

Item-non-response and Imputation of Labor Income in Panel Surveys:

A Cross-National Comparison based on SOEP, BHPS, HILDA

Joachim R. Frick (DIW Berlin, TU Berlin, IZA Bonn) &

Markus M. Grabka (DIW Berlin)

Official Statistics System (OSS) Seminar Series, Statistics New Zealand,

8 November 2007, Wellington

Motivation

- When analyzing substantive research issues ...
- ... the comparison across time and space ...
- ... requires consideration / harmonization of ...
 - research question [striving for best practice]
 - methodology [definition of dependent construct]
 - data production incl. post-survey treatment [how to deal with measurement error]

Structure of Presentation

- I. Background: INR in panel surveys
- II. Data – 3 panels included in CNEF
 - Questionnaires
 - Production of *aggregated* annual labor income
 - Probability and selectivity of Item-non-response (INR)
- III. Imputation techniques
- IV. Empirical application: A cross-national perspective
 - Incidence of INR
 - How does imputation affect inequality & mobility
 - INR in a longitudinal perspective
 - Imputed data in wage regressions
- V. Conclusion (from a cross-national perspective)

I. Background

Item-non-response (INR) in population surveys (esp. panel data):

- *Schräpler* (2003): complexity of surveyed construct
- *Hill & Willis* (2001): formulation of questions matters
- *Schräpler & Wagner* (2001): interviewer-respondent matching
- *Rendtel* (1995), *Riphahn & Serfling* (2003): interviewer change
- *Jarvis & Jenkins* 1998; *Biewen* 2001; *Frick & Grabka* 2005; *Riphahn & Serfling* 2005, *Hawkes & Plewis* 2006; *Wooden & Watson* 2006: INR strongly related to Income Inequality and mobility (e.g. higher refusals in tails of income distribution)
- *Lee et al* (2004): INR and UNR not independent (to be modelled together)
- *Loosfeldt et.al.* (1999): INR in t = predictor of UNR in $t+1$
- *Burton et al* (1999): *Cooperation continuum*
(complete answers — incomplete answers [INR] — no answer at all [UNR])

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, etc)
 - ✓ *multiple* imputation techniques

II. Data

- Germany: German Socio-Economic Panel Study, SOEP [1992-2004] ;
New Sample F [2000-2004]
- 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
(previous year's income from dependent and self-employment, incl. extra payments)

Questionnaire - BHPS

The last time you were paid, what was **your gross pay - that is including any overtime, bonuses, commission, tips or tax refund, but before any deductions** for tax, national insurance or pension contributions, union dues and so on?

IF 'DON'T KNOW / CAN'T REMEMBER' PROBE: 'Can you give me an approximate amount?'

ENTER TO NEAREST £: ASK E21 *IPAYGL*

Don't know..... 8 **GO TO E22**

Refused 9 **GO TO E31 (page 43)**

RESPONDENT TO CHECK PAY SLIP IF POSSIBLE

Questionnaire - HILDA

F19 Last financial year, what was your total wage and salary income from all jobs before tax or anything else was deducted? Do not include income from businesses. This should be gathered at F24, Enter annual amount

(whole \$) \$ _____ → F22

Don't know.....999999 → F20

F22 During the last financial year did you, at any time: work in your own business or farm; or were a silent partner in a partnership; or were a beneficiary of a trust (excluding those that are used just for investment purposes)?

Yes.....1

No..... 2 → F28a

F24 Excluding dividends, in the last financial year, what was your total income from wages and salary from these incorporated businesses before income tax was deducted? Please exclude wages and salary already reported. This includes trusts from F22 Enter amount (whole \$) \$

Recorded elsewhere.....9999998

Don't know.....9999999

F26a In the last financial year, did you have any unincorporated businesses?

Yes.....1

No..... 2 → F28a

Note: respondents cannot answer NO to both F26a and F23. If they do, query.

F26b What was your total share of profit or loss from your unincorporated businesses or farms before income tax but after deducting business expenses in the last financial year?

Enter amount (whole \$) → F27

Don't know 9999999 → F28a

Questionnaire - SOEP

Q76. We have already asked for your current income. In addition, please state what sources of income you received in the past calendar year 2001, independent of whether the income was received all year or only in certain months. Look over the **list of income sources and check all that apply. For all sources that apply please indicate **how many months you received this income in 2001** and how much this was on **average per month**. (Please state the gross amount which means not including deductions for taxes or social security).**

Source of income	Received in 2001	Months in 2001	Gross amount per month EURO
Wages or salary as employee (including wages for training, "Vorruhestand", wages for sick time ("Lohnfortzahlung"))	<input type="checkbox"/>	<input type="text"/> <input type="text"/>	<input type="text"/>
Income from self-employment, free-lance work	<input type="checkbox"/>	<input type="text"/> <input type="text"/>	<input type="text"/>
Additional employment	<input type="checkbox"/>	<input type="text"/> <input type="text"/>	<input type="text"/>

Did you receive any of the following additional payments from your employer last year (2001)? If yes, please state the gross amount.

- 13th month salary in total EURO
- 14th month salary in total EURO
- Additional Christmas bonus in total EURO
- Vacation pay in total EURO
- Profit-sharing, premiums, bonuses in total EURO
- Other in total EURO
- No, I received none of these

Production of *aggregated* annual labor income

- ⇒ Questionnaire: Single but complex question (BHPS) vs. detailed list of various income components (SOEP)
- ⇒ Monthly income components need to be aggregated to “annual measures”
 - SOEP: “average monthly amount” times number of months with receipt
 - BHPS / HILDA: financial year vs. calendar year
- ⇒ Aggregation across various income components for a given individual (e.g. individual labor income from dependent employment/self-employment/second job, x-mas bonus, holiday bonus, gratifications, etc.)
- ⇒ INR on any one of the input components *contaminates* the measure of “Total Annual Labor Income”

Panel survey methodology and potential relevance with respect to non-response

- Started in different years (SOEP 1984 / 2000; BHPS 1991; HILDA 2001)
- Additional sub-samples & respondents since start (SOEP/BHPS)
 - are long-term participants selective in their response behavior?
- *All* adult HH-members are interviewed (SOEP / BHPS / HILDA)
 - in principle, no proxy interviews by household head (as is the case in PSID)
- Interview mode: typically face-to-face interview, however:
PAPI / CAPI / self-admin. mix (SOEP), PAPI→CAPI (BHPS), PAPI / CAPI (HILDA)
 - does a long-term personal relationship between interviewer & respondent improve response quality ?

Estimating the probability for INR on labor income – Results from random effects probit models

	SOEP	HILDA	BHPS	SOEP-F
Age				
Age squared			+	
Male	++	---	++	
edu1==2	++	--	--	
edu1==3		---	---	
edu1==4		---	---	
Disability status		+++		
Married		---		
# HH members aged 0-14				
Metrop. area		---	---	---
Remote area	++	+	---	++
Tenure		---	---	
Tenure squared	+	+++	+++	
Foreigner	--			---
Public service		---	---	

	SOEP	HILDA	BHPS	SOEP-F
Firm size: small		+++		
Firm size: large				
East Germany	---	n.a.	n.a.	---
Months full-time (last year)	---	---	---	--
Months part-time (last year)	---	n.a.	n.a.	---
Months unempl. (last year)	+++	+++	--	
Left educ. last year			.	.
Self employed	+++	+++	+++	+++
Problems during Interview	+++	--		+++
# Interviews = 2	.	---		
# Interviews = 3+	---	---	---	---
Obs.	120818	35238	72696	22456
N	24178	10722	11134	7063
-2 Log-Likelihood	-36493.31	-5151.03	-24036.69	-8807.08
Pseudo-R-squared	.1254	.1609	.2120	.1261

Note: Time effects controlled, but not reported here. +++/-- 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.

