of endogeneity and the strength of the instrument, classical IV methods perform consid-

15 Here we used the method of two-stage least squares, which in this case of exact identification is equivalent to the method of limited information maximum likelihood (LIML).

erably well and appear to be well suited to solving the endogeneity problem between education and entrepreneurial performance.

6.3. Family background variables as instruments

The use of family background variables, such as parents' or spouse's education, has been criticized (Trostel et al. 2002, Psacharopoulos and Patrinos 2004). Scholars have argued that these variables do 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. Our results show that this criticism is unjustified. Although prior research indeed shows that family background variables have an effect on the respondent's income level and returns to education (Lam and Schoeni 1993, Altonji and Dunn 1996), our Bayesian analysis demonstrates that this is not a problem for the estimate of education in the entrepreneurial income equation: relaxing the strict validity assumption on the family background instruments does *not* lead to strongly different results. The results remain qualitatively the same 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, family background variables can be used 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 changes in schooling laws (Angrist and Krueger 1991), family background variables are available in many household surveys, including the German Socio-Economic Panel (SOEP) and the British Household Panel Survey (BHPS). Furthermore, family background variables are usually highly correlated with the respondent's level of education. Thus, it is possible to avoid the issue of having a weak instrument (Bound et al. 1995).

6.4. Bayesian methods in labor market research

This paper concerns an application where the use of Bayesian methods provides additional insights for labor market research. We use Bayesian methods to measure the degree of endogeneity between two variables in a very precise way, which would not have been possible with classical methods. The same approach is relevant in other areas of labor market research where endogeneity is suspected to be a problem, e.g., the effect of capital constraints on entrepreneurial income or occupational choice (Hurst and Lusardi 2004, Parker and Van Praag 2006). Furthermore, as Hoogerheide et al. (2007a, 2007b) show, Bayesian methods may be used to evaluate the strength of instruments in IV regressions. Furthermore, the fact that Bayesian methods do not rely on asymptotic theory and statistical tests to disprove a particular theory provides new perspectives; because it is not necessary to reach a particular significance level, the Bayesian approach is ideally suited to testing competing theories. Bayesian analysis simply makes a statement regarding which theory is more likely. For the same reason, Bayesian methods have an advantage over classical methods in cases where the sample is rather small or multicollinearity is an issue (both of which decrease significance levels). Finally, Bayesian methods may be used as a robustness check for results obtained with traditional methods.

6.5. The influence of education on entrepreneurial income

The literature on entrepreneurial income and returns to education is less straightforward than the much older and broader literature regarding employee income and returns to education (Bosma et al. 2004, Van der Sluis et al. 2008, Parker 2009). Parker demonstrates that less clear results may be expected in the entrepreneurial context (Parker 2009: chapter 13). He justifies this argument in a variety of ways, such as measurement issues (tax evasion, income under-reporting, high non-response rates, incomparable legal structures, nonpecuniary benefits, heterogeneities in entrepreneurial activities, etc.), the roles of inequality and volatility of entrepreneurial income, observational regularities that demonstrate that formal education is unrelated to higher levels of entrepreneurial success and conflicting predictions from human capital theory. Basing his analysis largely on Van

der Sluis et al.'s (2005, 2008) two meta-analyses of the numerous empirical findings, Parker concludes that the returns to education for entrepreneurs are high, "but much more careful econometric modeling of entrepreneurs' payoffs is needed to establish the robustness and generality of this finding" (Parker 2009: 382). Our paper contributes precisely to this research gap: we find significant and positive payoffs to education through careful use of new econometric techniques.

7. Conclusion

Education and human capital in general (including age and experience in addition to type and duration of schooling) are commonly analyzed variables in the entrepreneurship and labor economics literature. We used German data to establish the returns to education for entrepreneurs. We show that they are highly positive and that the standard OLS model produces different estimates than an IV model. The findings that the returns to education are positive for entrepreneurs and that endogeneity is an unresolved issue in the literature on returns to education are not new. Instead, we focused on three questions: whether endogeneity is really present, whether it matters and whether the selected instruments make sense. Using Bayesian methods, we find that the relationship between education and entrepreneurial income is indeed endogenous and that the impact of endogeneity on the estimated relationship between education and income is sizeable. We do so using family background variables and show that relaxing the strict validity assumption of these instruments does not lead to strongly different results, which is an important finding because family background variables are generally strongly correlated with education and are available in most datasets. Our approach is relevant beyond the field of the returns to education for income. It applies wherever the three questions arise, e.g., in research about entry into entrepreneurship (Bates 1995, Blanchflower, 2000, Van der Sluis et al. 2008), exit from entrepreneurship (Block and Sandner 2008, Van Praag 2003, Stam, Thurik and Van der Zwan 2010) or the roles of education and human capital in government-initiated start-up programs (Pfeiffer and Reize 2000, Sandner et al. 2008, Dencker et al. 2009). However, careful investigation of the three questions is also essential beyond the world of entrepreneurship research. From a practical perspective, our findings are of particular interest for policy makers and institutions that evaluate the benefits of entrepreneurial education programs (Oosterbeek et al. 2009). Due to the endogeneity problem discussed in this paper, the effect of those programs can be seriously under or overestimated.

References

Altonji, J.G., Dunn, T.A. 1996. The effects of family characteristics on the return to education. The Review of Economics and Statistics 78(4): 692-704.

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.

Audretsch, D.B. 2007. Entrepreneurship capital and economic growth, Oxford Review of Economic Policy 23(1): 63-78.

Audretsch, D.B., Thurik, A.R. 2001. What is new about the new economy: sources of growth in the managed and entrepreneurial economies, Industrial and Corporate Change 10(1): 267-315.

Bates, T. 1995. Self-employment entry across industry groups. Journal of Business Venturing 10(2): 143-156.

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.

Blanchflower, D. 2000. Self-employment in OECD countries. Labour Economics 7(5): 471-505.

Block, J., Sandner, P. 2008. Necessity and opportunity entrepreneurs and their duration in self-employment: evidence from German micro data. Journal of Industry, Competition and Trade 9(2): 117-137.
Block, J., Hoogerheide, L., Thurik, A.R. 2009. Education and entrepreneurial choice: an instrumental variables analysis. Tinbergen Institute Discussion Paper 2009-088/3.
Block, J., Wagner, M. 2010. Necessity and opportunity entrepreneurs in Germany: characteristics and earnings differentials. Schmalenbach Business Review, forthcoming.
Bosma, N., Van Praag, M., Thurik, A.R., De Wit, G. 2004. The value of human and social capital investments for the business performance of startups. Small Business Economics 23(3): 227-236.
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.
European Commission, 2003. Green Paper: Entrepreneurship in Europe. Brussels, Directorate General Enterprise and Industry.
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.
Grilo, I., Thurik, A.R. 2008. Determinants of entrepreneurial engagement levels in Europe and the US. Industrial and Corporate Change 17(6): 1113-1145.
Dencker, J.C., Gruber, M., Shah, S.C. 2009. Individual and opportunity factors influencing job creation in new firms. Academy of Management Journal 52(6): 1125-1147.
Hausman, J.A. 1978. Specification tests in econometrics. Econometrica 46(6): 1251-1271.
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.
Hurst, E., Lusardi, A. 2004. Liquidity constraints, household wealth, and entrepreneurship. Journal of Political Economy 112(2): 319-347.
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.
Koop, G. 2003. Bayesian Econometrics, Wiley: Chichester, UK.
Lam, D., Schoeni, R.F. 1993. Effects of family background variables on earnings and returns to schooling: evidence from Brazil. Journal of Political Economy 101(4): 710-740.
Lancaster, T. 2005. An introduction to modern Bayesian econometrics. Blackwell Publishing: Oxford.
Lazear, E.P. 2004. Balanced skills and entrepreneurship. American Economic Review Papers and Proceedings 94(2): 208-211.
Lazear, E.P. 2005. Entrepreneurship. Journal of Labor Economics 23(4): 649-680.
Meager, N. 1992. Does unemployment lead to self-employment Small Business Economics 4(2): 87-103
OECD, 2009. Education at a Glance. Paris.
Oosterbeek, H., Van Praag, M., IJsselstein, A. 2009. The impact of entrepreneurship education on entrepreneurship skills and motivation. European Economic Review, doi:10.1016/j.euroecorev.2009.08.002.
Parker, S.C. 2009. The economics of entrepreneurship, second edition. Cambridge University Press: Cambridge, UK..
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.
Pfeiffer, F., Reize, F. 2000. Business start-ups by the unemployed – an econometric analysis based on firm data. Labour Economics 7(5): 629-663.

