

Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Personal Distribution of Income

Joachim R. Frick & Markus M. Grabka

ABSTRACT: This paper deals with missing income data in panel surveys due to item-non-response and with longitudinal and cross-sectional imputation as a strategy to cope with this phenomenon. Using data from the German SOEP, we compare various measures of income inequality and mobility based only on truly observed information to those derived from all (i.e., observed and imputed) observations. We find an increase in inequality due to imputation which appears to be more relevant in the lower part of the income distribution. More important there is a robust picture about income mobility being significantly understated using truly observed information only, a finding which is conveyed by the cumulation of item-non-response over time. Finally, longitudinal analyses provide evidence for a positive inter-temporal correlation between item-non response and any kind of subsequent (item- and unit-) non-response.

The German Socio-Economic Panel Study



1. Motivation

- Item-non response is a major problem in population surveys, especially for income questions
 - reason: "don't know", refusal due to sensitivity
 - typically income from self-employment or capital income most affected (in case of SOEP ~12-25%)
 - child benefit (~1-3%) and old age pensions (~3-8%) have low shares of item-non-response
- SOEP specificities to be considered with respect to the probability of non-response:
 - varying panel duration for sub-samples
 - individual interview with all household members
 - interview situation
- Research questions:
 - Selectivity of item-non-response and imputation
 - Impact of imputation on income inequality and income mobility

Literature

Missing mechanisms:

- Rubin (1976): MAR, MCAR, MNAR

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

- Biewen (2001): higher refusals in tails of income distribution
- Schräpler (2003): complexity of surveyed construct matters
- Schräpler & Wagner (2001): interviewer-respondent matching
- Hill & Willis (2001): formulation of questions matters
- Rendtel (1995), Riphahn & Serfling (2003): interviewer change
- Loosfeldt et.al. (1999): item-non-response is a predictor for subsequent unit-non-response

Table 1: Estimating the Probability of Item-Non-Response in "Total Income", 1993-2002 (Coefficients from a Random Effects Probit Model)

Variable	Coeff.	Std.Err.	Mean	Variable	Coeff.	Std.Err.	Mean
# Interviews = 1	.4241**	(.0266)	.0905	Head: 1-11 months FT	.2063**	(.0239)	.0915
# Interviews = 2-4	.2521**	(.0190)	.1933	Head: 1-11 months UE	.2387**	(.0273)	.0729
Int. Mode : Self-compl	-.2656**	(.0169)	-.3289	Head: 12 months UE	-.1888**	(.0358)	-.0335
Int. Mode : CAPI	-.3421**	(.0223)	-.3202	Head: 1-11 months PT	.3178**	(.0305)	.0373
Age of Head < 25 yrs	.0672**	(.0323)	.0511	Head: 12 months PT	.1198**	(.0309)	.0521
Age of Head > 64 yrs	-.2897**	(.0313)	-.1960	Head: 1-11 months Pen.	.3244**	(.0434)	.0168
# Adults in HH = 2	.1554**	(.0205)	.5492	Head: 12 months Pen.	-.0575*	(.0276)	-.2680
# Adults in HH >= 3	.4114**	(.0256)	.1855	East Germany	-.1767**	(.0224)	-.2573
Kids in HH (<17 yrs)	.1611**	(.0181)	.3234	Comm. Size < 2.000	-.0341	(.0292)	.0944
Head: Female	-.1054**	(.0201)	-.3771	Bottom Decile	.1525**	(.0253)	.0778
Head: Foreigner	.0414	(.0300)	.1134	Top Decile	.1718**	(.0254)	.0907
Head: Univ / FHS	-.0411+	(.0226)	-.1801	Owner occupier	.1198**	(.0180)	.3957
Head: No voc. training	-.1339**	(.0225)	-.1825	Head: Low life satisf.	-.1174**	(.0284)	.0512
Constant	-.1598**	(.0343)	-	Overall Mean (DepVar)	-	-	.192

Note: Only one observation per HH (n=85103 / 17021 individuals). Year of observation controlled, but not shown in table. **p<.001; *p<.05; +p<.1 Pseudo R² = 0.109. -2 Log Likelihood: -41504.073; -2 Log Likelihood (Full Model): -36980.113; LR chi2(35) = 2465.65; Reference Categories: More than 4 Interviews, Mode: Interviewer present, Age of Head 26-64 years, single adult, Head has completed vocational training, Head was 12 months in fulltime employment in previous year. Source: SOEP 1993-2002 (pooled data); own calculations