III. Principles of imputation methods in 3 panels

- Imputation in panel studies profits from longitudinal information
 - use data from $[t-n, \dots, t-1, t+1, \dots, t+m]$ for imputation of INR in t (e.g. CHINTEX-Project, *Spiess & Göbel (2003)*)

- BHPS / HILDA / SOEP → all three panels use longitudinal data in their respective imputation process
 - see *Taylor et al. 2005, Starick 2005, Grabka & Frick 2005,*

HILDA and SOEP

- 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

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

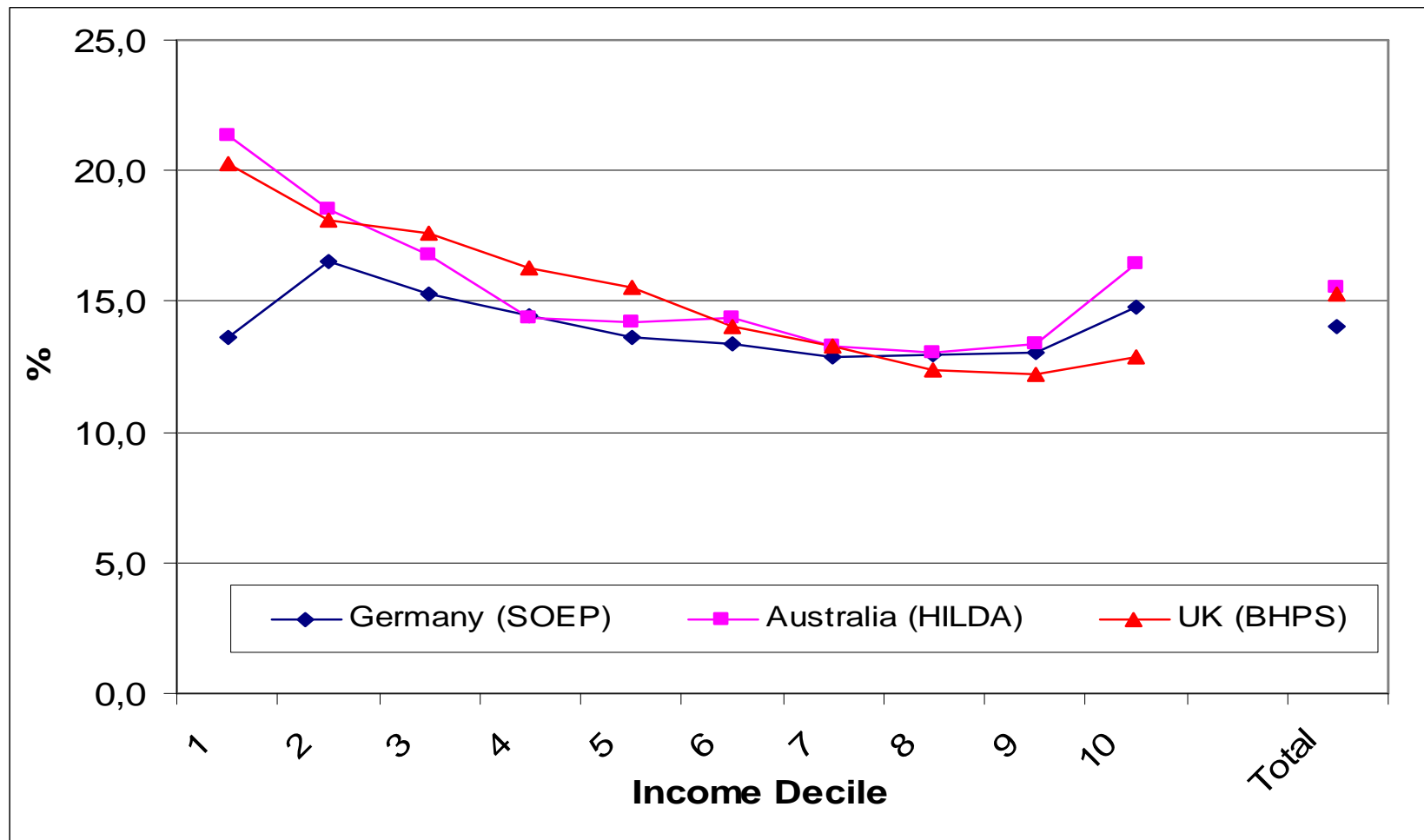
Does the choice of the imputation technique affect substantive research results – and thus cross-national comparability ?

- Robustness check by implementing the “row-and-column” imputation for BHPS
 - ✓ Only for individuals with longitudinal data on labor earnings (ca. 80%)
 - ✓ For all others, we retain the original BHPS imputation (regression based)
- ✓ Comparison of ...
 - ✓ HILDA – “Row-and-column” imputation
 - ✓ SOEP – “Row-and-column” imputation – all samples
 - ✓ SOEP – “Row-and-column” imputation – “fresh” sample F
 - ✓ BHPS – original imputation
 - ✓ BHPS – “Row-and-column” imputation

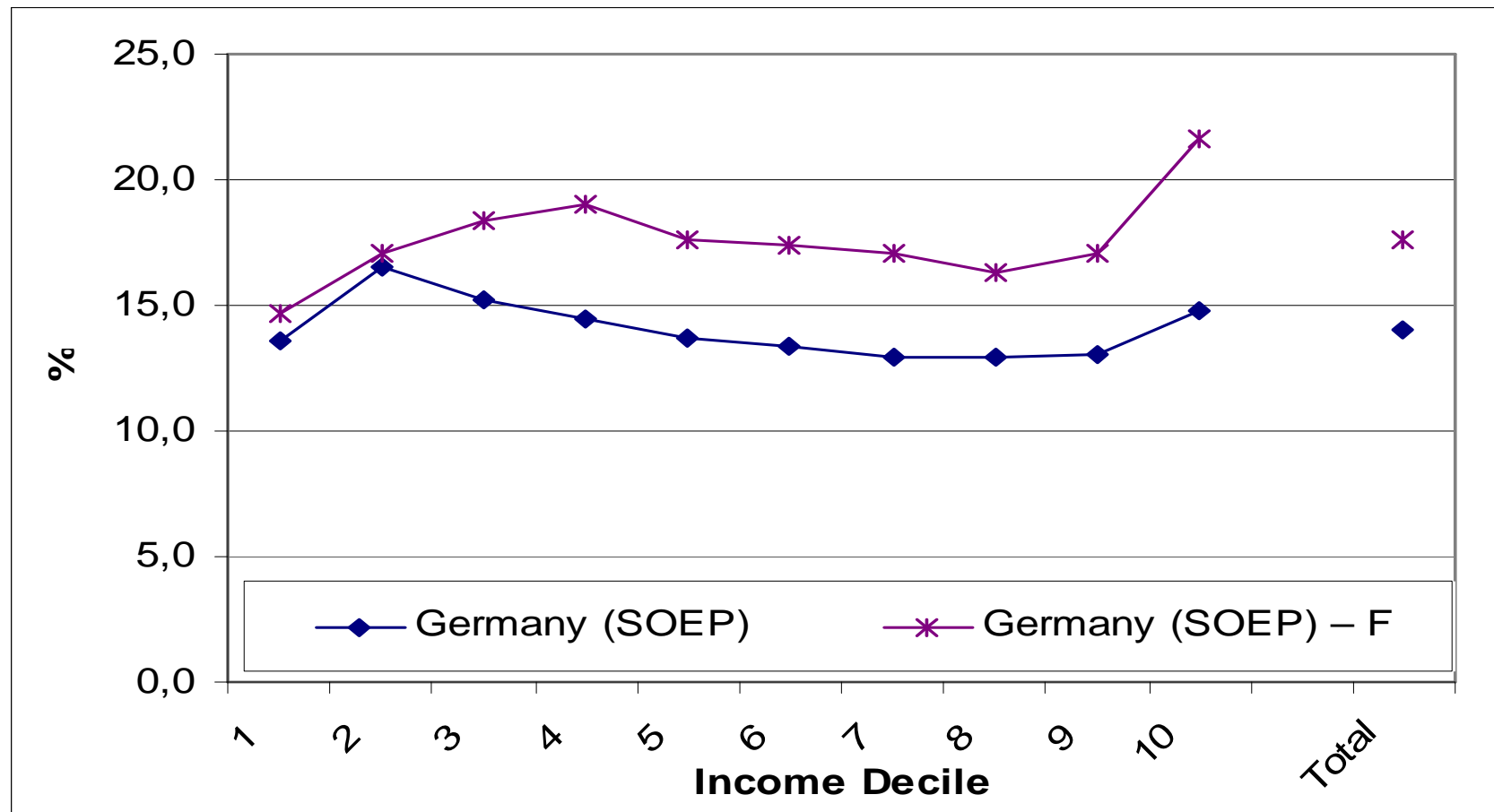
IV. Empirical application - A cross-national perspective

