DIW BERLIN

4.2: Correction for item non-response

Markus M. Grabka (DIW Berlin)

Meeting of providers of OECD income distribution data

Paris, 21./22. February 2013, OECD

1 Introduction

Types of Non response

- Unit non-response
- Partial Unit non-response

  non-response of at least one unit, or member, of an otherwise participating household
- Item Non-response

2 Causes of Non Response

- Complexity of surveyed construct (Schräpler 2003)
- Formulation of questions matters (Hill & Willis 2001)
- Interviewer-respondent matching, lacking intimicy (Schräpler & Wagner 2001)
- Interviewer change (Riphahn & Serfling 2003)
- Ignorance (Giami 2012)
- INR strongly related to Income Inequality and mobility (e.g. higher refusals in tails of income distribution) (Wooden & Watson 2006)
- INR and UNR not independent: INR in t = predictor of UNR in t+1 (Lee et al 2004, Loosfeldt et.al. 1999)
- ...

3 How to deal with INR?

Rubin (1976): Missing mechanism (MCAR, MAR, MNAR)

- case-wise deletion (only valid observations)
- weighting
- imputation
  - ✓ *single* imputation techniques
    - institutional imputation
    - expert imputation
    - mean substitution
    - cold and hot deck
    - regression-based
    - row-and-column-imputation using longitudinal data (Little & Su 1989)
    - etc.
  - ✓ *multiple* imputation techniques (e.g. MICE)

4 Some Pros and Cons of Imputation

- **Pro**
  - ✓ complete dataset with “full” information
  - ✓ reduce potential bias in survey estimates if MNAR
  - ✓ imputation makes analysis easier (less experienced users)

- **Cons**
  - ✓ imputation cannot eliminate all non-response bias
  - ✓ in case of variance reduction → “significant” results more likely
  - ✓ may produce additional bias (e.g., regression to the mean, non-harmonization across surveys)

5 Incidence of INR

Incidence of INR in selected SOEP income components

	1986 Sample A-B	1993 Sample A-C	2001 Sample A-F
	Share of persons with INR (%)		
Labor income from main job	5.0	5.8	9.2
Income from self-employment	17.9	14.4	25.7
Christmas Gratification	4.1	3.6	6.3
Pension Income (own)	8.8	5.7	2.7
Child Benefit	0.8	2.1	4.4
Interest & Dividends	19.0	12.3	13.9

Source: SOEP, Survey years 1986, 1993, 2001; unweighted results.