Psacharopoulos, G., Patrinos, A. 2004. Returns to investment in education: a further update. Education Economics 12(2): 111-134.
Reynolds, P.D., Camp, S.M., Bygrave, W.D., Autio, E., Hay, M. 2002. Global Entrepreneurship Monitor 2001 Executive Report. Babson College, London Business School.
Riley, J.G. 1979. Testing the educational screening hypothesis. Journal of Political Economy 87(5): S227-52.
Riley, J.G. 2002. Weak and strong signals. Scandinavian Journal of Economics 104(2): 213-236.
Rossi, P.E., Allenby, G.M, McCulloch, R. 2005. Bayesian Statistics and Marketing, Wiley: Chichester, UK.
Sandner, P.G., Block, J.H., Lutz, A. 2008. Determinanten des Erfolgs staatlich geförderter Existenzgründungen–eine empirische Untersuchung. Zeitschrift für Betriebswirtschaft 78(7/8): 753-777.
Stam, E., Thurik, A.R., Van der Zwan, P. 2010. Entrepreneurial exit in real and imagined markets. Industrial and Corporate Change 19: doi:10.1093/icc/dtp047.
The Economist (printed edition), 2009. Sources: cause and defect. August 13.
Trostel, P., Walker, I., Wooley, P. 2002. Estimates of the economic return to schooling for 28 countries. Labour Economics 9(1): 1-16.
Ucbasaran, D., Westhead, P., Wright, M. 2008. Opportunity identification and pursuit: does an entrepreneur's human capital matter? Small Business Economics 30(2): 153-173.
Van der Sluis, J., Van Praag, C.M., Vijverberg, W. 2005. Entrepreneurship selection and performance: a meta-analysis of the impact of education in less developed countries. World Bank Economic Review 19(2): 225-261.
Van der Sluis, J., Van Praag, C.M., Vijverberg, W. 2008. Education and entrepreneurship selection and performance: a review of the empirical literature. Journal of Economic Surveys 22(5): 795-841.
Van Praag, C.M. 2003. Business survival and success of young small business owners. Small Business Economics 21(1): 1-17.
Van Praag, C.M., Van Witteloostuijn, A., Van der Sluis, J. 2009. Returns for entrepreneurs versus employees. The effect of education and personal control on the relative performance of entrepreneurs vis-à-vis employees. Mimeo.
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.
Wagner, J. 2003. Testing Lazear's jack-of-all-trades view of entrepreneurship with German micro data. Applied Economics Letters 10(11): 687-689.
Webbink, D. 2005. Causal effects in education. Journal of Economic Surveys, 19(4): 535-560.
Wooldridge, J. 2002. Econometric analysis of cross section and panel data. MIT Press: London.

Table 1: Posterior results of random effects instrumental variables model

Dependent variable: Log (hourly earnings in self-employment)

	Mean and standard dev. of posterior distribution		Percentiles of posterior distribution			
Variables	**Mean**	**Std. dev**	**2.5%**	**97.5%**	**25%**	**75%**
Education	0.105	0.013	0.079	0.130	0.096	0.113
Experience	0.023	0.003	0.016	0.029	0.020	0.025
Experience²/10	-0.004	0.001	-0.005	-0.002	-0.004	-0.003
Unemployment duration	-0.034	0.015	-0.064	-0.004	-0.044	-0.023
Male	0.304	0.034	0.236	0.371	0.281	0.327
Married	0.044	0.027	-0.008	0.096	0.026	0.062
Non-German	0.107	0.071	-0.033	0.247	0.059	0.155
Wealth	0.037	0.005	0.028	0.046	0.034	0.040
West Germany	0.450	0.036	0.379	0.521	0.425	0.474
Year 1985	0.097	0.066	-0.031	0.226	0.052	0.141
Year 1986	0.127	0.068	-0.005	0.260	0.082	0.173
Year 1987	0.133	0.066	0.004	0.263	0.088	0.178
Year 1988	0.180	0.068	0.046	0.313	0.134	0.226
Year 1989	0.210	0.067	0.079	0.341	0.164	0.255
Year 1990	0.244	0.068	0.110	0.379	0.197	0.290
Year 1991	0.300	0.067	0.169	0.431	0.255	0.346
Year 1992	0.313	0.066	0.184	0.441	0.269	0.358
Year 1993	0.410	0.065	0.282	0.538	0.366	0.454
Year 1994	0.413	0.065	0.286	0.541	0.369	0.457
Year 1995	0.475	0.065	0.349	0.602	0.431	0.519
Year 1996	0.490	0.066	0.362	0.619	0.445	0.534
Year 1997	0.437	0.065	0.311	0.564	0.393	0.481
Year 1998	0.503	0.065	0.376	0.629	0.459	0.546
Year 1999	0.568	0.065	0.441	0.694	0.524	0.612
Year 2000	0.503	0.063	0.379	0.625	0.460	0.545
Year 2001	0.510	0.063	0.386	0.633	0.468	0.553
Year 2002	0.629	0.063	0.505	0.753	0.586	0.672
Year 2003	0.661	0.064	0.536	0.786	0.618	0.704
Year 2004	0.663	0.064	0.537	0.788	0.619	0.706
Agriculture	-0.502	0.065	-0.630	-0.374	-0.546	-0.458
Manufacturing	-0.031	0.040	-0.110	0.047	-0.058	-0.004
Retail	-0.104	0.040	-0.184	-0.025	-0.131	-0.076
Hotel and Restaurant	-0.180	0.067	-0.312	-0.048	-0.225	-0.135
Financial Services	0.148	0.053	0.044	0.252	0.112	0.184
Firm Services	0.013	0.038	-0.062	0.088	-0.013	0.039
Construction	-0.034	0.046	-0.124	0.056	-0.065	-0.003
Health	0.173	0.051	0.072	0.274	0.139	0.208
Transportation	0.014	0.069	-0.121	0.148	-0.032	0.061
Culture, Sports, and Leisure	0.013	0.070	-0.124	0.148	-0.034	0.060
ρ	-0.122	0.040	-0.199	-0.044	-0.149	-0.095

Notes: N = 8,288 observations on 2,280 individuals (period 1984-2004).
A non-informative prior was used for all coefficients of the unrestricted reduced form of this random effects panel data IV model; see appendix 3. 110,000 Gibbs draws have been generated (using pseudo random number generators in Matlab™); the first 10,000 draws have been discarded as a burn-in.
Instruments for education: *respondent's father's education* (category indicators)
Reference categories: *year 1984* and industry category *other*.
ρ = correlation (education, error term).

