Oliver Lipps

Income imputation in the Swiss Household Panel 1999-2007

Lausanne, January 2010

FORS Working Papers 2010-1

FORS Working Paper series

The FORS Working Paper series presents findings related to survey research, focusing on methodological aspects of survey research or substantive research. Manuscripts submitted are papers that represent work-in-progress. This series is intended to provide an early and relatively fast means of publication prior to further development of the work. A revised version might be requested from the author directly.

Further information on the FORS Working Paper Series can be found on www.fors.unil.ch.

How to cite this document:
Lipps, O. (2010). Income imputation in the Swiss Household Panel 1999-2007. *FORS Working Paper Series,* paper 2010-1. Lausanne: FORS.

Acknowledgements
This study has been realized using the data collected in the "Living in Switzerland" project, conducted by the Swiss Household Panel (SHP), which is based at the Swiss Centre of Expertise in the Social Sciences FORS, University of Lausanne. The project is financed by the Swiss National Science Foundation. I thank two colleagues for valuable comments on a former version of this paper.

ISSN 1663-523x (online)

FORS
c/o University of Lausanne, Vidy
1015 Lausanne
Switzerland
E-mail: paperseries@fors.unil.ch

Summary

This paper describes the methods used and the steps taken to impute missing income values in the Swiss Household Panel Survey (SHP). Missing values that result from both item- and unit-nonresponse are imputed. We impute income on the individual level, distinguishing between several income components.

The imputed item- and unit nonresponse income distributions are compared with the distributions of the validly reported cases. The level of imputed income from *employment* resulting from *item*-nonresponse is similar to that of the validly reported. Other imputed income components from *item*-nonresponse are generally slightly, imputed income from *unit*-nonresponse considerably higher than that from the validly reported cases. This result shows that imputing missing income may avoid biased level estimates. Income variations of the valid cases and the imputed cases are not different.

Keywords: item-nonresponse, unit-nonresponse, imputation, income

JEL-classification: C81, D31, I32

Income imputation in the Swiss Household Panel 1999-2007

Oliver Lipps[1]

1. Missing income and imputation in the SHP in brief

Item nonresponse occurs in surveys if respondents are not able or not willing to give a valid answer on survey questions. Mostly, difficult or sensitive questions such as income questions are concerned. In some cases, also technical errors might be responsible. Another source of nonresponse are individuals who do not give an interview at all (unit nonresponse). Both forms of nonresponse occur in both cross-sectional and panel surveys; in the latter in all or only some waves.

To impute missing income data in the Swiss Household Panel (SHP), we generally use both cross-sectional and longitudinal methods for all income variables.[2] For individuals with a never validly reported income component, a cross-sectional method must be used first: the income is "initialized" using a stochastic regression. We start by using all available relevant covariates, and impute the missing value in the first wave to be imputed with all covariates validly given. If one of the covariates is missing in all waves, we stepwise drop covariates according to significance. The last initialization step, which involves only a few missing values, is a median imputation. This procedure is repeated using a reversed order of waves, i.e., from the most recent to the first wave ("last-first") to be imputed. In case of a discrepancy between the first-last and the last-first initialized value, the value that is imputed in an earlier regression step is used. If the imputation step is the same, one of the two is randomly selected.

Once the income component is initialized or validly reported in at least one wave by all income eligible individuals, we impute missing income data in all other waves using longitudinal imputation methods. Similar methods to those performed in the German Socio-Economic Panel (SOEP) (Frick and Grabka 2004), or the Household, Income and Labour Dynamics in Australia (HILDA) Panel Survey (Starick and Watson 2007) are used, e.g. by also giving preference to the Little & Su (L&S) imputation technique.

[1] FORS, c/o University of Lausanne, Bâtiment Vidy, CH – 1015 Lausanne
oliver.lipps@fors.unil.ch

[2] Frick and Grabka (2004) give an overview of the commonly used single imputation techniques (p. 6 ff.).

Unlike the SOEP, which does not distinguish class variables to match recipients and possible donors at all, the HILDA uses age classes, and we use education as class variable.

We impute missing income values for all income components of all individuals, who report to having received income from this source without giving the amount or a guesstimate. In addition we impute missing income values asked in a proxy interview. If the likelihood is high that a *unit*-nonresponding individual earns income from a specific source, this is also imputed. Concluding an income from a specific source is possible, because in the SHP socio-demographic information such as sex, age, education, or especially occupational status of all household members is available from the household grid questionnaire. Being listed in the household grid is a necessary condition to be eligible for imputation.

2. Motivation for item and unit missing income imputation in the SHP

An obvious motivation to impute missing income data is that not doing so leads to a loss of prediction power due to listwise deletion of cases. Also, because of its membership in the Cross National Equivalent File (CNEF[3]), the Swiss Household Panel (SHP) must deliver imputed item-nonresponse income variables to the CNEF (Frick et al. 2007, Lipps and Kuhn 2009). The major motivation however is that using only validly reported income may bias both population level and variation estimates and model results.

Although weights are designed to help correct for unit-nonresponse, they do nothing in correcting for the bias due to item-nonresponse. However item-missing values are not missing completely at random. This makes imputation of item-missing income values necessary. In the SOEP, "ignoring cases with item-nonresponse tends to underestimate income levels as well as variance …. Additionally, in line with findings in the literature, item-nonresponse on income appears to be selective with respect to both tails of the income distribution, especially at the upper end" (Frick and Grabka 2004:20). Therefore, imputing missing income cases may give more realistic income level and variation measures. In longitudinal analyses, income mobility seems to be underestimated using only validly reported cases (Frick and Grabka 2007).

