Frick, Joachim R.; Grabka, Markus M.

Research Report

Missing income data in the German SOEP: Incidence, imputation and its impact on the income distribution

SOEP Survey Papers, No. 225

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Frick, Joachim R.; Grabka, Markus M. (2014) : Missing income data in the German SOEP: Incidence, imputation and its impact on the income distribution, SOEP Survey Papers, No. 225, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/104999

SOEP The German Socio-Economic Panel Study

225

SOEP Survey Papers

Series C - Data Documentations

SOEP — The German Socio-Economic Panel Study at DIW Berlin 2014

Missing Income Data in the German SOEP

Incidence, Imputation and its Impact on the Income Distribution

Joachim R. Frick and Markus M. Grabka

DIW SOEP

Running since 1984, the German Socio-Economic Panel Study (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:

Series A – Survey Instruments (Erhebungsinstrumente)

Series B – Survey Reports (Methodenberichte)

Series C – Data Documentations (Datendokumentationen)

Series D – Variable Descriptions and Coding

Series E – SOEPmonitors

Series F – SOEP Newsletters

Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at
http://www.diw.de/soepsurveypapers

Editors:
Prof. Dr. Gert G. Wagner, DIW Berlin and Technische Universität Berlin
Prof. Dr. Jürgen Schupp, DIW Berlin and Freie Universität Berlin

Please cite this paper as follows:

Joachim R. Frick, Markus M. Grabka. 2014. Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income Distribution. SOEP Survey Papers 225: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

Contact: DIW Berlin
SOEP
Mohrenstr. 58
10117 Berlin

Email: soeppapers@diw.de

JOACHIM R. FRICK
MARKUS M. GRABKA[1)]

MISSING INCOME DATA IN THE GERMAN SOEP: INCIDENCE, IMPUTATION AND ITS IMPACT ON THE INCOME DISTRIBUTION

Berlin, 2003

Reprint 2014

1) DIW Berlin, The Research Infrastructure 'Socio-Economic Panel (SOEP)', Berlin/Germany, mgrabka@diw.de

Missing Income Data in Panel Surveys:
Incidence, Imputation and its Impact on the Income distribution*

by Joachim R. Frick and Markus M. Grabka (DIW Berlin)

Summary:

This paper deals with the question of selectivity of missing data on income questions in large panel surveys due to item-non-response and with imputation as one alternative strategy to cope with this issue. In contrast to cross-section surveys, the imputation of missing values in panel data can profit from longitudinal information which is available for the very same observation units from other points in time. The "row-and-column imputation procedure" developed by Little & Su (1989) considers longitudinal as well as cross-sectional information in the imputation process. This procedure is applied to the German Socio-Economic Panel study (SOEP) when deriving annual income variables, complemented by purely cross-sectional techniques.
Based on the SOEP, our empirical work starts with a description of the overall incidence of imputation and its relevance given by imputed income as a percentage share of the total income mass: e.g. while 21 % of all observations have at least one missing income component of their pre-tax post-transfer income, 9 % of the overall income mass is imputed. However, this picture varies considerably for more recent sub-samples of the panel survey. Secondly, we analyze the respective impact of imputation on the personal distribution of income as well as on results of income mobility. When comparing income inequality measures 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 and this effect appears to be relevant in both tails of the distribution, although somewhat more prominent among higher incomes. Longitudinal analyses show firstly a positive correlation of item-non-response on income data over time, but also provide evidence of item-non-response as being a predictor of subsequent unit-non-response. Applying various income mobility indicators there is a robust picture about income mobility being understated using truly observed information only. Finally, multivariate models show that survey-related factors (number of interviews, interview mode) as well as indicators for variability in income receipt (due to increased complexity of household structure and income composition) are significantly correlated with item-non-response. In conclusion, our empirical results based on the German SOEP indicate the selectivity of item-non-response on income questions in social surveys and push the necessity for adequate imputation.

Keywords: Item-Non-Response, Imputation, Income Inequality
JEL-classification: C81, D31, I32

1 Introduction

A common phenomenon in population surveys is the failure to collect complete information on interesting characteristics at individual and household level. In general, one can differentiate between unit and item-non-response. Unit-non-response results in the lack of any information on a given observation and turns out to be the strongest type of refusal -- this issue is not being dealt with in this paper. If, however, only a subset of information is missing from an otherwise responding observation this failure is referred to as item-non response. The latter may be caused by a respondent's reservation to answer to a question that appears to be too sensitive (to him/her), or that affects confidentiality and privacy or simply from the fact that the correct answer is not known.

Reviewing some findings from the literature reveals a variety of approaches to tackle this phenomenon. Schräpler (2003) stresses the survey context as a direct cause for item-non-response, where increased complexity or increased heterogeneity of a surveyed construct like income will result in a higher share of item-non-response. Hill and Willis (2001) argue that response propensities depend on how the respective question is formulated.

Privacy and confidentiality issues are another motivation to keep things to oneself. There is a risk of disclosure of answers to third parties or social undesirability of the answers (Schräpler 2002). That might be one reason why such refusal is concentrated in the tails of the income distribution (e.g., Biewen 2001).

There is some empirical evidence, that the interview situation needs to be considered as well: Schräpler and Wagner (2001) provide evidence, that it is not only the individual respondent's characteristics that may be associated with item-non-response, but also interviewer-respondent matching effects based on age and sex. Rendtel (1995) found that in the case of a panel survey the change of a well known interviewer will also alter the willingness of respondents to cooperate in a survey. While supporting this finding for item-

non-response on income questions, Riphahn and Serfling (2003) find non-response behavior on wealth questions to be negatively correlated to an interviewer change. Finally, Loosveldt et al. (1999) indicate a positive correlation between item and subsequent unit-non-response.

According to Rubin (1976) the mechanisms leading to missing data can be classified into three subgroups:

- Missing Completely at Random (MCAR),
- Missing at Random (MAR), and
- Missing Not at Random (MNAR).

MCAR means that the missing data mechanism is unrelated to the variables under study, whether missing or observed. The observed values are a random sample of the underlying population and any analysis on complete cases yields the very same results as the full data set would have. An alternative and weaker version of the MCAR assumption is the Missing at Random (MAR) condition. The cause of the missing data is unrelated to the missing values, but may be related to the observed data. In other words MAR means that the missing values are related to either observed covariates or response variables. The third subgroup is Missing Not at Random (MNAR). MNAR occurs when the Missing mechanism depend on the actual value of the missing data. This is the most difficult condition to model for.

While MCAR and MAR may in principle be ignorable missing mechanisms, MNAR requires adequate statistical treatment. There are in principle three ways how to deal with missing values, (a) case-wise deletion, (b) weighting and (c) imputation. A case-wise deletion may be either list-wise (complete cases only) or pair-wise. This very commonly used technique implies that cases are deleted which contain missing data in the variables which are relevant for the analysis being carried out. This procedure can substantially lower the sample

size, leading to a severe lack of power and may cause a bias in the analysis, depending on the underlying missing mechanism.

The second strategy to deal with missing values is weighting, i.e. “correcting” for an under-representation of certain characteristics by increasing the population weight of those observations which in fact did participate in the survey, although they face a higher risk of non-response. A relevant outcome of this strategy is an increase in the variance, producing less precise estimates.

The third strategy is imputation, where single and multiple imputation methods are being discussed (c.f. Rubin 1987). In contrast to single imputation techniques which provide only one estimate for each missing component, multiple imputation techniques return *m* complete datasets by imputing *m* times. It is argued that multiple imputation yields improved estimators compared to case-wise deletion, or single ad-hoc imputation methods, however, this technique may be shunned by less sophisticated users of micro-data (c.f. Spiess & Goebel 2003; Scheffer 2002).

The advantage and also the aim of imputation is to complete a data set with "full" information for all observed individuals, which reduces bias in survey estimates and – from a point of a data user – also simplifies the analysis. Retaining all observations, independent from item-non-response, is supposed to yield an improved basis for (social) policy oriented analyses. However, it must be noted, that even a very sophisticated approach for substituting for non-response may not be sufficient to completely eliminate any bias resulting from it in the first place. As such, the adequate choice of the imputation technique is a problem in itself. Potential bias due to imputation may creep in due to “regression-to-the-mean effects” and a potential change in total variance -- most likely a decline -- can occur. This is of special relevance in surveys with a rather small number of observations or if small subgroups are affected above the average rate.

Finally, even if one successfully imputes missing values, data quality may still be influenced by several other distortions such as measurement error or even in case of "truly observed" cases, e.g., due to rounding (Hanisch and Rendtel 2002). "None of the approaches is always right or always wrong and it is important to understand the conditions under which each approach is preferred" (Binder 1996: 571).