2. Item-non-response & Imputation techniques applied to SOEP data

Imputation of annual income figures in SOEP is a two-step process:

- "Row-and-column" imputation (Little & Su 1989) using cross-sectional and longitudinal information as well as considering a stochastic component due to nearest neighbor matching
 - Imputed value $Z_{ij} = [r_i] * [c_j] * [Y_{ij} / (r_i * c_j)]$
 - $c_j = 19 * Y_j / S Y_k$ column effect (for each cross-section or wave of data, here 19)
 - $r_i = m_i^{-1} S(Y_{ij} / c_j)$ 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 $j = 1, \dots, 19$ 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
 - "Row-and-column" imputation fails, if no longitudinal information is available
- "Purely Cross-sectional imputation": choice of imputation model depends on the complexity of the income construct and the number of missing observations
 - Institutional imputation (e.g. Child Benefits)
 - Median-based (e.g. military service pay)
 - Median-Share based (e.g. Christmas bonus)
 - Regression-based (e.g. wage, interest & dividends)

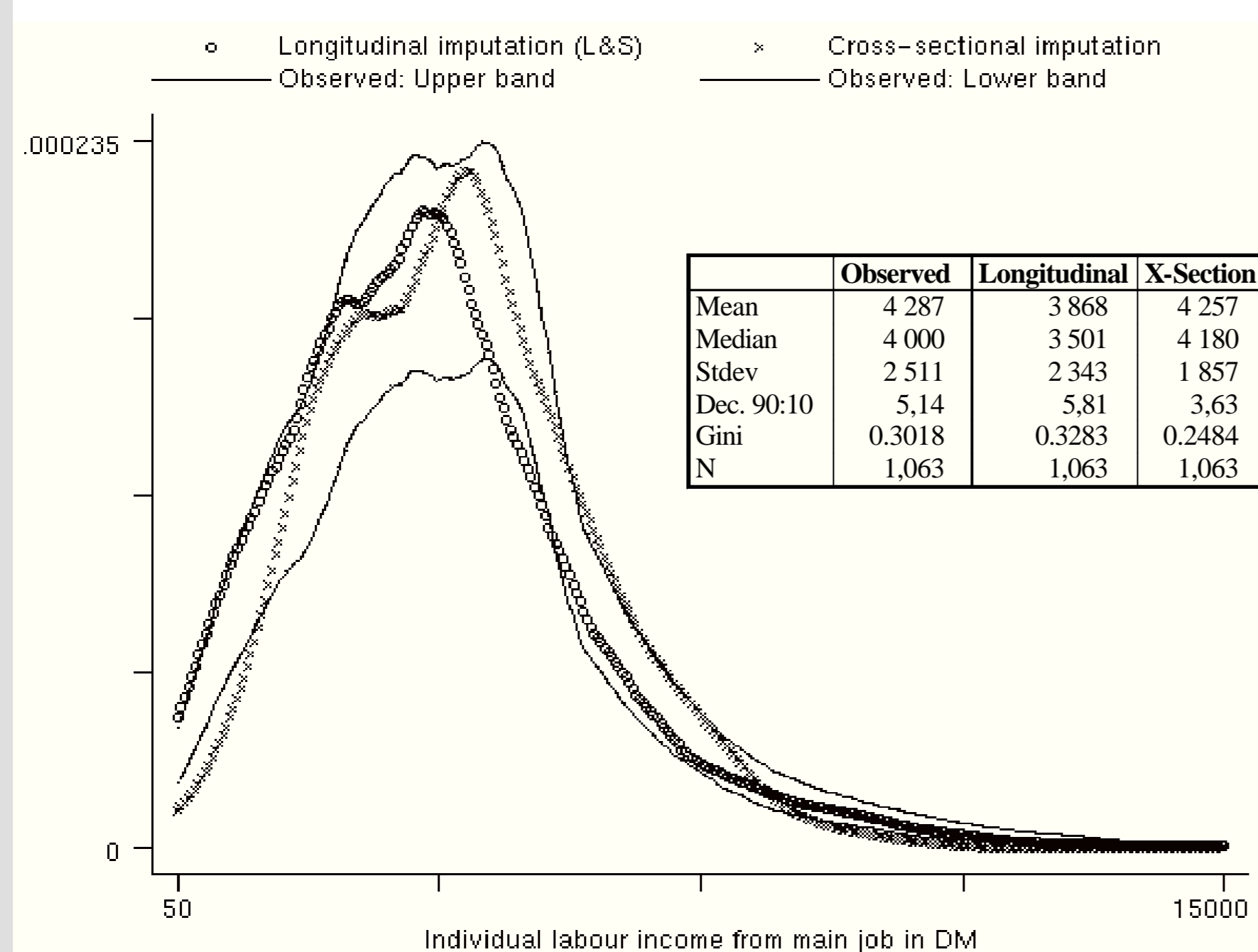
Table 2: Incidence of item-non-response for selected SOEP income components and applied imputation technique

