The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach

University of York

Department of Economics and Related Studies

2020.02.19

Paolo Brunori
University of Florence

Guido Neidhöfer
ZEW

Margaret Thatcher

First, that the pursuit of equality itself is a mirage. What's more desirable and more practicable [...] is the pursuit of equality of opportunity.

Speech to the Institute of SocioEconomic Studies
New York, September 15, 1975

Raul Castro

Socialismo significa justicia social e igualdad, pero igualdad de derechos, de oportunidades, no de ingresos.

Speech at the Asamblea Nacional del Poder Popular
La Habana, July 11, 2008

EOP

- why so successful?

  1. EOP = equality + freedom;

  2. EOP is sufficiently vague.

- our contribution: set a standard.

Literature

3 generations of contributions:

- theory: Rawls (1971), Dworkin (1981), Arneson (1989) and Cohen (1989), Fleurbaey (1994), Roemer (1998);
- IOP measurement: Bourguignon et al. (2007), Lefranc et al. (2009), Checchi and Peragine (2010), Almas et al. (2011), Ferreira and Gignoux (2011);
- econometric specification: Li Donni et al. (2015), Carrieri and Jones (2018), Brunori, Hufe and Mahler (2018).

Roemer's Model

$$y_i = g(C_i, e_i)$$

- y_i: individual's i outcome;
- C_i: circumstances beyond individual control;
- e_i: effort;
- no random component:

$$e_i = e_j \cap C_i = C_j \rightarrow y_i = y_j \ , \ \ \forall i, j \in 1, ..., n$$

Types, tranches and IOP

- Romerian *type*: individuals sharing same circumstances;
- effort *tranche*: individuals exerting the same effort;

$$\text{EOP} \iff e_i = e_j \rightarrow y_i = y_j \ , \ \forall i, j \in 1, ..., n$$

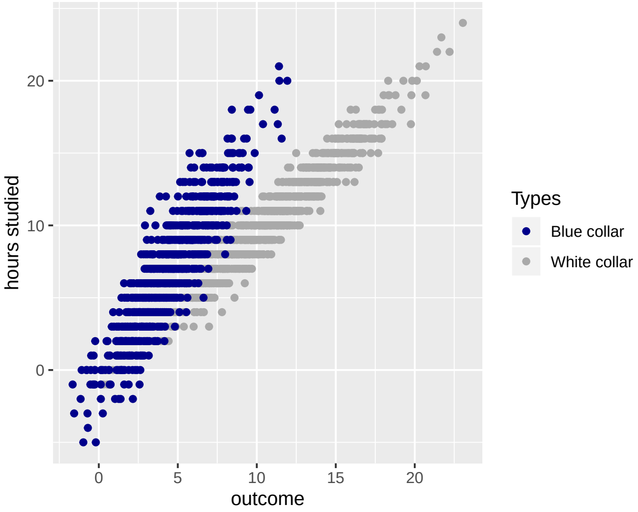
- Then: IOP = within-tranche inequality.

Effort

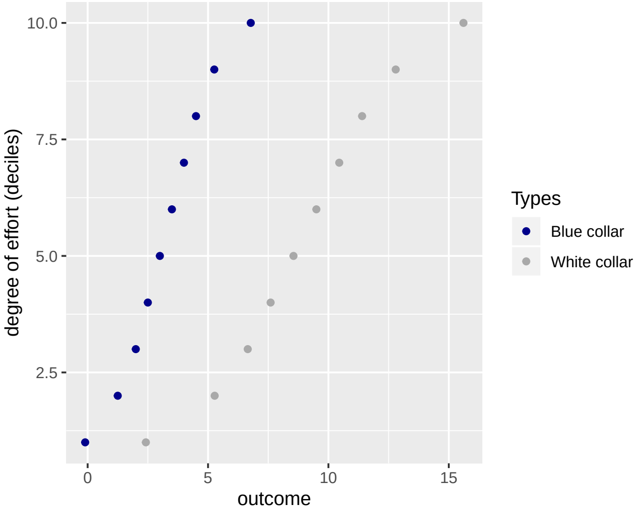
- Roemer's identification strategy, two assumptions:

    1. monotonicity: $\frac{\partial g}{\partial e} \geq 0$
    2. orthogonality: $e \perp\!\!\!\perp C$

Effort identification



Effort identification, cnt



Degree of effort

- with observable effort = quantile of the type-specific effort distribution;
- with unobservable effort = quantile of the type-specific outcome distribution.