This study deals with the question of how to deal with item-non-response on *income* in a large *panel survey* (the German Socio-Economic Panel Study, SOEP) and the underlying selectivity process for missing income data which is the empirical basis for analyses of the personal income distribution and poverty. Very often, this type of analysis relies on income aggregates rather than single income items which actually have been collected during the interview. A prominent example is welfare analysis, based on a household's post-government (i.e., post-tax post-transfer) income. The relevance of missing values in such an income construct is very much affected by the degree to which aggregation across persons is necessary: e.g. household labor income – being just one major component of market income – consists of all individual household members' labor income, which itself is the sum of labor income from first and second jobs, self-employment income, one-time allowances such as vacation or Christmas bonuses, etc.. Almost by definition, the risk of understating the true income from a given source is increasing with the number of items and household members across which income components have to be aggregated. If the underlying missing mechanisms do not follow a random process, all derived information on income inequality will be severely distorted. In case of a panel survey this will also hamper the analysis of income mobility.

The paper is organized as follows: After briefly outlining alternative imputation methods, chapter 2 deals with the incidence of item non-response in selected income components of the German SOEP and the applied imputation techniques. Chapter 3 compares results on income distribution and inequality based only on completely observed cases to

those derived from imputed cases: We analyze incidence and relevance of item non-response across the personal income distribution and investigate specific problems related to longitudinal data with respect to income mobility and subsequent unit non-response. Based on multivariate modeling we reinforce the allegation of selectivity in item-non-response on income and the necessity for an adequate treatment, e.g. by means of imputation. Chapter 4 concludes from a user's point of view.

2 Imputation techniques

2.1 Commonly used single imputation techniques

Single imputation techniques which are applied to cope with item-non-response in population surveys vary considerably in terms of the underlying complexity. The following is a selection of commonly used techniques:

- Logical Imputation: This imputation method is only feasible in cases where a straightforward link between a piece of missing information and at least one observed characteristic can be established, e.g. imputation of sex given by the first name.
- Expert Imputation: The expert imputation is commonly used in official statistics (e.g the German Income and Expenditure Survey (EVS), EUROSTAT), based on the extensive experience of such institution in collecting survey data. However, this approach is highly subjective and – in most cases – unreproducible for interested third parties. As such, this procedure most likely yields distortions in the data set (e.g. mean-bias).
- Mean or Median Substitution: The Mean Substitution is a very prominent example, which – almost by definition - creates distortion in variance. Median-based substitution may yield somewhat more conservative results and is less sensitive to outliers.

- Zero Substitution: This procedure gives a value of zero to the missing (income) information, which strongly underestimates the true value to an unpredictable extent as well as biases the true variance of the information of interest.
- Hot Deck Imputation: The idea of the hot deck-imputation is to randomly select an internal donor (i.e. donor is part of the same database as recipient) with a complete set of data whose values are then assigned to a recipient with missing data. The donor and recipients typically are matched according to certain characteristics such as sex, race, age, etc.. This technique may yield a reduction of variance.
- Cold Deck techniques are comparable to hot-deck procedures, however, donor's information is taken from external sources which may yield additional unpredictable distortion.
- Applying regression estimates on the basis of completely observed cases to otherwise comparable observations with missing data, the regression-based imputation techniques are certainly less selective than the above mentioned approaches. However, problems arise from regression to the mean effect, a potential distortion of variance and the normative decision on the choice of covariates. An improvement comes with the consideration of a randomized residual which may be added to the predicted value.
- The Row-and-Column Imputation as described by Little & Su (1989) is possible on the basis of longitudinal data only. It takes advantage of information on the very same individual over time by combining row (unit) and column (period/trend) information. In principle, the imputed value is the result of a combination of *row effect, column effect and a residual effect.* The column effects are given by: $c_j = (j * Y_j) / \Sigma Y_k$, i.e., these are calculated for each wave of data (e.g. 16 waves over the time period 1984-1999) where $j = 1, ..., 16$ and Yj is the sample mean income for year j. The row effects, $r_i = m_i^{-1} * \Sigma (Y_{ij} / c_j)$, are computed for each sample member. Y_{ij} is the income for individual i in year j and m_i is the number of recorded months. Sorting cases by r_i and

matching the incomplete case *i* with information from the nearest complete case, say *l*, yields the imputed value $i = [r_i] * [c_j] * [Y_{lj} / (r_l * c_j)]$. 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. Overall, the corresponding bias in variance appears to be somewhat less severe. However, 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.

2.2 Application to the SOEP

The German Socio-Economic Panel Study (SOEP) is the longest running, ongoing major panel study in Europe (see SOEP Group 2001, Haisken-DeNew and Frick 2002). Following is a brief sketch of survey specific features which may be relevant for the following analysis of item-non-response:

- all adult household members (aged 17 and over) are surveyed individually, plus one interview with household head (in contrast to the US PSID, where only the head is interviewed giving proxy information on all the other household members),
- in order to keep the survey sample a representative one[1] as well as to ensure sufficiently high numbers of observations, various new sub-samples have been incorporated after the initial start in 1984. For theses new observations – by definition – less longitudinal data is available.

[1] The fall of the Berlin wall in late 1989 and the subsequent unification of East and West Germany was in fact an enlargement of the survey territory which had been taken care of by the inclusion of a new sample in June 1990 (Sample C). Other than that, massive immigration since the initial sampling took place in 1983 caused the underlying population to change considerably. If immigrants form new households, their sampling probability is Zero, i.e., immigration is only considered in an ongoing panel study as long as people move into existing households. In order to cope with this phenomenon an "immigrant" sample (Sample D) has been established in 1994/95. Since then two new "refreshment samples" E and F have been started in 1998 and 2000, respectively. These new samples help to stabilize the number of observations and they serve as a basis for controlling panel effects as well (see Schupp and Wagner, 2002). The most recent extension is given by sample G (High Income Sample) which started in 2002; however, this data is not considered in this paper.

- interviews usually take place as face-to-face interviews with the interviewer filling in the questionnaire. In order to keep (unit- as well as item) response rates high, a personal relationship between interviewer and respondent appears to be very helpful. That is why the stability of the interviewer over time is considered to be crucial.

As is true for all population surveys, item-non-response on income questions is a major concern in the SOEP as well. The construction of any aggregated (annual) income variables needs to consider the complexity of household as well as income composition.[2] According to the underlying specificities of a given type of income, the respective variable is surveyed from different units of observation (individual vs. household) and across different dimensions of time (month vs. year). While in some cases a simple amount per year can be asked right away, in other cases a more appropriate way of surveying is to collect the number of months with receipt of a given type of income over the last calendar year which is multiplied with the respective average monthly amount in order to yield the annual amount.

As such, almost all income constructs need to aggregate information across various income sources (e.g. an individual's annual labor income is made up by up to 10 single income variables) as well as across all household members (e.g. a household's pre-tax income may rely on up to 46 different income components for a one-person household[3]). Obviously, the probability of an income measure being affected by item non-response and its resulting impact on the aggregated income measure increases with the number of the considered variables and the aggregation level (cf. Schräpler 2003).

---

[2] Annual income figures are most relevant for cross-national comparative (longitudinal) research given the different perceptions of the relevant time frame when asking for income in continental European countries as compared to Anglo-Saxon countries. The SOEP variables considered in this paper, where developed in the context of the Cross-National-Equivalent File (CNEF) which includes data for Great Britain, Canada, Germany and the USA (see Burkhauser et.al. 2001) .

[3] This includes up to 30 income components observed from individual at a monthly basis, up to 6 variables on annual basis as well as up to 5 household level variables each, for monthly as well as annual income components. Overall, this aggregated income is input for a model simulating taxes and social security contributions according to the rather complex German tax filing procedures (see Schwarze 1995)

The imputation of item-non-response related missing income data in the SOEP follows a two step procedure (cf. Grabka and Frick 2003).[4] The general principle is to apply the row and column imputation technique (hereafter L & S) whenever longitudinal income data is available, and to run purely cross-sectional imputation techniques otherwise. As a matter of fact, the empirical implementation of L & S in the case of SOEP fails in all those cases where a given income component is not observed in any other wave of data considered in the imputation process. This includes not only first time respondents, but also those observations for whom a given income variable has been surveyed for the very first time. In all of those cases there is a need for an alternative imputation procedure which is based on cross-sectional data only, i.e., on data observed from other units (individuals or household, respectively) in the very same wave. See Appendix for a complete overview of the techniques applied for the various SOEP income variables.

- Following logical imputation, institutional or external information is used to impute missing amounts of those income components which are perfectly related to otherwise observed information, e.g. child benefit which is fixed per child, direct housing support for owner occupiers which is related to the number of children and the construction year of the building, as well as nursing care insurance which is fixed to the observed needs.
- Median Substitution takes place for income components which are of minor relevance in terms of the number of affected cases ($n < 10$) as well as with respect to the level (e.g. military service pay, maternity benefit). Median Substitution for Subgroups is performed for e.g. housing benefit for owner occupiers by household size.
- Median Share Substitution is chosen if a link between two income variables can be established, e.g. the median share of the monthly labor earnings and the Christmas bonus in the private sector in Germany is about 35%. Any observation with a missing Christmas

[4] For the application of imputation approaches in other surveys see e.g., Nicoletti and Peracci (2003) for the ECHP, Cao (2001) for the Health and Retirement Survey (HRS), and Taylor et al. (1998) for the BHPS

bonus in the private sector is assigned an imputed value given by the individually observed labor income times the (median) share of 35%. This allows for more variation of the imputed income values than single median substitution would do.