With the exception of labor income, all (imputed) income variables in the SHP delivered to the CNEF are aggregated on the household level. This renders the imputation of missing *units* (individuals) important to avoid underestimation of these aggregated sums. (Partial household) unit nonresponse is a problem particularly in centralized

[3] Apart from the SHP, current members of the CNEF are the SOEP, the HILDA, the U.S. Panel Study of Income Dynamics (PSID), the British Household Panel Survey (BHPS), and the Canadian Survey of Labour and Income Dynamics (SLID).

telephone household surveys such as the SHP and mostly affects individuals other than household reference persons (Lipps 2009). Frick et al. (2009) suggest the following measures to deal with unit-nonresponse:

"(a) Ignoring the fact that a household member (and its income information) is missing, thus assuming the non-responding individual's income is zero ...

(b) Adjusting the calculation of the equivalence scale by ignoring the person's contribution to household income as well as to household needs, thus in principle ignoring the person's existence…

(c) Eliminating all households observed with [nonresponding individuals] …, thus assuming that these households are missing completely at random" (p. 2).

(d) Imputing unit-nonresponding income values.

Frick et al. (2009) find that applying (a) to (c) results in "a systematic downward bias in level and development of income inequality and relative poverty whereas income mobility will be overstated" (p. 31). They conclude that imputation of various components instead of only adjusting the income measure may be considered advantageous.

3. Income components imputed

All income components to be imputed are collected in the individual questionnaire in the SHP. Some income concepts and therefore income questions and calculation algorithms have changed since the start of the SHP in 1999 (Kuhn 2008) and are fully comparable only since 2002. Since we impute and construct all income components from 1999 on, using different algorithms, care must be taken when comparing income across waves until 2001, and from a wave before 2002 with one after 2001.

Imputed income variables comprise the following income sources (annual amounted), constructed from the original income variables asked (Kuhn 2008):[4]

1. Income from employment: net ("i$$empyn")[5]
2. Income from independent work: net ("i$$indyn")
3. (old age) pension: annual income ("i$$oasiy")
4. Invalidity pension: ("i$$aiy")
5. Income from pension insurance: ("i$$peny")
6. Income from unemployment fund: ("i$$uney")

[4] In the waves before 2002, we impute other (aggregated) income variables, e.g., total working income (wyn). Due to non comparability across the waves analysed here, they are not listed.
[5] In the SHP notation, "i" is the first letter of income variable names; "$$" denotes the survey year (from 99 on).

7. Income from social assistance: ("i$$wely")
8. Grants, scholarship: ("i$$gray")
9. Income from other institutions, organizations: ("i$$insy")
10. Income from family allowances: ("i$$famy")[6]
11. Income from people in private households (outside the household): ("i$$pnhy")
12. Yearly income from other sources: ("i$$osy")

Codes to be imputed are[7]

-8 (other error)

-7 (filter error)[8]

-5 (irregular, difficult to say)

-1 (does not know)

-2 (does not want to say)

In the SHP 21,732 individuals in 8,529 households were ever listed between 1999 and 2007 (including children), of whom 18,320 ever responded, either by completing an own individual questionnaire, or by means of a proxy. In the following Table 1, we list the number[9] of missing values by nonresponse category (item-, or unit-), the number of individuals that need initialization, the number of possible donors (i.e., those with validly reported nonzero income), and a variation (standard-deviation/mean) characteristic for the income variables. Missing values and donors are summed over the respective waves. Note that wave specific item and unit nonresponse is exclusive. We also calculate pearson correlations of income with age-group and education (compare Table 1 in Starick and Watson (2007)), to find income discriminating variables, available from the household grid. These characteristics help to find the most suitable method for imputation. Note that due to comparability, of the 9 waves from 1999-2007 only descriptive statistics from 2002 until 2007 are listed. Unit-nonrespondents are assumed to earn income according to their occupational status and number of children, both available from the household grid. E.g. if the unit-nonrespondent is actively occupied, s/he is supposed to earn positive wages, which are imputed by distinguishing part or full time employment. Similarly, while for (old age) retired people the components of old age social security pensions (oasiy and peny) are imputed, for "other" retired people invalidity pensions (aiy) are imputed. Unemployed individuals are attributed

[6] Family allowances are asked separately only from 2004 on (Kuhn 2008).
[7] -4 denotes „no income from the respective income source.
[8] does not occur.
[9] Note that we do not list the percentages because absolute numbers of donors and recipients provide better insight about the feasibility of the imputation procedures.

unemployment assistance. Child allowances are allocated (to households) according to the number of children.

The number of recipients is comparatively high for income from social assistance (wely). “Unit-nonrespondents” in this category include those who refuse to give information on any source of income and for whom it is not clear from the available information which income source they touch. To impute income for these individuals, rather than to assign a fixed minimum income, we use donors who earned “income from social assistance”.[10]

[10] We are aware that this procedure is based on very strong assumptions. In addition, this artificially blows up the unit nonrespondents in this income category.

Table 1: Nonresponse characteristics of the income variables in the SHP 2002-2007. Descriptive statistics (shaded right) are averaged over waves.

	Missings		*Of which:*	*Potential Donors*			
Income source:	*N Item NR*	*N Unit NR*	*N to be initialized*	*Number donors*	*Std-dev. / Mean*	*Corr with age*[11]	*Corr with Educat*[12]
employment	*1689*	*7,744*	*2,806*	*23,561*	*.84*	*.24*	*.31*
indep. work	*862*	*0*	*325*	*3,769*	*2.02*	*.11*	*.12*
old age pension	*516*	*1,581*	*627*	*6,209*	*.46*	*.10*	*.02*
invalidity pension	*145*	*234*	*164*	*1,080*	*.66*	*-.07*	*.10*
pension insurance	*485*	*1,537*	*763*	*3,859*	*1.34*	*-.06*	*.14*
unemployment fund	*73*	*212*	*186*	*880*	*.94*	*.29*	*.14*
social assistance	*52*	*678*	*428*	*417*	*1.11*	*-.02*	*-.03*
grants, scholarship	*74*	*0*	*46*	*415*	*1.12*	*.17*	*.12*
other institutions	*113*	*0*	*91*	*1211*	*2.60*	*.08*	*.04*
family allowances[13]	*329*	*1,754*	*845*	*3757*	*.77*	*.07*	*.01*
private transfer (ext.)	*304*	*0*	*171*	*3420*	*2.61*	*.22*	*.16*
other sources	*622*	*0*	*448*	*3,498*	*4.93*	*.01*	*.05*

If an income variable is imputed for both nonresponse components, we find that more missing cases stem from unit nonresponding individuals than from item nonresponding cases.