Table 2: Posterior results of random effects model
Dependent variable: Log (hourly earnings in self-employment)

	Mean and standard dev. of posterior distribution		Percentiles of posterior distribution			
Variables	**Mean**	**Std. Dev.**	**2.5%**	**97.5%**	**25%**	**75%**
Education	0.060	0.003	0.053	0.066	0.057	0.062
Experience	0.021	0.002	0.017	0.026	0.020	0.023
Experience²/10	-0.004	0.000	-0.005	-0.003	-0.004	-0.004
Unemployment duration	-0.041	0.010	-0.060	-0.021	-0.047	-0.034
Male	0.325	0.022	0.281	0.368	0.310	0.340
Married	0.042	0.018	0.007	0.079	0.030	0.055
Non-German	0.081	0.047	-0.011	0.174	0.049	0.112
Wealth	0.040	0.003	0.033	0.046	0.037	0.042
West Germany	0.436	0.023	0.390	0.481	0.420	0.452
Year 1985	0.099	0.050	0.000	0.197	0.065	0.133
Year 1986	0.132	0.052	0.029	0.234	0.097	0.167
Year 1987	0.141	0.051	0.040	0.239	0.106	0.175
Year 1988	0.188	0.053	0.084	0.291	0.153	0.224
Year 1989	0.220	0.051	0.120	0.319	0.186	0.255
Year 1990	0.256	0.052	0.152	0.358	0.221	0.291
Year 1991	0.312	0.051	0.212	0.411	0.277	0.346
Year 1992	0.325	0.049	0.228	0.422	0.292	0.358
Year 1993	0.424	0.049	0.328	0.519	0.391	0.457
Year 1994	0.428	0.048	0.333	0.522	0.395	0.461
Year 1995	0.493	0.048	0.399	0.586	0.461	0.526
Year 1996	0.510	0.048	0.415	0.605	0.478	0.543
Year 1997	0.459	0.047	0.366	0.552	0.427	0.491
Year 1998	0.527	0.047	0.435	0.619	0.495	0.559
Year 1999	0.592	0.047	0.500	0.684	0.561	0.624
Year 2000	0.532	0.045	0.443	0.620	0.501	0.562
Year 2001	0.541	0.045	0.453	0.630	0.511	0.572
Year 2002	0.667	0.045	0.578	0.755	0.636	0.697
Year 2003	0.701	0.045	0.612	0.790	0.671	0.732
Year 2004	0.704	0.045	0.615	0.792	0.673	0.734
Agriculture	-0.539	0.043	-0.624	-0.454	-0.568	-0.509
Manufacturing	-0.034	0.028	-0.090	0.021	-0.053	-0.015
Retail	-0.111	0.028	-0.166	-0.057	-0.130	-0.092
Hotel and Restaurant	-0.213	0.046	-0.304	-0.123	-0.244	-0.182
Financial Services	0.150	0.036	0.078	0.221	0.126	0.175
Firm Services	0.029	0.027	-0.024	0.082	0.011	0.048
Construction	-0.037	0.031	-0.099	0.025	-0.058	-0.016
Health	0.210	0.035	0.142	0.279	0.187	0.234
Transportation	-0.006	0.048	-0.101	0.088	-0.039	0.026
Culture, Sports, and Leisure	0.054	0.049	-0.043	0.150	0.021	0.087

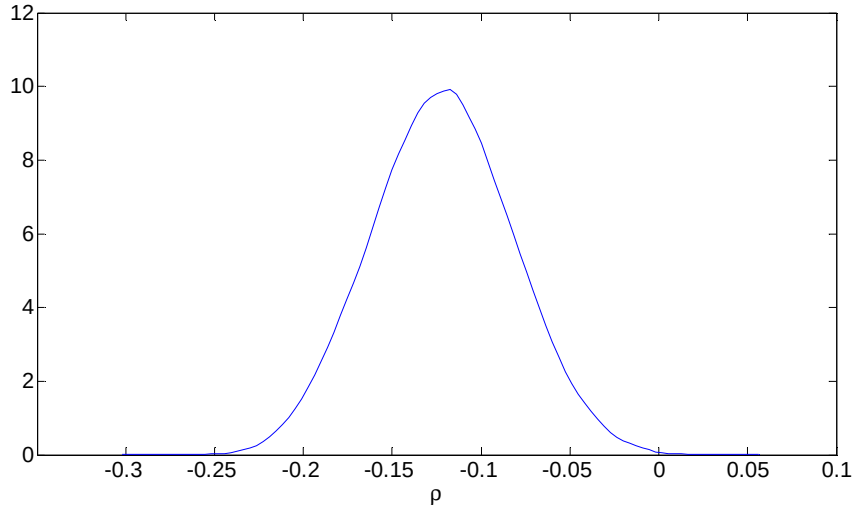
Notes: N = 8,288 observations on 2,280 individuals (period 1984-2004).
A non-informative prior was used for all coefficients of this random effects panel data model. 110,000 Gibbs draws have been generated (using the pseudo random number generators in Matlab™); the first 10,000 draws have been discarded as a burn-in.
Reference categories: *year 1984* and industry category *other*.

Table 3: Posterior results for β in random effects instrumental variables model under Normal prior $N(0,\tau^2)$, and truncated normal prior restricted to positive values, for the ratio of father's education's effect over own education's effect on income.
Dependent variable: Log (hourly earnings in self-employment)

Variables			Mean	Std. dev	2.5%	97.5%
Education	Normal prior	$\tau = 0$	0.105	0.013	0.079	0.130
	For	$\tau = 0.05$	0.111	0.018	0.078	0.150
	γ / β	$\tau = 0.10$	0.113	0.023	0.072	0.158
	Truncated	$\tau = 0$	0.105	0.013	0.079	0.130
	normal prior	$\tau = 0.05$	0.102	0.014	0.073	0.129
	for γ / β	$\tau = 0.10$	0.098	0.015	0.069	0.126

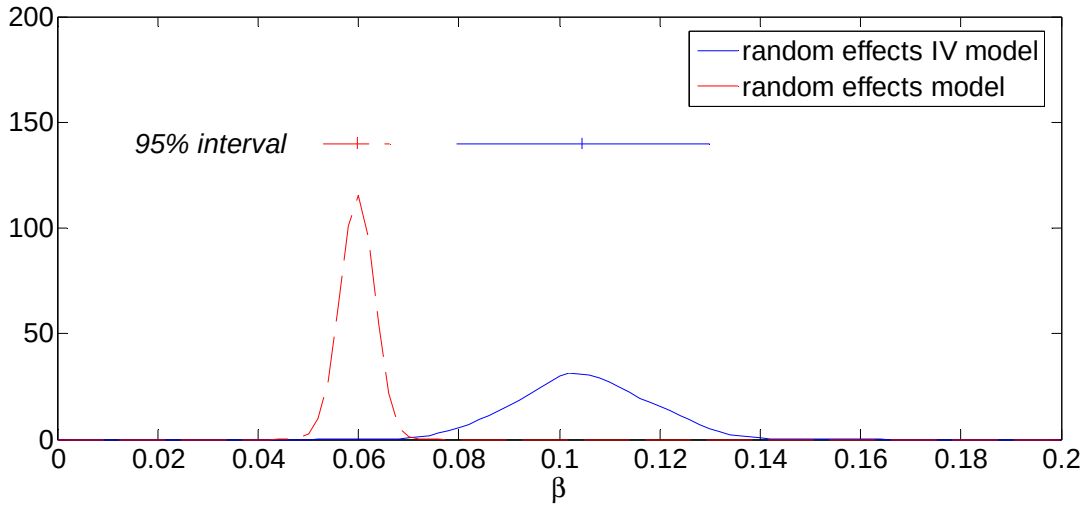
Notes: τ has the interpretation of the standard deviation of the ratio of the father's education effect over own education effect on log hourly earnings in self-employment. For the [truncated] normal prior, $\tau = 0.10$ corresponds with a 95% prior belief that an extra year of father's secondary education has a *direct* effect on income (*in addition to* the effect that is captured in own education and controls) between (approximately) -20% [0%] and 20% of own education's effect, which seems a rather conservative assumption. Posterior estimation results for β are robust with respect to deviations in the assumption of exactly valid instruments. The conclusion that endogeneity matters does not change if we change the validity assumption of the instruments somewhat.

