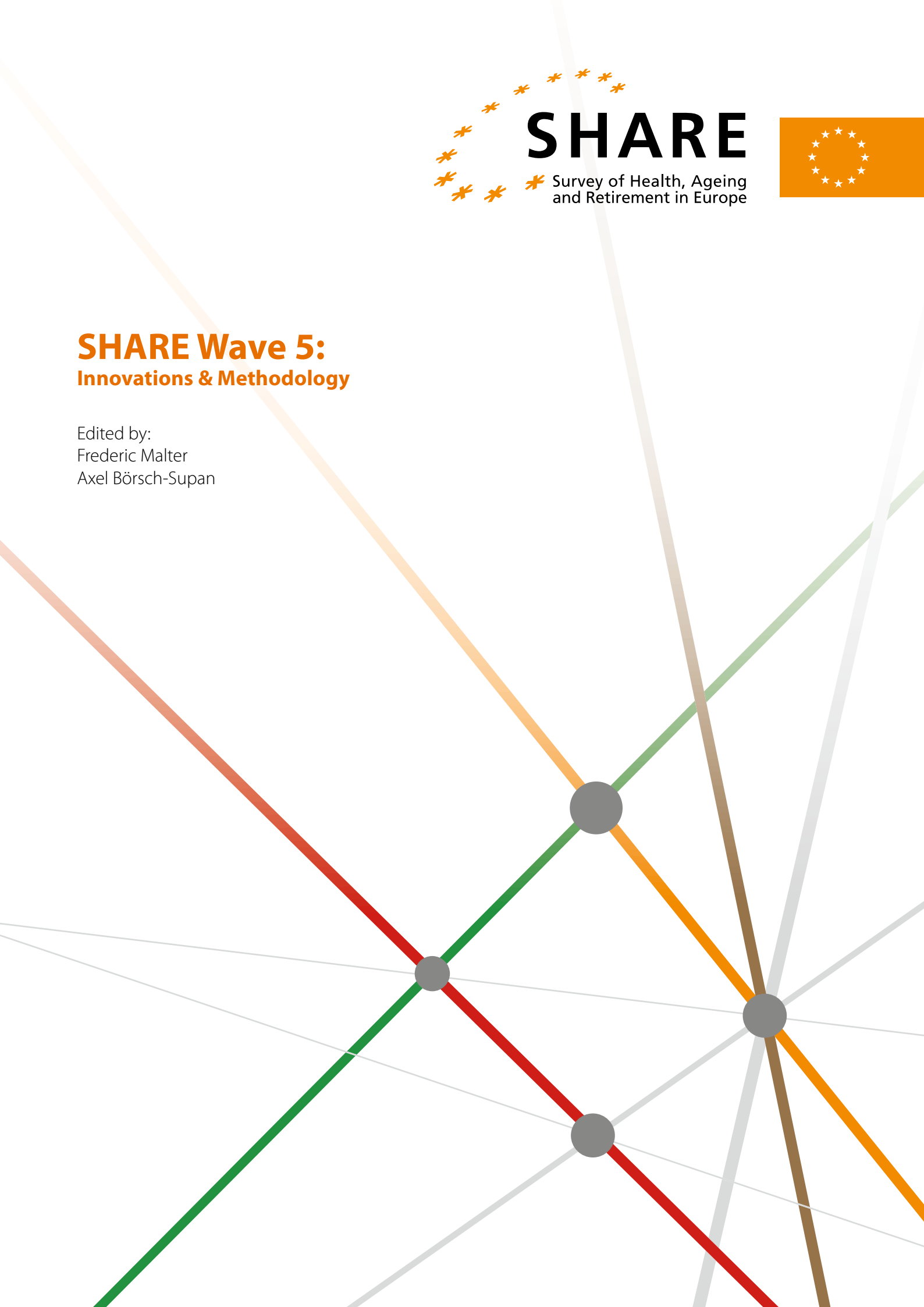




## **SHARE Wave 5: Innovations & Methodology**

Edited by:  
Frederic Malter  
Axel Börsch-Supan





# **SHARE Wave 5:**

## **Innovations & Methodology**

## SHARE Wave 5: Innovations & Methodology

Edited by:  
Frederic Malter  
Axel Börsch-Supan

### **Authors:**

Mauricio Avendano  
Axel Börsch-Supan  
Johanna Bristle  
Martina Celidoni  
Enrica Croda  
Dominika Duda  
Sabine Friedel  
Christian Hunkler  
Hendrik Jürges  
Thorsten Kneip  
Julie Korbmacher  
Ulrich Krieger  
Anne Laferrère  
Giuseppe De Luca  
Frederic Malter  
Maurice Martens  
Michał Myck  
Monika Oczkowska  
Claudio Rosetti  
Gregor Sand  
Daniel Schmidutz  
Morten Schuth  
Elisabetta Trevisan  
Melanie Wagner  
Iggy van der Wielen  
Arnaud Wijnant

Published by:

Munich Center for the Economics of Ageing (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)

Amalienstrasse 33

80799 München

Tel: +49-89-38602-0

Fax: +49-621-38602-390

[www.mea.mpisoc.mpg.de](http://www.mea.mpisoc.mpg.de)

Layout and printing by:

VALENTUM KOMMUNIKATION GMBH

Bischof-von-Henle-Str. 2b

93051 Regensburg

© Munich Center for the Economics of Ageing, 2015

Suggested citation:

**Malter, F. and A. Börsch-Supan (Eds.) (2015). *SHARE Wave 5: Innovations & Methodology*. Munich: MEA, Max Planck Institute for Social Law and Social Policy.**