- Incidence of INR
- Imputation and Income Inequality & Mobility
 - How would the picture look like if we used the observed data only ?
 - Comparing results obtained from “All cases” to results based on “observed cases” only.
- Imputation and Wage (Quantile) Regressions

Labor Income: Incidence of INR by income deciles in SOEP, HILDA, and BHPS (1) (conditional on survey imputation)

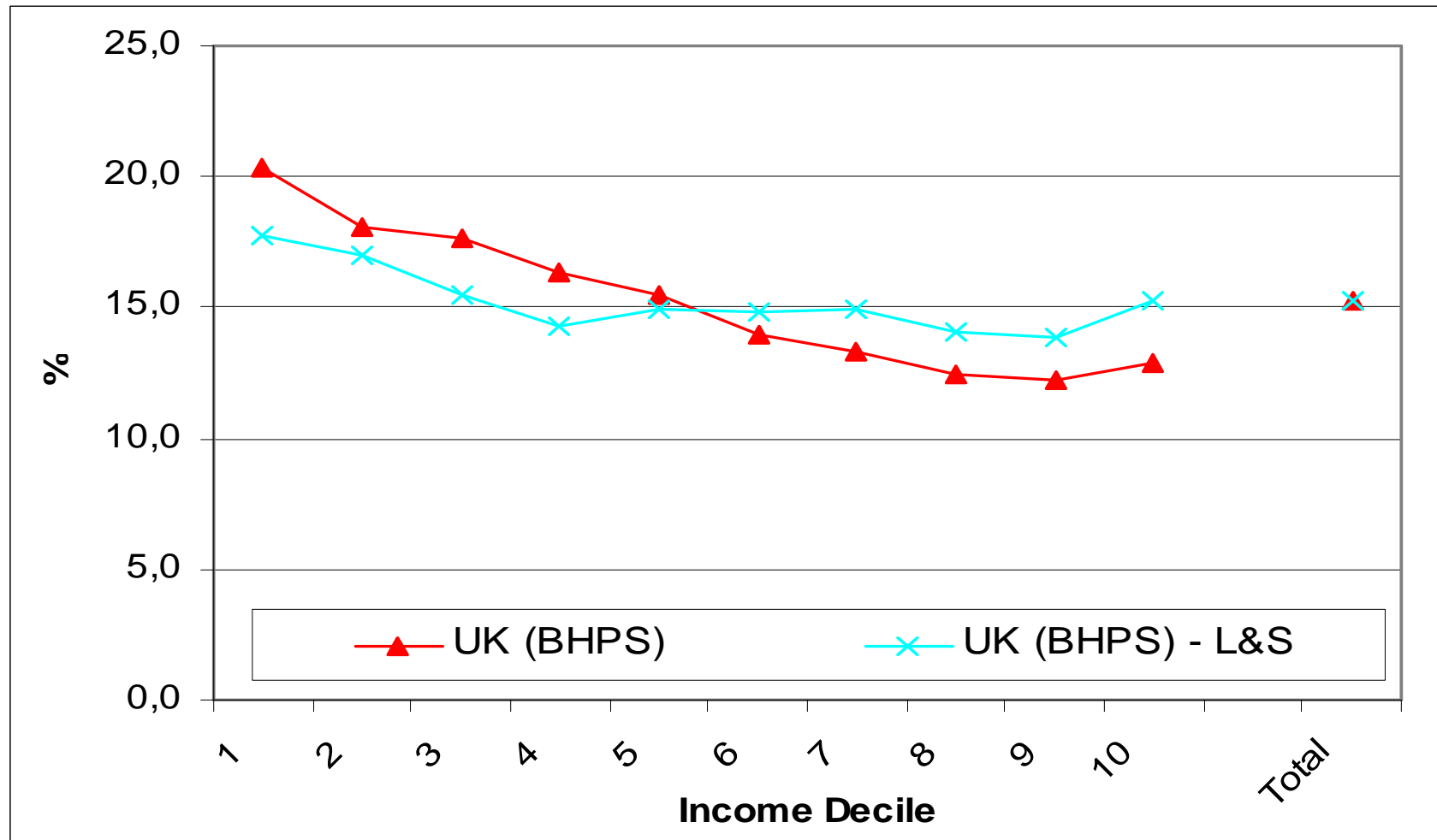


Labor Income: Incidence of INR by income deciles in SOEP vs SOEP-F (→ panel maturity matters)