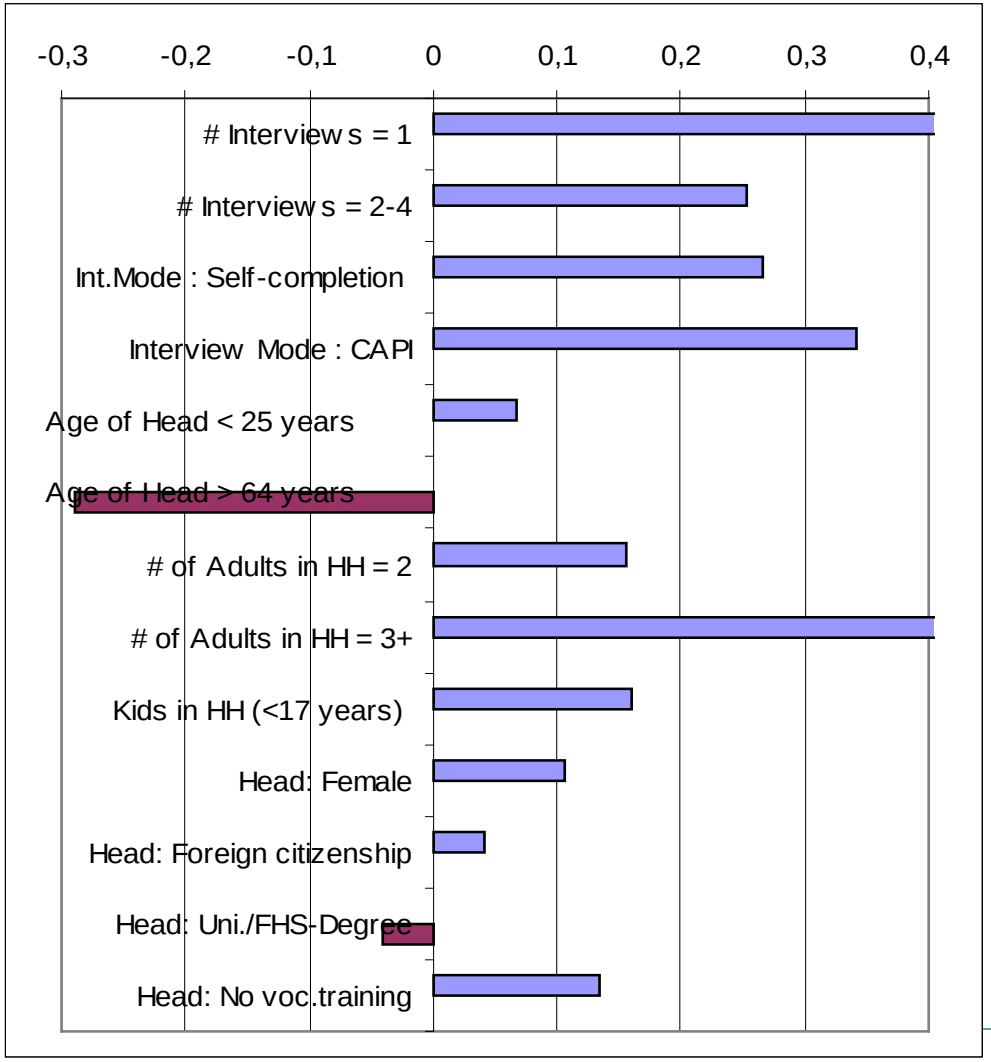
6 Selectivity of INR

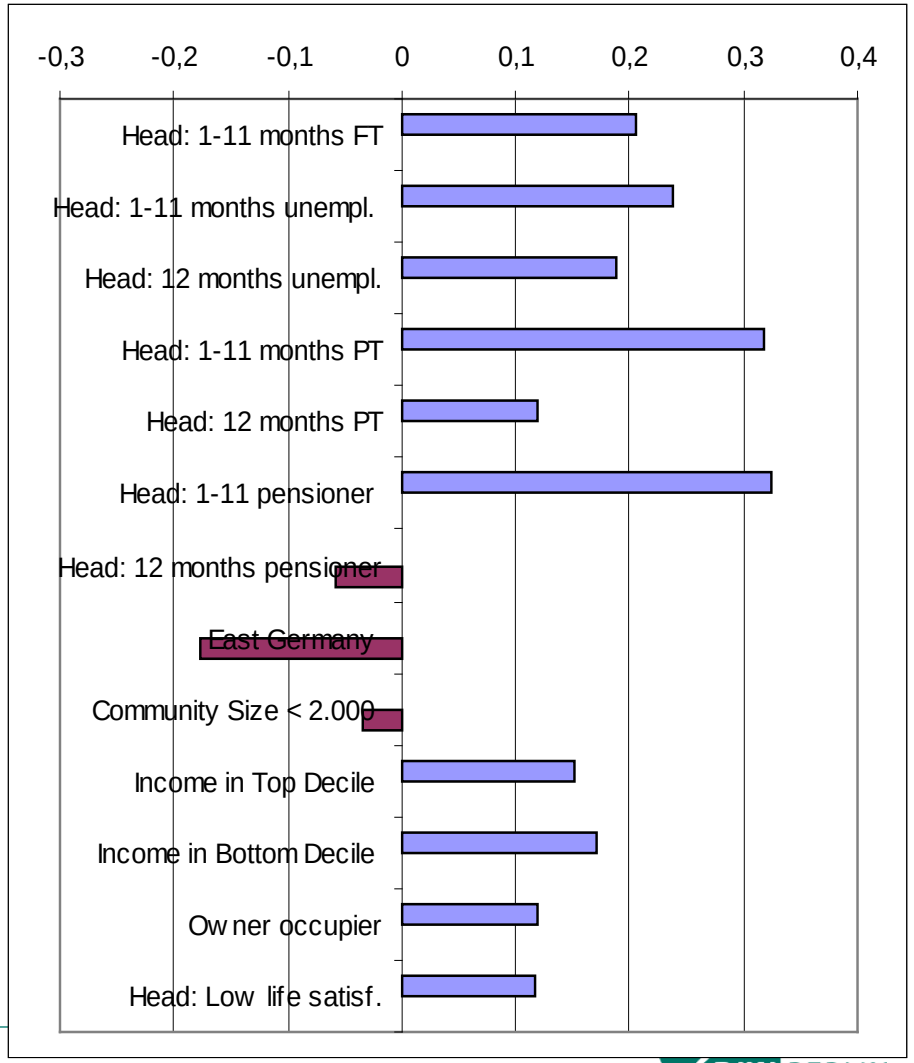
Estimating the probability for INR on *total annual household income* (pre-tax, post-transfer)

→ Results from a Random Effects Probit Model 1993-2002

7 Selectivity of INR

Probability of INR in "Total Income", 1993-2002





Source: SOEP.