- Regression-based substitution is used for more complex income constructs e.g. "interest and dividends" or "individual labor income from first job"; in the latter case Mincer-type wage regressions are applied for imputation purposes (cf. Grabka and Frick 2003).

The share of missing values varies considerably across different types of income (Table 1); e.g. while "child benefits" and "Christmas bonuses" are observed for more than 96% of the relevant population, the picture is quite different for "interest and dividends" and "income from self-employment" which show response rates of only 80% to 85%. It appears, that the latter two incomes are not only more complex to capture, but that they are more sensitive as well. This may also be conveyed by the fact that these income components are less stable over time. The resulting missing values are pre-dominantly imputed by means of L & S, however, it is again the more volatile components which demand purely cross-sectional imputation. This may be the case partly because these items are observed less often per individual, partly because these individuals may have a higher probability to attrite from the panel survey leaving less observations for the imputation routines to draw from.

It may be interpreted from Table 1, that more recent years show an increase in item-non-response. However, Table 2 illustrates that (for the case of "individual labor income") this effect comes from the inclusion of the very large sub-sample F (starting year 2000) which contributes about 40% of all observations in 2001. The older samples A-E have a considerably lower share of item-non response which is also rather stable over time (less than 10%), while it can be assumed that in sample F a trustful relationship between interviewer and

---

where "hot deck" techniques as well as regression based imputation methods are applied (see also http://www.iser.essex.ac.uk/bhps/doc/index.html [Section V]).

respondents still needs to be established.[5] As a result of this, item-non-response in sample F is about 50% higher than in the older samples. Additionally, for this very young sample, there is no sufficient longitudinal data to draw from for the application of the L & S procedure, i.e., the share of those missing values that are imputed by means of purely cross-sectional techniques is rather high.

In order to shed some more light on the quality of these two procedures, Figure 1 offers kernel density estimates for both types of imputations (L & S vs. cross-sectional regression-based approach) in comparison to the results for the "truly" observed population. These calculations are based on a random sample of approx. 1000 observations for which a positive value has been observed and which provide longitudinal information as a prerequisite for the L & S procedure, i.e. we can compare the results of our various imputation techniques to the truly observed information. Mean and Median of the cross-sectional imputation procedure are in better compliance with those of the observed distribution than is the case for the L & S procedure. Although both techniques at first glance appear to yield rather reliable results, the distribution of the cross-sectional procedure is more "out-of-bounds" when comparing both imputation results to the upper and lower bounds of a 2-Sigma confidence band of the observed distribution, something that is due to the clearly understated variation. Another impressive illustration is incorporated in the significant understatement of the Gini coefficient: While the L & S procedure overstates inequality by about 9%, the cross-sectional approach understates the Gini by about 18%.

Concluding from this, one may argue that the L & S procedure, taking advantage from the individual's own record over time, yields more reliable imputation results than a purely cross-sectional approach does.[6] Although L & S is our preferred approach there is obvious need for a purely cross-sectional imputation in case of lacking longitudinal data. Following

[5] However, we can not rule out at this point that the lower item-non-response in samples with a longer SOEP-history are simply the result of selective attrition.

from this, in the remainder of this paper we will interpret all imputed values together ("imputed cases") to be matched with the observed population ("observed cases") in order to show the impact of item-non-response and imputation on the overall distribution ("all cases") of selected income aggregates.

While the exercise in Figure 1 was based on a sub-sample of hypothetically imputed cases with actually observed "income from first job" in order to illustrate differences in the imputation techniques, Figure 2 gives kernel density estimates based on the complete samples of observed and factually imputed populations, respectively. Here, the latter group encompasses *both* imputation techniques as described above. However, when comparing these distributions one should keep in mind that – in contrast to Figure 1 – the underlying populations may differ.[7] Overall, the distributions of observed and imputed values appear to be rather congruent with the latter yielding somewhat lower mean and median, and slightly higher values for inequality as given by the Decile Ratio 90:10 and the Gini coefficient.

3 Income distribution analyses based on observed and imputed income data

The analysis of the personal distribution of income is a major welfare oriented application of population surveys. Typically, post-tax post-transfer income ("Post-Government Income") is used to address questions on income inequality and poverty. However, in the following section we will also deal with pre-tax post-transfer income ("Total income") in order to better control for an eventual distorting influence from taxes and social

[6] This finding is in line with those of Spiess and Goebel (2003) based on survey and register data for Finland.

[7] Also the number of observations for the two graphs differ considerably: approx. 1,000 for the imputed population and 11,000 for actually observed population, respectively.

security contributions, which are hard to survey in Germany, and therefore need to be simulated in the SOEP.

3.1 Imputation and income inequality

A comparison of basic statistics (Mean, Median and Standard Deviation) for post-government income for actually observed and imputed cases (see left panel of Table 3) shows income levels and variation to be higher among the imputed population which translates accordingly into the result for the overall population ("All cases"). Using observed and imputed cases yields a mean value of 60.050 DM which is 3.3 % higher than the values resulting from only using the actually observed cases. This picture is in principle the same for "Total Income", i.e. tax simulation does not alter our findings. More important for welfare analysis is *equivalent* income, taking into account differential needs for households of different size and structure (see right panel of Table 3). By applying a rather simple equivalence scale given by the square root of household size, we find in principal the same pattern but the deviation between the results for "observed cases" vs. "all cases" is considerably reduced due to the implicit redistribution.

Extending the focus on established measures of inequality (Gini, MLD, SCV, Decile Ratios, and Poverty Rate) the basic finding is that imputation slightly increases inequality and poverty – a clear indication that the underlying imputation procedures are not mean-biased (see Table 4). Figure 3 illustrates this effect by means of kernel density estimates for observed and imputed cases – the overall distribution of the combined population ("All cases") appears to be somewhat flatter than the one based only on observed values.

This finding is complemented in a very illustrative way by Figure 4, which supports the hypothesis that item-non-response is more prominent in the tails of the income

distribution. This is true in terms of the simple incidence but also in terms of relevance of the resulting imputation, measured as the share of post-government income coming from imputed values. A most prominent example is given by the top Decile where we observe for almost 30% of the population at least one missing income component – which after imputation accounts for as much as 14% of the equivalent post-government income of these persons. Overall, the incidence of item-non-response, i.e. the share of the population with at least one missing income component, is about 21% while the relevance of imputation in terms of imputed income mass as a share of total income accounts for as much as 9%. Looking only at the population with at least one imputed income component, the amount of income being imputed as a share of their overall post-government income is as much as 43%.

3.2 Imputation and Income mobility

Longitudinal data is a prerequisite for (income) mobility analyses and is clearly affected by *any* non-response behavior. In Figure 6a we check the probability of unit- as well as item-non-response in t1 (here year 2001) conditional on Post-Government Income quintiles in t0 (here year 2000). Besides the above mentioned U-shaped pattern of item-non-response across the income distribution, it appears that unit-non-response is more prominent in the lowest quintile while item-non-response is found more frequently among high income households. Figure 6b further differentiates the population by imputation status in t0 yielding two most important effects: firstly, there is a highly significant positive correlation of item-non-response over time. Secondly, item non-response on income is a clear indication for future refusal.

Table 5 summarizes results on the impact of imputation on income mobility as given by various well-known mobility indices. Given that all depicted measures yield very comparable results, we exemplary present the results according to the non-directional measure by Fields and Ok (1999): based on “all cases” we find a mobility value for post-government

income over the period 2000 to 2001 of 0.210. However, the group specific values for "observed cases" and "imputed cases" differ considerably with 0.185 and 0.259, respectively.

Conditional on our normative assumptions regarding the imputation techniques applied to SOEP, our results on income *inequality* indicate that an analysis based on observed values only may be biased in terms of income levels and distribution. Income *mobility* analysis provide the indication that persons with rather instable occupational status and resulting volatile income compositions tend to have higher (item-)non-response.

4 Estimating the probability for item-non-response

It has been argued above (section 2.2) that late entrants into a panel survey yet may lack a trustful relationship to the interviewer which *ceteris paribus* yields higher item-non-response on income questions. Figure 5 gives incidence and relevance of item-non-response for the separate sub-samples of SOEP (as given in observation year 2001) also indicating the respective starting year: there appears to be a clear picture that the most recent samples drawn in 1998 and 2000 have an incidence of missing income data which is about twice as high as that for the East German sub-sample C which started in 1990. In terms of relevance this difference is even more pronounced with the new samples incorporating about 11% of imputed income mass, compared to only 4% in sample C. With respect to both indicators samples A "West Germans", B "Foreigners", and D "Immigrants" take an intermediate position.

Obviously, there is need to control in a multivariate approach whether sampling, surveying or individual characteristics are relevant correlates of (item-)non-response affecting the picture in Figure 5. Table 6 presents results from a Random-Effects Probit model estimating the probability of item-non-response for at least one of the income components

included in "Total Income" (standard statistics are given in the right column) over a ten-year period, 1993-2002. By pooling individual observations across time we have the chance to control for otherwise unobserved heterogeneity. Given that the dependent variable as well as almost all RHS-variables are constructed at the level of household, they are constant across all members of a household and therefore not independent from each other. That is why we choose household heads to be representatives of the household they live in and only estimate the model on this reduced population.

