4 Consistency Checks and Editing

4.1 Introduction

Data editing is understood to mean the ex post amendment of electronic records of observations collected through individual interviews so as to correct any errors or logical inconsistencies that may have occurred during the survey, as well as the aggregation of information that was recorded via proxies, typically with a view to keeping the questionnaire as clear and user-friendly as possible. The editing process is thus essential for improving the quality and consistency of the datasets.[1]

The raw data of surveys do not always contain the information that the questions were intended to elicit. As respondents in the HFCS occasionally either experienced difficulties in understanding the questions asked or had insufficient knowledge on the substance of the survey, they may sometimes have provided inaccurate information. At the same time, interviewers may have recorded information incorrectly (see also chapter 3), or data may have been processed inaccurately. In the HFCS, great importance was attached to minimizing such errors.

This chapter provides insights into the consistency analyses and edits performed for the HFCS in Austria, starting with information on the number of edits performed (section 4.2) and followed by explanations on the consistency checks conducted during and after the interviews (sections 4.3 and 4.4). Furthermore, we outline the flags used to highlight ex post adjustments of the collected observations (section 4.5), provide a detailed account of ex post editing (section 4.6) and describe formating and editing after multiple imputations (section 4.7). The chapter ends with concluding remarks (section 4.8).

4.2 Number and Type of Edits

All in all, around 22,000 of the close to 841,000 observations collected through the HFCS were edited, i.e. a mere 2.6% of all data points are amended (see the "All edits" column in table 1).

Table 1

Number and Type of Edits

	All[1]	Type of edits		
		based on expert analysis[2]	based on proxies[3]	deleted values[4]
Total observations[5]	840,714	840,714	840,714	840,714
Number of edits	21,837	6,867	13,767	1,203
Share of edited observations in total observations (%)	2.6	0.8	1.6	0.1

Source: HFCS Austria 2010, OeNB.

[1] *All edits.*
[2] *Number or percentage share of edits made on the basis of expert assessments.*
[3] *Number or percentage share of edits made on the basis of other survey information (e.g. verbatim records).*
[4] *Number or percentage share of deleted observations.*
[5] *Observations refer only to observed information. Observations that are filter missings are excluded.*

[1] *See e.g. Kennickell (2011) and Bledsoe and Friess (2002) for information on the editing measures used in the Federal Reserve's Survey of Consumer Finances.*

The three columns next to "All edits" indicate the different types of edits. Editing resulted in changes to the collected values in the case of less than 6,900 observations. These changes involved primarily inconsistent values that were corrected through following investigations and/or other information or were deleted and imputed on the basis of the imputation model. Two-thirds of all of the amendments, i.e. about 13,800 observations, could be deduced from the verbatim records and by using the responses given to user-friendly questions on, say, respondents' life assurance contracts or their total annual net income. All in all, some 1.6% of all observations were amended through edits of this kind. This low figure can be seen as an indication of the successful design of the questionnaire and the special training of the interviewers (chapters 2 and 3). In 1,203 cases, finally, i.e. for only 0.1% of all observations, the values collected were deleted and replaced with the filter missing (" . ") – among other things because a head variable was edited.

A case in point[2] would be the duplicate recording of income from pensions, first under "received employee income" and second under "received income from public pensions." Here, the head variable "received employee income" (PG0100) was changed to "No" and the value (PG0110) entered under this income variable was deleted because the respective income figure had already been correctly recorded under the pension income variable (PG0300 and PG0310).

4.3 Checks for Consistency during the Interviews

The HFCS was based on Computer-Assisted Personal Interview (CAPI). The CAPI format has a number of advantages over the use of paper-based questionnaires or phone-based interviews. The interviewer used a laptop on which the survey software had been installed and was guided through the questionnaire on screen. The information collected was checked for appropriateness and consistency as it was being entered. Any questions of clarification that the respondents might have raised could be resolved immediately either by the interviewer or with the aid of the explanatory documentation at hand, so that errors could be avoided already at the time the data were recorded.

However, checks for consistency in the course of an interview are subject to limitations with regard to both substance and number. An excessive number of consistency checks during an interview would make it exceedingly long. The respondents are "worn down" and the standard of the data collected would decline. Interviews might even have to be broken off in individual cases.