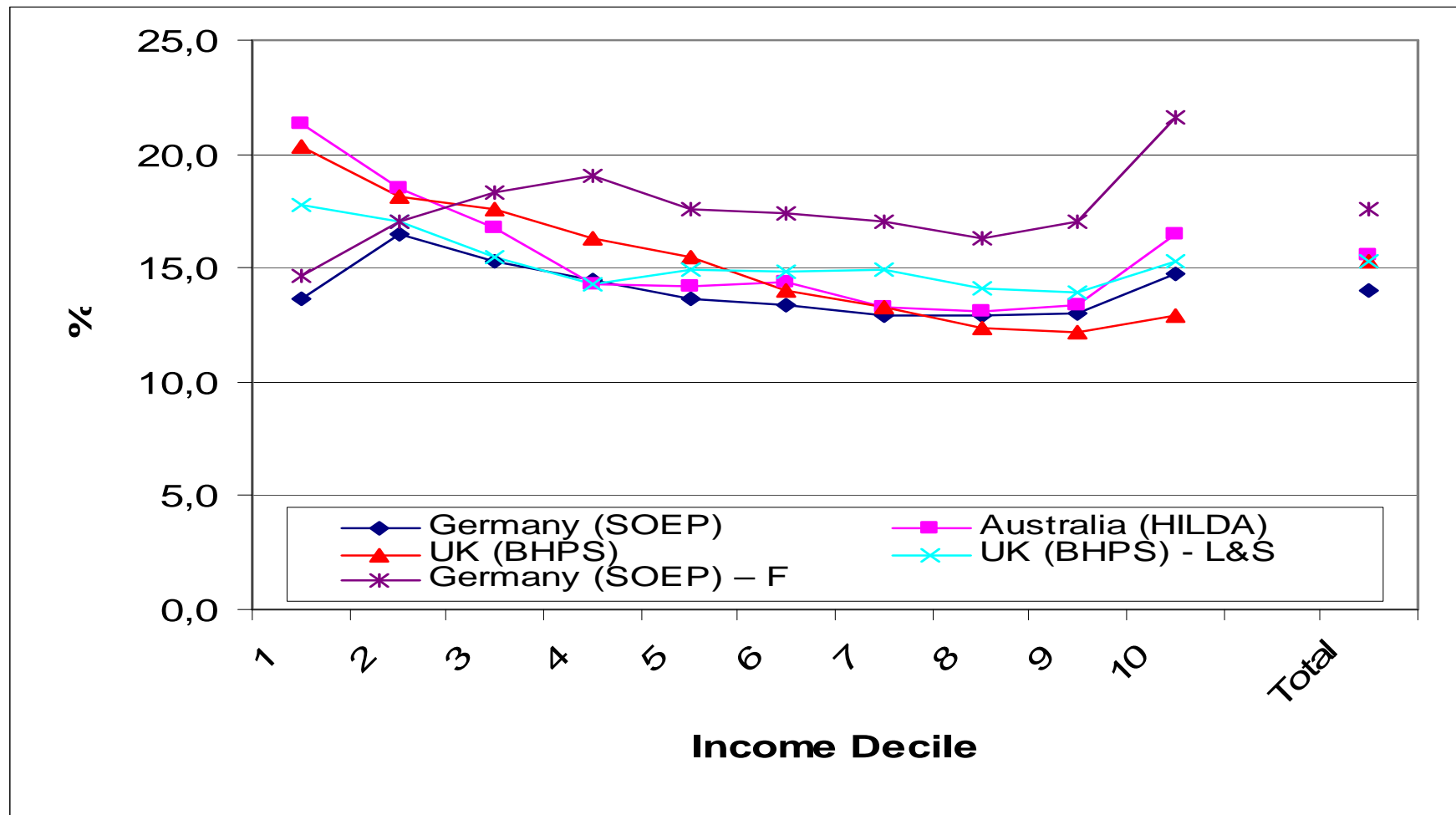


Labor Income: Incidence of INR by income deciles in BHPS vs. BHPS-LS

(→ imputation technique matters)



Labor Income: Incidence of INR by income deciles in SOEP, HILDA, and BHPS (2) (conditional on imputation)



Imputation and labor income inequality

(→ using only observed cases understates inequality)

	Germany (SOEP)		Australia (HILDA)		UK (BHPS)		UK (BHPS)-L+S		Germany SOEP-F	
	"All cases"	Deviation: "All" vs. "Obs."	"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	24695	+1,6
Median	21940	-0,6	23375	-2,3	11360	-2,7	11553	-1,1	21781	0,0
Income inequality										
Theil 0 (MLD)	0,4096	+1,0	0,4587	+4,5	0,4425	+8,6	0,4211	+3,4	0,4467	+0,1
Gini	0,4141	+1,0	0,4273	+1,9	0,4280	+1,7	0,4268	+1,4	0,4336	+0,8
Half-SCV	0,3488	+3,5	0,4456	+1,5	0,4709	+4,9	0,4652	+3,6	0,3858	+5,9
Decile ratio 90:10	13,7	+0,4	14,9	+5,2	12,7	+6,1	12,4	+3,9	15,4	+0,2
Decile ratio 90:50	2,1	+0,9	2,2	+2,3	2,3	+1,8	2,3	+1,1	2,2	+1,4
Decile ratio 50:10	6,5	-0,6	6,8	+2,7	5,4	+4,2	5,4	+3,2	7,0	-1,4
Average N <i>per cross-section</i>	10773	+13,4	9082	+15,3	5002	+18,1	5002	+18,1	6790	+20,4

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

Shaded cells indicate statistically significant deviations (HILDA and SOEP: Random group approach; BHPS: bootstrapping).

Imputation and labor income mobility

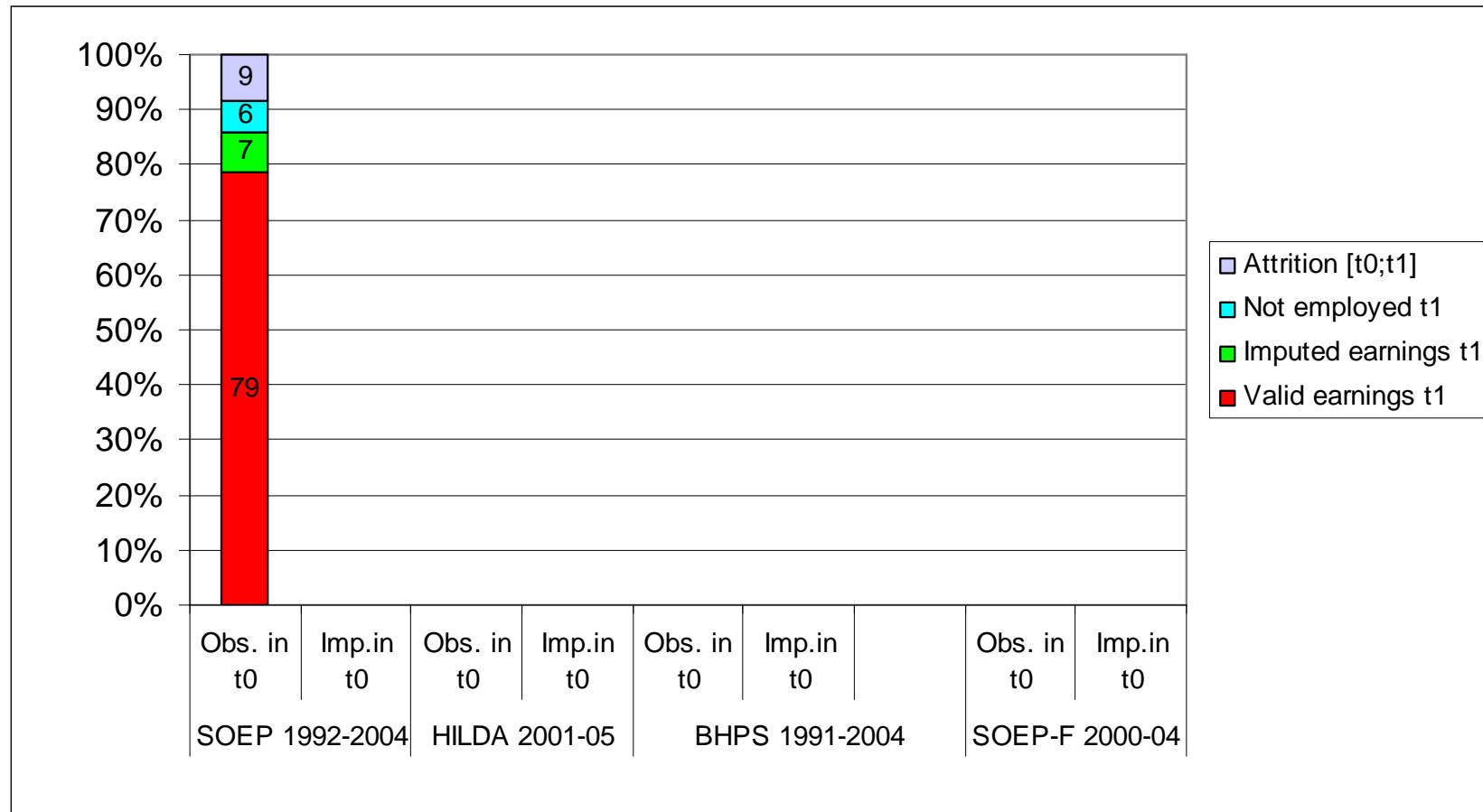
(→ using only observed cases understates mobility)