		1986	1993	2001
		Sample A-B	Sample A-C	Sample A-F
- in % -				
Labor income from first job	Observed cases	95.0	94.2	90.8
	Imputed cases	5.0	5.8	9.2
	• Longitudinal	4.1	5.1	5.4
	• X-Sectional	0.9	0.7	3.8
Income from self-empl.	Observed cases	82.1	85.6	74.3
	Imputed cases	17.9	14.4	25.7
	• Longitudinal	11.5	9.5	12.2
	• X-Sectional	6.4	4.9	13.5
Child Benefit	Observed cases	99.2	97.9	95.6
	Imputed cases	0.8	2.1	4.4
	• Longitudinal	0.7	1.9	2.7
	• X-Sectional	0.1	0.2	1.7

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

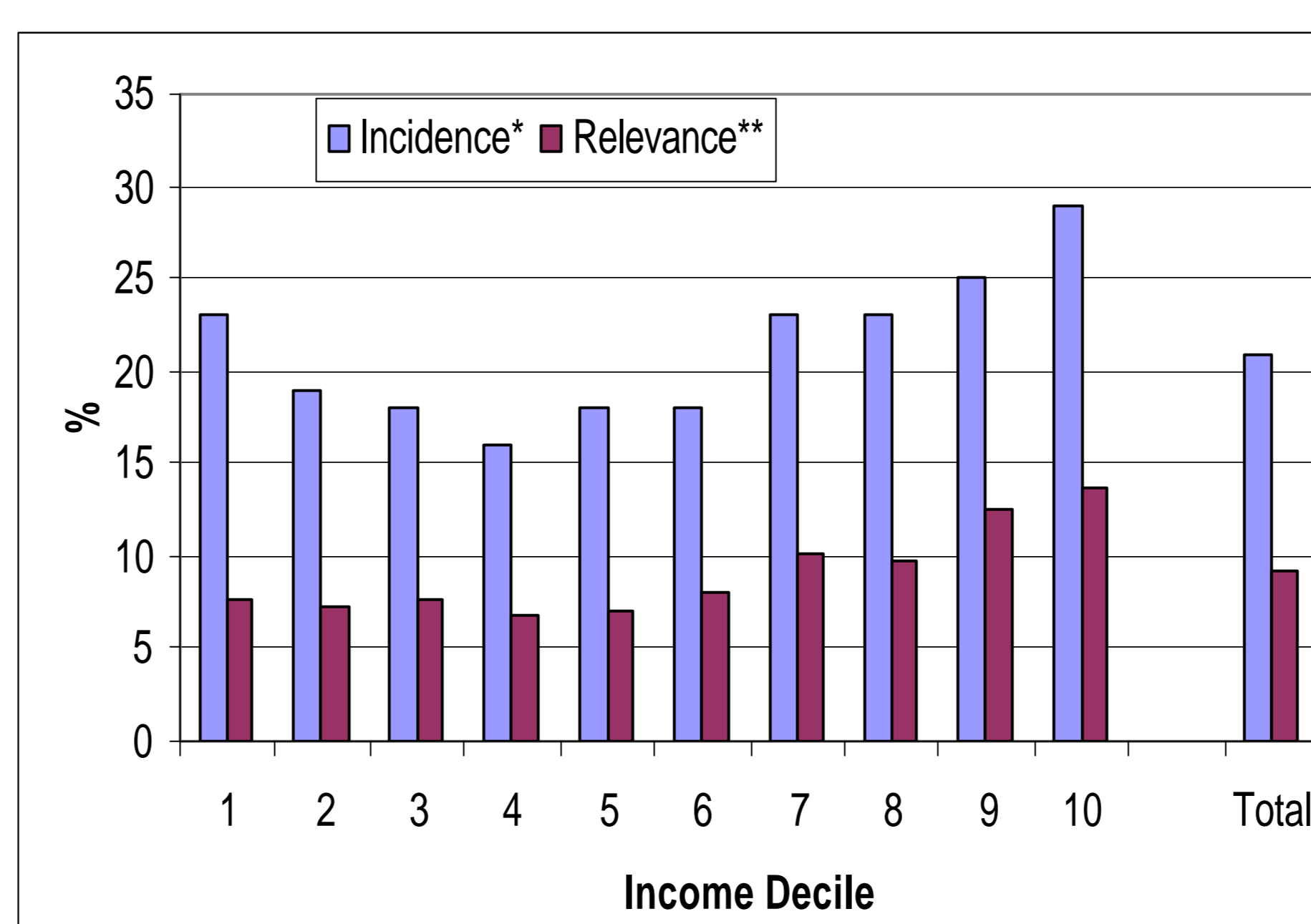
3. Imputation and income inequality

Figure 1: Kernel Density Estimates for "Individual Labor income from main job": alternative imputations vs. observed data



Note: Calculations are based on a random sample of 1063 observations for which a positive value has been observed and who provide longitudinal information as a prerequisite for the Little & Su procedure. Bands indicate a 2-Sigma confidence interval. Bandwidth is set to 50.