Restrictions of substance arise from the fact that the details against which answers are meant to be checked for consistency need to have been collected in the first place. These limitations do not apply to simple consistency checks linked to specific predefined benchmarks: Whenever certain limits are exceeded or undercut, pop-up warnings appear that allow the entry to be checked right away. Yet the information necessary for more complex consistency checks often does not become available until the latter stages of the interview as it first needs to be generated through other sets of questions.

[2] *Examples given in this chapter are indented for ease of reference.*

The digitalized version of the questionnaire used for the HFCS provided for more than 150 consistency checks,[3] typically in the form of "soft" checks: Whenever a test criterion was violated, a window popped up on the laptop screen to draw attention to the inconsistency of the answer given.

If a household with a disposable net monthly income of EUR 1,000 indicated, for instance, that – in addition to consumption expenses totaling EUR 900 – it had typically supported nonhousehold members with EUR 200 per month in the past year, the following highlighted pop-up appeared in the questionnaire input mask:

"The sum of total consumption expenditure and regular remittances to nonhousehold members exceeds the household's total net income. Are the figures correct? Please confirm that they are, or amend the figure(s) as necessary."

As the figures collected may have referred to different moments in time, or the remittances might have been financed through sales of assets, or as the household's income might have dropped as a result of one or more members losing their job, the correctness of the figures provided could not be ruled out. In such a case, the interviewer would ask the respondent to confirm or correct the household's total income, its remittances and its consumption expenditure.

Other consistency checks programmed into the digitalized version of the HFCS questionnaire in Austria made it possible to proceed with the survey only if an answer that had been recognized as being incorrect or inconsistent was amended accordingly. However, these so-called "hard" checks were only used in cases where a particular answer could definitely be ruled out.

If a person stated that, say, he/she had lived in Austria for 40 years but gave his/her age as 30, the following highlighted error message would appear:

"The respondent has been living in Austria for longer than his/her age allows. This is not possible. Please correct the information as necessary."

Thus, proceeding with the CAPI questionnaire required changing the age given to at least 40 years, or reducing the period of residence in Austria to 30 years or less (or an alternation of both variables).

4.4 Postinterview Checks for Consistency

4.4.1 Analysis of the Data by Experts

During the fieldphase of the HFCS in Austria, the data on households deemed to be final by the survey company were forwarded to the OeNB in seven installments. Upon receipt, all data were analyzed by experts right away.[4] On the one hand, these analyses served to improve the consistency of the data recorded for each household. On the other hand, they were used to check the survey software (in particular, to review the programming of the questionnaire) and the mechanisms used by the survey company to process the data.

[3] *A list of all the consistency checks that were programmed into the digitalized version of the questionnaire can be found in the online appendix.*

[4] *The evaluation was carried out with the aid of external sources of data such as the OeNB's 2008 Household Survey of Housing Wealth and the EU-SILC (conducted by Statistics Austria).*

The datasets for households actually interviewed and those for households that refused to participate where analyzed on a case-by-case basis. This made it possible to assess and optimize the success of interviewers in convincing households to participate. Thus it was almost impossible for interviewers to cherry-pick "easy" or more readily accessible households, which would have created a bias towards certain households (e.g. housewife and pensioner bias) and distorted the data accordingly. The interviewers were aware of the fact that the list of addresses was limited to the 4,436 households of the gross sample. This made it possible to ensure that interviewers would not select the less difficult households and then move on to a new set of addresses. The incentive for interviewers to handle the strictly limited address material as efficiently as possible was supported through a merit pay system and the relatively high effort that was required from interviewers to participate in the survey. Furthermore, area managers were advised to avoid allocating new households to interviewers before the latter had made an adequate effort to survey the households they had been assigned. The decision to exclude subsequent draws (substitute households) is among the key criteria for a successful survey, and moreover essential for ensuring the representativeness of the sample (Vehovar, 1999).

Initial analysis of the information on individual households during the field phase covered the information provided on geographical location and structure, financial and real assets, debt and income, whether households had come to ownership of property by inheritance or gift, comments made by households or interviewers, as well as the date, time and duration of the interviews. This set of information enabled an initial assessment of the quality of the interview. The microdata on every single household were checked for consistency of substance and reviewed by at least two analysts. Issues requiring clarification were discussed by the whole team, which then decided on the way forward.

In addition, this stage of the process was also used to assess the interviewers (see also chapter 3) and to draw their attention to errors or incomprehensibilities. The shortcomings identified in this process were often of a minor nature, but three interviewers whose results were not up to the required standards (e.g. regarding nonresponse) were excluded.