[11] Age classes are in 10 year groups.
[12] Education measured in three (about equally sized) levels.
[13] Variable available from 2004 on.

4. Imputation methods

In this section, the different *longitudinal* methods and the stochastic *cross-sectional* regression used to impute missing income values are described. The longitudinal imputation methods L&S, its extended variant, and the simple carryover method are described first. If the income component is never validly reported, longitudinal methods fail to provide a positive imputation value (Frick and Grabka 2004). In these cases, cross-sectional imputation methods must be used first. We describe the cross-sectional regression method used to "initialize" the income component. As is usual, we assume that the income missing mechanism responsible is MAR (missing at random). This means, the missing data are at random once controlled for observed variables. All imputation procedures are programmed in STATA©.

4.1. Little & Su method

The L&S imputation technique, also known as the "row and column" imputation procedure (Frick and Grabka 2004), considers longitudinal as well as cross-sectional information in the imputation process. The imputed value is the result of a combination of a row effect, a column effect and a residual effect. The column (year) effects are given by $c_j = \frac{\overline{Y_j}}{\overline{Y}}$, where j = 1, ..., m [number of years], $\overline{Y_j}$ is the sample mean income for year j, and $\overline{Y}$ is the mean of $\overline{Y_j}$ over all j. The column effect c_j can be interpreted as the inflation factor in year j. The row (person) effects, $r_i = \frac{\sum \frac{Y_{ij}}{c_j}}{m_i}$, are computed for each sample member i. Y_{ij} is the income for individual i in year j and m_i is the number of recorded waves. r_i corresponds to i's mean expected income Sorting cases by r_i and matching the incomplete case i with information from the nearest complete case, say l (the donor), yields the imputed value $\tilde{Y}_{ij} = [r_i] * [c_j] * \left[\frac{Y_{lj}}{r_l * c_j} \right]$. The three terms in brackets represent the row, column, and residual effects. The first two terms estimate the predicted mean, and the last term is the stochastic component of the imputation from the matched case. Again, it must be noted that this approach fails to provide a positive imputation value if only cross-section information is available for a given individual.

4.2. Extended Little & Su method

The extended L&S technique with imputation class (Starick and Watson 2007) distinguishes donors and recipients by taking into account common characteristics. Since donors and recipients should have similar characteristics that are associated with

the variables being imputed, we calculate the correlation between age-group and income component, and education and income component, see the last two columns in Table 1. Unlike Starick and Watson (2007), we use education for the extended L&S technique. This is because not only does age have a high correlation with some income components like unemployment benefits, but even more so does education, e.g., with income from employment. We can thus expect more similarities between donors and recipients by looking for donors within the same education group.

4.3. Carryover procedure

If reported information from another wave is available, the closest reported value is imputed without modification by the carryover method.[14] Note that Starick and Watson (2007) use only the wave before the missing as imputation candidates ("last value carried forward"). In our version, we start with the missing value's next wave, and proceed with the previous wave, if the value from the next wave is not valid. We use values from more distant waves if the components from closer ones are all missing or otherwise not applicable. Unlike Starick and Watson (2007), we do not use the random carryover method that draws one of two possible neighboring values at random.

4.4. Imputation of individuals without income information: cross-sectional "initialization"

All imputation methods described above require that the income component is validly reported in at least one wave. If it is missing in all waves, it needs to be "initialized" first. This is done by means of a cross-sectional stochastic regression based imputation technique[15]. We use as many covariates as possible for the initial regressions, and drop covariates subsequently. Generally, we use similar covariates as Grabka and Frick (2003) to impute the different income components. In each regression step, we regress the income component on all covariates using the first wave to be imputed, and proceed using the next wave, until the value is imputed or still missing values make a reduction of the number of covariates necessary.

Specifically, we proceed as follows, separately for each income component for each eligible individual:

1. Check, if income component needs unit-nonresponse imputation. If yes, also include unit-nonresponding cases to the imputation dataset.[16]
2. Check, if income component is validly reported in at least one wave. If yes, use the appropriate longitudinal imputation method ((extended) L&S, Carryover).

[14] We will consider an inflation factor in the next program version, since e.g., old age pensions increase at a comparable rate.
[15] We use the procedure "uvis" in STATA.
[16] The dataset that underlies the imputation has as many records as individuals, and stores wave-specific income variables and covariates in columns.

Also use the longitudinal imputation once a value is "initialized", that means, imputed in one wave according to steps 3.

3. If income component is never validly reported, initialize. This means:
    - Check possible covariates for the regression imputation. Include also other income components that are (already) available[17].
    - Regress on the whole set of relevant covariates using the reported cases, starting with the first wave to be imputed. If no covariate is missing, use the regression based predicted value as initialization.
    - If any covariate is missing in the first wave, use the second wave to be imputed, etc., until the last wave to be imputed.
    - If any covariate is missing in all waves, drop covariates according to significance, and start again with the first wave. Proceed with increased wave/ dropped covariates until there are any missing values in the last covariate(s).
    - If all significant covariates contain missing values, use a median imputation in the final step.
    - Repeat the whole "initialization" procedure starting with the last wave to be imputed, until the first wave.
    - To decide whether the initialized value from the first ("left to right") or the second ("right to left") regression imputation procedure is used, check which procedure finds a valid value in an earlier regression step. This is the finally initialized value. If both procedures deliver a valid value at the same step, randomly select one of the two values.