8 Principles of Imputation

- Using longitudinal information in the Imputation process

  from previous & future interviews

  $[t_0, \ldots, t_{-1}, t_{+1}, \ldots, t_m]$ for imputation of INR in t

- EU-financed CHINTEX-Project (Change from input to ex-post harmonization)

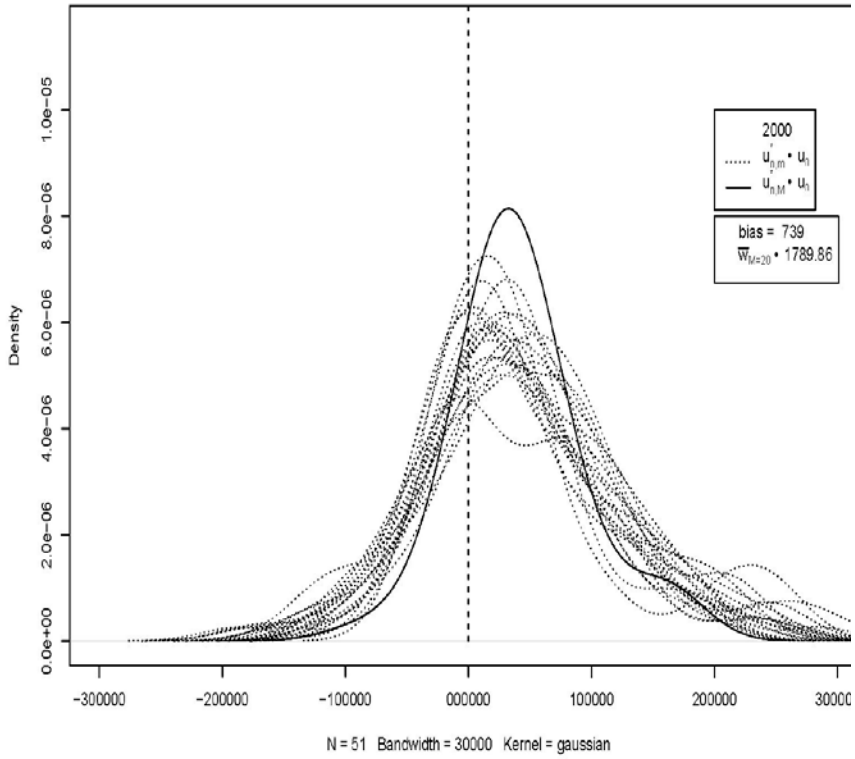
  ECHP: existing imputation mainly based on x-sectional methods

  Spiess & Göbel (2003): longitudinal imputation improves data quality significantly (multiple imputation techniques)