ISBN 978-3-00-049309-6

# CONTENTS

<b>1</b>	<b>SHARE Wave 5: Balancing innovation and panel consistency</b>	<b>8</b>
	Axel Börsch-Supan and Frederic Malter, Munich Center for the Economics of Aging MEA at the Max Planck Institute for Social Law and Social Policy (MPISOC)	
<b>2</b>	<b>Questionnaire innovations in the fifth wave of SHARE</b>	<b>15</b>
	<b>2.1 Questionnaire development in the fifth wave of SHARE</b>	<b>16</b>
	Frederic Malter, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	
	<b>2.2 Measuring early childhood circumstances in SHARE Wave 5: A “mini childhood” module</b>	<b>18</b>
	Mauricio Avendano, London School of Economics and Political Science & Harvard School of Public Health Enrica Croda, Department of Economics, Ca’ Foscari University of Venice	
	<b>2.3 Innovations for better understanding deprivation and social exclusion</b>	<b>29</b>
	Michał Myck, Monika Oczkowska and Dominika Duda, Centre for Economic Analysis (CenEA), Szczecin	
	<b>2.4 Health care utilization and out-of-pocket expenses</b>	<b>37</b>
	Hendrik Jürges, University of Wuppertal	
	<b>2.5 Identifying second-generation migrants and naturalized respondents in SHARE</b>	<b>43</b>
	Christian Hunkler, Thorsten Kneip, Gregor Sand and Morten Schuth, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	
	<b>2.6 SHARE questionnaire encyclopaedia (or “question-by-question manual” or “Q-by-Q”)</b>	<b>48</b>
	Anne Laferrère, National Institute for Statistics and Economic Studies (INSEE), Paris Frederic Malter, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	
<b>3</b>	<b>Software innovations in SHARE Wave 5</b>	<b>51</b>
	Maurice Martens, Iggy van der Wielen, Arnaud Wijnant, CentERdata, Tilburg University Gregor Sand, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	

<b>4</b>	<b>A note on record linkage in SHARE</b>	60
	Julie M. Korbmacher, Daniel Schmidutz, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	
<b>5</b>	<b>Interviewing interviewers: The SHARE interviewer survey</b>	67
	Julie M. Korbmacher, Sabine Friedel, Melanie Wagner, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC) Ulrich Krieger, University of Mannheim	
<b>6</b>	<b>Sample design and weighting strategies in SHARE Wave 5</b>	75
	Giuseppe De Luca, University of Palermo Claudio Rossetti, LUISS Guido Carli Frederic Malter, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	
<b>7</b>	<b>Item nonresponse and imputation strategies in SHARE Wave 5</b>	85
	Giuseppe De Luca, University of Palermo Martina Celidoni, University of Padua Elisabetta Trevisan, University of Padua & Netspar	
<b>8</b>	<b>Fieldwork monitoring and survey participation in fifth wave of SHARE</b>	101
	Thorsten Kneip, Frederic Malter, Gregor Sand, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	
<b>9</b>	<b>Access to SHARE data and citation rules</b>	158
	Daniel Schmidutz, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	
<b>10</b>	<b>Measuring interview length with keystroke data</b>	165
	Johanna Bristle, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC)	

## 7 Item nonresponse and imputation strategies in SHARE Wave 5

*Giuseppe De Luca, University of Palermo*

*Martina Celidoni, University of Padua*

*Elisabetta Trevisan, University of Padua & Netspar*

### 7.1 Introduction

Nonresponse is a serious problem that affects most empirical studies based on survey data. A distinction is usually made between two types of nonresponse. The first – unit nonresponse – occurs when eligible sample units fail to participate in a survey because of noncontact or explicit refusal to cooperate (see Chapter 8). The second – item nonresponse – emerges when responding units do not provide useful answers to particular items of the questionnaire as it is often the case with income, wealth and consumption expenditure items. The potential implications of the two types of nonresponse are similar, namely selectivity bias and loss of precision. The key difference is that for unit nonresponse all items of the questionnaire are missing, while for item nonresponse missing observations are confined to specific items of the questionnaire. Such distinction has therefore relevant implications for the auxiliary information that can be used in ex-post adjustment procedures. For unit nonresponse, the auxiliary information is necessarily confined to that obtained from the sampling frame or the data collection process (in SHARE, that's age, gender and regional NUTS1 indicators), whereas for item nonresponse the additional information collected during the entire interview process can be used.

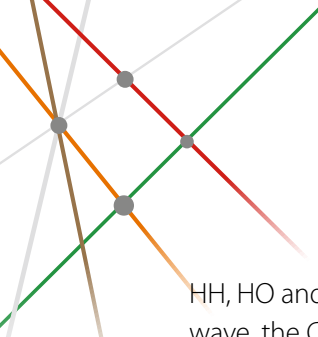
This chapter focuses on item nonresponse in the fifth wave of SHARE and the imputation strategies adopted to fill-in the missing values. The main features of the SHARE interviews and the prevalence of missing data are briefly discussed in Sections 7.2 and 7.3, respectively. In Section 7.4, we describe the strategies adopted to handle some practical issues faced in the construction of the imputation database. A non-technical description of the imputation procedure used in Wave 5 is given in Section 7.5. Except for minor differences in the underlying raw data, this procedure is very close to that used for Release 1.1 of Wave 4 data (publicly available since March 2013). Both procedures present however some important innovations with respect to the imputation strategies exploited for Release 2.4 of Wave 1 and Wave 2 data (publicly available since March 2011, see Christelis, 2011). Harmonized imputations for all waves of SHARE are planned to be delivered in the near future.

### 7.2 Features of the SHARE interview in Wave 5

The way the data are collected and the complexity of the questionnaire are known to be key determinants of non-sampling errors such as unit and item nonresponse and measurement errors. The data collection mode adopted in SHARE is the Computer Assisted Personal Interview (CAPI).

To reduce the burden of the interview process, some modules were asked to only one person per household. The so-called family respondent answered questions about children and help received (CH module and part of the SP module). Questions about financial items, total household income, incomes of other non eligible household members, housing, and household consumption expenditures (FT, AS,





HH, HO and CO modules) were instead answered by the so-called financial respondent. Since the second wave, the CAPI questionnaire also included skip-patterns for time-invariant variables of respondents who have already participated to previous waves. For these respondents, relevant time-invariant variables were directly preloaded in the interview instrument using the information provided in the previous waves.

Two additional dimensions of the complexity of the interview process were question wording and time reference period. Due to the nature of the topics investigated by SHARE, the wording of some questions was necessarily sensitive. Examples include some questions about physical health (“In which organ or part of the body have you had a cancer?”), mental health (“In the last month, have you felt that you would rather be dead?”), or economic issues (“Thinking of your household’s total monthly income, would you say that your household is able to make ends meet...”). Despite the sensitive wording, the fraction of missing values on this type of closed-ended questions was generally low. Large amounts of missing data occurred instead for monetary variables such as incomes, assets, and consumption expenditures which were collected through retrospective and open-ended questions that were sensitive and difficult to answer precisely.