Income mobility measure	Germany (SOEP)		Australia (HILDA)		UK (BHPS)		UK (BHPS)-L+S		Germany SOEP – F	
	“All cases”	Deviation: “All” vs. “Obs.”	“All cases”	Deviation: “All” vs. “Obs.”	“All cases”	Deviation: “All” vs. “Obs.”	“All cases”	Deviation: “All” vs. “Obs.”	“All cases”	Deviation: “All” vs. “Obs.”
Quintile matrix mobility: normalized avg. jump	0,179	+19,3	0,212	+15,5	0,183	+30,7	0,183	+31,7	0,182	+22,1
Fields & Ok: percentage income mobility	24,38	+29,1	28,81	+17,5	25,42	+47,0	24,74	+43,1	26,78	+30,7
Shorrocks: using Gini coeff.	0,029	+19,8	0,045	+30,8	0,028	+40,3	0,025	+27,8	0,0302	+26,4
Average N <i>per 2-wave balanced panel</i>	9878	+30,8	7474	+19,9	4389	+37,7	4389	+37,7	4928	+42,7

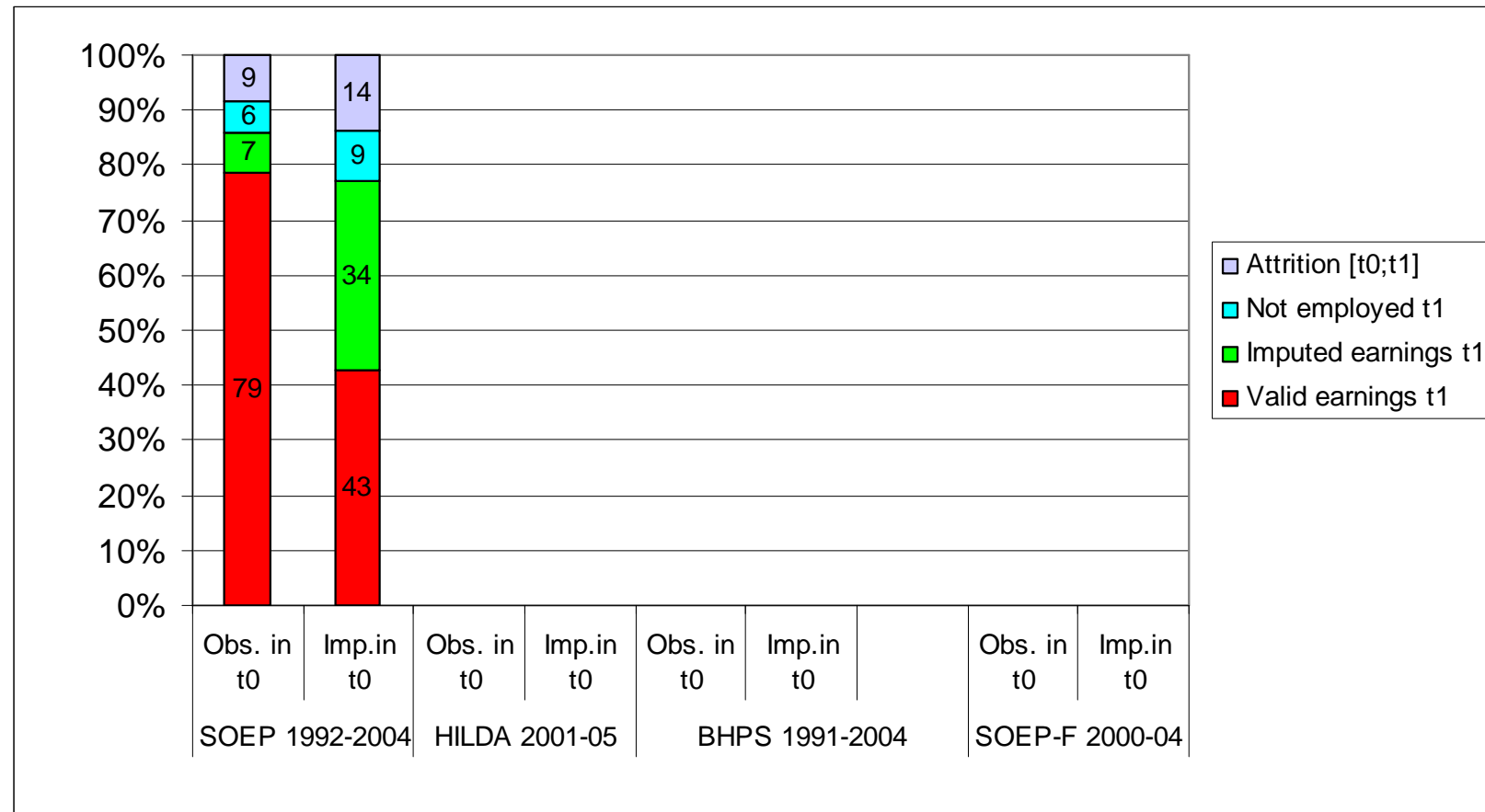
Source: SOEP survey years 1992-2004; HILDA survey years 2001-2005; BHPS survey years 1991-2004.

 Shaded cells indicate statistically significant deviations (HILDA and SOEP: Random group approach; BHPS: bootstrapping).

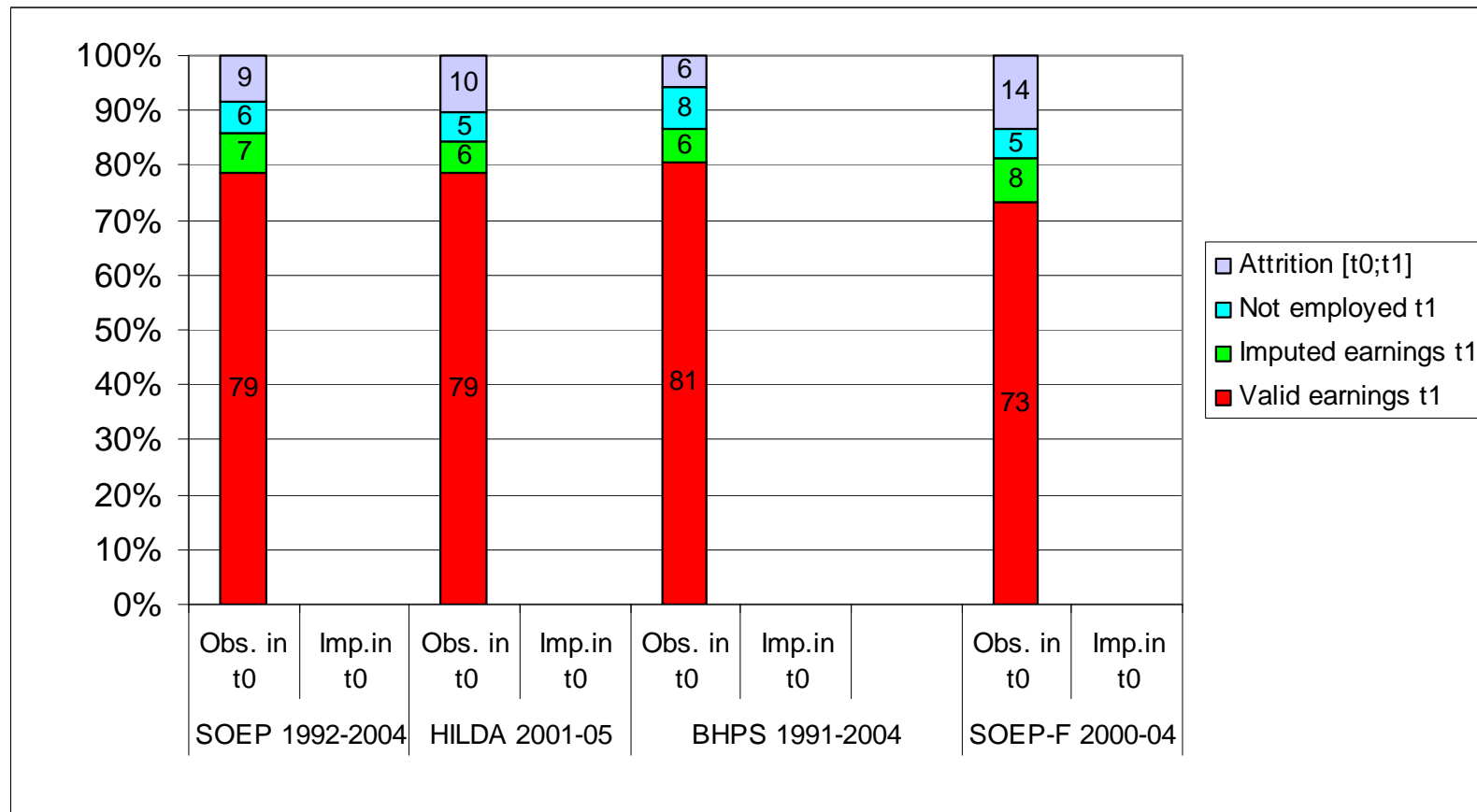
INR in a longitudinal Perspective: The Case of individual labor earnings