Figure 1: Correlation of the error terms (posterior density of ρ)



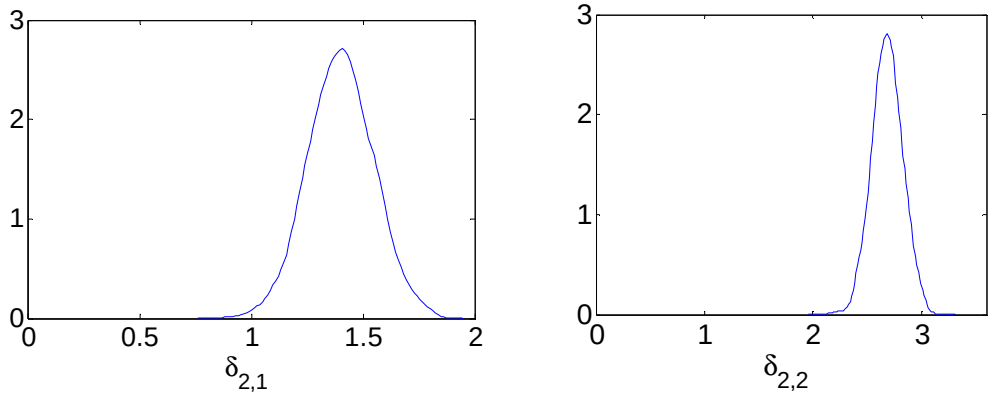
Notes: The figure shows the posterior density of ρ = correlation (education, error term) in a random effects IV model. The 95% posterior interval of ρ lies between -0.199 and -0.044, so that the value $\rho = 0$ is rejected. We conclude that education is an endogenous variable.

Figure 2: Illustration of the substantial difference between outcomes from random effects panel data models with and without IV



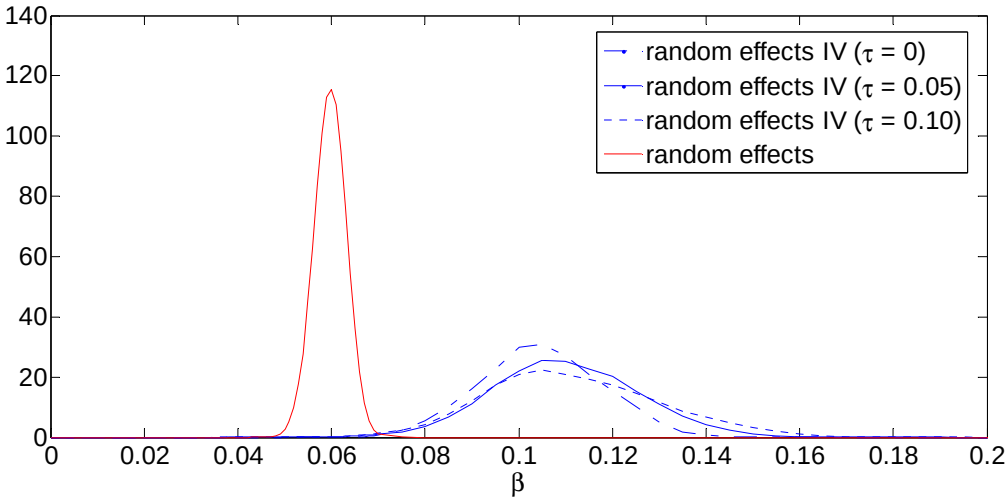
Notes: Posterior distributions of β in random effects model and random effects IV model: posterior densities - with posterior 95% interval (interval between 2.5% and 97.5% quantiles) and posterior mean indicated. The posterior mean of β is substantially higher in the random effects IV model. In fact, the posterior 95% intervals do not overlap, which indicates that results from both models strongly differ. We conclude that taking into account endogeneity (via the use of an IV model) matters.

Figure 3: Posterior densities of $\delta_{2,1}$ and $\delta_{2,2}$ in random effects IV model



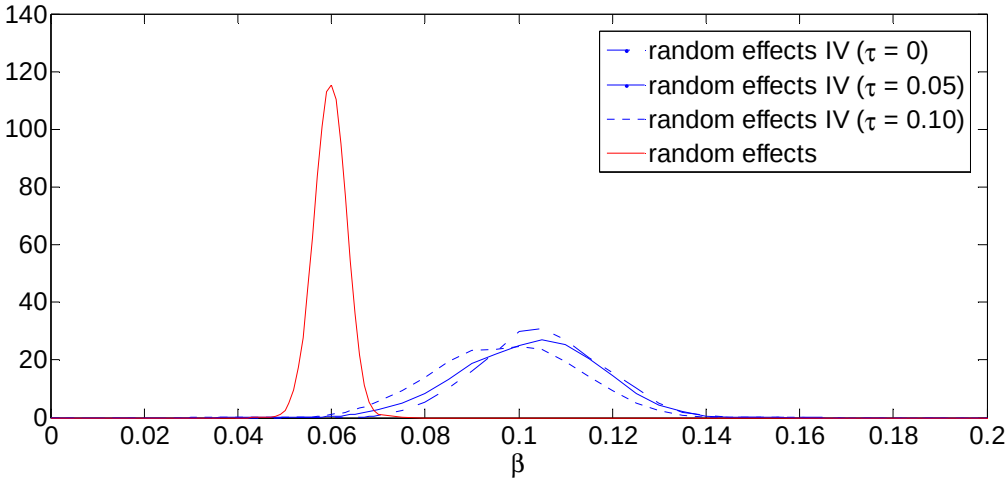
Notes: $\delta_{2,1}$ and $\delta_{2,2}$ are the effect of the 0/1 instruments indicating whether father's secondary education falls in category 2 (primary and secondary education spell of approximately 10 years) or category 3 (approximately 13 years) on the respondent's education , where the reference category 1 refers to approximately 9 years. The effect of father's education is clearly non-zero, so that we do not face the problem of *weak instruments* here.