The time reference period of monetary variables varied considerably depending on the question being asked. Questions about employment incomes and financial transfers refer to the last calendar year, questions about consumption expenditures refer to a typical month, and questions about assets refer to the current situation at the time of the interview. For questions about pensions, regular transfers, rent payments, and repayments of loans and mortgages, the period covered by a typical payment was asked after asking for the average amount of the last payments.

In case of initial nonresponse to open-ended questions for monetary variables, the respondent was asked a sequence of unfolding-bracket (UB) questions aimed to recover partial information on the missing monetary amount. Specifically, the respondent was asked whether the amount was larger than, smaller than, or about equal to three predefined thresholds defined at the country level. The threshold in the first UB question was assigned randomly and the sequence of UB questions either stops or continues with the next threshold depending on the answer given to the previous questions. The information collected through the sequence of UB questions can be an approximate point estimate (i.e. about equal to one of the three thresholds) or an interval estimate. The sequence of UB questions was uninformative only if the respondent did not give a substantial answer (i.e. neither ‘Refuse’ nor ‘Don’t know’) to the first question of the sequence.

### **7.3 Prevalence of missing data**

As in the previous waves, most of the variables collected in the fifth wave of SHARE were only affected by small amounts of missing data (usually lower than 5%). Non-negligible amounts of missing data occurred instead for monetary variables about incomes, assets and consumption expenditures. Figure 7.1 shows the cross-country distribution of the item nonresponse rates for six monetary variables that are generally affected by a large amount of missing data: annual income from employment (EP205), regular payments from public old age pensions (EP078\_1), value of the house (HO024), expenditure on food consumed at home (CO002), amount hold in bank accounts (AS003) and liabilities (AS055). For this set of variables, the cross-country average of item nonresponse ranges between a minimum of 9 percent for regular payments from public old age pensions to a maximum of 36 percent for amount hold in bank accounts. However, item nonresponse seems

to be country-specific: Denmark and Sweden, for example, show low percentages of missing data for most of the variables considered (usually lower than 10%). In contrast, Spain, Slovenia, Luxemburg and Israel exhibit item nonresponse rates that are considerably higher than the average. There, item nonresponse becomes particularly worrisome for some wealth components with more than 60 percent of the data missing.

Although questionnaire design and sample management system are standardized across countries in order to ensure an ex-ante harmonization of the national data, this between-country variability in item nonresponse may reflect the impact of other cross-country differences in fieldwork procedures (e.g. reputation and quality of the national survey agencies, experience, education and training of the interviewers) as well as differences in the composition of the national samples and the compliance behavior of the national target populations towards the survey requests.

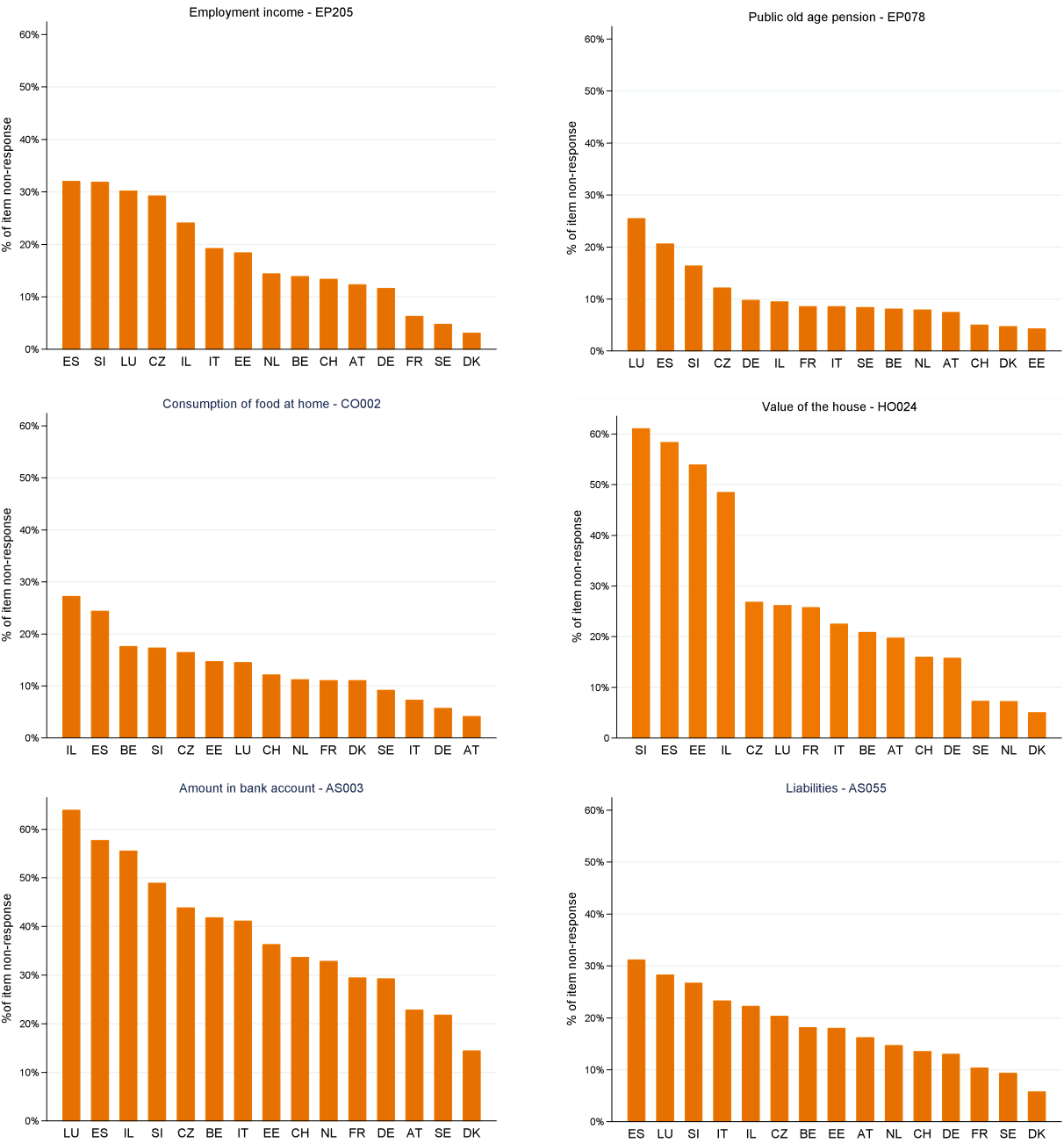


Figure 7.1: Percentage of missing values for some monetary variables by country



## 7.4 Practical decisions about imputations

Handling item nonresponse in a cross-national, multi-disciplinary and longitudinal survey like SHARE is a challenging task that involves many different decisions that have to be balanced against each other. In this section we therefore describe the key steps that were necessary to construct the imputation model. Since many of the practical issues addressed here unavoidably affect the outcomes of this model, we found it important to inform data users of the rational driving the construction of the SHARE public-use imputation dataset.