4.4.2 Follow-Up Investigations

If the data analysis did not lead to the root of a data problem, households were contacted again by the survey company to clarify uncertainties and ensure that data were recorded correctly. A typical case of a data problem that was easy to spot and did not require follow-up investigations was rewriting a negative current account balance as a positive liability (account overdraft) while setting the current account balance to zero (see also section 4.6). This was simply a matter of adhering to the recording conventions for such liabilities. Decisions on follow-up investigations were always guided by the principle that any ex post editing of data and additional burdens on participating households should be kept to a minimum. All in all, a follow-up was necessary on specific details of some 400 households. Many of the unusual results (e.g. particularly high amounts of assets) were confirmed or else corrected in the course of the follow-up investigations.

4.4.3 Investigation of Outliers

Particular attention was paid within the scope of the individual analyses to the recognition and processing of outliers (exceptionally high or low values), which were recorded above all for financial variables, the size the household income or the size of the dwelling. Any outliers that were not removed from the dataset were actually not the result of interview errors but largely confirmed in the follow-up. It therefore seems appropriate in future studies based on HFCS data not to generally exclude outliers from the analysis, but rather to incorporate them in computations through the use of suitable methods.

4.4.4 Technical Review of Filters and Consistency

On top of the consistency checks programed into the digitalized version of the questionnaire and the analysis of the data by experts, the field phase also included detailed automated consistency checks of the data of all households.

All hard checks were applied repeatedly to the observations, for instance, in order to assess whether respondents might have given answers that precluded moving on to subsequent questions, thus requiring changes. Wherever the programming of individual hard checks was found to be faulty, the polling firm was informed so that the fault could be corrected.

The technical review also covered the questionnaire's complete set of filters to prevent programming errors from leading to extensive and costly follow-up interviews. Comprehensive tests of the questionnaire's programming prior to the start of the field phase as well as a pilot survey of 50 households led to the identification and correction of minor programming errors. For instance, whenever two members of a household refused to reveal their age, the second individual's answers to most of the personal questions were suppressed initially (see section 4.6.2.12).[5] These filter checks also made it possible to ensure that the coding of variables was consistent throughout the questionnaire.[6]

4.5 Flags

All edits (and imputations – see chapter 5) can be identified with flag variables, which document in detail how the individual HFCS observations were established (see table 2 for a list of the flags used to classify the observations). To comply with international provisions, we aggregated a number of flags to ensure that the datasets are comparable at the international level (section 4.7).

Group I

The flags allocated to group I were used to indicate that the data points had been compiled. Specifically, all the values documented with the survey software during the interview were assigned flag 1, while all filter missing observations (".") were flagged with a 0. Information recorded in loops (section 4.6.2.4) was – where necessary – moved in the iteration of a loop and assigned flag 2. This means that flag 2 observations are included in the dataset exactly as they were recorded but carry a different iteration number.

[5] *This problem was discussed and resolved soon after the receipt of the first installment of the survey data, so that it remains limited to only a few households.*

[6] *All HFCS variables were assigned value labels that explain the coding. The coding of the individual variables is also included in the questionnaire (available in the online appendix).*

Table 2

Flags Used in the HFCS in Austria

Group I	0	Not applicable (i.e. skipped due to routing)
	1	Recorded as collected, complete observation
	2	Recorded as collected, but moved in iteration
Group II	1050	Not imputed, originally "Don't know"
	1051	Not imputed, originally "No answer"
	1052	Not imputed, originally not collected due to missing answer to a previous question
	1053	Not imputed, originally collected from a range or from brackets
	1054	Not imputed, collected value deleted or value not collected due to a CAPI or interviewer error
	1055	Not imputed, edited to "missing" due to incorrect answer to a higher-order question
Group III	2050	Missing, set as "missing" for anomynization purposes
	2051	Missing, set as "missing" because data were not collected
Group IV	3050	Edited, set to modified value as considered incorrect or unreliable
	3051	Edited, adjusted on the basis of other information obtained in the (national) survey
	3052	Edited, adjusted on the basis of the verbatim records
	3053	Edited, set as "missing" (".")
Group V	4050	Imputed, originally "Don't know"
	4051	Imputed, originally "No answer"
	4052	Imputed, originally not collected due to missing answer to a previous question
	4053	Imputed, originally collected from a range or from brackets
	4054	Imputed, collected value deleted or value not collected due to a CAPI or interviewer error
	4055	Imputed, originally not recorded due to incorrect answer to a higher-order question

Source: HFCS Austria 2010, OeNB.