With respect to survey characteristics, we control for number of interviews in the panel survey as well as for the interview mode. According to our model, the bivariate results presented in Figure 5 may be mostly driven by the number of given interviews rather than by affiliation to a specific sample. In their first interview the chance for item-non-response is considerably higher than that of the reference group of persons with 5 or more interviews. Although still significant, this effect is already clearly reduced for participants who give their second to forth interview.

Also the interview mode appears to be relevant: we find CAPI[8] interviews as well as self-completed questionnaires to result in significantly higher item-non-response than the reference group where the interviewer filled in the questionnaire.[9]

A major hypothesis with respect to item-non-response is linked to the overall complexity of income and household composition, i.e. the number of income sources and their volatility are assumed to be positively related to item-non-response. In fact, we do find significant positive effects for increasing number of adults and the existence of children to contribute to increasing income complexity. Older household heads who can be assumed to

[8] Computer Assisted Personal Interviewing (CAPI) was introduced to the SOEP in 1998 together with the start of sub-sample E. Meanwhile this mode is also widely used in the older sub-samples A through D as well..

[9] This result for CAPI may be surprising in light of the general argument that this mode improves (income) data quality and reliability. Our finding, however, could result from CAPI being more successful in identifying whether a person receives a certain type of income than self-completers, however, still failing to collect the exact amount of income, i.e., in case of self-completers there would even be no information about the need to impute.

make their living on stable income flows, show a significant lower probability of item-non-response while for younger heads there is a significant but considerably smaller positive correlation. Interpreting "Belonging to the Top Income Decile" and "Owner Occupier" as indicators of increased complexity and diversified portfolio structures we also find the expected positive effects. But this is also true for poor households belonging to the bottom Decile: this finding re-iterates the results of Figure 4, according to which the distribution of item-non-response across the income distribution is U-shaped.

Furthermore, we control for occupational variability at monthly level over the previous calendar year, leaving the assumed stable group of heads with 12 months of fulltime employment as reference group. We compare those to heads who spend the complete 12 months in the same status of being "pensioner", "registered unemployed" or "part-time employed" and those who experienced occupational mobility by spending between one and eleven months in one such status. Here the overall result is that heads who have a rather instable occupational status are significantly more often associated with income related item-non-response than those who were permanently in the very same position.[10] Most interesting, we find permanent pensioners to be even negatively related to non-response compared to those who entered the retirement phase.

Skimming the remaining control variables, household heads in East Germany appear to be more willing to answer to income questions. On the other hand female head, heads without a completed vocational training as well as those with low life satisfaction tend to have higher item non-response. While the underlying mechanism for non-response in case of women[11] and the low educated may result from a "Don't know", it may be more of a general reluctance to cooperate if one is unsatisfied. Finally, our controls for observation years

[10] We cannot rule out, that these results are somewhat driven by *heapening* effects according to which respondents who spend most of the previous year's months in a given status may tend to overstate this by giving 12 months instead.

[11] This effect for women is in line with the findings by Schräpler (2003).

indicate *ceteris paribus* a significant reduction in item-non-response in recent years which may be contributed to improvements in the questionnaire.

Summing up, we find clear indication for the selectivity of item-non-response on income questions for survey related as well as for substantive variables. As such, any income analysis which is based on cases with observed income information is effectively selecting on household size, income level and composition and – even more important – on observations with reduced income variability.

5 Concluding Summary

This study deals with the question of how to cope with item-non-response on income questions in a large panel survey (the German SOEP) and the underlying selectivity process for missing income data.

As in any other survey, incidence (and relevance) of item non-response differ remarkably across income components: e.g. while child benefits exhibit only less than 3% missing data, this share is as high as 20% for interest and dividends.[12] In this paper, we stress the need for imputation due to the underlying selectivity processes of item-non-response on income questions in social surveys. Concluding from multivariate regression analysis, we find an increasing complexity of a given household's income structure and composition, the interview situation and the panel history of a given individual to be strongly correlated with the probability for item non-response.

Imputation is the preferred way to adjust for this in case of annual income figures in the SOEP, with a mix of various imputation routines being applied: The row-and-column imputation procedure as suggested by Little and Su (1989) is used whenever longitudinal information is available, and purely cross-sectional imputation otherwise. Using longitudinal

information in the imputation procedure appears to yield more reliable results in terms of preserving variance than cross-sectional imputation methods do, which is in line with findings by Spieß and Goebel (2003).[13]

Results based on annual income variables for SOEP using various standard inequality indicators show that ignoring cases with item-non-response ("case-wise deletion") tends to underestimate income levels as well as variance, conditional on the imputation methods applied. Additionally, in line with findings in the literature, item-non-response on income appears to be selective with respect to both tails of the income distribution, especially at the upper end. Results from income mobility analyses provide evidence for a positive correlation between item- and subsequent unit-non-response.

Given the normative decisions involved in the imputation process there are two central propositions to providers of survey data from a "user's point of view": first, imputed values in micro data must be flagged in order to allow to differentiate those from truly observed information and secondly, all imputation procedures must be documented in a comprehensible way.[14]

Future research in this area will have to consider the extension towards multiple imputation techniques which may also help to cope with partial unit-non-response, i.e. non-responding individuals in otherwise responding households, which are most likely to yield an underestimate of the true income aggregate at the household level.[15]

---

[12] From a survey methodological point of view it is important for panel research that more recent entrants into a longitudinal population are less likely to answer to income questions.

[13] It should be noted that the single imputation techniques currently applied to the SOEP income information probably underestimate the true variance and as such there may be demand for more complex variance estimation methods (e.g., jackknife estimators).

[14] In case of the annual income figures included in the SOEP data distribution, there is one imputation flag for each income aggregate which gives the percentage share of income that has been imputed.

[15] This phenomenon may not appear relevant if *all* income information is collected at the household level or – like in case of the PSID – by means of proxy interviews from one respondent per household. However, this approach has its limitations in relying completely on the willingness to cooperate and the information available to the responding person.

References

Biewen, M. (2001): Item non-response and inequality measurement: Evidence from the German earnings distribution, Allgemeines Statistisches Archiv 85, 409-425.

Binder, D.A. (1996): Comment. Journal of the American Statistical Association, Vol. 91, No. 434, 510-512.

Buck, N.; Nicolleti, C.; McCulloch, A. and Burton, J. (2003): Report on attrition analysis and item non-response. CHINTEX Deliverable No. 6, o.O.: Mimeo.

Burkhauser, R.V., B. Butrica, M.C. Daly, and D.R. Lillard, "The Cross-National Equivalent File: A product of cross-national research," in I. Becker, N. Ott, and G. Rolf, eds., Soziale Sicherung in einer dynamischen Gesellschaft. Festschrift für Richard Hauser zum 65. Geburtstag, Campus, Frankfurt/New York, 354-376, 2001.

Butrica, B. A. 1997: Imputation methods for filling in missing values in the PSID-GSOEP Equivalent File 1980-1994, Cross-National Studies in Aging. Program Project Paper, Center for Policy Research, The Maxwell School. Syracuse, NY: Syracuse University

Cao, H. (2001) IMPUTE: A SAS Application System for Missing Value Imputations With Special Reference to HRS Income/Assets Imputations. HRS/AHEAD Documentation Report, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, July 2001. (http://hrsonline.isr.umich.edu/docs/userg/dr-007.pdf accessed September 23, 2003).

Fields, G. S. and E. Ok (1999), Measuring movement of incomes, Economica, 66(264), 455–471.

Grabka, M. M. and J. R. Frick (2003), Imputation of Item-Non-Response on Income Questions in the SOEP 1984–2002, DIW Research Notes No. 29, DIW Berlin.

Hanisch, J. and U. Rendtel (2002): Quality of income data from panel surveys with respect to rounding. Chintex working paper Nr. 6.

Haisken-DeNew, J. P. and J. R. Frick (Eds.) (2003): DTC – Desktop Companion to the German Socio Economic Panel Study (SOEP), Updated to Wave 19 (S), DIW Berlin.

Hill, D. and R. J. Willis (2001): Reducing Panel Attrition: A Search for Effective Policy Instruments, Journal of Human Resources 36(3), 416-438.

Little, R.J.A. and Rubin D.B. (1987) Statistical Analysis with Missing Data, John Wiley & Sons: New York.

Little, R.J.A. and Su, H.-L. (1989): Item Non-Response in Panel Surveys. In: Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. P. (eds.): Panel Surveys. John Wiley, New York: 400-425.

Loosveldt, G., J. Pickery, and J. Billiet, 1999, Item non-response as a predictor of unit non-response in a panel survey, Paper presented at the International Conference on Survey Non-response, Portland Oregon, USA.

Nicoletti, C. and Peracci, F. (2003): Imputation procedures and the quality of income information in the ECHP. Working Paper.