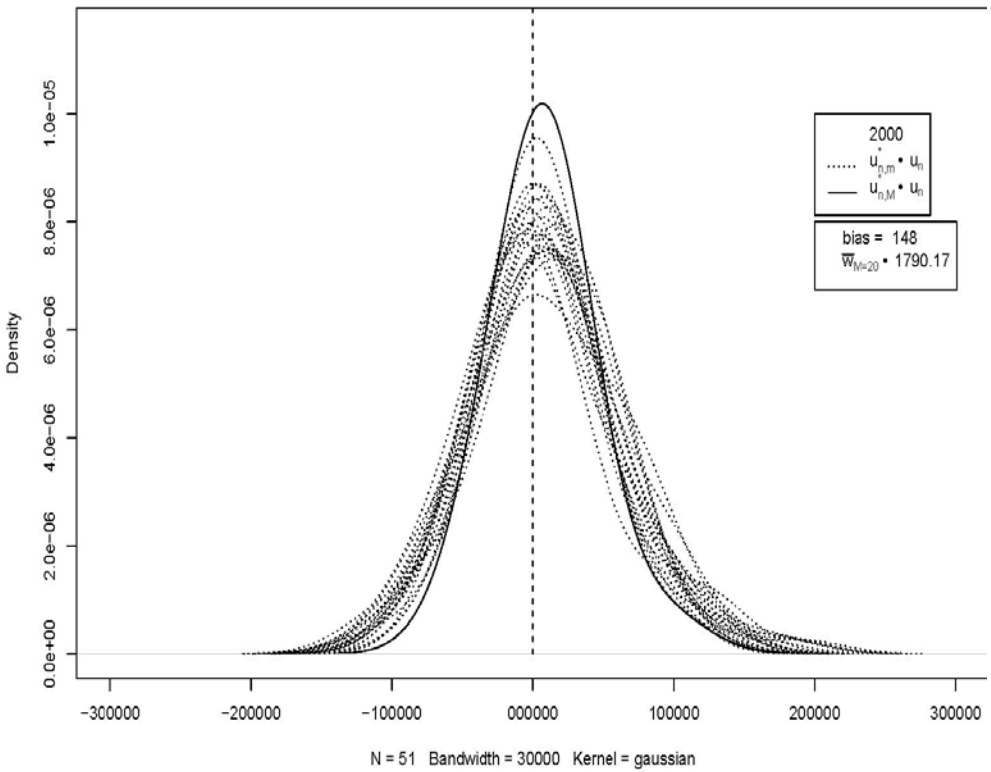
9 Principles of Imputation

Difference between "Imputed values and register data"
Annual Gross Labor Income, Finland 2000 (Spiess & Göbel 2003)

Current ECHP-imputation



CHINTEX-imputation



DIW BERLIN

10 Does imputation help to draw the right picture ?

- Equivalent Post-Government income

  (Equiv. scale = SQRT HH-Size)

- Income Distribution, Inequality, Poverty
- Income Mobility

- Comparing results from analysis based on

  "Observed cases only" vs. "All cases (observed + imputed)"

Does imputation help to draw the right picture ?

Equivalent Post-Gov't Income, Inequality and Poverty 2001

	Imputation Status			Deviation: "All" vs. "Observed" (%)
	"All cases"	"Observed cases"	"Imputed cases"	
Mean	**36 760** *[35 977; 37 366]*	**36 152** *[35 385; 36 750]*	**39 015** [37 234; 40 221]	**+1,7**
Median	**33 334** *[33 175; 33 689]*	**32 797** *[32 203; 33 343]*	**35 720** *[33 912; 37 860]*	**+1,6**
MLD	**0.1350** *[0.1266; 0.1405]*	**0.1283** *[0.1169; 0.1346]*	**0.1577** *[0.1444; 0.1740]*	**+5,2**
Gini	**0.2698** *[0.2616; 0.2771]*	**0.2641** *[0.2552; 0.2741]*	**0.2855** *[0.2674; 0.2886]*	**+2,2**
SCV	**0.2977** *[0.2580; 0.3170]*	**0.2961** *[0.2440; 0.3217]*	**0.2958** *[0.2379; 0.3237]*	**+0,5**
90:10	**3.44** *[3.29; 3.66]*	**3.32** *[3.16; 3.54]*	**3.89** *[3.58; 4.05]*	**+3,6**
90:50	**1.77** *[1.69; 1.85]*	**1.75** *[1.70; 1.85]*	**1.79** *[1.69; 1.88]*	**+1,1**
50:10	**1.94** *[1.90; 2.01]*	**1.89** *[1.85; 1.95]*	**2.18** *[2.04; 2.41]*	**+2,6**
Poverty Rate	**14.3** *[13.1; 15.4]*	**14.1** *[12.3; 15,5]*	**15.2** *[14.9; 15,9]*	**+1,4**
N (x-section 2001)	27 915	22 399	5 516	+24,6

Values in brackets give a 93% confidence interval based on variance estimation according to the Random Group Approach.