Group II

Recorded observations that were incomplete or inadequate were assigned group II flags. Such observations include cases where the respondent was unable or refused to answer the question (entries of "Don't know" or "No answer"), or proved unable to give an exact figure and provided a range instead. Included here are also observations that were not available on account of edits of either the variable in question or a head variable (flags 1054 and 1055). Observations with group II flags were not imputed (chapter 5).

Group III

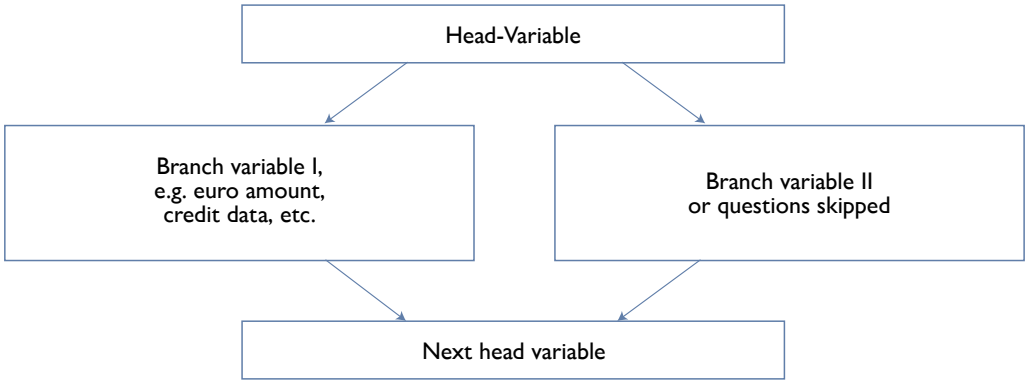
Observations and/or variables that were not recorded or that were recorded but subsequently deleted from the datasets on account of anonymization rules were assigned group III flags.

Group IV

Flags belonging to group IV indicate ex post edits of an observed value. The following types of ex post edits can be distinguished: edits required by logical inconsistencies were flagged with 3050; observations that were adjusted using other information obtained in the survey (e.g. that on life assurance contracts – see section 4.6.2.9 for details) were flagged with 3051; observations that were adjusted on the basis of verbatim records (see section 4.6.2.3) were encoded with 3052; and edits to delete a value and set the observation set to filter missing (as in the case of duplicate entries) were flagged with 3053. All observations that were corrected through follow-up investigations were assigned flag 3050.

Chart 3

Sequence of Questions



Source: HFCS Austria 2010, OeNB.

Group V

Flags in group V mirror those of group II. If it was possible to impute missing values, the first digit of the flag was changed to 4. For instance, if respondents had provided a range rather than an exact figure, which was subsequently imputed, this observations was flagged with 4053 after multiple imputations. This ensures that all information can be tracked even after imputations.

Chart 3 indicates how questions were typically structured in the HFCS questionnaire. Let us take employee income to give an example of the tree of questions[7] and the use of flags.

> The head variable for recording employee income serves to ascertain whether or not a household has an income of this kind (to be answered with "Yes" or "No"). If the question was answered with "Yes," the amount was recorded and the interview continued with the next head variable in the questionnaire – in this case, the question on self-employment income. If a household had no income of this kind, or if the respondent failed to provide the necessary information (answering the question with "Do not know" or "No answer"), the interview continued with the question on self-employment income (the next head variable). Depending on which answers were given, all the observations recorded were initially flagged with with 1 or 0. If it was found later (e.g. as a result of information on the employment situation in the household) that the question on employee income had been incorrectly answered with "No," the initial response was corrected and flagged as "Edited, set to modified value as considered incorrect or unreliable" (3050) and the corresponding field for entering the amount was released for imputation. Following imputation, the value was then reflagged as "Imputed, not recorded due to incorrect answer of a higher-order question" (4055).
>
> If, for instance, the question on the educational attainment of a member of the household (variable (A)PA0200) had been answered by selecting the category "Other qualifications" and if that answer was subsequently found

[7] *See chapter 2 for details on the structure of the questionnaire.*

to match one of the predefined categories, the observation was flagged as "Edited, adjusted on the basis of the verbatim records" (3052) in the flag variables of the personal dataset.