Rendtel, U. (1995): Panelausfälle und Panelrepräsentativität. Campus Verlag, Frankfurt/Main - New York.

Riphahn, R. T. and O. Serfling (2002): Item Non-Response on Income and Wealth Questions. IZA Working paper nr. 573.

Riphahn, R. T. and Oliver Serfling (2003): Heterogeneity in Item Non-Response on Income and Wealth Questions., In: Schmollers Jahrbuch, Jg. 123, Vol. 1, p. 95-107.

Rubin, D.B. (1976). Inference with missing data. Biometrika, 63, 581-592.

Rubin, D.B. (1987): Multiple Imputation for Non-Response in Surveys. John Wiley & Sons, New York

Schafer, J. et al. (1993): Multiple Imputation of Missing Data in NHANES III, Proceedings of the 1993 Annual Research Conference, United States Bureau of the Census, Washington, DC.

Scheffer, J. (2002): Dealing with Missing Data, Res. Lett. Inf. Math. Sci. (2002) 3, 153-160.

Schräpler J.-P. (2002): Respondent Behavior in Panel Studies - A Case Study for Income-Nonresponse by means of the German Socio-Economic Panel (GSOEP). In DIW-Discussion Paper, No. 299.

Schräpler, J.-P. (2003): Gross Income Non-response in the German Socio-Economic Panel - Refusal or Don't Know? In: Schmollers Jahrbuch, Jg. 123, Vol. 1, S. 109-124.

Schupp, J. and G. G. Wagner (2002): Maintenance of and Innovation in Long-term Panel Studies: The Case of the German Socio-Economic Panel (GSOEP). In: Allgemeines Statistisches Archiv, Vol. 86(2), pp 163-175.

Schwarze, J., Simulating German Income and Social Security Tax Payments Using the GSOEP, Syracuse University, Syracuse, NY: Cross-National Studies in Ageing Project Paper No. 19., 1995.

SOEP Group (2001): "The German Socio-Economic Panel (SOEP) after more than 15 years – Overview," Vierteljahrshefte zur Wirtschaftsforschung, 70(1), 7-14.

Schräpler, J.-P. and G. G. Wagner (2001): Das Verhalten von Interviewern - Darstellung und ausgewählte Analysen am Beispiel des "Interviewerpanels" des Sozio-ökonomischen Panels, *Allgemeines Statistisches Archiv* 85, 45-66.

Spieß, M. and J. Goebel (2003): A comparision of different imputation strategies with respect to income related questions. Paper presented on the Chintex final conference "Harmonisation of Surveys and Data Quality", Wiesbaden, 26 and 27 May 2003.

Taylor, M. F. et al, eds., British Household Panel Survey User Manual. Introduction, Technical Reports and Appendices, Colchester: University of Essex, ESRC, 1998.

Table 1: Incidence of item-non-response for selected SOEP income components and applied imputation technique

		1986 Sample A-B	1993 Sample A-C	2001 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
	• Little&Su	4.1	5.1	5.4
	• X-Section	0.9	0.7	3.8
X-Mas Gratification	Observed cases	95.9	96.4	93.7
	Imputed cases	4.1	3.6	6.3
	• Little&Su	3.2	3.0	2.6
	• X-Section	0.9	0.6	3.7
Income from self-employment	Observed cases	82.1	85.6	74.3
	Imputed cases	17.9	14.4	25.7
	• Little&Su	11.5	9.5	12.2
	• X-Section	6.4	4.9	13.5
Pension Income (own)	Observed cases	91.2	94.3	97.2
	Imputed cases	8.8	5.7	2.7
	• Little&Su	4.6	4.0	0.2
	• X-Section	4.2	1.7	2.5
Interest & Dividends	Observed cases	81.0	87.7	86.1
	Imputed cases	19.0	12.3	13.9
	• Little&Su	7.9	5.6	2.4
	• X-Section	11.1	6.7	11.5
Child Benefit	Observed cases	99.2	97.9	95.6
	Imputed cases	0.8	2.1	4.4
	• Little&Su	0.7	1.9	2.7
	• X-Section	0.1	0.2	1.7

Note: "Little & Su" results are based on the "row-and-column procedure" as described in Little & Su (1989); "X-Section" gives results for those observations which are imputed by cross-sectional techniques, only.
Source: SOEP, Survey years 1986, 1993, 2001; unweighted results.

Table 2: 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	2001 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
• Little&Su	6.2	8.1	6.6	7.4	5.2
• X-Section	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.

Table 3: Selected income aggregates by imputation status, 2001

	Imputation Status	Post-Gov't Income	Total Income	Post-Gov't Income (equivalized)	Total Income (equivalized)
Mean	All cases	60.050	82.761	36.760	50.515
	• Observed cases	58.040	79.827	36.039	49.388
	• Imputed cases	66.612	92.339	39.113	54.196
	Deviation "All" vs. "Observed"(%)	3,5	3,7	2,0	2,3
Median	All cases	54.678	71.848	33.334	44.167
	• Observed cases	52.931	69.322	32.797	43.201
	• Imputed cases	61.308	82.205	35.327	47.933
	Deviation "All" vs. "Observed"(%)	3,3	3,6	1,6	2,2
Stddev	All cases	59.930	96.266	33.644	56.383
	• Observed cases	57.209	92.239	32.535	55.062
	• Imputed cases	67.597	107.886	37.073	60.465
	Deviation "All" vs. "Observed"(%)	4,8	4,4	3,4	2,4
N	All cases	29.306			
	• Observed cases	22.937			
	• Imputed cases	6.369			
	Deviation "All" vs. "Observed"(%)	27,8			

Note:
"All cases" is based on both, observed and imputed observations. "Observed cases" is based only on observations without any item-non-response for any of the income components considered in the respective income aggregate. "Imputed cases" is based on observations with at least one imputed income component.
Source: SOEP, Survey year 2001 (samples A-F), weighted results.

Table 4: Imputation and Income Inequality, 2001

	Imputation Status			
	"All cases"	"Observed cases"	"Imputed cases"	Deviation: "All" vs. "Observed" (%)
MLD	0,1350	0,1281	0,1550	+5,4
Gini	0,2698	0,2634	0,2858	+2,4
SCV	0,2977	0,2961	0,2958	+0,5
90:10	3.44	3.30	3.82	+4.2
90:50	1.77	1.73	1.81	+2.3
50:10	1.94	1.91	2.10	+1.6
Poverty Rate (PL=60% Median)	14.3	14.2	14.7	+0.7

Note:
Basis of these calculations is equivalent annual post-government income. "All cases" is based on both, observed and imputed observations. "Observed cases" is based only on observations without any item-non-response for any of the income components considered in the respective income aggregate. "Imputed cases" is based on observations with at least one imputed income component.
Source: SOEP, Survey year 2001 (samples A-F), weighted results.

Table 5: Imputation and Income Mobility, 2001

	Imputation Status			
	"All cases"	"Observed cases"	"Imputed cases"	Deviation: "All" vs. "Observed" (%)
Quintile Matrix Mobility				
→ Average jump	0,467	0,413	0,584	+13,1
→ Normalized avg. jump	0,187	0,165	0,234	+13,3
Fields & Ok (1996):				
→ `Percentage` income mobility	18,36	15,99	22,88	+14,8
Fields & Ok (1999)				
→ Non-directional	0.210	0.185	0.259	+13,5
Shorrocks (1987)				
→ Using Gini	0,03504	0,02974	0,04333	+17,8

Note:
Basis of these calculations is equivalent annual post-government income. "All cases" is based on both, observed and imputed observations. "Observed cases" is based only on observations without any item-non-response for any of the income components considered in the respective income aggregate. "Imputed cases" is based on observations with at least one imputed income component.
Source: SOEP, Survey year 2001 (samples A-F), weighted results.