5. Which longitudinal method for which variable?

Starick and Watson (2007) report from a simulation study, that for cross-sectional estimates, carryover methods often perform the best, but perform poorly on the distributional accuracy of change between waves. The L&S method usually provides a reasonable compromise between the accuracy of level estimates versus estimates of change, particularly for respondents. Where there is a reasonably good correlation between the imputation class variable (age or education) used in the L&S method and the variable being imputed, the L&S variant that uses imputation class performs better than the basic L&S. However, when the imputation class variable is only weakly associated with the variable to be imputed, the basic L&S method performs better, especially when the donor pool is small. They find in addition that the carryover methods are more likely to understate change and overstate correlation between waves. With respect to *cross-survey* robustness of results from different imputation methods, Frick and Grabka (2007) compare imputed values from the SOEP, the

[17] This requires carefully analyzing the optimal order of income imputation.

HILDA, and the British Household Panel Survey (BHPS). While for the SOEP and the HILDA the L&S imputation method was traditionally used, for the BHPS both cross-sectional methods (traditionally used to impute BHPS missing income values) was tested against the L&S (new) method. They find that using L&S technique also for the BHPS produces remarkably similar (structural) results. Giving priority to the longitudinal L&S method is certainly in line with the harmonization efforts put forward by Frick and Grabka.

Following Starick's and Watson's (2007) recommendations, we use the longitudinal method listed in

Table 2 to impute item- or unit- missing income values.

Table 2: Longitudinal Imputation method and unit-nonresponse Imputation used.

Income component *Income source:*	*Longitudinal Imputation Method*	*Unit-Nonrresponse Imputation*
employment: net (empyn)	*Ex. L&S (Education)*	*yes*
independent work: net (indyn)	*Ex. L&S (Education)*	*no*
(old age) Pension: annual income (oasiy)	*Carryover*	*yes*
invalidity pension (aiy)	*Carryover*	*yes*
pension insurance (peny)	*Ex. L&S (Education)*	*yes*
unemployment fund (uney)	*L&S*	*yes*
social assistance (wely)	*L&S*	*yes*
grants, scholarship (gray)	*L&S*	*no*
institutions, organizations (insy)	*L&S*	*no*
family allowances (famy)[13]	*Carryover*	*yes*
people from other private households (pnhy)	*Ex. L&S(Education)*	*no*
other sources (osy)	*L&S*	*no*

6. Income Distributions with and without imputed Values

In this section, we compare the imputed income variables, distinguishing the three disjoint missingness classes:

- Validly reported income values
- Imputed item-nonresponse values
- Imputed unit-nonresponse values

We generally drop the upper 1% percentile from the data (from each wave for the income from employment variable; from the pooled waves for all other variables). In the kernel density estimates graphs, the curves of the income variable densities by missingness class are drawn. In the tables, we list the medians, the standard deviations and the sample sizes, before we graph them to facilitate interpretation.

6.1. Income from employment (empyn)

The sample size of the (imputed) income from employment (empyn) allows for analyses separated by wave.

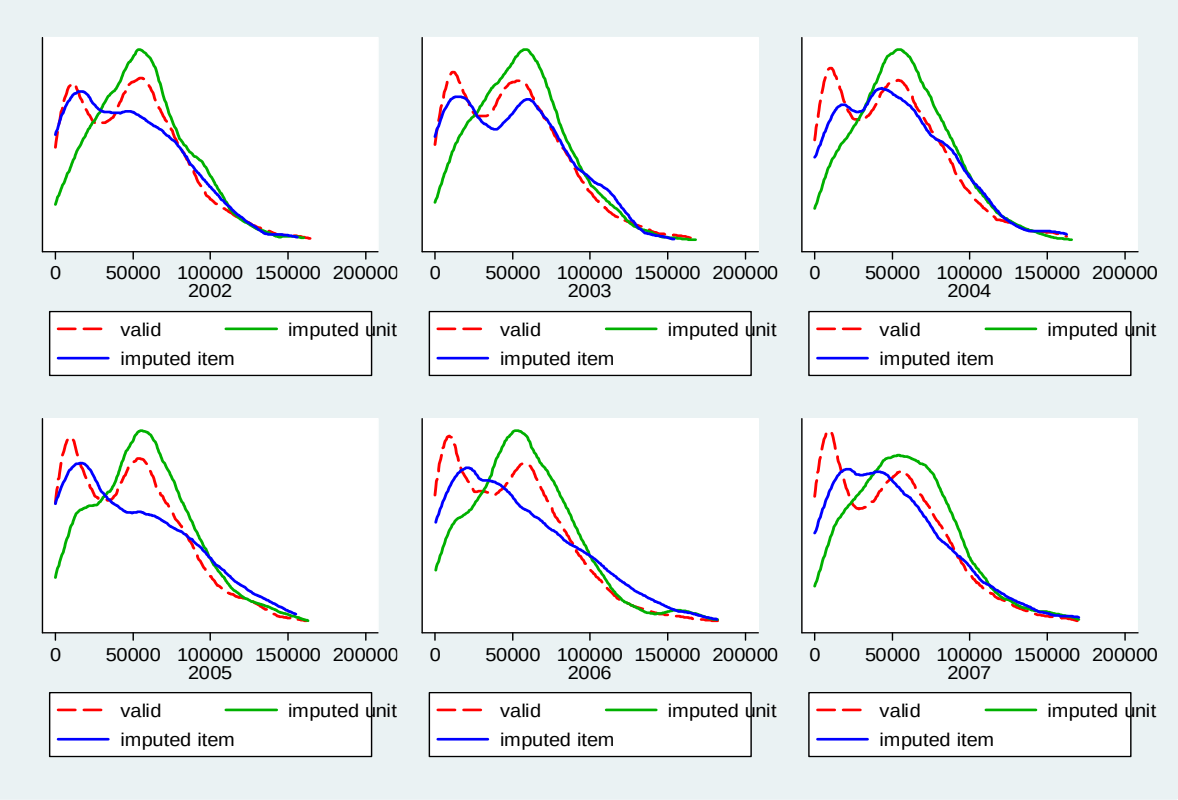

Figure 1: Densities of income from employment in the SHP 2002-2007, by missingness class and year.