Shaded cells indicate statistically significant deviations

12 Does imputation help to draw the right picture ?

Income Mobility and Imputation 2000-2001

Index	Imputation Status			Deviation "All" vs. "Observed" (%)
	"All cases"	"Observed cases"	"Imputed cases"	
Quintile Matrix Mobility	**0.467**	**0.413**	**0.584**	**+13.1**
Average jump	*[0.452; 0.483]*	*[0.394; 0.437]*	*[0.541; 0.618]*	
Quintile Matrix Mobility	**0.187**	**0.165**	**0.234**	**+13.3**
Normalized average jump	*[0.181; 0193]*	[0.157; 0.175]	*[0.216; 0.247]*	
Fields & Ok (1996)	**18.36**	**15.99**	**22.88**	**+14.8**
Percentage income mobility	*[17.23; 19.43]*	[15.34; 16.78]	*[21.52; 23.78]*	
Fields & Ok (1999)	**0.210**	**0.185**	**0.259**	**+13.5**
Non-directional	*[0.200; 0.220]*	[0.167; 0.203]	*[0.244; 0.271]*	
Shorrocks (1987)	**0.0829**	**0.0748**	**0.0955**	**+10.8**
using Theil Coefficient	*[0.0709; 0.095]*	*[0.0629; 0.0811]*	*[0.0814; 0.1040]*	
N (balanced panel)	*26 609*	*18 201*	*8 408*	*+46.2*

Values in brackets give a 93% confidence interval based on variance estimation according to the Random Group Approach.

13 Summary

- Item-non-response is highly selective
- Case-wise deletion yields to biased results
- Imputation provides effective means to cope with INR
- Empirical results suggest to make use of longitudinal data
- Relevance of INR (and imputation) varies significantly across income distribution
- Panel Research: INR is positively correlated with any type of non-response in subsequent waves
- Consideration of multiple Imputation, to cope with the uncertainty embedded in the imputation process, however data structure is more complex
- Harmonization of Imputation technique across countries matters (Frick & Grabka 2010)

Thank you for your attention.

DIW Berlin — Deutsches Institut für Wirtschaftsforschung e.V.
Mohrenstraße 58, 10117 Berlin
www.diw.de

Editor
Dr. Markus M. Grabka
(mgrabka@diw.de)

Harmonization of imputation techniques

- Germany: Socio-Economic Panel Study, **SOEP** [1992-2004]
- Australia: Household, Income and Labour Dynamics in Australia, **HILDA** [2001-2005]
- UK: British Household Panel Survey, **BHPS** [1991-2004]

Variable of Interest: Individual Annual Labor Earnings

Imputation strategy: SOEP & HILDA

- 1. Step: “row-and-column-imputation” (*Little & Su* 1989):
  + nearest neighbor matching including an error term
  - SOEP: entire population
  - HILDA: within several age groups

- 2. Step (in case of lacking longitudinal information)
  - SOEP: Hot deck + randomly chosen error term
  - HILDA: nearest neighbor regression method

- Data Quality Check (simulation studies):
  - *Starick* (2005): L&S in case of HILDA yields more reliable results than hot deck
  - *Frick&Grabka* (2005): similar finding for SOEP → longitudinal imputation superior to purely cross-sectional imputation

Imputation strategy: BHPS

- regression based predictive mean matching (PMM), *Little* (1988) = Hot Deck
- shifting 3-year windows → 14 different regression models
- Advantage: a truly observed value is used as basis for imputation plus error term
- "Quality": R^2 in the first 3 waves varies between 0.78 and 0.94

Does the choice of the imputation technique affect substantive research results – and thus cross-national comparability ?

- ✓ Robustness check by implementing the "row-and-column" imputation for BHPS
- ✓ Comparison of …
  - HILDA "Row-and-column" imputation
  - SOEP "Row-and-column" imputation
  - BHPS (1) Original imputation
  - **BHPS (2) "Row-and-column" imputation**

Imputation and labor income inequality

(→ using only observed cases understates inequality)

	Germany (SOEP)		Australia (HILDA)		UK (BHPS)		UK (BHPS)-L+S	
	"All cases"	Deviation: "All" vs. "Obs."	"All cases"	Deviation: "All" vs. "Obs."	"All cases"	Deviation: "All" vs. "Obs."	"All cases"	Deviation: "All" vs. "Obs."
Mean	24408	0,0	27349	-1,0	13621	-1,8	13849	-0,2
Median	21940	-0,6	23375	-2,3	11360	-2,7	11553	-1,1
Income inequality								
Theil 0 (MLD)	0,4096	+1,0	0,4587	+4,5	0,4425	+8,6	0,4211	+3,4
Gini	0,4141	+1,0	0,4273	+1,9	0,4280	+1,7	0,4268	+1,4
Half-SCV	0,3488	+3,5	0,4456	+1,5	0,4709	+4,9	0,4652	+3,6
Decile ratio 90:10	13,7	+0,4	14,9	+5,2	12,7	+6,1	12,4	+3,9
Decile ratio 50:10	6,5	-0,6	6,8	+2,7	5,4	+4,2	5,4	+3,2
Average N *per cross-section*	***10773***	*+13,4*	***9082***	*+15,3*	***5002***	*+18,1*	***5002***	*+18,1*

Source: SOEP survey years 1992-2004; HILDA survey years 2001-2005; BHPS survey years 1991-2004.