This flag system allows the origins of every single observation in the HFCS to be tracked. No flags were used to encode the variables for identifying households and persons, nor were the country codes and the iteration number of the imputation flagged. The flags described here provide for a more detailed breakdown by category than those incorporated in the international HFCS dataset that can be obtained from the ECB. For reasons of international consistency, the flags were aggregated prior to submission to the ECB (section 4.7).

4.6 Ex Post Editing

4.6.1 Case-by-Case Review

A detailed case-by-case review of all households allowed inconsistencies to be identified and eliminated through follow-up investigations and ex post editing. Respondents' answers were checked for plausibility against known benchmarks, as derived from descriptive statistics (e.g. on the average income) compiled on the basis of already completed HFCS interviews and external sources of data. Moreover, the review process heavily relied on proxies that recorded values in aggregated form and/or in a variety of other ways.

Both interviewers (chapter 3) that produced nonstandard results and follow-up investigations by the survey company were reviewed with particularly great care. Analyses by experts were generally used to resolve the following issues through ex post edits:

- Double entries: Cases where an inheritance, for instance, was recorded both under "Household main residence inherited" and in the "Inheritances and gifts" chapter, or where the same income was recorded in two different income categories, had to be corrected.
- Missing or additional "zeros": In a few cases interviewers added or left out a zero by accident when recording amounts; this had to be amended accordingly.
- Real estate ownership patterns in the context of the intra-household distribution of this asset item: A few cases where the household's financially knowledgeable representative had taken the question on the ownership of the main residence to relate to himself/herself as an individual, rather than – as intended – to the household as such, had to be corrected.
- Implausible values: Values that remained implausible after follow-up investigations had to be changed into missing data entries that were subsequently imputed.
- Erroneous entries by interviewers: For example, the value 10 (for October 2011) that had been entered as the contact month for an interview conducted in 2011, the outcome of which had been submitted in February, had to be changed to read 1 (for January 2011) because the interview phase had ended in June.

Such edits related to the whole questionnaire, not just to individual variables. Amendments to recorded data were kept to a minimum and – wherever follow-up investigations and/or the use of proxies (such as verbatim records) failed to provide further information – inconsistent observations were changed to missing and released for imputation. Inconsistent or implausible observations were processed with great care and only deleted if there was virtually no doubt about the inconsistency.

4.6.2 Structural Editing

4.6.2.1 Cleaning

When answering the HFCS questions, respondents occasionally gave inaccurate answers but subsequently corrected those answers as they proceeded through the questionnaire. These corrections also necessitated a change in the sequence of questions following the initial question because the new answers called for different filter settings. The initially "wrong" path through the questionnaire, however, remained in place for transparency reasons and had to be cleaned up ex post.

4.6.2.2 Currency Conversion

Respondents could give amounts in different currencies (chapter 2). The edits set out below relate both to entries of exact amounts and to ranges indicated by the respondents (entries for predefined ranges had to be in euro).

Typically, amounts were given either in euro or in Austrian schillings. In particular, the value of the main residence (both the purchase price and the current value) was often given in Austrian schillings. All Austrian schilling amounts were subsequently converted into euro at the irrevocably fixed conversion rate of EUR 1 = ATS 13.7603.[8] In addition, some amounts were given in Deutsche mark. In such cases, too, the irrevocably fixed conversion rate set by the ECB was used for conversion into euro, namely EUR 1 = DEM 1.95583.[8]

In the case of foreign currency loans, amounts were also given in Japanese yen and Swiss francs. The value of the amount outstanding at the time of the interview was converted into euro on the basis of the average of the exchange rates recorded in 2010, while the total at the time of borrowing was converted at the average of the exchange rates recorded in the year in which the loan was taken out, with the exchange rates published on the OeNB's website[9] being used as reference values.

4.6.2.3 Verbatim Records

For many questions, respondents were given the option of choosing the category "Other" and providing a verbatim response, essentially with a view to making the questionnaire as user-friendly as possible. Thus, recording of the verbatim description was allowed if it was not possible to assign a respondent's answer to a predefined category during the interview. The verbatim entries were used to assign answers to specific categories ex post, as proved to be possible in most cases. Wherever this could not be done, the initial categorization of the variable as "Other" was retained. All observations subjected to ex post edits on the basis of verbatim records were flagged with flag 3052 (see section 4.5 for details on the flags used).