INR in a longitudinal Perspective: The Case of individual labor earnings

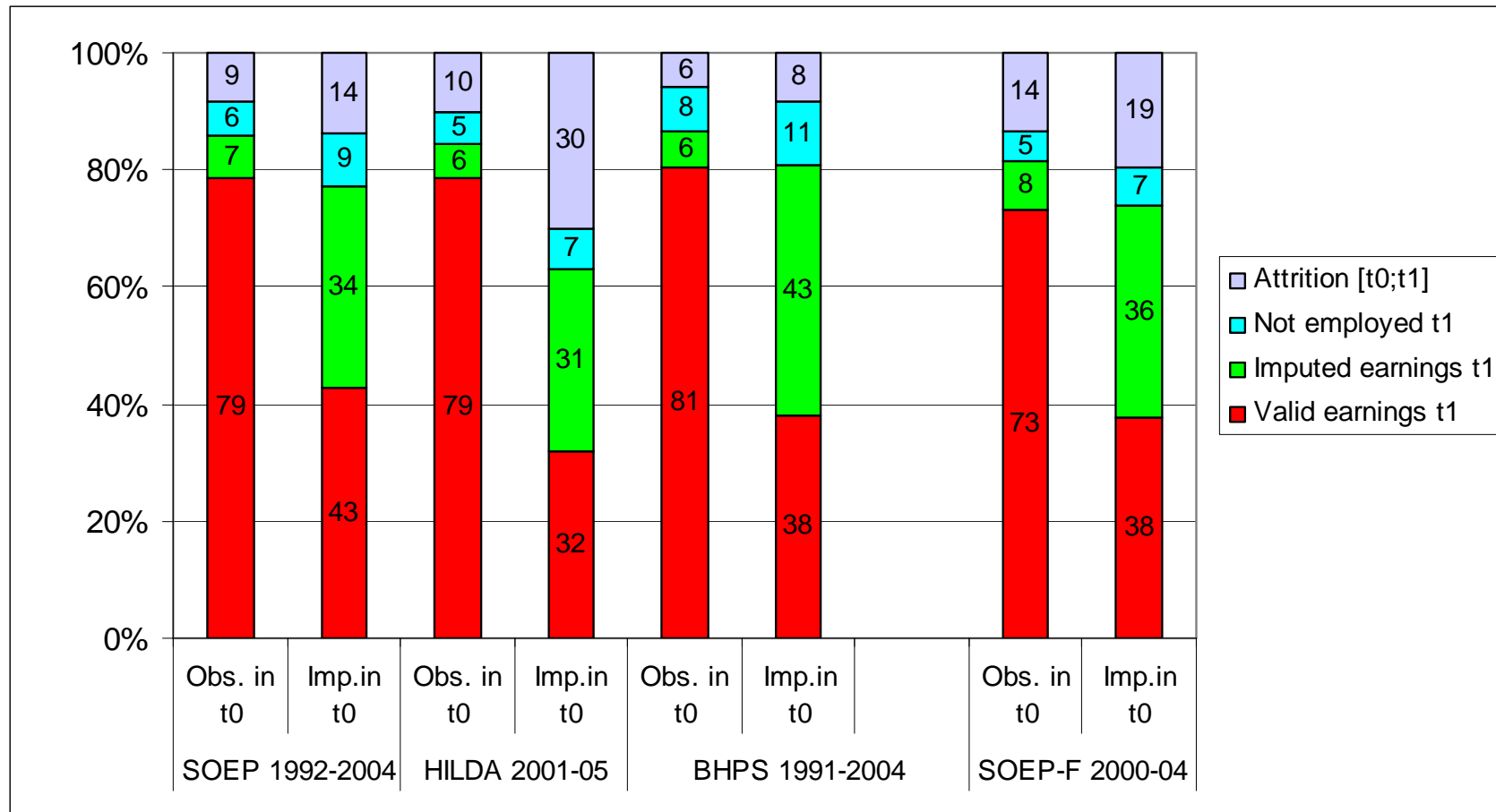


INR in a longitudinal Perspective: The Case of individual labor earnings



INR in a longitudinal Perspective: The Case of individual labor earnings

(→ “cooperation continuum” does exist in all panels)



Results from fixed-effects Wage Regression (1)

	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.
OSS-Seminar, Statistics New Zealand, Wellington, 8 November 2007

Results from fixed-effects Wage Regression (2)

	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^{z*}		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.
OSS-Seminar, Statistics New Zealand, Wellington, 8 November 2007

Results from fixed-effects Wage Regression (3)

	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.
OSS-Seminar, Statistics New Zealand, Wellington, 8 November 2007

Results from fixed-effects Wage Regression (4)

	Germany (SOEP)		Australia (HILDA)		UK (BHPS)		UK (BHPS) – “L&S”		Germany (SOEP) – Sample F	
	observed	all cases	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**		0,042**
Constant	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++
Observations	119030	134337	35661	38681	62049	72729	62049	72904	20355	24392
N (Persons)	24183	25487	11097	11522	10352	11137	10352	11138	6797	7448
R-squared	0,49	0,45	0,22	0,17	0,52	0,44	0,52	0,37	0,38	0,34

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.
OSS-Seminar, Statistics New Zealand, Wellington, 8 November 2007

Results from quantile wage regressions (1)

	Germany (SOEP)			Australia (HILDA)			UK (BHPS)			UK (BHPS - "L+S")			Germany (SOEP-F)		
	p25	p50	p75	p25	p50	p75	p25	p50	p75	p25	p50	p75	p25	p50	p75
Age	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++
Age squared	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Female with kid(s)	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Male with kid(s)	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++
Married			---	---	---	---	---	---	---	---	---	---			
Disability Status		+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++		++	++
Metrop. area	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	++	+++	+++
Remote area	---	---	---	---	---	---	---	---	---	---	---	---	---	---	-
Educational level	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++
East Germany	---	---	---	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	---	---	---
Self employed	---	--	+++	---	---		---	---		---	---		---		+++
Retired	---	---	---		+		---	---		---	---		---	---	---
Left education	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Months Unempl.	---	---	---	---	--	---	+++	++		+++	++	---	---	---	---
Months FT	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++
Months PT	+++	+++	+++	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	+++	+++	+++
Imputed Labor Y	-0,027**	0,004	0,046**	-0,195**	-0,017	0,078**	-0,119**	-0,076**	-0,036**	-0,139**	-0,004	0,066**	-0,016	0,013+	0,052*
Constant	1,454**	2,155**	2,665**	-0,960**	0,446**	1,652**	0,532**	1,141**	1,720**	0,537**	1,251**	1,935**	1,239**	1,996**	2,447**
Observations	139351			38681			72729			72904			25634		
R-squared	0,477	0,395	0,349	0,264	0,208	0,168	0,331	0,279	0,243	.307	.256	.222	0,470	0,396	0,345

Population: working age: 20-60 (Germany), 20-65 (Australia and UK).