### *Dimensionality of imputation model*

Due to the large number of variables collected in SHARE Wave 5, the first issue was how to select a feasible subset of core variables that accommodates a wide variety of analyses that data users might want to perform. Preliminary choices regarding the dimensionality of the imputation model are particularly important in the context of multivariate imputation procedures that attempt to preserve the correlation structure of the imputed variables. Unlike univariate imputation procedures, these methods require that multiple variables are imputed simultaneously on the basis of some Markov Chain Monte Carlo (MCMC) technique. The problem is that as the number of variables to be imputed jointly increases, these iterative techniques often require significant effort in programming and fine tuning. A compromise between generality and complexity of the imputation model was therefore needed. Our strategy to deal with this problem was as follows. First, we selected a rather large number of variables expected to be relevant for the key purposes of the survey. Second, to simplify complexity of the imputation model, our multivariate imputation procedure was employed only for a smaller subset of variables with relevant fractions of missing data (see Section 7.5). Furthermore, this procedure was restricted to aggregated subsets of income, wealth and consumption expenditure items only.

### *Data standardization*

After selecting a set of core variables to be imputed, we constructed for each of them a binary eligibility indicator which identified those respondents eligible to answer that specific question by taking into account possible inconsistencies in the raw data, country-specific deviations from the generic version of the CAPI questionnaire, branching, skip patterns and proxy interviews. For open-ended questions on monetary variables, which are usually preceded by one or more ownership questions, we also constructed a set of binary ownership indicators to identify a subset of eligible respondents with a non-zero monetary amount. Conditional on eligibility and ownership, non-zero values of monetary variables were converted (if needed) in annual Euro amounts to avoid differences in the time reference period of each question and the national currencies of non-Euro countries.

### *Outliers*

We symmetrically trimmed two percent of complete cases from the country-specific distribution of annual Euro amounts to exclude outliers that may have a disproportional influence on survey statistics. This implies that, in addition to non-substantial answers (“Don’t know” and “Refusal”), we also imputed outliers in the tails of the distribution of each monetary variable.

### *Logical constraint*

Complete cases and imputed values were required to satisfy a set of logical constraints on ownership of the variables included into the imputation model which helped to avoid unreasonable combinations of the imputed data. For example, the ownership indicators of some financial assets (bonds, stocks and mutual funds) are set to zero (no ownership) if it is known that the household does not own a bank account.

### *Preserving the partial information from sequences of UB questions*

Another useful source of information to reduce uncertainty on missing values of monetary variables is given by the sequence of answering UB questions. Table 7.1 and Table 7.2 show that, in several cases, this survey instrument allows recovering helpful information for more than 50 percent of the initial missing data. As mentioned before, the information derived from UB questions can be of two types: approximate point estimates (1) or interval estimates (2). In the first case, missing amounts are directly imputed using the thresholds selected by the respondents throughout the sequence of UB questions. In the second case, UB interval estimates are combined with the additional information from logical constraints and percentiles of the country distribution to shrink as much as possible the bounds placed on missing data.

**Table 7.1: Fraction of point estimates provided by the sequences of UB questions as percent of initially missing data**

Country	Income from employment	Public old age pension	Expenditure on food consumed at home	Value of the house	Amount in bank account	Liabilities
Austria	0.21	0.22	0.41	0.33	0.17	0.08
Germany	0.10	0.17	0.36	0.3	0.17	0.09
Sweden	0.21	0.32	0.37	0.13	0.13	0.14
Netherlands	0.19	0.29	0.35	0.21	0.12	0.03
Spain	0.18	0.24	0.37	0.33	0.17	0.16
Italy	0.28	0.19	0.39	0.35	0.25	0.21
France	0.13	0.26	0.43	0.30	0.19	0.14
Denmark	0.14	0.23	0.5	0.14	0.12	0.08
Switzerland	0.19	0.22	0.39	0.24	0.22	0.25
Belgium	0.19	0.17	0.36	0.35	0.12	0.1
Israel	0.08	0.11	0.19	0.09	0.1	0.11
Czech Republic	0.24	0.25	0.37	0.36	0.19	0.15
Luxembourg	0.10	0.06	0.17	0.08	0.11	0.04
Slovenia	0.22	0.12	0.42	0.23	0.14	0.12
Estonia	0.18	0.49	0.35	0.26	0.22	0.19
Total	0.18	0.24	0.37	0.27	0.17	0.13

**Table 7.2: Fraction of interval estimates provided by the sequences of UB questions as percent of initially missing data**

Country	Income from employment	Public old age pension	Expenditure on food consumed at home	Value of the house	Amount in bank account	Liabilities
Austria	0.43	0.32	0.27	0.46	0.37	0.47
Germany	0.50	0.47	0.31	0.40	0.34	0.44
Sweden	0.45	0.36	0.31	0.39	0.33	0.33
Netherlands	0.41	0.39	0.25	0.38	0.27	0.30
Spain	0.24	0.23	0.18	0.29	0.20	0.36
Italy	0.25	0.28	0.23	0.31	0.24	0.27
France	0.55	0.43	0.35	0.46	0.45	0.61
Denmark	0.33	0.23	0.25	0.45	0.27	0.29
Switzerland	0.45	0.23	0.26	0.37	0.29	0.31
Belgium	0.44	0.32	0.38	0.39	0.37	0.49
Israel	0.34	0.31	0.24	0.52	0.29	0.33
Czech Republic	0.35	0.27	0.32	0.29	0.24	0.35
Luxembourg	0.44	0.50	0.53	0.64	0.32	0.31
Slovenia	0.32	0.43	0.19	0.26	0.18	0.20
Estonia	0.43	0.23	0.37	0.34	0.33	0.38
Total	0.39	0.32	0.29	0.38	0.30	0.38