3-step estimation

1. identification of Romerian types;
2. measurement of degree of effort exerted;
3. (Roemer) IOP = within-tranche inequality

Roemerian types

- two empirical issues of Roemer's theory:

    1. unobservable circumstances (underfitted model);

    2. sparsely populated types (overfitted model).

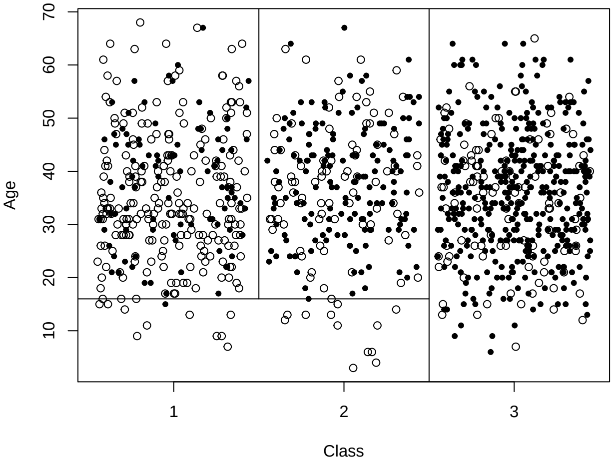
- bias-variance trade-off → downward - upward bias;

- preferred IOP estimates: min MSE.

Romerian types, cnt

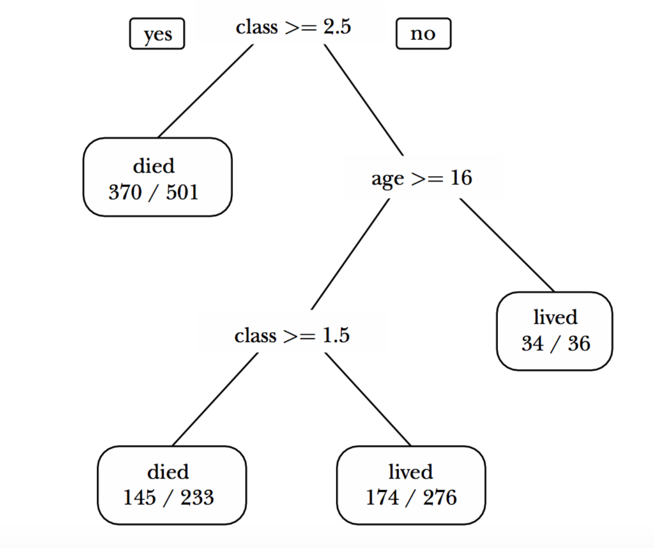
- we use regression tree to identify types;
- partition the space of regressors into non-overlapping regions (Morgan and Sonquist,1963; Breiman et al.,1984)
- the population is divided into non-overlapping subgroups
- prediction of each observation is the the mean value of the dependent variable in the group

What is a tree? cnt.



Source: Varian, 2014

What is a tree? cnt.



Source: Varian, 2014

What is a tree? cnt.

- overfitted models explain perfectly in-sample (high in-sample IOP);
- but perform poorly out-of-sample (low out-of-sample IOP);
- different restrictions to prevent overfitting lead to different types' partition.

Conditional inference trees

- we use *conditional inference trees* (Hothorn et al., 2006);
- splitting are based on a sequence of statistical test;
- Brunori, Hufe, Mahler (2018): when IOP measurement is understood as a prediction problem they outperform standard methods in identifying types.

The algorithm

- choose α
- $\forall p$ test the null hypothesis of independence:
  $H^{C_p} = D(Y|Cp) = D(Y), \forall C_p \in \mathbf{C}$
- if no (adjusted) p-value $< \alpha \rightarrow$ exit the algorithm
- select the variable, $C^\star$, with the lowest p-value
- test the discrepancy between the subsamples for each possible binary partition based on $C^\star$
- split the sample by selecting the splitting point that yields the lowest p-value
- repeat the algorithm for each of the resulting subsample

Effort

- recall: IOP quantifies to what extent individuals exerting the same degree of effort obtain the same outcome;
- standard approach: choose an arbitrary number of quantiles;
- low efficiency and limited comparability across studies.

Bernstein polynomials

- approximate the ECDF with a polynomial;
- for any quantile $\pi \in [0, 1]$ we can predict the expected outcome in all types;
- we use Bernstein polynomials.