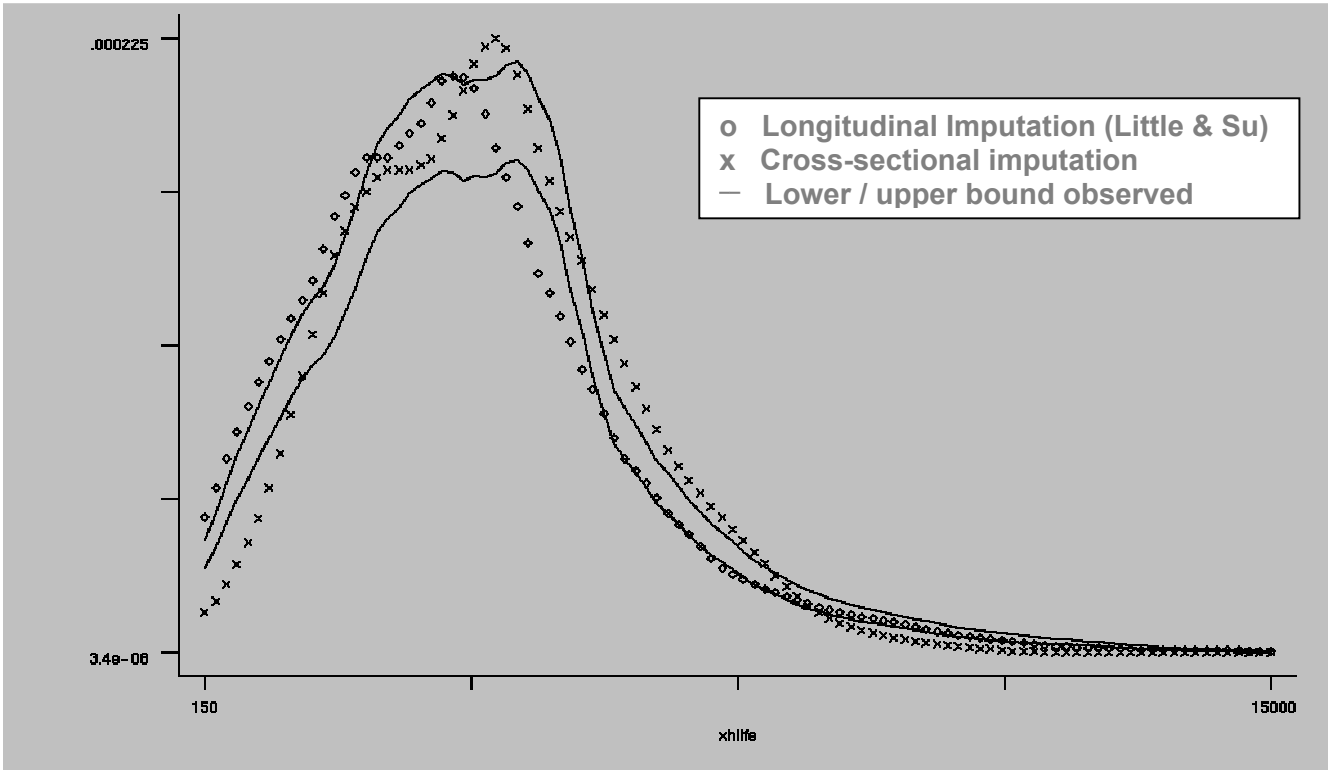
Table 6: Estimating the Probability of Item-Non-Response in "Total Income", 1993-2002 (Results from a RE-Probit Model)

Variable	Coefficient	(Std.Error)	Mean (Stdev)
# Interviews = 1	.4241**	*(.0266)*	.0905 (.2869)
# Interviews = 2-4	.2521**	*(.0190)*	.1933 (.2435)
Interview Mode : Self-completion	.2656**	*(.0169)*	.3289 (.4698)
Interview Mode : CAPI	.3421**	*(.0223)*	.1302 (.3365)
Age of Head < 25 years	.0672*	*(.0323)*	.0511 (.2203)
Age of Head > 64 years	–.2897**	*(.0313)*	.1960 (.3970)
# Adults in Household = 2	.1554**	*(.0205)*	.5492 (.4975)
# Adults in Household = 3 and more	.4114**	*(.0256)*	.1855 (.3887)
Children in Household (<17 years)	.1611**	*(.0181)*	.3234 (.4677)
Head: Female	.1054**	*(.0201)*	.3771 (.4846)
Head: Foreign citizenship	.0414	*(.0300)*	.1134 (.3171)
Head: University / FHS-Degree	–.0411+	*(.0226)*	.1801 (.3843)
Head: No vocational training	.1339**	*(.0225)*	.1825 (.3862)
Head: 1-11 months fulltime employed	.2063**	*(.0239)*	.0915 (.2884)
Head: 1-11 months reg. unemployed	.2387**	*(.0273)*	.0729 (.2600)
Head: 12 months reg. unemployed	.1888**	*(.0358)*	.0335 (.1801)
Head: 1-11 months part-time employment	.3178**	*(.0305)*	.0373 (.1895)
Head: 12 months part-time unemployment	.1198**	*(.0309)*	.0521 (.2222)
Head: 1-11 months pensioner	.3244**	*(.0434)*	.0168 (.1287)
Head: 12 months pensioner	–.0575*	*(.0276)*	.2680 (.4429)
East Germany	–.1767**	*(.0224)*	.2573 (.4372)
Community Size < 2.000 inhabitants	–.0341	*(.0292)*	.0944 (.2924)
Total Income in bottom Decile	.1525**	*(.0253)*	.0778 (.2679)
Total Income in top Decile	.1718**	*(.0254)*	.0907 (.2872)
Owner occupier	.1198**	*(.0180)*	.3957 (.4890)
Head: Low life satisfaction	.1174**	*(.0284)*	.0512 (.2205)
Observation Year=1994	.0001	*(.0295)*	.0788 (.2695)
Observation Year=1995	–.0641*	*(.0296)*	.0811 (.2730)
Observation Year=1996	.0425	*(.0292)*	.0803 (.2718)
Observation Year=1997	.1013**	*(.0291)*	.0795 (.2705)
Observation Year=1998	.0528+	*(.0290)*	.0893 (.2852)
Observation Year=1999	.0435	*(.0290)*	.0862 (.2806)
Observation Year=2000	–.1467**	*(.0289)*	.1543 (.3613)
Observation Year=2001	–.2042**	*(.0278)*	.1394 (.3463)
Observation Year=2002	–.1353**	*(.0278)*	.1338 (.3404)
Constant	–1.598**	*(.0343)*	–
Mean (Dep. Variable)	–	–	.1910 (.3931)

Note: Only one observation per household (n=85103 / 17021 individuals); Pseudo R^2 = 0.109.
-2 Log Likelihood: -41504.073; -2 Log Likelihood (Full Model): -36980.113 ; LR chi2(35) = 2465.65;
**: p<.001; *: p<.05; +: p<0.1
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.

Figure 1: Kernel Density Estimates for "Individual Labor income from first job": Comparing exemplary results from alternative imputation techniques with observed values

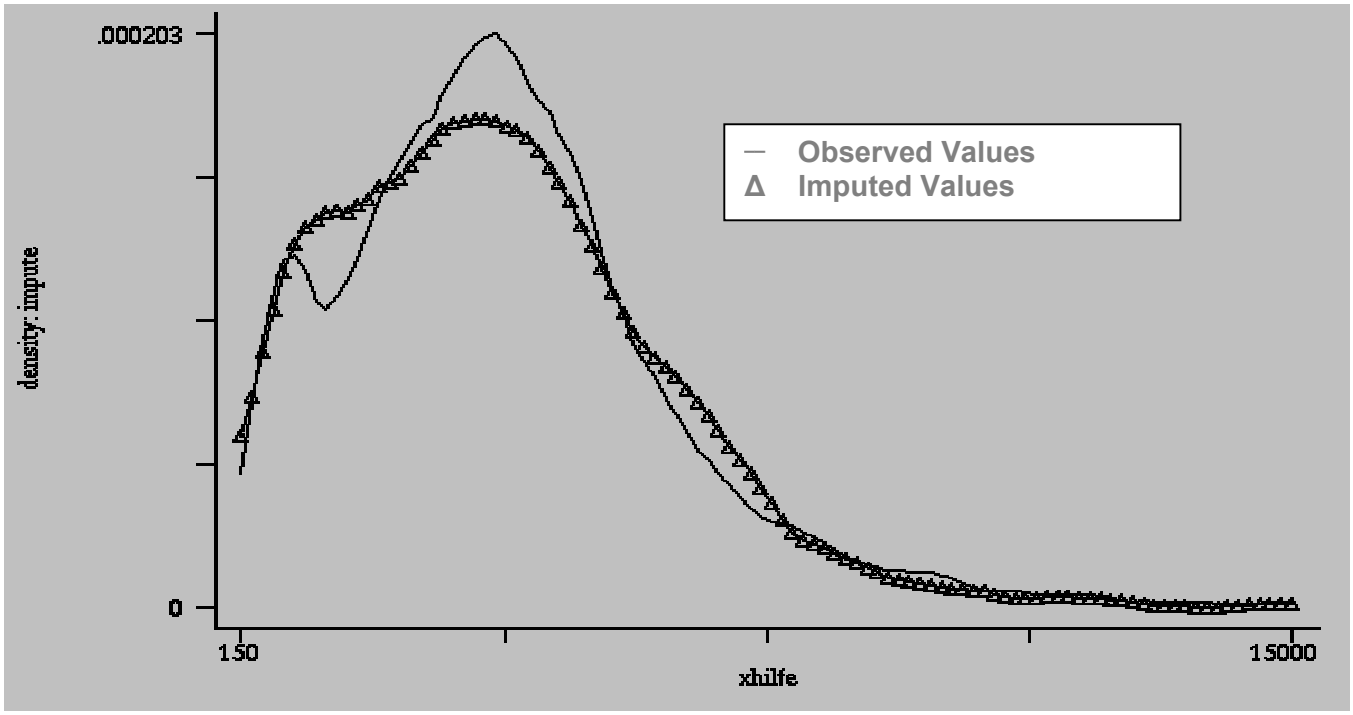


	Observed	Little & Su	X-Section
Mean	4 286	3 867	4 257
Median	4 000	3 501	4 180
Stdev	2 510	2 342	1 857
Dec. 90:10	5,14	5,81	3,63
Gini	0.3019	0.3284	0.2485
N	1,086	1,086	1,086

Note: Calculations are based on a random sample of 1086 observations (prgroup = 4) for which a positive value has been observed and who provide longitudinal information as a prerequisite for the Little & Su procedure.