### *Aggregation*

After exploiting the information available for each item, we reduced the number of monetary variables that had to be imputed jointly by aggregating 55 items on income, wealth and consumption expenditure into 17 aggregated variables. Each aggregated variable is obtained by summing two or more original items as illustrated in Table 7.3. Notice that the choice of aggregating such long list of income, wealth and expenditure items into a considerably smaller subset of key variables was considered a reasonable strategy to reduce the computational complexity of the imputation model. However, the use of aggregated variables is not a panacea. This simplification has both theoretical and practical implications. From a theoretical viewpoint, aggregation corresponds to imposing linear restrictions on the imputation model and this may undermine validity of the analyses that users can perform on the basis of imputed data (see, for example, Rubin, 1996). From a practical viewpoint, the SHARE public-use data only contain imputations for the chosen set of aggregated variables, but not for their particular components. In addition, special attention was needed to deal with country-specific deviations from the generic version of the CAPI questionnaire and the preservation of the partial information available for missing aggregated values. The last issue was particularly important because, when aggregating several items, it was often the case that only some of them were missing. Moreover, logical constraints and sequences of UB questions may provide interval information on the missing observations of each item. Thus, even if aggregated variables are regarded as missing, the available information for the single components can be used to define bounds for missing aggregated values.

**Table 7.3: Aggregate variables in Wave 5**

Aggregate variables	Components	Variable name
Regular payments from public old age, early retirement, survivor and war pensions	Public old age pension	EP078_1
	Public old age supplementary pension	EP078_2
	Public early retirement pension	EP078_3
	Main public survivor pension	EP078_7
	Secondary public survivor pension	EP078_8
	Public war pension	EP078_9
Regular payments from private occupational pensions	Occupational old age pension from last job	EP078_11
	Occupational old age pension from second job	EP078_12
	Occupational old age pension from third job	EP078_13
	Occupational early retirement pension	EP078_14
	Occupational disability or invalidity insurance	EP078_15
	Occupational survivor pension	EP078_16
Regular payments from disability pensions and benefits	Main public disability insurance pension	EP078_4
	Secondary public disability insurance pension	EP078_5
Regular payments of other private pensions	Regular life insurance payments	EP094_1
	Regular private annuity or personal pension payments	EP094_2
	Long-term care payments from private insurance	EP094_5
Regular payments from private transfers	Alimony	EP094_3
	Regular payment from charities	EP094_4
Lump-sum payments from public old age, early retirement, survivor and war pensions	Lump-sum payments from public old age pension	EP082_1
	Lump-sum payments from public old age supplementary pension	EP082_2
	Lump-sum payments from public early retirement pension	EP082_3
	Lump-sum payments from main public survivor pension	EP082_7
	Lump-sum payments from secondary public survivor pension	EP082_8
	Lump-sum payments from public war pension	EP082_9
Lump-sum payments from private occupational pensions	Lump-sum payments from occupational old age pension from last job	EP082_11
	Lump-sum payments from occupational old age pension from second job	EP082_12
	Lump-sum payments from occupational old age pension from third job	EP082_13
	Lump-sum payments from occupational early retirement pension	EP082_14
	Lump-sum payments from occupational disability or invalidity insurance	EP082_15
	Lump-sum payments from occupational survivor pension	EP082_16
Lump-sum payments from disability pensions and benefits	Lump-sum payments from main public disability insurance pension	EP082_4
	Lump-sum payments from secondary public disability insurance pension	EP082_5

**Table 7.3: Aggregate variables in Wave 5 (continued)**

Aggregate variables	Components	Variable name
Lump-sum payments of other private pensions	Lump-sum payments from life insurance	EP209_1
	Lump-sum payments from private annuity or personal pension	EP209_2
	Lump-sum payments from long-term care private insurance	EP209_5
Lump-sum payments from private transfers	Lump-sum payments from alimony	EP209_3
	Lump-sum payments from charities	EP209_4
Rent and home-related expenditures	Amount rent paid	HO005
	Other home-related expenditures	HO008
Income from rent or sublet	Income from sublet	HO074
	Income from rent of real estate	HO030
Income from other household members	Other household members' net income	HH002
	Other household members' net income from other sources	HH011
Bond, stock and mutual funds	Government/corporate bonds	AS007
	Stocks	AS011
	Mutual funds	AS017
Savings in long term investments	Individual retirement accounts from respondent	AS021
	Individual retirement accounts from partner	AS024
	Contractual savings	AS027
	Whole life insurance holdings	AS030
Paid out-of-pocket for outpatient care	Paid out-of-pocket for doctor visits	HC083
	Paid out-of-pocket for dental care	HC093
Paid out-of-pocket for nursing home and home-based care	Paid out-of-pocket for home-based care	HC129
	Paid out-of-pocket for nursing home	HC097

## 7.5 The imputation procedure used in SHARE

The imputation procedure used in Wave 4 and Wave 5 exhibited some important innovations with respect to the procedure adopted in Wave 1 and Wave 2. Two differences were particularly striking. First, as discussed in the previous section, some items were now imputed in aggregate terms to simplify the computational burden of the imputation model. For similar reasons, separate imputations for longitudinal and refreshment subsamples were no longer considered and lagged variables from previous waves were not used as predetermined predictors any more. The second important difference is that we handle the problem of non-responding partners (NRPs) differently, namely the fact that only one of the two partners may have agreed to be interviewed. Unlike the strategy adopted in the first two waves, NRPs are now viewed as a problem of unit nonresponse (not item nonresponse) due to the limited information available to cope with this type of nonresponse error. Our imputation procedure provides only an indirect estimate of the income from NRPs to avoid understating total household income when only one of the two partners was interviewed. As discussed at length at the end of this section, the strategy used to recover this information exploits the distinction between couples with and without NRPs and additional information obtained from a one shot question on monthly household income (HH017).