Bernstein polynomials

- Sergei Bernstein (1912)
- mathematical basis for curves' approximation in computer graphics
- outperform competitors (kernel estimators) in approximating distribution functions (Leblanc, 2012)

Bernstein polynomial of degree 4

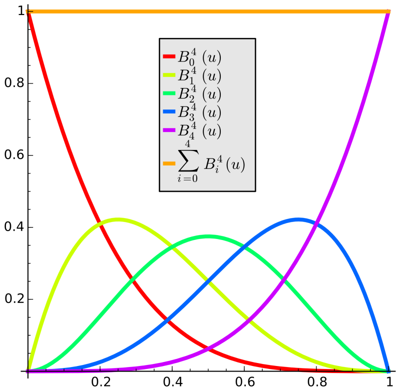
$$B_4(x) = \sum_{v=0}^{4} \beta_v b_{v,4}$$

where β_vs need to be estimated and the Bernstein basis polynomial $b_{v,k}$ is:

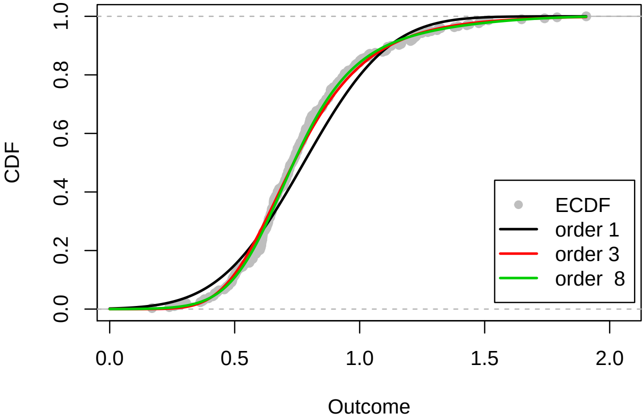
$$b_{v,k} = \binom{k}{v} x^v (1-x)^{k-v}$$

$$\begin{aligned} b_{0,4} &= (1-x)^4 \\ b_{1,4} &= 4x(1-x)^3 \\ b_{2,4} &= 6x^2(1-x)^2 \\ b_{3,4} &= 4x^3(1-x) \\ b_{4,4} &= x^4 \end{aligned}$$

Bernstein polynomials, cnt



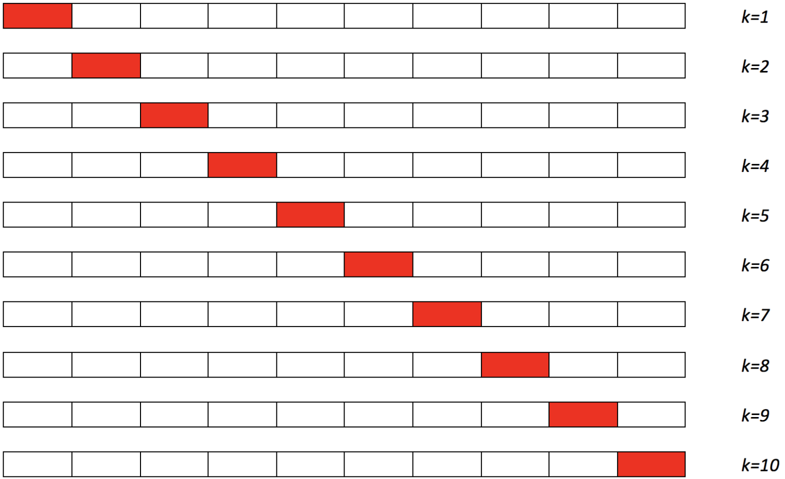
ECDF approximation by Bernstein polynomials



Choice of the polynomial's degree

- the polynomial is estimated with the *mlt* algorithm written by Hothorn (2018);
- out-of-sample log-likelihood to select the most appropriate order of the polynomial;
- out-of-sample log-likelihood is estimated by 10-fold cross validation;