Figure 4: Posterior density for β in random effects IV model under Normal prior $N(0,\tau^2)$ for the ratio of father's education's effect over own education's effect on income, and posterior density for β in random effects model (without IV)



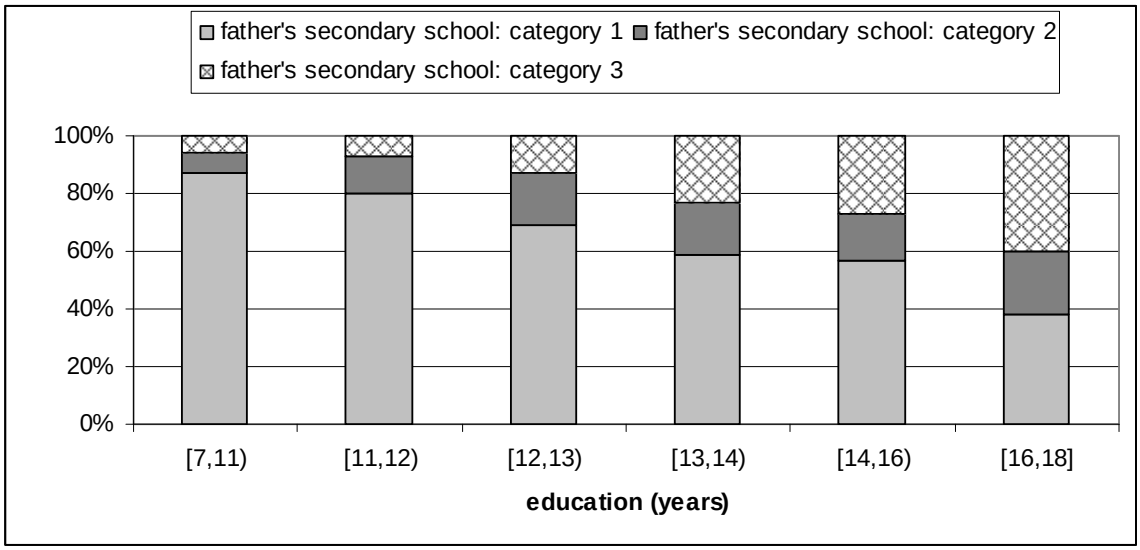
Notes: τ has the interpretation of the standard deviation of the ratio of father's education's effect over own education's effect on log hourly earnings in self-employment. $\tau = 0.10$ corresponds with a 95% prior belief that an extra year of father's secondary education has a *direct* effect on income (*in addition to* the effect that is captured in own education and controls) between (approximately) -20% and 20% of own education's effect, which seems a rather conservative assumption. The graphs show that posterior estimation results for β are robust with respect to deviations in the assumption of exactly valid instruments. The conclusion that endogeneity matters does not change if we change the validity assumption of the instruments.

Figure 5: Posterior density for β in random effects IV model under truncated Normal prior $N(0,\tau^2)$, truncated to positive values, for the ratio of father's education's effect over own education's effect on income, and posterior density for β in random effects model (without IV)



Notes: τ has the interpretation of the standard deviation of the ratio of father's education's effect over own education's effect on log hourly earnings in self-employment. $\tau = 0.10$ corresponds with a 95% prior belief that an extra year of father's secondary education has a *direct* effect on income (*in addition to* the effect that is captured in own education and controls) between 0% and (approximately) 20% of own education's effect, which seems a rather conservative assumption. The graphs show that posterior estimation results for β are robust with respect to deviations in the assumption of exactly valid instruments. The conclusion that endogeneity matters does not change if we change the validity assumption of the instruments.

Figure 6: Illustration of the effect of the respondent's father's education instruments on the respondent's years of education



Note: father's secondary school instruments have explanatory power for education on different levels. A high father's education (category 3) implies a higher probability of medium versus low level of education, and a higher probability of high versus medium level of education.

Appendix 1: description of variables

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
Year 1984-2004	Dummies for years 1984-2004
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)
	Continuous variables and ordinal variable
Entrepreneurial 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	Ordinal variable including the following secondary school certificates: "Hauptschule" (approx. 9 yrs.), "Realschule" (approx. 10 yrs.), "Fachhochschulreife" (approx. 12 yrs.), "Abitur" (approx. 13 yrs.)
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)

Appendix 2: classical regression

Table A2: Results of classical regression with dependent variable log (hourly earnings in self-employment)

	OLS Regression		Two-stage Least Squares (2SLS) Instrumental Variables Regression [1]	
Variables	**Coefficient (SE)**		**Coefficient (SE)**	
Education	0.060 (0.005)	***	0.113 (0.016)	***
Experience	0.021 (0.004)	***	0.023 (0.004)	***
Experience²/10	-0.004 (0.000)	***	-0.004 (0.000)	***
Unemployment duration	-0.042 (0.016)	***	-0.026 (0.015)	*
Male	0.327 (0.034)	***	0.293 (0.034)	***
Married	0.044 (0.027)		0.038 (0.026)	
Non-German	0.077 (0.072)		0.116 (0.070)	*
Wealth	0.038 (0.005)	***	0.036 (0.005)	***
West Germany	0.434 (0.036)	***	0.466 (0.036)	***
Year 1985	0.099 (0.066)		0.101 (0.067)	
Year 1986	0.131 (0.068)	*	0.133 (0.069)	*
Year 1987	0.139 (0.066)	**	0.138 (0.068)	**
Year 1988	0.185 (0.068)	***	0.185 (0.070)	***
Year 1989	0.218 (0.067)	***	0.216 (0.069)	***
Year 1990	0.253 (0.069)	***	0.248 (0.070)	***
Year 1991	0.311 (0.067)	***	0.304 (0.068)	***
Year 1992	0.325 (0.066)	***	0.316 (0.067)	***
Year 1993	0.423 (0.065)	***	0.415 (0.067)	***
Year 1994	0.428 (0.065)	***	0.416 (0.067)	***
Year 1995	0.492 (0.065)	***	0.477 (0.066)	***
Year 1996	0.509 (0.065)	***	0.492 (0.067)	***
Year 1997	0.459 (0.064)	***	0.437 (0.066)	***
Year 1998	0.526 (0.064)	***	0.501 (0.066)	***
Year 1999	0.593 (0.064)	***	0.566 (0.066)	***
Year 2000	0.532 (0.063)	***	0.503 (0.064)	***
Year 2001	0.542 (0.063)	***	0.508 (0.065)	***
Year 2002	0.664 (0.063)	***	0.625 (0.065)	***
Year 2003	0.699 (0.063)	***	0.655 (0.065)	***
Year 2004	0.702 (0.063)	***	0.656 (0.065)	***
Agriculture	-0.520 (0.065)	***	-0.461 (0.068)	***
Manufacturing	-0.037 (0.040)		-0.007 (0.041)	
Retail	-0.106 (0.040)	***	-0.073 (0.042)	*
Hotel and Restaurant	-0.207 (0.068)	***	-0.128 (0.071)	*
Financial Services	0.142 (0.053)	***	0.177 (0.053)	***
Firm Services	0.023 (0.038)		0.006 (0.039)	
Construction	-0.042 (0.046)		-0.006 (0.046)	
Health	0.199 (0.051)	***	0.135 (0.055)	**
Transportation	-0.004 (0.069)		0.054 (0.071)	
Culture, Sports, and Leisure	0.042 (0.069)		0.016 (0.070)	
Intercept	0.125 (0.108)		-0.593 (0.230)	***
N obs. (individuals)	8,288 (2,280)		8,288 (2,280)	
Obs. per group: min., avg., max.	1; 3.6; 20		1; 3.6; 20	
R^2 within, between, overall	0.053; 0.356; 0.306		0.050; 0.330; 0.281	
Wald Chi²	1,536.25 ***		1,576.80 ***	

Notes: * p<0.05 ** p<0.01 *** p<0.001
SE=Robust standard errors
[1] Instrument for education: *respondent's father's education*
(F-test for significance of the instrument: F(1,8286)=767.49 (p<0.001); R²=0.162)
Durbin-Wu-Hausman-test for endogenity: p<0.001
Reference categories: *year 1984* and industry category *other*

Appendix 3: Bayesian analysis of an instrumental variables model for panel data

The model consists of two equations, both describing panel data with (random) individual effects:

$$y_{it} = x_{it}\beta + w_{it}'\delta_1 + \alpha_{1,i} + \varepsilon_{it} \qquad (\text{i} = 1,2,\ldots, \text{n}_{\text{individuals}}\ ; \text{t} = 1,2,\ldots, \text{n}_{\text{obs,i}}) \qquad \text{(A1)}$$

$$x_{it} = z_{it}'\delta_2 + \alpha_{2,i} + v_{it} \qquad (\text{i} = 1,2,\ldots, \text{n}_{\text{individuals}}\ ; \text{t} = 1,2,\ldots, \text{n}_{\text{obs,i}}) \qquad \text{(A2)}$$

with y_{it} = log(income) of individual i at time t ;
x_{it} = education of individual i at time t ;
w_{it} = control variables for individual i at time t ;
$\alpha_{1,i}$, $\alpha_{2,i}$ = individual effect of individual i ;
ε_{it} , v_{it} = error term (in addition to individual effect) for individual i at time t ;
z_{it} = instruments (education of father of individual i, and control variables).