Source: SOEP, Survey year 2001 (samples A-F), weighted results.

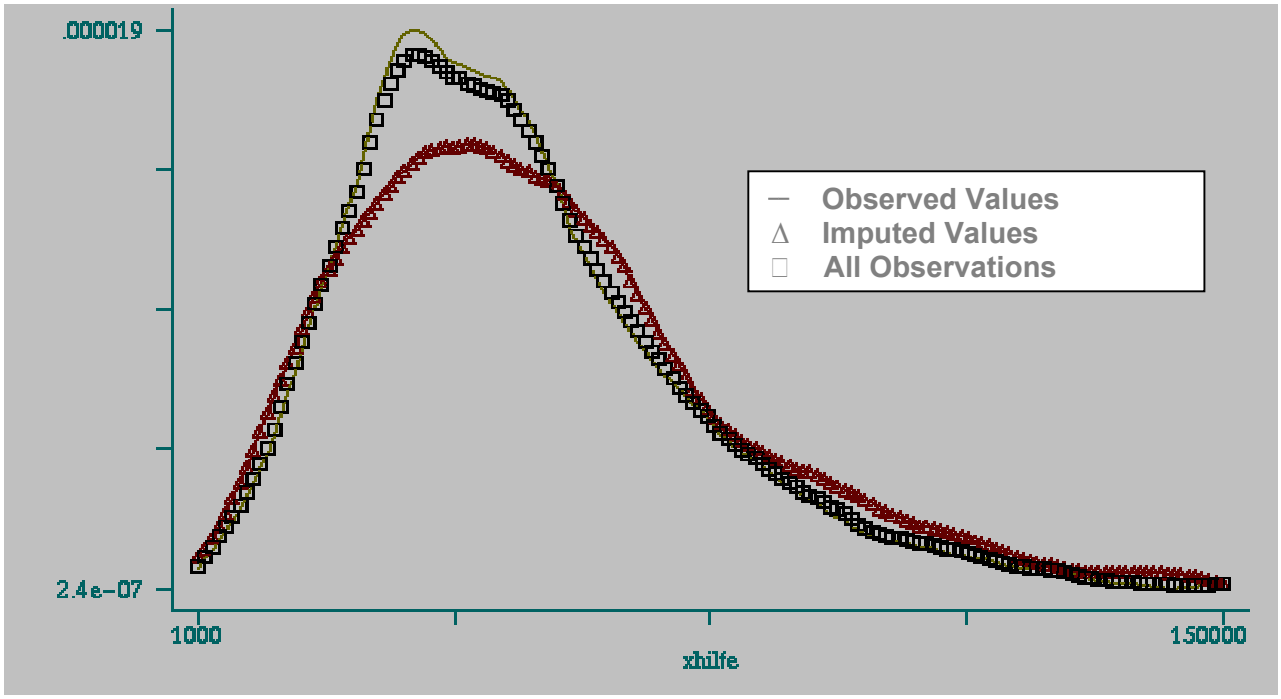
Figure 2: Kernel Density Estimates for "Individual Labor income from first job": Imputed and actually observed values



	Observed	Imputed
Mean	4 074	3 873
Median	3 800	3 602
Stdev	2 621	2 423
Dec. 90:10	6,36	7,17
Gini	0.3295	0.3360
N	11,148	1,124

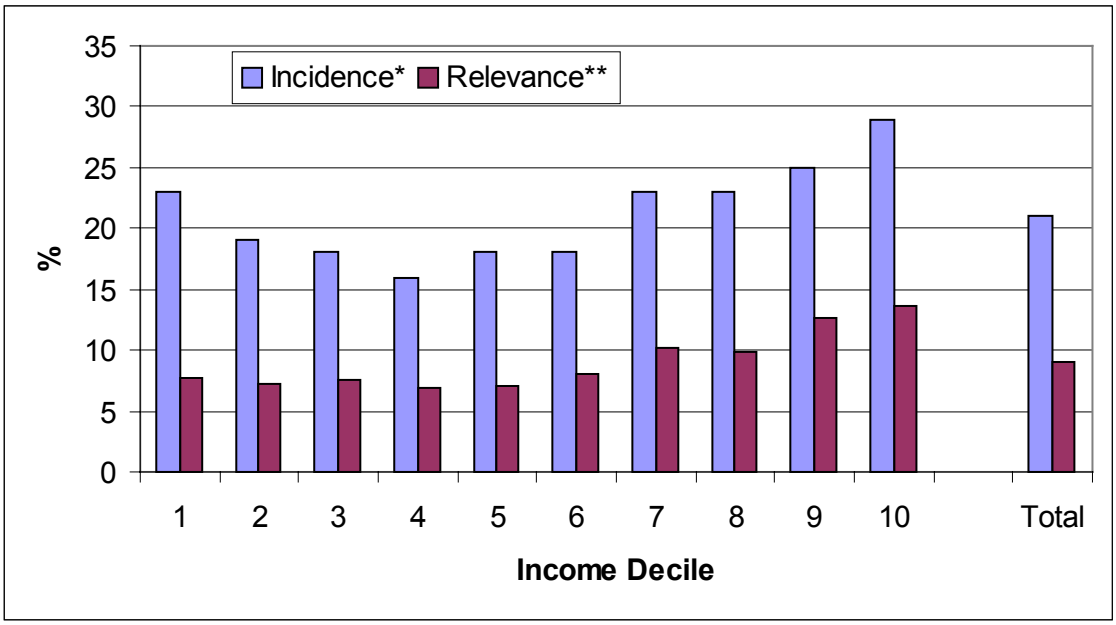
Source: SOEP, Survey year 2001 (samples A-F), weighted results.

Figure 3: Kernel Density Estimates for equivalent "Post-Government Income" by imputation status



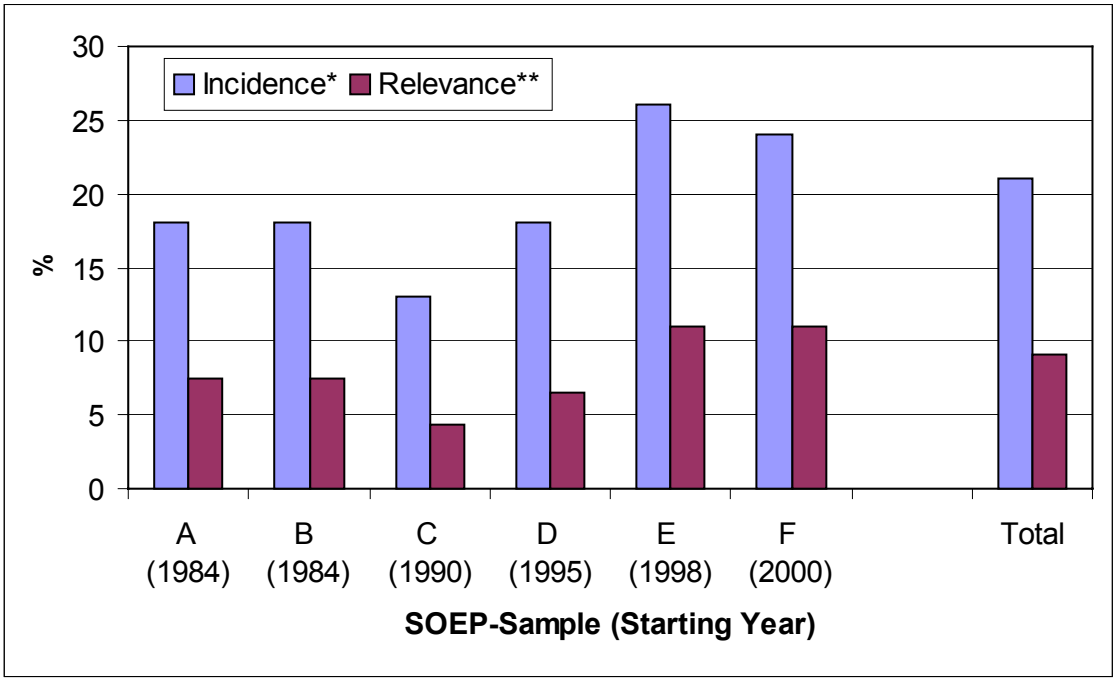
Source: SOEP, Survey year 2001 (samples A-F), weighted results.

Figure 4: Incidence and Relevance of Item-Non-Response in equivalent "Post Government Income" by Income Deciles



Note: * 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". Equivalence Scale = Square Root of Household Size.
Source: SOEP, Survey year 2001 (samples A-F), weighted results.

Figure 5: Incidence and Relevance of Item-Non-Response in equivalent "Post Government Income" by SOEP-Sample



Note: * 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". Equivalence Scale = Square Root of Household Size.
Source: SOEP, Survey year 2001 (samples A-F), weighted results.