k-fold cross validation



10-fold Cross Validation

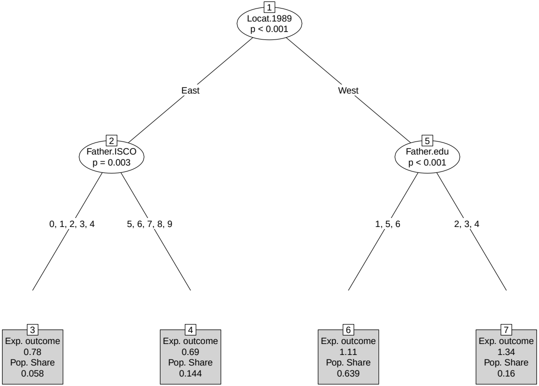
IOP estimation

- Shape of all type-specific distribution functions $\rightarrow$ distribution of EOP violations
- $IOP = Gini\left(\frac{y_i}{\mu_j}\right)$, μ_j expected outcome at percentile j;
- no longer need to choose a particular number of effort quantiles;
- number of quantiles varies to maximize estimate reliability.

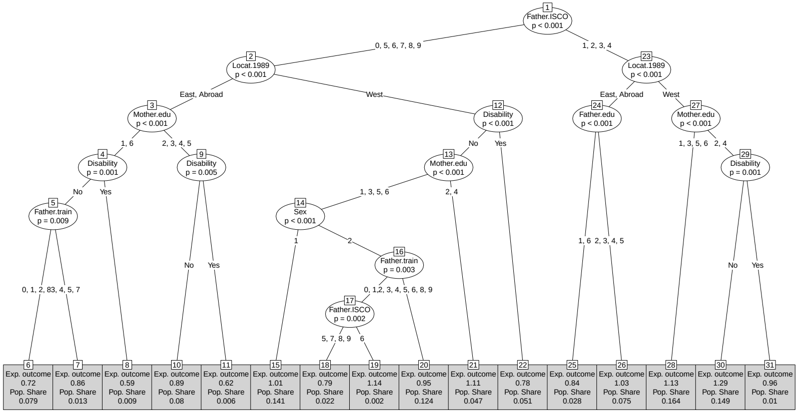
Data

- SOEP (v33) including all subsamples apart from the refugee samples;
- 25 waves 1992-2016;
- adult individuals (30-60);
- circumstances considered: migration background, location in 1989, mother's education, father's education, father's occupation, father's training, month of birth, disability, siblings;
- outcome: 'age-adjusted' household equivalized income