While the validly reported income from employment is bimodal in all waves, this is not necessarily the case for the imputed item and never for the imputed unit responding incomes. Imputed income curves from the item-nonrespondents, although close to that from the validly reporting cases, is somewhat "smoothened", and slightly underrepresent the lowest income groups. The imputed unit-nonrespondent´s incomes generally have a smaller variance. This is due to a comparatively strong underrepresentation of the lower income groups.

Table 3 depicts the medians, standard deviations and sample sizes of the income from employment, by missingness class and wave.

Table 3: Median and Standard Deviation: income from employment (in Swiss Franks) SHP 2002-2007.

	1. valid reported			2. imputed item nonresp.			3. imputed unit nonresp.		
	Median	Std.dev	N	Median	Std.dev	N	Median	Std.dev	N
2002	48,600	33,883	3,297	43,071	34,903	293	54,688	30,585	1,063
2003	48,000	34,541	3,076	48,576	35,470	189	55,889	30,914	876
2004	47,840	34,068	4,710	49,348	35,201	446	55,898	30,244	1,853
2005	47,390	34,829	3,978	43,664	39,358	247	55,832	32,558	1,426
2006	47,390	36,118	3,998	45,478	40,833	245	55,396	34,289	1,246
2007	46,800	35,757	4,274	46,020	37,197	242	57,238	33,144	1,210

To facilitate interpretation, we graph the medians and the standard deviations:

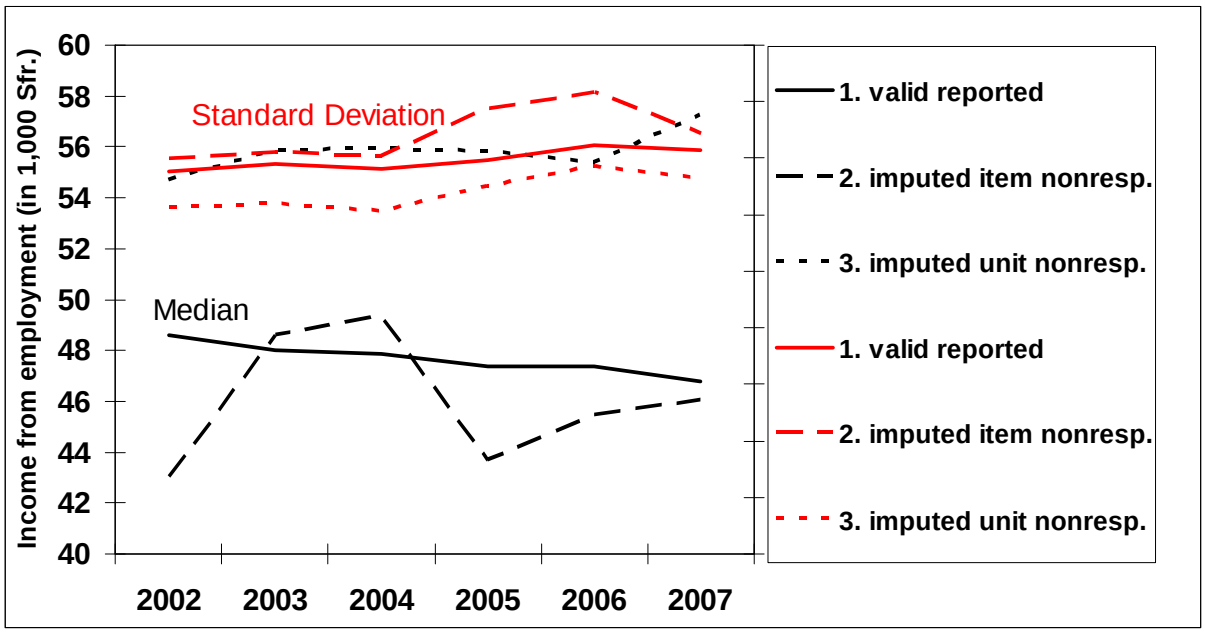

Figure 2: Median (black) and Standard Deviation (red) for income from employment, SHP 2002-2007.

Expectedly from Figure 1 and Table 3, while the level (black: median) of the validly reported incomes and the item-nonresponse imputed incomes are similar, we find a higher level among the imputed unit-nonresponding individuals in all years. The variation (red: standard deviation) across the missingness classes is the same over the years.

6.2. Income from other sources with imputed unit-nonresponse

In addition to income from employment, unit-nonresponding individuals are imputed for the following income sources:

- come from invalidity pension (aiy)
- Income from unemployment fund (uney)
- Income from social assistance (wely)
- Income from old age pension (oasiy)
- Income from pension insurance (peny)
- Income from family allowances (famy)

Similarly to income from employment, we first depict the graph with the density curves of these income components by missingness class in Figure 3. Note that we now pool the data over all waves between 2002 and 2007.