Similarly to the previous procedure, variables of Wave 5 were imputed by univariate or multivariate methods depending on the prevalence of missing values. Simple univariate methods, such as hot-deck and regression imputations, were used when the fraction of missing values was lower than 5 percent for the entire sample and lower than 10 percent at the country level. Variables with fractions of missing values above these thresholds were instead imputed jointly by the fully conditional specification (FCS) method (van Buuren et al., 1999; Raghunathan et al., 2001), an iterative imputation procedure. More precisely the FCS method imputes multiple variables iteratively via a sequence of univariate imputation models, one for each imputation variable, using as predictors all variables except the one being imputed. Despite a lack of rigorous theoretical justification (see, for example, Arnold et al., 1999, 2001; van Buuren et al. 2006; van Buuren, 2007), the FCS method is one of the most popular multivariate imputation procedures used in practice due to its flexibility in handling complicated data structures. Recent comparisons of the FCS method with other multivariate imputation methods can be found in Lee and Carlin (2010) and references therein.

### ***Univariate imputations***

This set of imputations was performed in an early stage separately by country. We first imputed socio-demographic characteristics such as age and education that were affected by a small fraction of missing values so that these variables could then be used as exogenously observed predictors in the imputation of the other variables. Our set of predictors for hot-deck imputations typically included gender, age group, years of education and self-reported health. For some variables additional predictors were also used. For example, we also employed the number of children when imputing the number of grandchildren and an indicator for being a patient in a hospital overnight during the last year when imputing health-related variables. Variables that were known to be logically related, such as respondent's weight, height and body mass index, were imputed simultaneously by hot-deck.

### ***Multivariate imputations***

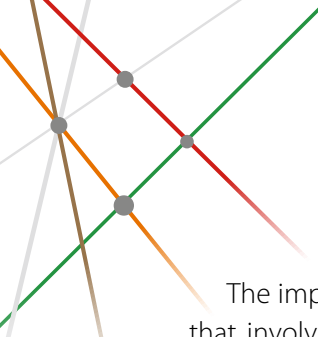
FCS imputations were performed separately by country and household type to allow for heterogeneity across these different groups. The household types considered were singles and third respondents<sup>1</sup> (sample 1), couples with both partners interviewed (sample 2), and all couples – with and without NRPs (sample 3). Notice that sample 2 is embedded into sample 3. This overlapping partitioning of the sample was introduced to estimate total household income in couples with NRPs. The basic idea was that we could first impute total household income of couples belonging to sample 2. In sample 3, couples with both partners interviewed could then be used as valid observations to impute total household income of couples with NRPs. Before providing additional details on this aspect of the new imputation procedure, we discuss other important features of the FCS method.

The set of variables imputed jointly by the FCS method was country- and sample-specific, but it usually consisted of monetary variables only. In addition to the above criterion, we also required that the sample used in the estimation step of the FCS method includes at least 100 donor observations in sample 1 and 150 donor observations in samples 2 and 3. Monetary variables that did not satisfy this additional requirement were imputed first and then used as observed predictors in the imputation of the other variables.

---

<sup>1</sup> Third respondents are singles living with a couple, e.g. parents or relatives. Usually, these are respondents who entered in the sample at the time of Wave 1, when all household members over 50 years were interviewed





The imputation of each monetary variable was always carried out on the basis of a two-part model that involved a probit model for ownership and a regression model for the amount conditional on ownership. To account for skewness in the right tails of these distributions, strictly positive variables were transformed in logarithms. Instead, variables that may also take negative values, such as income from self-employment, bank account, and value of own business, were transformed using the inverse hyperbolic sine transformation. The set of exogenous predictors was also sample-specific. For singles and third respondents, it included gender, age, years of education, self-perceived health, number of children, number of chronic diseases, score of the numeracy test, employment status and willingness to answer. For couples with both partners interviewed, we used a larger set of predictors that also included the mentioned variables for the partner of the designated respondent. For couples with NRPs, the predictors referring to the NRPs were confined to age and years of education only. In few cases where the number of observations available for the estimation step was lower than 30, missing values were imputed on the basis of a smaller subset of predictors (gender, age, years of education and self-reported health only). Imputed monetary values were always constrained to fall within the individual-level bounds that incorporated the partial information available on missing observations. As discussed in Section 7.4, these bounds summarized the information obtained from percentiles of the country distribution, logical constraints on ownership and amount, sequences of UB questions, and the partly observed items of aggregate variables in an explicit and applicable form.

For monetary variables imputed jointly by the FCS method, the sequence of univariate imputations was performed in a similar fashion. The main difference was that, in addition to the above set of exogenous predictors, the prediction equation of each item included imputed values of all monetary variables except the one being imputed. Furthermore, the imputation process was repeated several times until the iterative algorithm reached a stationary distribution<sup>2</sup>. The set of monetary variables was excluded only in the first iteration in order to initialize the starting values of the algorithm.

Particular attention was devoted to the imputation of total household income because SHARE provides two alternative measures of this variable. The first measure (“thinc”) could be obtained by a suitable aggregation at the household level of all individual income components<sup>3</sup>, while the second (“thinc2”) could be obtained from the one-shot question on monthly household income (HH017). The choice between these two alternative measures is not obvious. On the one hand, there is evidence that asking about an exhaustive list of disaggregated income components may lead to a more accurate measure of total household income than asking about a single one-shot question (see, for example, Browning et al. 2003 for a related issue in the context of consumption expenditure questions). According to this viewpoint, thinc could be preferred to thinc2. On the other hand, however, the aggregation of a larger number of income components usually leads to a considerably larger amount of missing data. In addition, the aggregated measure of total household income could be underestimated because of the NRPs problem. Based on these considerations, we believe that none of the two measures of total household income could be strictly preferred to the other and thus we let the users decide which of the two measures was more suitable for their research questions. Moreover, the availability of these two alternative measures may greatly improve the imputation process because each measure could contribute relevant information on the missing values of the other measure. Our procedure to impute these two measures of total household income consisted of three stages.

---

<sup>2</sup> As discussed in Christelis (2011), convergence of the algorithm is assessed by the Gelman-Rubin criterion (Gelman and Rubin 1992; Gelman et al. 2004) applied to the mean, the median and the 90th percentile of the five imputed distributions of each monetary variable. Convergence is also assessed for generated variables such as total household income (thinc), total household expenditure (thexp) and household net worth (hnetw). After an initial set of 7 burn-in iterations, this criterion suggests that convergence is usually achieved for most of the statistics considered before reaching the pre-specified maximum number of 30 iterations.

<sup>3</sup> This is the measure of total household income that is comparable with that provided in the imputation datasets of Wave 1 and Wave 2.