The total number of observations is $n_{total} = \sum_{i=1}^{n_{individuals}} n_{obs,i}$. The individual effects $\alpha_i = (\alpha_{1,i}, \alpha_{2,i})'$ are independently normally distributed: $\alpha_i \sim N(\mu_\alpha, \Sigma_\alpha)$. The error terms $(\varepsilon_{it}, v_{it})$ are independently normally distributed $(\varepsilon_{it}, v_{it})' \sim N(0, \Sigma_{\varepsilon,v})$, independently from the α_i $(i = 1,2,\ldots, n_{individuals})$. Both $(\varepsilon_{it}, v_{it})$ and α_i are independent from w_{jt}, z_{jt} (i,j = 1,2,…., $\text{n}_{\text{individuals}}$; t = 1,2,…., $\text{n}_{\text{obs,i}}$). Define $\Psi_\alpha = \Sigma_\alpha^{-1}$ and $\Psi_{\varepsilon,v} = \Sigma_{\varepsilon,v}^{-1}$. Define *y*, *x*, *w*, *z*, α as the vectors and matrices consisting of y_{it}, x_{it}, α_i, w_{it}, z_{it}, respectively.

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

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

The *complete data* likelihood is:

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

with

$$p(y, x \mid \alpha, w, z, \theta) = (2\pi)^{-n_{total}} |\Psi_{\varepsilon,v}|^{n_{total}/2} \times$$
$$\exp\left[-\frac{1}{2}\sum_{i=1}^{n_{individuals}} \begin{pmatrix} y_{it} - x_{it}\beta - w_{it}'\delta_1 - \alpha_{1,i} \\ x_{it} - z_{it}'\delta_2 - \alpha_{2,i} \end{pmatrix}' \Psi_{\varepsilon,v} \begin{pmatrix} y_{it} - x_{it}\beta - w_{it}'\delta_1 - \alpha_{1,i} \\ x_{it} - z_{it}'\delta_2 - \alpha_{2,i} \end{pmatrix}\right]$$

$$p(\alpha \mid w,z,\theta) = (2\pi)^{-n_{individuals}} \mid \Psi_\alpha \mid^{n_{individuals}/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n_{individuals}}(\alpha_i - \mu_\alpha)'\Psi_\alpha(\alpha_i - \mu_\alpha)\right]$$

The posterior density kernel is:

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

$$\mid \Psi_{\varepsilon,v} \mid^{(n_{total}-3)/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n_{individuals}}\begin{pmatrix} y_{it} - x_{it}\beta - w_{it}'\delta_1 - \alpha_{1,i} \\ x_{it} - z_{it}'\delta_2 - \alpha_{2,i}\end{pmatrix}'\Psi_{\varepsilon,v}\begin{pmatrix} y_{it} - x_{it}\beta - w_{it}'\delta_1 - \alpha_{1,i} \\ x_{it} - z_{it}'\delta_2 - \alpha_{2,i}\end{pmatrix}\right]\times$$

$$\mid \Psi_\alpha \mid^{(n_{individuals}-3)/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n_{individuals}}(\alpha_i - \mu_\alpha)'\Psi_\alpha(\alpha_i - \mu_\alpha)\right]\exp\left(-\frac{1}{2}\frac{(\beta - \mu_{\beta,prior})^2}{\sigma^2_{\beta,prior}}\right)$$

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 (1981)) to simulate draws from the posterior distribution, iteratively sampling from the full conditional posteriors:

(i) $\Psi_{\varepsilon,v} \mid y,x,w,z,\alpha,\theta_{-\Psi_{\varepsilon,v}} \sim$

$$Wishart\left(n_{total}, \left[\sum_{i=1}^{n_{individuals}}\sum_{t=1}^{n_{obs,i}}\begin{pmatrix} y_{it} - x_{it}\beta - w_{it}'\delta_1 - \alpha_{1,i} \\ x_{it} - z_{it}'\delta_2 - \alpha_{2,i}\end{pmatrix}\begin{pmatrix} y_{it} - x_{it}\beta - w_{it}'\delta_1 - \alpha_{1,i} \\ x_{it} - z_{it}'\delta_2 - \alpha_{2,i}\end{pmatrix}'\right]^{-1}\right)$$

(ii) $\Psi_\alpha \mid y,x,w,z,\alpha,\theta_{-\Psi_\alpha} \sim Wishart\left(n_{individuals}, \left[\sum_{i=1}^{n_{individuals}}(\alpha_i - \mu_\alpha)(\alpha_i - \mu_\alpha)'\right]^{-1}\right)$

(iii) $\mu_\alpha \mid y,x,w,z,\alpha,\theta_{-\mu_\alpha} \sim N\left(\frac{1}{n_{individuals}}\sum_{i=1}^{n_{individuals}}\alpha_i, \frac{1}{n_{individuals}}\Sigma_\alpha\right)$

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

$$V_{\beta,\delta_1} = \left[\begin{pmatrix}\left(\sigma^2_{\beta,prior}\right)^{-1} & 0 \\ 0 & 0\end{pmatrix} + (\sigma^2_{\varepsilon|v})^{-1}\sum_{i=1}^{n_{individuals}}\sum_{t=1}^{n_{obs,i}}\begin{pmatrix} x_{it}^2 & x_{it}w_{it}' \\ x_{it}w_{it} & w_{it}w_{it}'\end{pmatrix}\right]^{-1}$$

$$\mu_{\beta,\delta_1} = V_{\beta,\delta_1}\left[\begin{pmatrix}\left(\sigma^2_{\beta,prior}\right)^{-1}\mu_{\beta,prior} \\ 0\end{pmatrix} + (\sigma^2_{\varepsilon|v})^{-1}\sum_{i=1}^{n_{individuals}}\sum_{t=1}^{n_{obs,i}}\begin{pmatrix} x_{it} \\ w_{it}\end{pmatrix}\left(y_{it} - \alpha_{1,i} - \mu_{\varepsilon_{it}|v_{it}}\right)\right]$$

where $\mu_{\varepsilon_{it}|v_{it}} = (x_{it} - z_{it}'\delta_2 - \alpha_{2,i})\sigma_{\varepsilon,v}/\sigma_v^2$ and $\sigma^2_{\varepsilon|v} = \sigma^2_\varepsilon - \sigma^2_{\varepsilon,v}/\sigma^2_v$ with

$$\Sigma_{\varepsilon,v} = \begin{pmatrix}\sigma^2_\varepsilon & \sigma_{\varepsilon,v} \\ \sigma_{\varepsilon,v} & \sigma^2_v\end{pmatrix}.$$