Shaded cells indicate statistically significant deviations

2 Results from fixed-effects wage regression

	Germany (SOEP)	
	observed	all cases
Age	+++	+++
Age squared	---	---
Female with kid(s)*	---	---
Male with kid(s) *	+++	+++
Disability Status *		--
Married *		
Metrop. area *	+++	+++
Remote area*		
Intermed. education*	--	--
Upper education*		
Highest educ. level*	+++	+++
East Germany*	---	---
Self employed*	--	
Became retired*		
Left education *	---	---
Unempl. (last year) *	---	---
Months FT (last year)	+++	+++
Months PT (last year)	+++	+++
Imputed Labor Y*		**0.064*** **
Constant	+++	+++
Observations	119030	134337
N (Persons)	24183	25487
R-squared	**0,49**	**0,45**

Population: working age: 20-60 (Germany), 20-65 (Australia and UK). (*) indicates dummy variables.
Note: Time effects controlled, but not reported. Significance level: +++/--- sig. at 1%; ++/-- sig. at 5%; +/- sig. at 10%.
Source: SOEP survey years 1992-2004; HILDA survey years 2001-2005; BHPS survey years 1991-2004.

2 Results from fixed-effects wage regression

	Germany (SOEP)		Australia (HILDA)	
	observed	all cases	observed	all cases
Age	+++	+++	+++	+++
Age squared	---	---	---	---
Female with kid(s)*	---	---	---	---
Male with kid(s) *	+++	+++	---	---
Disability Status *		--	-	-
Married *			+++	+++
Metrop. area *	+++	+++	+	+
Remote area*				
Intermed. education*	--	--	++	++
Upper education*			+++	+++
Highest educ. level*	+++	+++	+++	+++
East Germany*	---	---	**n.a.**	**n.a.**
Self employed*	--		---	
Became retired*			+++	+++
Left education *	---	---	--	--
Unempl. (last year) *	---	---	---	---
Months FT (last year)	+++	+++	+++	+++
Months PT (last year)	+++	+++	**n.a.**	**n.a.**
Imputed Labor Y*		**0.064*** **		**0,052** **
Constant	+++	+++	+++	+++
Observations	119030	134337	35661	38681
N (Persons)	24183	25487	11097	11522
R-squared	**0,49**	**0,45**	**0,22**	**0,17**

Population: working age: 20-60 (Germany), 20-65 (Australia and UK). (*) indicates dummy variables.
Note: Time effects controlled, but not reported. Significance level: +++/--- sig. at 1%; ++/-- sig. at 5%; +/- sig. at 10%.
Source: SOEP survey years 1992-2004; HILDA survey years 2001-2005; BHPS survey years 1991-2004.

2 Results from fixed-effects wage regression

	Germany (SOEP)		Australia (HILDA)		UK (BHPS)	
	observed	all cases	observed	all cases	observed	all cases
Age	+++	+++	+++	+++	+++	+++
Age squared	---	---	---	---	---	---
Female with kid(s)*	---	---	---	---	---	---
Male with kid(s) *	+++	+++	---	---	--	--
Disability Status *		--	-	-	-	--
Married *			+++	+++	+++	+++
Metrop. area *	+++	+++	+	+	+++	+++
Remote area*					+	++
Intermed. education*	--	--	++	++		
Upper education*			+++	+++	+++	+++
Highest educ. level*	+++	+++	+++	+++	+++	+++
East Germany*	---	---	**n.a.**	**n.a.**	**n.a.**	**n.a.**
Self employed*	--		---		---	---
Became retired*			+++	+++	---	---
Left education *	---	---	--	--	---	---
Unempl. (last year) *	---	---	---	---	+++	+++
Months FT (last year)	+++	+++	+++	+++	+++	+++
Months PT (last year)	+++	+++	**n.a.**	**n.a.**	**n.a.**	**n.a.**
Imputed Labor Y*		**0.064*** **		**0,052** **		**-0,042** **
Constant	+++	+++	+++	+++	+++	+++
Observations	119030	134337	35661	38681	62049	72729
N (Persons)	24183	25487	11097	11522	10352	11137
R-squared	**0,49**	**0,45**	**0,22**	**0,17**	**0,52**	**0,44**

Population: working age: 20-60 (Germany), 20-65 (Australia and UK). (*) indicates dummy variables.
Note: Time effects controlled, but not reported. Significance level: +++/--- sig. at 1%; ++/-- sig. at 5%; +/- sig. at 10%.
Source: SOEP survey years 1992-2004; HILDA survey years 2001-2005; BHPS survey years 1991-2004.