Figure 6: Income Mobility and Imputation: The case of "Post-Government Income" 2000-2001

Figure 6a:

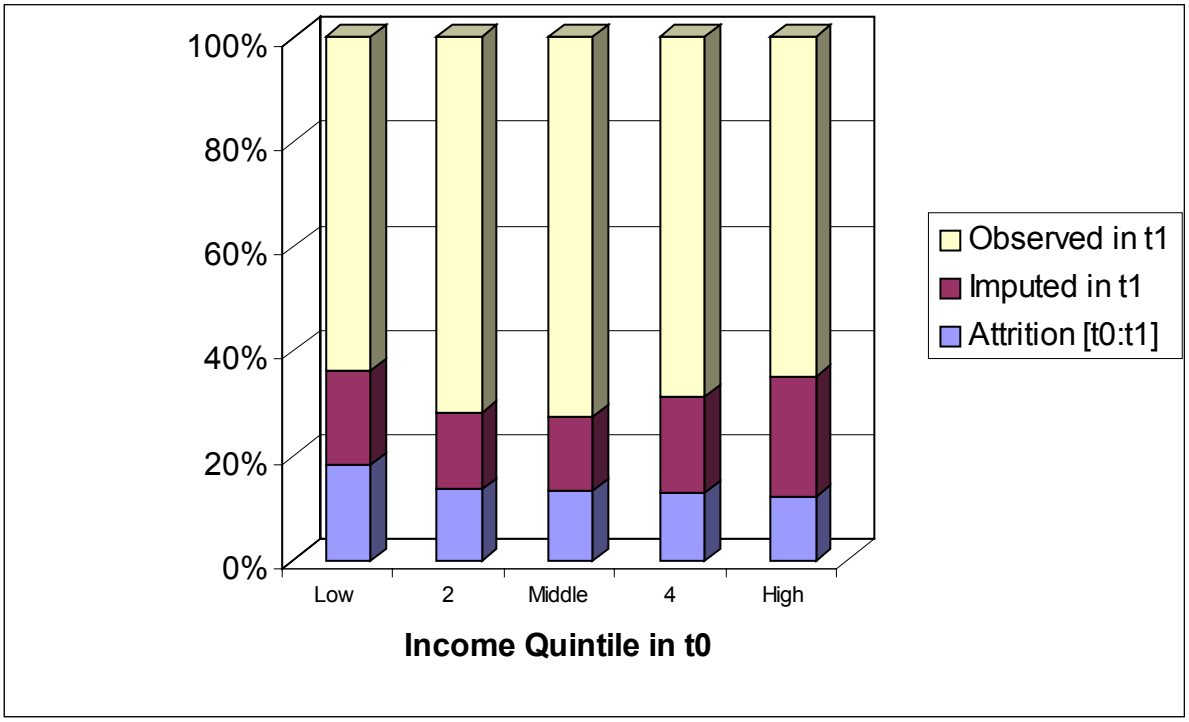
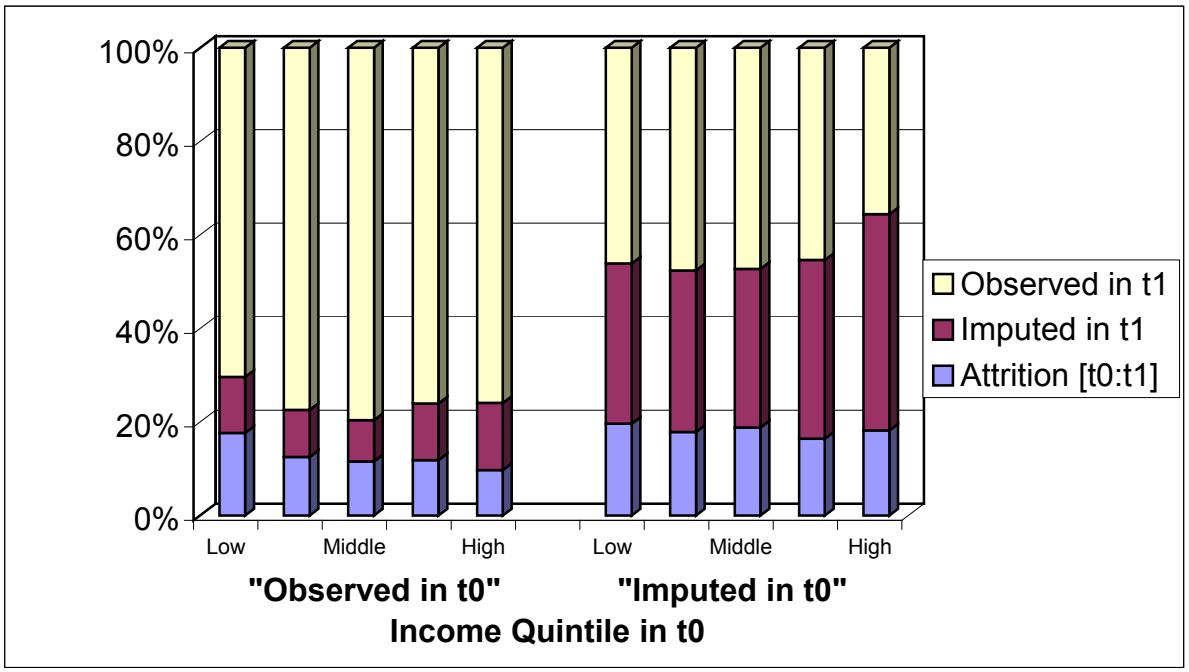


Figure 6b:



Source: SOEP, Survey years 2000-2001 (samples A-E), weighted results.

APPENDIX :

Aggregated SOEP annual income variables, the underlying original survey information from SOEP income components, and the respective type of imputation in case of missing values due to item-non-response[16]

Income Aggregate	**Input (=original survey information with aggregation level)**	**primary imputation technique[1)]**	**secondary imputation technique**
I11101$$ (Household Pre-Government Income)	sum (I11103$$ + I11104$$ + I11106$$ + I11117$$)	see respective input variables below	see respective input variables below
I11103$$ (Household labour income) [10 inputs]	• Aggregated Household Labour Income • first job • second job • self-employment • 13th monthly payments • 14th monthly payments • Christmas bonuses • vacation/holiday pay • profit sharing, premiums • other bonuses • military service pay	 L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S	 R M-G R M-S M-S M-S M-S M-S M-S M-S M
I11104$$ (Household asset income) [3 inputs]	• income from rent and lease • minus operating & maintenance costs • interest & dividends	L & S L & S L & S	M-G M-G R
I11106$$ (Household private transfers) [2 inputs]	• Aggregated Household private transfers • individual private transfers • alimony	 L & S L & S	 M M
I11117$$ (Household private pensions) [6 inputs]	• Aggregated Household private pensions • own company retirement plan • own pension for public employees • other own pension • widow company retirement plan • widow pension for public employees • other widow pension	 L & S L & S L & S L & S L & S L & S	 R R R R R R

1) L&S = Little & Su 1989
M = Median Substitution
M-G = Median Substitution by Subgroups
M-S = Median share Substitution
R = Regression based imputation
Fixed = Fixed amounts

[16] These income aggregates are available in the Cross-National Equivalent File (CNEF) together with comparably harmonized data for Canada (SLID), the UK (BHPS), and the USA (PSID). For a detailed description see Burkhauser et.al. (2001) and http://www.human.cornell.edu/pam/gsoep/equivfil.cfm

Income Aggregate	Input (=original survey information with aggregation level)	primary imputation technique[1)]	secondary imputation technique
I11102$$ (Household Post-Government Income)	= sum (I11101$$ + I11107$$ + I11108$$ - I11109$$)	see respective input variables below	see respective input variables below
I11107$$ (Household total public transfers) [6 inputs]	• Aggregated Household private pensions <u>Individual level:</u> • higher education grants • maternity benefits • unemployment benefit • unemployment assistance • subsistence allowances • early retirement benefit	 L & S L & S L & S L & S L & S L & S	 M-G M M-S M-S M-S M-S
[6 inputs]	<u>Household level:</u> • housing benefit for renters • housing benefits for owner-occupiers • child benefit • social assistance • special help income • nursing care insurance • direct housing support for owners	 L & S L & S L & S L & S L & S L & S L & S	 R M-G Fixed M-G M-G Fixed Fixed
I11108$$ (Household Social Security Pensions) [12 inputs]	• Aggregated Household social security Pensions • own GRV pension • own minors pension • own civil servants pension • own war victims pension • own farmers pension • own accident insurance pension • widow GRV pension • widow minors pension • widow civil servants pension • widow war victims pension • widow farmers pension • widow accident insurance pension	 L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S L & S	 R R R R R R R R R R R R R
I11109$$ (Household federal Taxes and SSC)	complete imputation based on a micro-simulation programme	completely simulated	
I11105$$ (Imputed rental value)	complete imputation based on a hedonic regression estimation	completely simulated	
I11118$$ (Windfall income)	revenues from inheritances, lotteries, etc. (> 5000 DM / 2.500 EUR)	M	