(v) $\delta_2 \mid y, x, w, z, \alpha, \theta_{-\delta_2} \sim N\left(\mu_{\delta_2}, V_{\delta_2}\right)$ with $V_{\delta_2} = \left[\left(\sigma^2_{v|\varepsilon}\right)^{-1} \sum_{i=1}^{n_{individuals}} \sum_{t=1}^{n_{obs,i}} z_{it} z_{it}' \right]^{-1}$

$$\mu_{\delta_2} = V_{\delta_2} \left[(\sigma^2_{v|\varepsilon})^{-1} \sum_{i=1}^{n_{individuals}} \sum_{t=1}^{n_{obs,i}} z_{it} \left(x_{it} - \alpha_{2,i} - \mu_{v_{it}|\varepsilon_{it}} \right) \right]$$

where $\mu_{v_{it}|\varepsilon_{it}} = (y_{it} - x_{it}\beta - w_{it}'\delta_1 - \alpha_{1,i})\, \sigma_{\varepsilon,v} / \sigma^2_{\varepsilon}$ and $\sigma^2_{v|\varepsilon} = \sigma^2_v - \sigma^2_{\varepsilon,v} / \sigma^2_{\varepsilon}$.

(vi) $\alpha_i \mid y, x, w, z, \theta \sim$

$$N\left(\left[\Psi_\alpha + n_{obs,i}\, \Psi_{\varepsilon,v}\right]^{-1} \left[\Psi_\alpha\, \mu_\alpha + \Psi_{\varepsilon,v} \sum_{t=1}^{n_{obs,i}} \begin{pmatrix} y_{it} - x_{it}\beta - w_{it}'\delta_1 \\ x_{it} - z_{it}'\delta_2 \end{pmatrix} \right], \left[\Psi_\alpha + n_{obs,i}\, \Psi_{\varepsilon,v}\right]^{-1} \right)$$

where conditionally on data and θ, the α_i ($i = 1,2,\ldots,n_{individuals}$) are independent.

Our Gibbs sampling approach combines Rossi et al. (2005) and an extension of Koop (2003). Given α, we draw θ following Rossi et al. (2005, chapter 7). Given θ, we draw α following an extension of Koop (2003, section 7.3) of Bayesian individual effects to the bivariate case.

Assuming that the true model is (A1)-(A2), the endogeneity of education x_{it} in (A1) is reflected by the correlation between x_{it} and the 'error term' $\alpha_{1,i} + \varepsilon_{it}$:

$$\text{corr}(x_{it}, \alpha_{1,i} + \varepsilon_{it}) = \frac{\text{cov}(x_{it}, \alpha_{1,i} + \varepsilon_{it})}{\sqrt{\text{var}(x_{it})\,\text{var}(\alpha_{1,i} + \varepsilon_{it})}}$$

where

$$\text{cov}(x_{it}, \alpha_{1,i} + \varepsilon_{it}) = \text{cov}(z_{it}'\delta_2 + \alpha_{2,i} + v_{it},\, \alpha_{1,i} + \varepsilon_{it}) = \text{cov}(\alpha_{2,i}, \alpha_{1,i}) + \text{cov}(v_{it}, \varepsilon_{it}) = \sigma_{\alpha_1,\alpha_2} + \sigma_{\varepsilon,v}$$

$$\text{var}(x_{it}) = \text{var}(z_{it}'\delta_2 + \alpha_{2,i} + v_{it}) = \delta_2'\,\text{var}(z_{it})\delta_2 + \sigma^2_{\alpha_2} + \sigma^2_v$$

$$\text{var}(\alpha_{1,i} + \varepsilon_{it}) = \sigma^2_{\alpha_1} + \sigma^2_{\varepsilon}$$

with $\text{var}(z_{it})$ the sample covariance matrix of the z_{it}.

Appendix 4: Plausible exogeneity: instruments that are plausibly (approximately) valid

In the model (A1)-(A2), there are $\tilde{k}$ instruments $\tilde{z}_{it}$ that are excluded from (A1), i.e. appearing in z_{it} but not in w_{it}. That is, in (A1)-(A2) we assume that $\tilde{z}_{it}$ has no *direct* effect on y_{it}, only via x_{it}. The $\tilde{z}_{it}$ satisfy the *exclusion* restriction, being exactly exogenous (in the sense that $\text{cov}(\tilde{z}_{it}, error_{it}) = 0$, where the 'error term' is defined as $error_{it} = \alpha_{1,i} + \varepsilon_{it}$).

Conley et al. (2008) present an alternative approach to inference for IV models with instruments whose validity is debatable. They provide an operational definition of plausibly (or approximately) exogenous instruments, and a Gibbs sampling method for posterior results that are consistent with instruments being only plausibly exogenous.

We consider an extension of (A1)-(A2) in the line of Conley et al. (2008), where the instruments $\tilde{z}_{it}$ have a (small) direct effect γ on y_{it}:

$$y_{it} = x_{it}\beta + \tilde{z}_{it}'\gamma + w_{it}'\delta_1 + \alpha_{1,i} + \varepsilon_{it} \qquad (\text{i} = 1,2,\ldots, \text{n}_{\text{individuals}}; \text{t} = 1,2,\ldots, \text{n}_{\text{obs,i}})$$

Defining the ratio $\tilde{\gamma} = \gamma / \beta$, this amounts to

$$y_{it} = x_{it}\beta + \tilde{z}_{it}'\tilde{\gamma}\,\beta + w_{it}'\delta_1 + \alpha_{1,i} + \varepsilon_{it} \qquad (\text{i} = 1,2,\ldots, \text{n}_{\text{individuals}}; \text{t} = 1,2,\ldots, \text{n}_{\text{obs,i}}) \qquad (\text{A1'})$$

We consider two prior specifications for $\tilde{\gamma}$:

(I) a normal prior distribution $\tilde{\gamma} \sim N(\mu_{\tilde{\gamma},prior}, \Sigma_{\tilde{\gamma},prior})$

(II) a truncated normal distribution $\tilde{\gamma} \sim TN_{[\tilde{\gamma} \subset A]}(\mu_{\tilde{\gamma},prior}, \Sigma_{\tilde{\gamma},prior})$, that is $N(\mu_{\tilde{\gamma},prior}, \Sigma_{\tilde{\gamma},prior})$ truncated to a subspace A.

We specify a proper, informative prior for $\tilde{\gamma}$, as is required to make inference possible (see Conley et al. (2008)). We specify $\mu_{\tilde{\gamma},prior} = 0$ and $\Sigma_{\tilde{\gamma},prior}$ a diagonal matrix.