Note: Time effects controlled, but not reported. +++/--- 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.

Results from quantile wage regressions (2): 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.

Results from quantile wage regressions (2): 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.

V. Conclusion (1)

- Item-non-response is selective (no MCAR mechanism in all 3 countries)
 - ✓ recommendations based on empirical analyses ignoring cases with missing income data are biased (e.g. Incidence of low wage and wage mobility are understated)
 - ✓ quantile wage regressions:
 - relevance of INR (and imputation) varies significantly across income distribution
 - imputation technique matters
- Imputation provides an effective means to cope with selective INR
 - ✓ maintaining variance is most important
 - ✓ strong suggestion: make use of longitudinal data, if at all possible

V. Conclusion (2)

- Panel Research: INR is positively correlated with any type of non-response in subsequent waves (→ support of “cooperation continuum” !)
 - ✓ similar results across panels
 - ✓ BUT: imputation technique matters !!!! (need to flag imputes!)

- Strong Suggestion: cross-national harmonization of post-data collection treatment of INR (by means of imputation routines)

comments welcome ...

Joachim R. Frick

jfrick@diw.de

Appendix

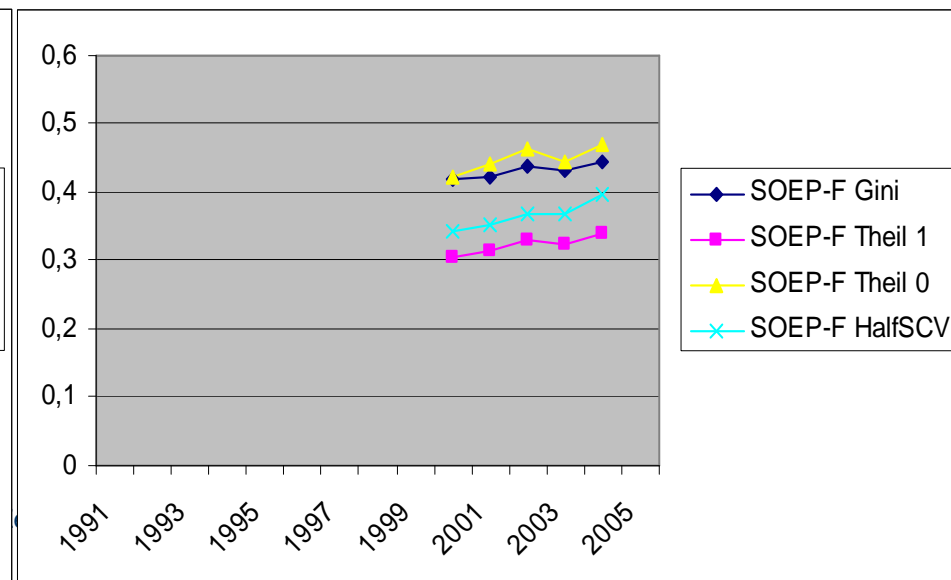
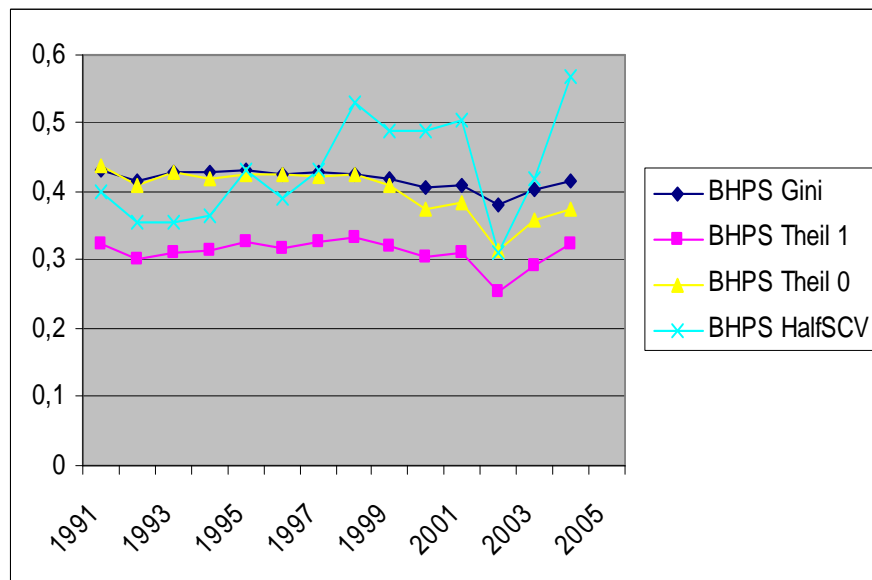
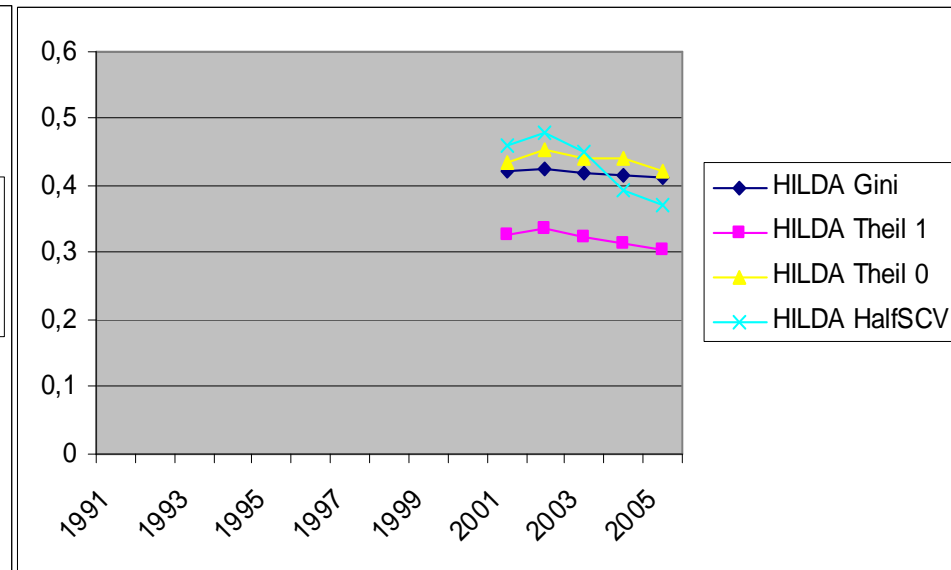
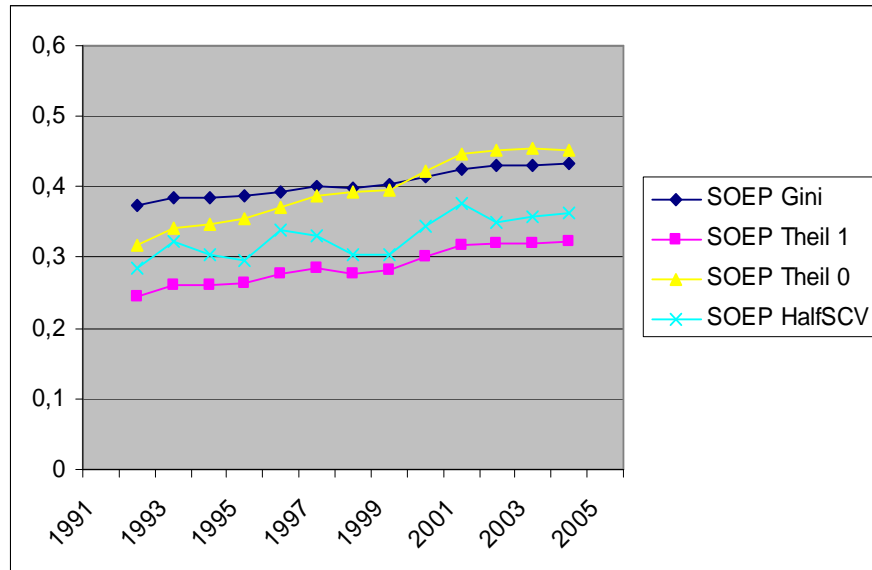
Imputation techniques applied in SOEP (1)