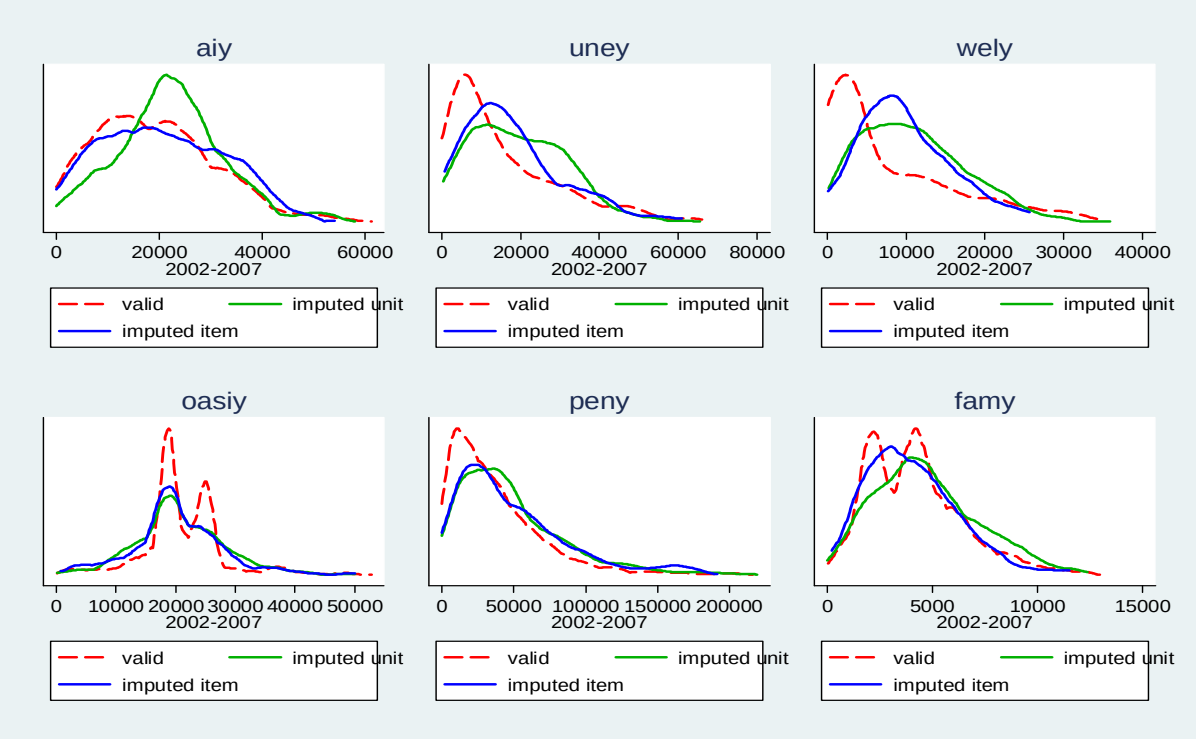

Figure 3: Densities of unit-imputed income components other than from employment in the SHP 2002-2007 (Family allowances: 2004-2007), by missingness class.

(Wave pooled) income curves from different missingness classes have different shapes. While the item and unit imputed income curves are rather similar with the exception of those from invalidity pensions (aiy), the validly reported income curve coincides surprisingly well with the imputed item-nonresponse curve for this component. Again, there is the tendency that low income cases are underrepresented in the imputed income variables.

Table 4: Median and Standard Deviation: unit-imputed income components other than from employment (in Swiss Franks) SHP 2002-2007.

	1. valid reported			2. imputed item nonresp.			3. imputed unit nonresp.		
	Median	Std.dev	N	Median	Std.dev	N	Median	Std.dev	N
aiy	18,000	11,889	1,069	19,856	11,791	144	22,105	10,412	232
uney	11,060	14,909	870	16,307	13,295	73	18,656	12,900	211
wely	4.200	8,374	412	8,459	5,735	51	10,018	6,532	673
oasiy	19,625	6,784	6,146	19,200	7,862	513	19,683	7,438	1,571
peny	28,800	32,692	3,813	37,285	38,600	478	40,849	36,269	1,533
famy	4,080	2,337	3,724	3,840	2,343	325	4,357	2,513	1,737

Figure 4 graphs the measures for the components:[18]

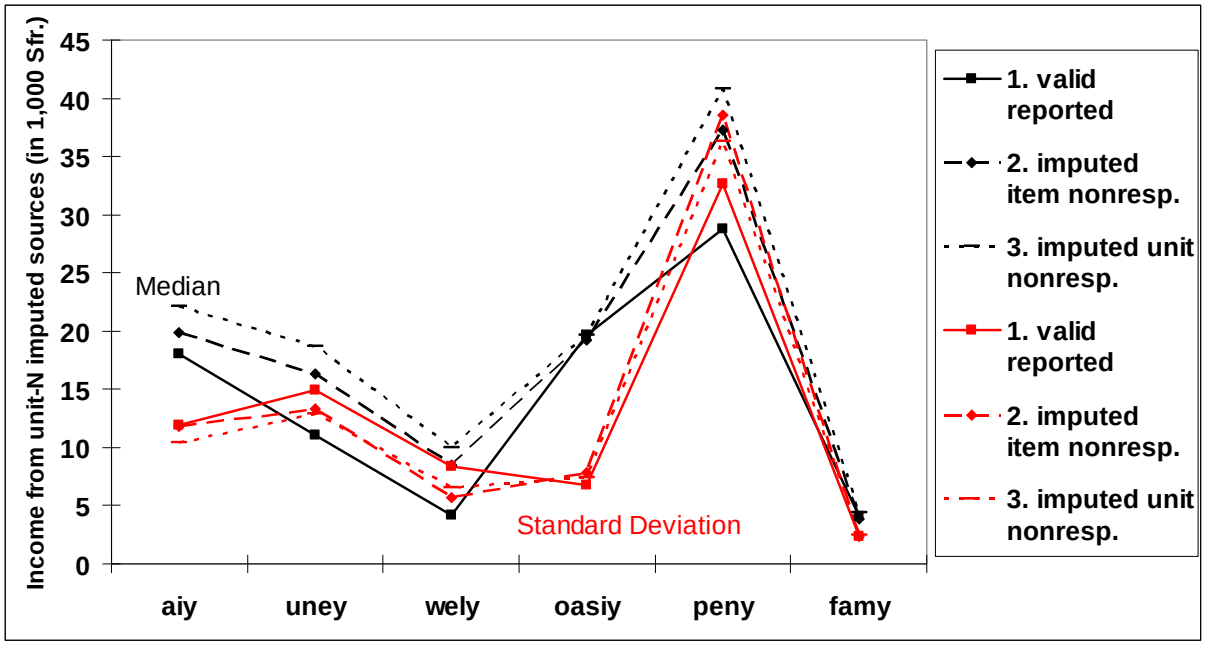

Figure 4: Median (black) and Standard Deviation (red), for unit-nonresponse imputed income components other than from employment, SHP 2002-2007.