- **Stage 1 (singles and 3<sup>rd</sup> respondent).** We imputed all monetary variables by the FCS method discussed before. At the end of each iteration, we also computed total household income (thinc), household net worth (hnetw) and total household expenditure (thexp) by suitable aggregations of the imputed income, wealth and expenditure items. We finally imputed the second version of total household income (thinc2) using total household income (thinc), household net worth (hnetw), total household expenditure (thexp), and characteristics of the household respondent as predictors. The imputed values of thinc2 were constrained to fall in the bounds derived from the sequence of UB questions for HH017.
- **Stage 2 (couples with both partners interviewed).** We used an imputation strategy similar to that adopted in stage 1, but with a larger set of predictors that also includes characteristics of the partner of the designed respondent.
- **Stage 3 (all couples – with and without NRPs).** Imputed values of all variables for the subsample of couples with both partners interviewed were obtained from stage 2. In stage 3, these couples entered the imputation sample only as observations available for the imputation of missing values on the other subsample of couples with NRPs. Similarly to the previous stages, we first imputed all monetary variables for the responding partners by standard implementation of the FCS method. Unlike stage 2, the predictors referring to the NRPs now consisted however of age and years of education only. At the end of each iteration, we also imputed total household income (thinc2) using household net worth (hnetw), total household expenditure (thexp) and characteristics of the responding partner as predictors and bound information derived from the sequence of UB questions for HH017. For all couples with NRPs, we finally imputed the total household income (thinc) using the second version of total household income (thinc2), household net worth (hnetw), total household expenditure (thexp) and characteristics of the responding partner as predictors, couples with two partners interviewed as observations available for the estimation step, and the imputed sum of incomes of the responding partner as lower bound.

To allow data users to take into account the additional variability generated by the imputation process, we provide five imputations of the missing values. These multiple imputations were constructed through five independent replicates of imputation procedure discussed above. Notice that neglecting this additional source of uncertainty by selecting only one of the five available replicates may lead to misleadingly precise estimates. The list of variables included in the SHARE public-use imputation dataset of Wave 5 is presented in Table 7.4. For each imputed variable we also provide a flag variable (named as `variablename_f`) which summarizes the status of the imputation process as illustrated in Table 7.5.

To conclude, we would like to point out that imputations are not the same as missing variable values. Although the use of imputed data is a quite common empirical strategy for handling missing data problems, validity of the underlying assumptions should not be taken for granted. Validity of the so-called fill-in approach (i.e. the simple approach of fill-in the missing values with imputations) is indeed based on two important conditions. The first is that the model used to create the imputations is correctly specified, including the assumptions on the assumed missing-data mechanism. The second is that the imputation model is congenial in the sense of Meng (1994), i.e. the imputation model cannot be more restrictive than the model used to analyze the filled-in data. Uncongeniality may occur, for

instance, when the model of interest and the imputation model are based either on different parametric assumptions or on different sets of explanatory variables. When these two conditions hold, the use of imputed data protects data users from potential nonresponse bias and loss of precision. However, the fill-in approach may also lead to biased estimates whenever the imputation model is either incorrectly specified or uncongenial (see, for example, Dardononi et al., 2011, 2014). Judgements on the validity of these assumptions in the context of concrete research questions remain a researcher's duty. To our experience, comparing the outcomes from different approaches for problems of item nonresponse (such as complete data analysis, simple and generalized missing indicator approaches, and sample selection models) may give important hints on the robustness of findings.

**Table 7.4: List of variables included in the imputation dataset of Wave 5**

Variable name	Description	Questionnaire
mergeid	Person ID	
implicat	Implicat number	
hhidcom5	Household ID Wave 5	
cvid	Wave specific person identifier	
cvidp	Wave specific person identifier of spouse/partner	
country	Country identifier	
language	Language of questionnaire	
htype	Household type	
fam_resp	Family respondent	
fin_resp	Financial respondent	
hou_resp	Household respondent	
excrate	Exchange rate	
nursinghome	Living in nursing home	MN024
hysize	Household size	
single	Single	
couple	Couple	
partner	Partner in the couple	
p_nrp	Partner of non responding partner	
sample1	Imputation sample for single	
sample2	Imputation sample for couples with two partners interviewed	
sample3	Imputation sample for all couples	
ydip	Earnings from employment	EP205
yind	Earnings from self-employment	EP207
ypen1	Annual old age, early retirement pensions, survivor and war pension	EP078_1-2-3-7-8-9
ypen2	Annual private occupational pensions	EP078_11-16
ypen3	Annual disability pension and benefits	EP078_4-5
ypen4	Annual unemployment benefits and insurance	EP078_6

**Table 7.4: List of variables included in the imputation dataset of Wave 5 (continued)**

Variable name	Description	Questionnaire
yopen5	Annual payment from social assistance	EP078_10
yreg1	Other regular payments from private pensions	EP094_1-2-5
yreg2	Other regular payment from private transfer	EP094_3-4
ylsum1	Lump sum payments for old age, early retirement, survivor and war pension	EP082_1-2-3-7-8-9
ylsum2	Lump sum payments for private occupational pension	EP082_11-16
ylsum3	Lump sum payments for disability pension and benefits	EP082_4-5
ylsum4	Lump sum payments for unemployment benefits and insurance	EP082_6
yslum5	Lump sum payments for social assistance	EP082_10
yslum6	Lump sum payments for other private pension	EP209_1-2-5
yslum7	Lump sum payments for other private transfer	EP209_3-4
rhre	Annual rent and home-related expenditures	HO005, HO008
home	Value of main residence	HO024
mort	Mortgage on main residence	HO015
ores	Value of other real estate – Amount	HO027
ysrent	Annual income from rent or sublet	HO074, HO030
yaohm	Annual income from other household members	HO002, HO011
fahc	Annual food at home consumption	CO002
fohc	Annual food outside home consumption	CO003
hprc	Annual home produced consumption	CO011
bacc	Bank accounts	AS003
bsmf	Bond, stock and mutual funds	AS007, AS011, AS017
slti	Savings for long-term investments	AS021, AS023, AS27, AS030
vbus	Value of own business	AS042
sbus	Share of own business	AS044
car	Value of cars	AS051
liab	Financial liabilities	AS055
yibacc	Interest income from bank accounts	
yibsmf	Interest income from bond, stock and mutual funds	
thinc	Total household net income - version A	
thinc2	Total household net income - version B	HH017
thexp	Total household expenditure (sum of rhre, fahc, fohc and hprc)	
hrass	Household real assets (home*perho/100+vbus*sbus/100+car+ores - mor)	
hgfass	Household gross financial assets (sum of bacc, bsmf and slti)	
hnfass	Household net financial assets (hgfass - liab)	
hnetw	Household net worth	