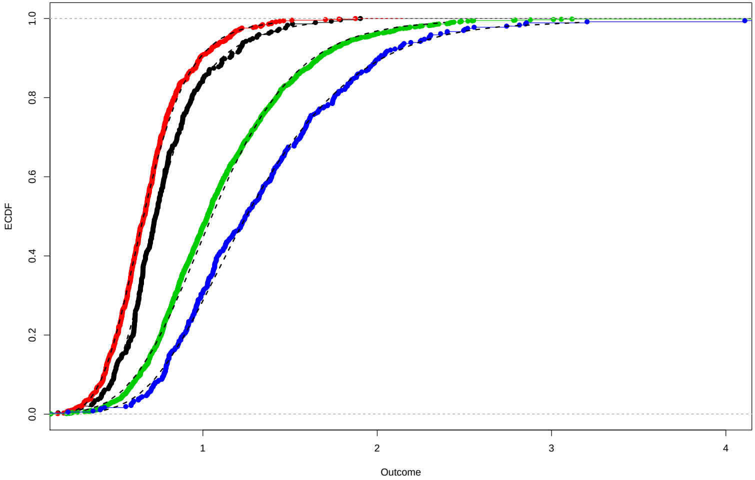
Opportunity tree in 1992



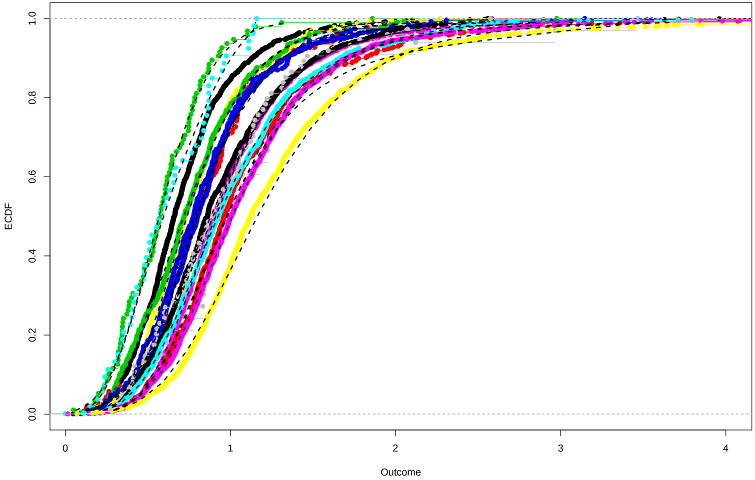
Opportunity tree in 2016



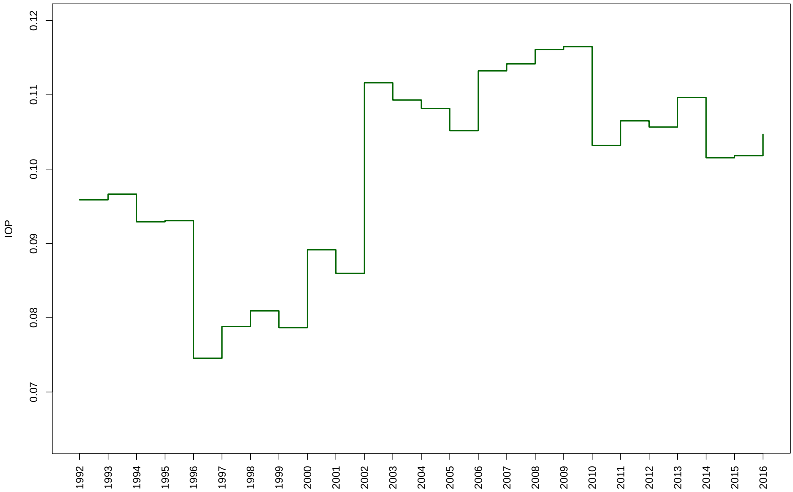
IOP in 1992



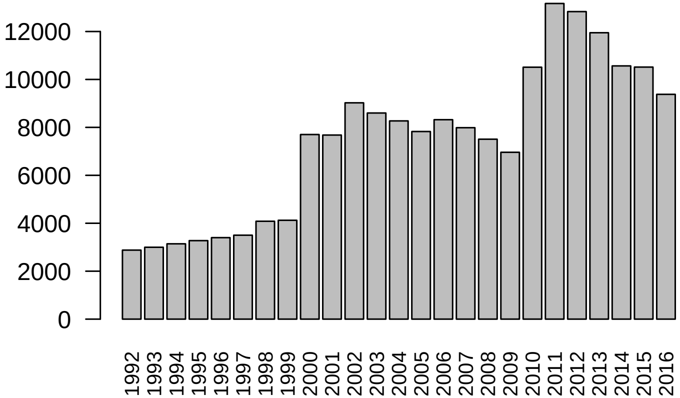
IOP in 2016



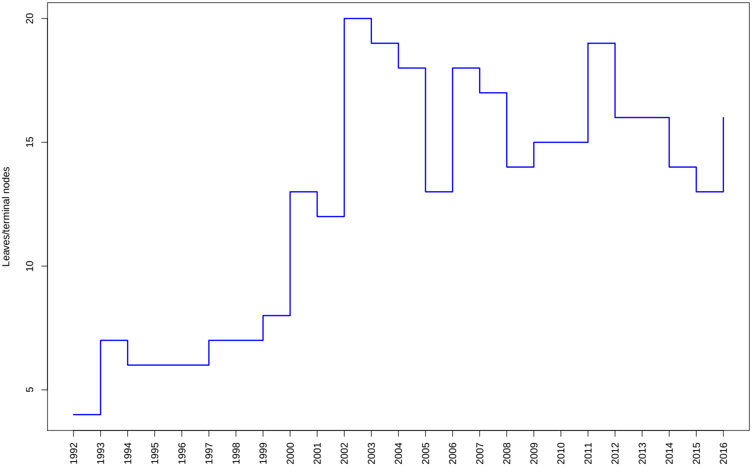
IOP trend 1992-2016



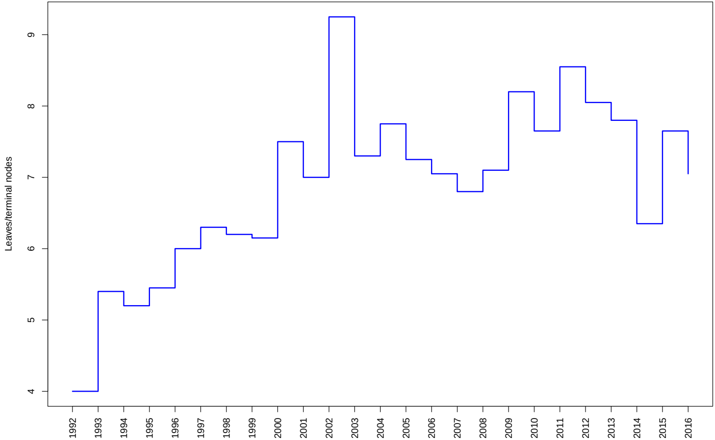
Sample size 1992-2016



Number of types 1992-2016



Mean number of types (same sample size) 1992-2016



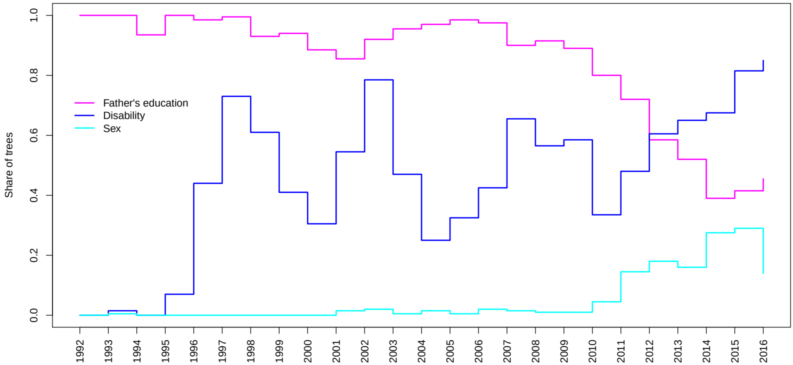
Mean IOP trend 1992-2016 (same sample size)

Confidence bounds are the 0.975 and 0.025 quantiles of the distribution of IOP estimates.

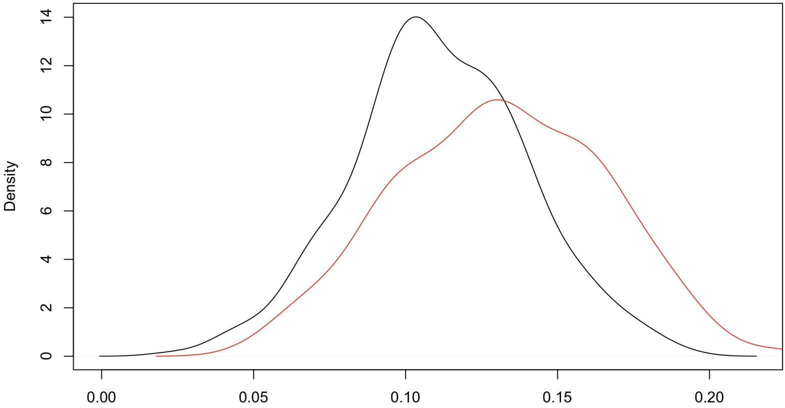
Summary

- wan approach to estimate IOP fully consistent to Roemer's theory;
- effort identification method maximizes efficiency and comparability;
- since 1992 in Germany the opportunity structure has become more complex;
- IOP declined after reunification and increased with *Hartz reforms*;
- is today about 10% higher than in 1992.