2 Results from fixed-effects wage regression

	Germany (SOEP)		Australia (HILDA)		UK (BHPS)		UK (BHPS) – "L&S"	
	observed	all cases	observed	all cases	observed	all cases	observed	all cases
Age	+++	+++	+++	+++	+++	+++	+++	+++
Age squared	---	---	---	---	---	---	---	---
Female with kid(s)*	---	---	---	---	---	---	---	---
Male with kid(s) *	+++	+++	---	---	--	--	--	--
Disability Status *		--	-	-	-	--	-	
Married *			+++	+++	+++	+++	+++	+++
Metrop. area *	+++	+++	+	+	+++	+++	+++	+++
Remote area*					+	++	+	
Intermed. education*	--	--	++	++				
Upper education*			+++	+++	+++	+++	+++	+++
Highest educ. level*	+++	+++	+++	+++	+++	+++	+++	+++
East Germany*	---	---	**n.a.**	**n.a.**	**n.a.**	**n.a.**	**n.a.**	**n.a.**
Self employed*	--		---		---	---	---	---
Became retired*			+++	+++	---	---	---	---
Left education *	---	---	--	--	---	---	---	---
Unempl. (last year) *	---	---	---	---	+++	+++	+++	+++
Months FT (last year)	+++	+++	+++	+++	+++	+++	+++	+++
Months PT (last year)	+++	+++	**n.a.**	**n.a.**	**n.a.**	**n.a.**	**n.a.**	**n.a.**
Imputed Labor Y*		**0.064*** **		**0,052** **		**-0,042** **		**0,047** **
Constant	+++	+++	+++	+++	+++	+++	+++	+++
Observations	119030	134337	35661	38681	62049	72729	62049	72904
N (Persons)	24183	25487	11097	11522	10352	11137	10352	11138
R-squared	**0,49**	**0,45**	**0,22**	**0,17**	**0,52**	**0,44**	**0,52**	**0,37**

Population: working age: 20-60 (Germany), 20-65 (Australia and UK). (*) indicates dummy variables.
Note: Time effects controlled, but not reported. Significance level: +++/--- sig. at 1%; ++/-- sig. at 5%; +/- sig. at 10%.
Source: SOEP survey years 1992-2004; HILDA survey years 2001-2005; BHPS survey years 1991-2004.

2 Quantile wage regressions: Coefficients for "imputation flag“

	p25	p50	p75
Germany (SOEP)	-0,027**	0,004	+0,046**
Australia (HILDA)	-0.195**	-0.017	+0.078**
UK (BHPS)	-0.119**	-0.076**	-0.036**

Population: working age: 20-60 (Germany), 20-65 (Australia and UK).
Note: Controls include age, sex, kids in HH, marital status, health status, region, education, (change in) employment status, unemployment experience, time effects.
+++/--- sig. at 1%; ++/-- sig. at 5%; +/- sig. at 10%.
Source: SOEP survey years 1992-2004; HILDA survey years 2001-2005; BHPS survey years 1991-2004.

2 Quantile wage regressions: Coefficients for "imputation flag“

	p25	p50	p75
Germany (SOEP)	-0,027**	0,004	+0,046**
Australia (HILDA)	-0.195**	-0.017	+0.078**
UK (BHPS)	-0.119**	-0.076**	-0.036**
UK (BHPS - "L+S")	-0.139**	-0.004	+0.066**

Population: working age: 20-60 (Germany), 20-65 (Australia and UK).
Note: Controls include age, sex, kids in HH, marital status, health status, region, education, (change in) employment status, unemployment experience, time effects.
+++/--- sig. at 1%; ++/-- sig. at 5%; +/- sig. at 10%.
Source: SOEP survey years 1992-2004; HILDA survey years 2001-2005; BHPS survey years 1991-2004.