**Table 7.4: List of variables included in the imputation dataset of Wave 5 (continued)**

Variable name	Description	Questionnaire
gender	Gender	DN042
age	Age in 2010	DN003
age_p	Age of partner in 2010	DN003
yeduc	Year of education	DN041
yeduc_p	Year of education of partner	EX102
sphus	Self-perceived health - US scale	PH003
mstat	Marital status	DN014
nchild	Number of children	CH001
ngcchild	Number of grandchildren	CH201
gali	Limitation with activities	PH005
chronic	Number of chronic diseases	PH006
symptoms	Number of symptoms	PH010
bmi	Body mass index	PH012, PH013
weight	Weight	PH012
height	Height	PH013
mobility	Mobility limitations	PH048
adl	Limitations with activities of daily living	PH049_1
iadl	Limitations with instrumental activities of daily living	PH049_2
esmoked	Ever smoked daily	BR001
drinking	More than 2 glasses of alcohol almost everyday	BR019
phactiv	Physical inactivity	BR015
meals	Number of meals every day	BR025
orienti	Score of orientation in time test	CF003 - CF006
memory	Score of memory test	CF103
wllft	Score of words list learning test - trial 1	CF104_* - CF107_*
wllst	Score of words list learning test - trial 2	CF113_* - CF116_*
fluency	Score of verbal fluency test	CF010
numeracy1	Score of first numeracy test	CF012 - CF015
numeracy2	Score of second numeracy test	CF108 - CF112
eurod	EURO depression scale	MH002 - MH017
doctor	Seen/Talked to medical doctor	HC002
hospital	In hospital last 12 months	HC012
thospital	Times being patient in hospital	HC013
nhospital	Total nights stayed in hospital	HC014
sn_num	Number of people within social network	SN013
sn_sat	Satisfaction with social network	SN012
cjs	Current job situation	EP005
pwork	Did any paid work	EP002
empstat	Employee or self-employed	EP009
lookjob	Looking for job	EP337

**Table 7.4: List of variables included in the imputation dataset of Wave 5 (continued)**

Variable name	Description	Questionnaire
rhfo	Received help from others (how many)	SP002, SP005, SP007
ghfo	Given help to others (how many)	SP008, SP011, SP013
ghih	Given help in the household (how many)	SP018
rhih	Received help in the household (how many)	SP020
gfg	Number of given financial gifts 250 or more	FT002, FT007_*
rfg	Number of received financial gifts 250 or more	FT009, FT014_*
otr	Owner, tenant or rent free	HO002
perho	Percentage of house owned	HO070
fdistress	Household able to make ends meet	CO007
lifesat	Life satisfaction	AC012
lifehap	Life happiness	AC022
naly	Number of activities last year	AC035_*
saly	Satisfied with no activities	AC038
willans	Willingness to answer	IV004
clarify	Respondent asked for clarifications	IV007
undersq	Respondent understood questions	IV008
hrsc	Help needed to read showcards	IV018
nomxyear	Nominal exchange rate	
pppxyear	PPP adjusted exchange rates	
currency	Currency in which amounts are denominated	

**Table 7.5: Description of flag variable associated to imputations**

Varname_f	Label	Description
-99	Missing by design	Missing values depends from skip patterns in the questionnaire
1	Not designed resp	Missing values depends on the type of respondents designed to respond
2	No ownership	No declared ownership
3	Regular obs.	Regular observation
4	Imp: ub point	Imputation based on specific declared amounts in the unfolding brackets routing
5	Imp: ub range	Imputation is based on unfolding brackets range information
6	Imp: ub incomplete	Imputation is based on unfolding brackets partial information
7	Imp: ub uninformative	Unfolding brackets uninformative
8	Imp: ownership	Ownership has been imputed
9	Imp: amount	Imputed amount
10	Imp: outlier LB	Imputed value if lower than LB
11	Imp: outlier UB	Imputed value if lower than UB
12	Imp: aggregate	Imputation of the corresponding aggregate variable, see table 2
13	Imp: NRP	(only for thinc)
14	Imp: missing value	(only for explanatory variables imputed ex-ante by hot-deck)



## References

Arnold, B.C., Castillo, E. & Sarabia, J.M. (1999). *Conditional specification of statistical models*. New York: Springer.

Arnold, B.C., Castillo, E. & Sarabia, J.M. (2001). Conditionally specified distributions: An introduction. *Statistical Science*. 16, pp. 249-274.

Browning, M., Crossley, T.F. & Weber, G. (2003). Asking consumption questions in general purpose surveys. *The Economic Journal*. 133, p. 540-567.

Christelis, D. (2011). Imputation of missing data in Waves 1 and 2 of SHARE. *SHARE WP Series*, 01-2011.

Dardononi, V., Modica, S. & Peracchi, F. (2011). Regression with imputed covariates: a generalized missing-indicator approach. *Journal of Econometrics*. 162, pp. 362-368.

Dardononi, V., De Luca, G., Modica, S. & Peracchi, F. (2014). Model averaging estimation of generalized linear models with imputed covariates. *Journal of Econometrics*, in press. doi: 10.1016/j.jeconom.2014.06.002.

Gelman, A. & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*. 7, pp. 457-511.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian data analysis*. Second Edition. Boca Raton, FL: Chapman and Hall.

Lee, K.J. & Carlin, J.B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*. 171, pp. 624-632.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*. 9(4), pp. 538-558.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. 27, pp. 85-95.

Rubin, D.B. (1996). Multiple imputations after 18+ years. *Journal of the American Statistical Association*. 91, pp. 473-489.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*. 16, pp. 219-242.

Van Buuren, S., Brands, J.P.L., Groothuis-Oudshoorn, C.G.M. & Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*. 76, pp. 1049-1064.

Van Buuren, S., Boshuizen, H.C. & Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*. 18, pp. 681-694.