[18] Note that the faint lines connecting the income sources are for better readability.

Only the two measures for income from old age pensions (oasiy) and from family allowances (famy) are the same for all missingness classes, due to their small variations. For all other median values, the following order is respected: validly reported lowest, imputed item-nonresponse second, and imputed unit-nonresponse highest.

6.3. Income from sources without imputed unit-nonresponse

For the following income variables, unit-nonresponse is not imputed:

- Income from independent work (indyn)
- Income from Grants, scholarship (gray)
- Income from other institutions or organizations (insy)
- Income from people in private households (outside the household) (pnhy)
- Income from other sources: (osy)

In Figure 5, we depict kernel density estimates for these income sources, again dropping the upper 1% percentile.

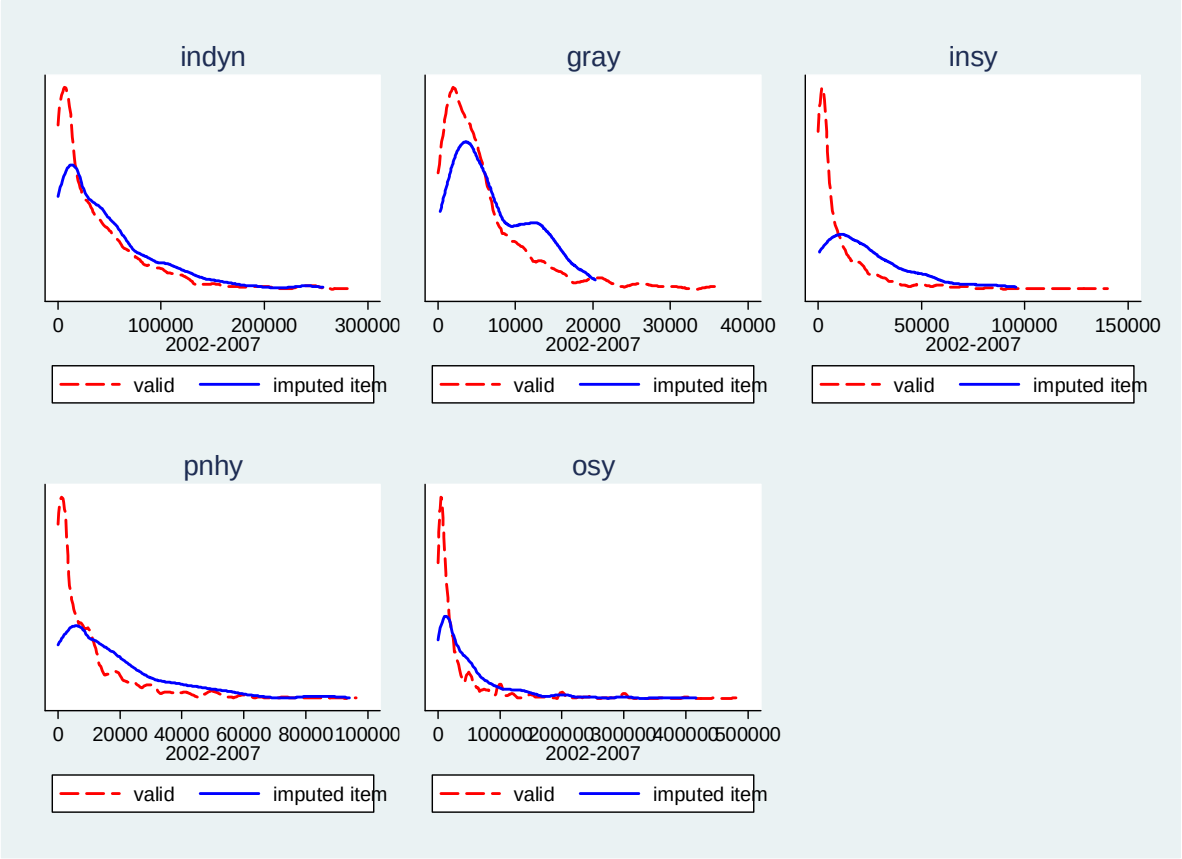

Figure 5: Densities of income from sources that are not unit nonresponse imputed in the SHP 2002-2007 (aggregated), by missingness class.

Again, low income cases are underrepresented in the imputed item income variables, relative to the validly reported cases.

Table 5: Median and Standard Deviation: not unit-imputed income components in the SHP 2002-2007 (Family allowances: 2004-2007), by missingness class (in Swiss Franks).

	1. valid reported			2. imputed item nonresp.		
	Median	Std.dev	N	Median	Std.dev	N
indyn	24,000	46,638	3,728	35,127	48,479	859
gray	4,000	6,020	411	5,173	5,114	74
insy	4,720	19,130	1,201	18,185	21,313	113
pnhy	5,000	15,660	3,385	12,310	17,389	304
osy	12,000	62,702	3,469	29,983	59,950	615

Figure 6 depicts the level and variation differences across the missingness classes for each component:

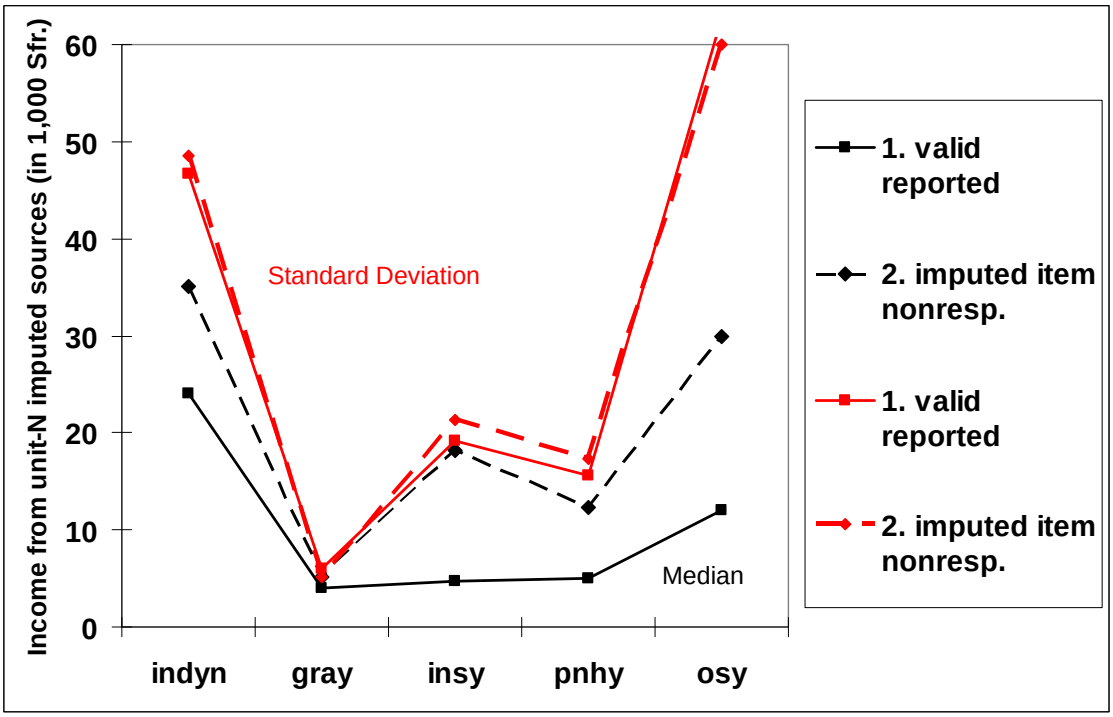

Figure 7: Median (black) and Standard Deviation (red), for not unit-nonresponse imputed income components, SHP 2002-2007.

Also with respect to the income variables that are not unit-nonresponse imputed, with the exception of income from grants (gray), the imputed item-nonresponse cases always have a much higher level than the validly reported cases. The variances are again about the same.

7. Summary

In the paper we describe the methods used to impute different income components in the Swiss Household Panel, and compare the results of the imputed cases with the validly reported values. We use a two stage approach: for each individual entitled to earn income from a specific income component, we first make sure that positive income is available in at least one wave; i.e., validly reported. If this is not the case, we use a cross-sectional stochastic regression-based imputation procedure to "initialize" the income component for all eligible individuals in one wave. Given the income component is present in at least one wave, we use appropriate longitudinal imputation procedures. Depending on the component, we use the stochastic Little & Su method (standard or extended version, by education), or the simple carryover method.

To check which consequences the imputation has on cross-sectional measures, we analyze level and variation of each income component, distinguished by the following missingness classes: validly reported, imputed value from item-nonresponse, and imputed value from unit-nonresponse. It turns out that while the *levels* of the imputed item-nonresponses are generally higher than the levels of the validly reported values, the imputed unit-nonresponses are considerably higher for almost all income components. The *variations* of the imputed values are mostly the same as those of the validly reported values. These findings prove the need to impute missings of both item- and unit nonresponding income values. The first is necessary to avoid bias from underestimated levels because item-missing cases (that have a higher income) are ignored. The second is necessary to avoid underestimated household income when income from all household members is aggregated.

Especially neglecting missing income from unit-nonresponse is a problem. Frick et al. (2009) show that – among the possible treatments of unit-nonresponse – unlike imputing income from unit-nonresponding cases, ignoring partial unit nonresponse, adjusting equivalence scales, or deleting partial unit-nonresponding households or individuals is not an option. Further research using data from the SHP could go beyond pure level or variation measures but compare income equality and mobility aspects that result by either imputing unit-nonresponding household members or just using reported income to adjust for the missing information.

8. References

Frick, J., S. Jenkins, D. Lillard, O. Lipps and M. Wooden (2007) The Cross-National Equivalent File (CNEF) and its Member Country Household Panel Studies. *Schmollers Jahrbuch* **4/2007**

Frick, J. and M. Grabka (2004) Missing Income Data in Panel Surveys: Incidence, Imputation and its Impact on the Income distribution. *DIW Discussion Papers* **376**, Berlin

Frick, J. and M. Grabka (2007) Item Non-Response and Imputation of Annual Labor Income in Panel Surveys from a Cross-National Perspective. *IZA Discussion Paper* **3043**

Grabka, M. and J. Frick (2003) Imputation of Item Non-Response on Income Questions in the SOEP. *DIW Research Notes* **29**, Berlin

Frick, J., M. Grabka and O. Groh-Samberg (2009) Imputation of Annual Income in Household Panel Surveys with partially non-Responding Households. Paper presented at the second ESRA conference, 29/6-3/7, Warsaw

Kuhn, U. (2008) Collection, construction and plausibility checks of income data in the Swiss Household Panel. Swiss Household Panel Working Paper 1_2008, Lausanne

Lipps, O. (2009) Attrition of Households and Individuals in Panel Surveys. *SOEPpapers* **164**

Lipps, O. and U. Kuhn (2009) Codebook for CNEF variables in the SHP (1999 - 2007). *Swiss Household Panel Working Paper* **5_09**, Lausanne

Starick, R. and N. Watson (2007) Evaluation of Alternative Imputation Methods: The HILDA Experience. HILDA, June 2006

Watson, N. (2004) Income and Wealth Imputation for Waves 1 and 2. *HILDA Project Technical Paper Series* **3/04**, University of Melbourne