Figure 2: Incidence and Relevance of Item-Non-Response in equivalent Post Gov't Income by Income Deciles 2001



* Incidence = Population Share with at least one imputed income component included in "Post Government Income".
** Relevance = Imputed Income as a share of "Post-Government Income".

Table 3: Equivalent Post-Gov't Income, Inequality and Poverty 2001

	Imputation Status			Deviation: "All" vs. "Observed" (%)
	"All cases"	"Observed cases"	"Imputed cases"	
Mean	36 760	36 152	39 015	+1,7
Median	33 334	32 797	35 720	+1,6
MLD (bottom-sensitive)	0.1350	0.1283	0.1577	+5,2
Gini	0.2698	0.2641	0.2855	+2,2
SCV (top-sensitive)	0.2977	0.2961	0.2958	+0,5
90:10	3.44	3.32	3.89	+3,6
90:50	1.77	1.75	1.79	+1,1
50:10	1.94	1.89	2.18	+2,6
Poverty Rate	14.3	14.1	15.2	+1,4
N (x-section 2001)	27 915	22 399	5 516	+24,6

Values in brackets give a 93% confidence interval (based on Random Group Approach). Colored areas indicate statistically significant differences.

4. Imputation and income mobility

Figure 3a: Income Mobility and Imputation: The case of equivalent Post-Gov't Income, 2000-2001