- “Row-and-column” imputation (*Little & Su, 1989*)
 - ✓ using cross-sectional and longitudinal information (also considering a stochastic component due to nearest neighbor matching)
 - ✓ imputed value $Z_{ij} = [r_i] * [c_j] * [Y_{ij} / (r_i * c_j)]$
 - ✓ $c_j = 23 * Y_j / \sum Y_k$ column effect (for each x-section or wave of data, here 23)
 - ✓ $r_i = m_i^{-1} \sum (Y_{ij} / c_j)$ row effect (individual’s longitudinal information)
 - ✓ Sorting cases by r_i and matching the incomplete case i with information from the nearest complete case, say l , yields the imputed value Z_{ij}
 - where $j = 1, \dots, 23$ waves
 - Y_j is the sample mean income for year j
 - Y_{ij} is the income for individual i in year j and
 - m_i is the number of recorded years
- ⇒ imputation fails if no longitudinal data is available

Imputation techniques applied in SOEP (2)

- “Purely cross-sectional” imputation depending on the complexity of the income construct and the number of missing observations
 - ✓ Median-based (e.g. military service pay)
 - ✓ Median-Share based (e.g. Christmas bonus)
 - ✓ Regression-based incl. error term (e.g. wages)
- x-sectional imputation only, if there is *no* longitudinal data

Inequality of individual labor earnings (valid observations, only)

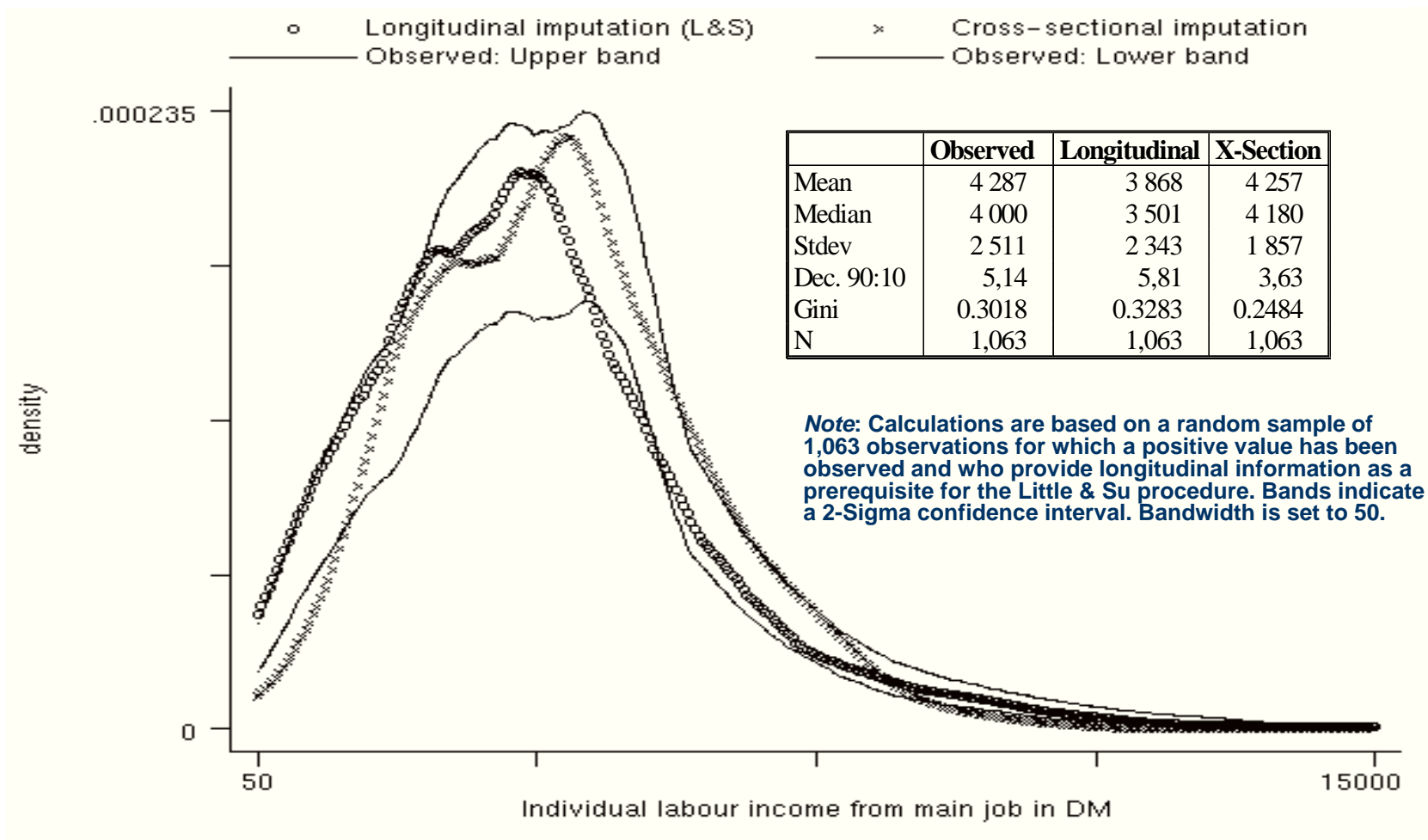


The case of SOEP: Imputation of aggregated “individual labor income including extra payments”¹

	1986 Sample A-B	1993 Sample A-C	2001 Sample A-F	2001	
				Sample A-E	Sample F
Observed cases	92.5	91.0	88.4	90.7	84.7
Imputed cases	7.5	9.1	11.7	9.3	15.4
thereof (by predominant technique):					
• Longitudinal	6.2	8.1	6.6	7.4	5.2
• X-Sectional	1.3	1.0	5.1	1.9	10.1

¹: excluding income from second job, self-employment income and military service pay.
Source: SOEP, Survey years 1986, 1993, 2001; unweighted results.

The case of SOEP: Kernel Density Estimates for "Individual Labor income from main job": alternative imputations vs. observed data



Conclusion from a (SOEP) user's point of view

- ✓ current SOEP release up to wave V, 2005 (in €)
 - \$PEQUIV: annual individual labor income aggregates (I11110\$\$) with corresponding imputation flag (I11210\$\$)
 - \$PEQUIV: income components (e.g. IJOB1\$\$) with corresponding imputation flags (e.g. FJOB1\$\$)
 - \$PGEN: current monthly gross and net labor income with corresponding imputation flags (LABGRO\$\$, LABNET\$\$; IMPGRO\$\$, IMPNET\$\$)

- ✓ documentation / publication
 - *Grabka* (2006): \$PEQUIV-Documentation, DIW-Data Doc. #12
 - *Frick & Grabka* (2005): Item-non-response on income questions in panel surveys, *Allgemeines Statistisches Archiv*, 89: 49-60