In the model (A1')-(A2) the Gibbs steps are as follows. Conditionally on $\tilde{\gamma} = \gamma / \beta$ and β, δ_1, we perform steps (i)-(iii), (v)-(vi) with $y_{it} - x_{it}\beta - z_{it}'\tilde{\gamma}\,\beta - w_{it}'\delta_1$ instead of $y_{it} - x_{it}\beta - w_{it}'\delta_1$. Conditionally on α and all other parameters, including $\tilde{\gamma}$, the full conditional posterior of β, δ_1 becomes:

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

$$V_{\beta,\delta 1} = \left[\begin{pmatrix} \left(\sigma^2_{\beta,prior}\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix} + (\sigma^2_{\varepsilon|v})^{-1} \sum_{i=1}^{n_{individuals}} \sum_{t=1}^{n_{obs,i}} \begin{pmatrix} (x_{it} + z_{it}'\tilde{\gamma})^2 & (x_{it} + z_{it}'\tilde{\gamma}) w_{it}' \\ (x_{it} + z_{it}'\tilde{\gamma}) w_{it} & w_{it} w_{it}' \end{pmatrix} \right]^{-1}$$

$$\mu_{\beta,\delta 1} = V_{\beta,\delta 1}\left[\begin{pmatrix}\left(\sigma^2_{\beta,prior}\right)^{-1}\mu_{\beta,prior}\\ 0\end{pmatrix} + (\sigma^2_{\varepsilon|v})^{-1}\sum_{i=1}^{n_{individuals}}\sum_{t=1}^{n_{obs,i}}\begin{pmatrix}x_{it}+z_{it}'\tilde{\gamma}\\ w_{it}\end{pmatrix}\left(y_{it}-\alpha_{1,i}-\mu_{\varepsilon_{it}|v_{it}}\right)\right]$$

This amounts to (iv) with $x_{it}+z_{it}'\tilde{\gamma}$ instead of x_{it}. Conditionally on α and all other parameters, including β, the full conditional posterior of $\tilde{\gamma}$ is:

(vii') $\tilde{\gamma} \mid y,x,w,z,\alpha,\theta_{-\tilde{\gamma}} \sim N\left(\mu_{\tilde{\gamma}},V_{\tilde{\gamma}}\right)$ with

$$V_{\tilde{\gamma}} = \left[\Sigma^{-1}_{\tilde{\gamma},prior} + \left(\sigma^2_{\varepsilon|v}\right)^{-1}\beta^2\sum_{i=1}^{n_{individuals}}\sum_{t=1}^{n_{obs,i}}\tilde{z}_{it}\tilde{z}_{it}'\right]^{-1}$$

$$\mu_{\tilde{\gamma}} = V_{\tilde{\gamma}}\left[\Sigma^{-1}_{\tilde{\gamma},prior}\,\mu_{\tilde{\gamma},prior} + (\sigma^2_{\varepsilon|v})^{-1}\beta\sum_{i=1}^{n_{individuals}}\sum_{t=1}^{n_{obs,i}}\tilde{z}_{it}\left(y_{it}-\beta x_{it}-w_{it}'\delta_1-\alpha_{1,i}-\mu_{\varepsilon_{it}|v_{it}}\right)\right]$$

under prior $\tilde{\gamma} \sim N(\mu_{\tilde{\gamma},prior},\Sigma_{\tilde{\gamma},prior})$. Under prior $\tilde{\gamma} \sim TN_{[\tilde{\gamma}\subset A]}(\mu_{\tilde{\gamma},prior},\Sigma_{\tilde{\gamma},prior})$, a truncated (to subspace A) normal posterior distribution results: $\tilde{\gamma} \mid y,x,w,z,\alpha,\theta_{-\tilde{\gamma}} \sim TN_{[\tilde{\gamma}\subset A]}\left(\mu_{\tilde{\gamma}},V_{\tilde{\gamma}}\right)$.

Assuming that the true model is (A1')-(A2), the endogeneity of education x_{it} in (A1) is reflected by the correlation between x_{it} and the 'error term' $\tilde{z}_{it}'\tilde{\gamma}\beta+\alpha_{1,i}+\varepsilon_{it}$:

$$\text{corr}(x_{it},\tilde{z}_{it}'\tilde{\gamma}\beta+\alpha_{1,i}+\varepsilon_{it}) = \frac{\text{cov}(x_{it},\tilde{z}_{it}'\tilde{\gamma}\beta+\alpha_{1,i}+\varepsilon_{it})}{\sqrt{\text{var}(x_{it})\,\text{var}(\tilde{z}_{it}'\tilde{\gamma}\beta+\alpha_{1,i}+\varepsilon_{it})}}$$

where

$$\begin{aligned}
\text{cov}(x_{it},z_{it}'\tilde{\gamma}\beta+\alpha_{1,i}+\varepsilon_{it}) &= \text{cov}(z_{it}'\delta_2+\alpha_{2,i}+v_{it},\, z_{it}'\tilde{\gamma}\beta+\alpha_{1,i}+\varepsilon_{it})\\
&= \tilde{\delta}_2'\,\text{var}(\tilde{z}_{it})\,\tilde{\gamma}\beta+\text{cov}(\alpha_{2,i},\alpha_{1,i})+\text{cov}(v_{it},\varepsilon_{it})\\
&= \tilde{\delta}_2'\,\text{var}(\tilde{z}_{it})\,\tilde{\gamma}\beta+\sigma_{\alpha_1,\alpha_2}+\sigma_{\varepsilon,v}\\
\text{var}(x_{it}) &= \text{var}(z_{it}'\delta_2+\alpha_{2,i}+v_{it})\\
&= \delta_2'\,\text{var}(z_{it})\delta_2+\sigma^2_{\alpha_2}+\sigma^2_v\\
\text{var}(\tilde{z}_{it}'\tilde{\gamma}\beta+\alpha_{1,i}+\varepsilon_{it}) &= \beta^2\tilde{\gamma}'\,\text{var}(\tilde{z}_{it})\tilde{\gamma}+\sigma^2_{\alpha_1}+\sigma^2_{\varepsilon}
\end{aligned}$$

with $\tilde{\delta}_2$ the subvector of δ_2 with elements corresponding to $\tilde{z}_{it}\subset z_{it}$; $\text{var}(z_{it})$, $\text{var}(\tilde{z}_{it})$ the sample covariance matrices of the z_{it}, $\tilde{z}_{it}$, respectively.

We specify the parameters of the prior for $\tilde{\gamma}$, the values of $\mu_{\tilde{\gamma},prior}$ and $\Sigma_{\tilde{\gamma},prior}$, as follows. Note that $\tilde{\gamma}_1$ and $\tilde{\gamma}_2-\tilde{\gamma}_1$ are interpreted as the ratio of the effect of 1 extra year (=10-9 years) and 3 extra years (=13-10 years) of father's secondary education over 1 extra year of own education. That is, $\tilde{\gamma}_1$

and $(\tilde{\gamma}_2 - \tilde{\gamma}_1)/3$ are ratios of effects for 1 year of father's secondary education versus own education. We specify (independent) N(0,τ^2) priors for $\tilde{\gamma}_1$ and $(\tilde{\gamma}_2 - \tilde{\gamma}_1)/3$, which is equivalent to assuming

$$\begin{pmatrix} \tilde{\gamma}_1 \\ \tilde{\gamma}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tau^2 \Sigma^*_{\tilde{\gamma}, prior} \right) \quad \text{with} \quad \Sigma^*_{\tilde{\gamma}, prior} = \begin{pmatrix} 1 & 1 \\ 1 & 10 \end{pmatrix}. \tag{A3}$$

As an alternative, we specify a truncated version of (A3), restricted to those values of $(\tilde{\gamma}_1, \tilde{\gamma}_2)$ with both $\tilde{\gamma}_1$ and $\tilde{\gamma}_2 - \tilde{\gamma}_1$ non-negative, such that the effect of father's education is restricted to have the same sign as own education (typically positive). We consider τ = 0.05 and τ = 0.10. For the normal prior, the choice of τ = 0.10 corresponds with a 95% prior belief that an extra year of father's secondary education has a *direct* effect on income (*in addition to* the effect that is captured in own education and controls) between (approximately) -20% and 20% of own education's effect. For the truncated normal prior, the specification of τ = 0.10 reflects a 95% prior belief that this is between 0% and (approximately) 20%.