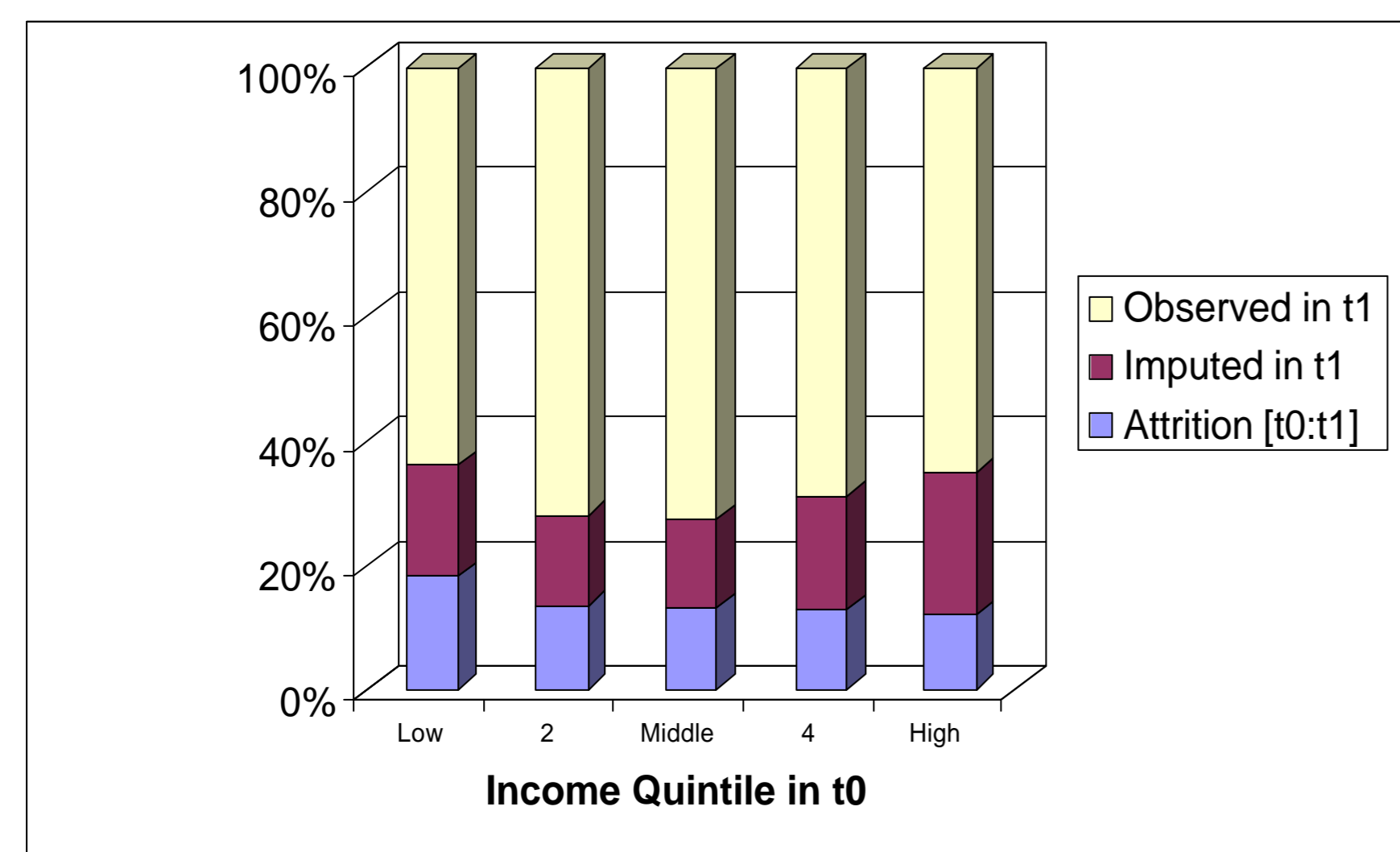


Figure 3b: Income Mobility and Imputation: The case of equivalent Post-Gov't Income by Imputation status, 2000-2001