Share of trees that use fathers education, disability and sex to obtain Romerian type



Distribution of bootstrap estimates



Mother/father raining

mtraining / ftraining

cod.	Berufsbildung M/V	Vocational Training M/F
1	Keine Ausbildung	No vocational degree
2	Berufliche Ausbildung	Vocational Degree
3	Gewerbliche oder Landwirtschaftliche Leh	Trade or Farming Apprentice
4	Kaufm.L.,Bfs,Handel	Business
5	Gesundheitswesen, FS,Techn.-o.Meisters	Health Care or Special Technical School
6	Beamtenausbildung	Civil Service Training
7	FHS,Ingeniuerschule	Tech Engineer School
8	Hochsch.,Universit. (In- und Ausland)	College, University (in GER or Abroad)
9	Sonstige Ausbildung	Other Training

Mother/father education

fsed / msed

cod.		Schulbildung Vater / Mutter	Father/Mother Education
	1	[1] Hauptschule	Lower Secondary
	2	[2] Realschule	Intermediate Secondary
	3	[3] Fachoberschule	Technical School
	4	[4] Abitur	Upper Secondary
	5	[5] sonstiger Abschluss	Other School Degree
	6	[6] Kein Abschluss	No School Degree
	7	[7] Keine Schule besucht	School not attended