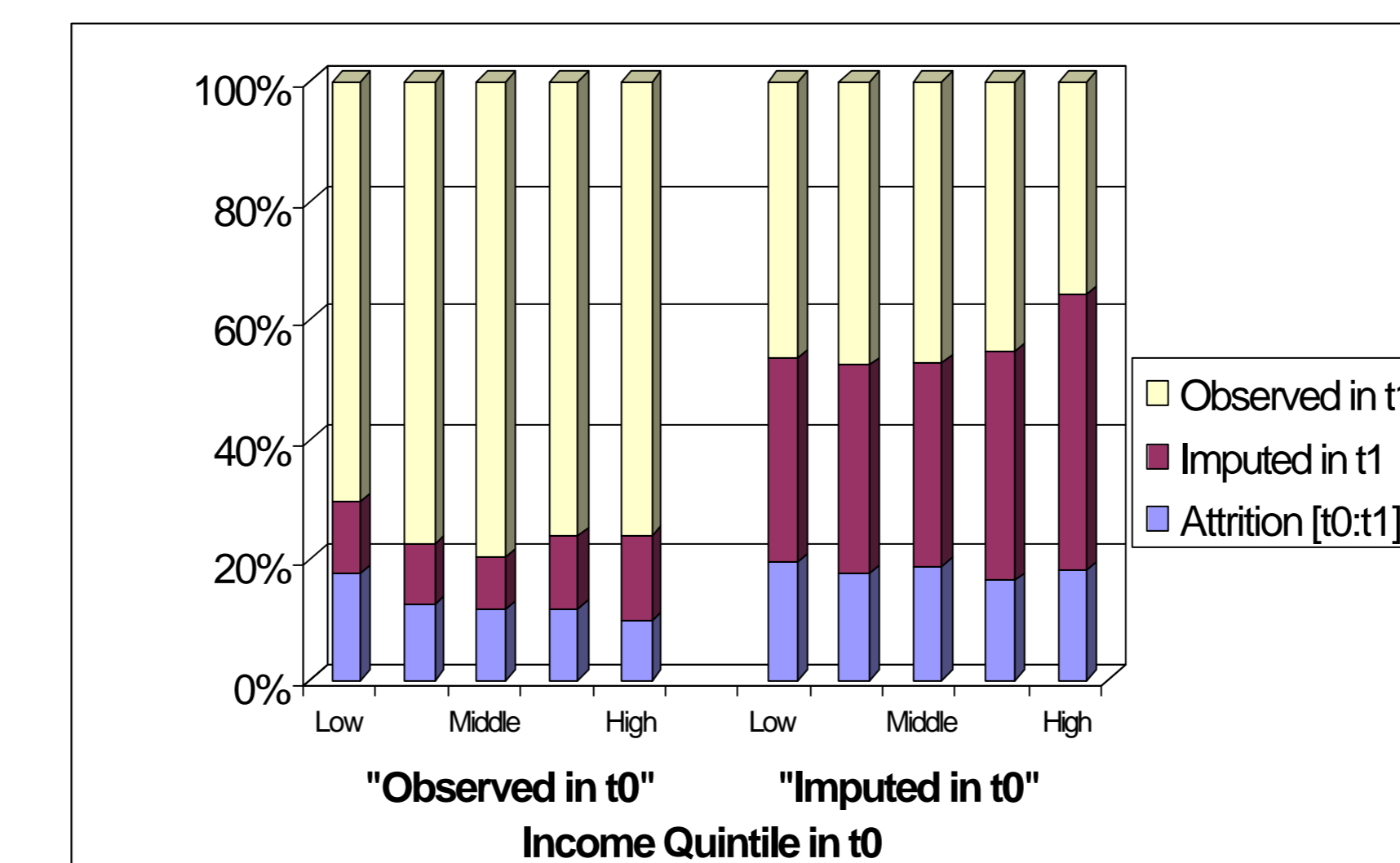


Table 4: Income Mobility indices and Imputation, 2000-2001

	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	0.437	0.618	+13.3
Normalized average jump	0.187	0.165	0.234	+14.8
Fields & Ok (1996)	18.36	15.99	22.88	+13.5
Percentage income mobility	17.23	19.43	23.78	+10.8
Fields & Ok (1999)	0.210	0.185	0.259	+13.5
Non-directional	0.200	0.203	0.271	+10.8
Shorrocks (1987)	0.0829	0.0748	0.0955	+10.8
using Theil Coefficient	0.0709	0.0851	0.1040	+10.8
N (balanced panel)	26 609	18 201	8 408	+46.2

Values in brackets give a 93% confidence interval (based on Random Group Approach). Colored areas indicate statistically significant differences.

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

- Item-non-response is highly selective
- Imputation provides effective means to cope with this problem
 - Empirical results suggest to make use of longitudinal data, if at all possible
- Item-Non-response and Imputation are positively linked
 - to income inequality and
 - to income mobility
- Panel Research: Item-Non-Response is positively correlated to (item and unit-) non-response in subsequent waves!
- Data providers must document imputation & flag imputes
 - documentation
 - CNEF: www.human.cornell.edu/pam/gsoep/equivfil.cfm
 - Grabka & Frick (2003) DIW-Research Note #29
 - SOEP release up to wave T, 2003 (\$PEQUIV-files)
 - income aggregates given as I111xx\$\$
 - imputation flags in variables I112xx\$\$ (imputed income as a % of the income aggregate)
 - imputed versions of all single surveyed income components at annual level (plus imputation flags)
 - all annual income data 1984-2003 expressed in EURO
- Future Research: implementation of multiple imputation in SOEP: Little & Su procedure may be extended by matching to k neighbors (e.g. k=5) instead of only nearest one
 - Full version of this paper.
- Frick, Joachim R. and Grabka, Markus M. (2003): Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income distribution. DIW-Discussions Papers No. 376, October 2003 (update April 2004), Berlin: DIW Berlin.