4.6.2.4 Navigation of Loops

As outlined in greater detail in section 2.6.1, some pieces of information were recorded in loops, which means that the interviewer ran through an identical set of questions for a sequence of items owned by the households. Information on the following items was collected in the form of loops:

[8] *For the irrevocably fixed euro conversion rates, see www.oenb.at/isaweb/report.do?lang=EN&report=2.12 (accessed on January 22, 2013).*

[9] *For the effective exchange rate indices of the euro, see www.oenb.at/isaweb/report.do?lang=EN&report=2.16 (accessed on January 22, 2013).*

- Mortgages on the main residence
- Real estate assets apart from the main residence
- Mortgages on such other property
- Unsecured loans
- Businesses owned by the household
- Inheritances and gifts

In the following we provide an explanation of the edits which were required because of loop questioning.

Recording Sequence

The sequence of items that were covered in loops followed a predefined order. With regard to mortgages backed by the primary residence, for instance, the first iteration of questions related to the mortgage with the highest amount outstanding, the second iteration of the loop to the mortgage with the second-highest outstanding amount and the third iteration to the third-highest loan amount outstanding. Some respondents did not always adhere to this sequence. Such cases were recoded in the course of the editing process – with the exception of the loop questions on inheritances, for which no recoding was carried out because respondents had been asked to go through the loops of questions on inherited wealth in descending order of relevance for the household's current wealth situation. At the same time, they were instructed to indicate only amounts as transferred rather than current amounts. Over the period between the time an item was inherited and the interview, certain inheritances can gain (or lose) more in value than others, or inherited residential property, for instance, might already have been passed on to children, so that it no longer has an impact on the household's wealth situation.

Regarding flags, every variable within a loop that was replaced with observations recorded for the same variable in another loop was flagged with 2 (section 4.5). Wherever a filter missing for one of the variables within a loop was replaced with the filter missing for the same variable in another loop, flag 0 "Not applicable; skipped due to rerouting" was used.

Skipping of Questions

In order to avoid breaking off an interview in mid-loop, respondents were allowed to skip parts of loop questions and to proceed directly with the summary questions, where either the residual sum total of the not yet recorded loans and/or businesses (more than three loans or businesses) or the sum total of all (up to three) loans and/or businesses was recorded. If questions within the loop for inheritances and gifts were skipped, information on the sum total of all inheritances was always requested in the summary question. As the summary questions of all sections of the dataset to be sent to the ECB were supposed to cover only any items that went beyond the first three itemized loans, real estate assets, inheritances and gifts, the relevant summary responses had to be edited accordingly. For ease of reference, edits are described in the following on the basis of the section of the questionnaire dealing with unsecured loans (section 2.5).

In the 21 cases in which a household had taken out only one unsecured loan, and had skipped questions within a loop, the type of edit depended on whether the respondent had indicated the outstanding amount (i) only in answer to the summary question, or (ii) both when running through

the first loop of questions and in answering the summary question, or (iii) neither during the first loop of questions nor in answer to the summary question. If the respondent had indicated the outstanding amount only in answer to the summary question – possibility (i) – this amount was entered as the answer to the appropriate question (in the first loop) and the entry under the summary question was edited to filter missing. If the respondent had indicated identical amounts – possibility (ii) – in answer to the loop question and under the summary question, the latter was edited to filter missing since it was a duplicate entry.[10] Where no amount was given at all, neither within the loop nor in the summary – possibility (iii) – solely the summary question was edited to filter missing.

In cases where a household had taken out two unsecured loans, and had skipped questions within a loop,[11] the type of edit depended on whether the respondent had (i) specified the highest loan outstanding and indicated an aggregate amount in answer to the summary question; or (ii) indicated outstanding loans in answer to both loops of questions and the summary question; or (iii) specified amount solely in the answer to the summary question; or (iv) given no amounts at all, neither in the answers to the loop item questions nor in the answer to the summary question. In the case of possibility (i), the amount outstanding of the lower of the two loans was taken to be the difference between the amount given in the answer to the summary question and that given under the first loop, with that amount then being entered accordingly. This was, however, done only if the sum total of the two loans outstanding exceeded the amount outstanding of the first loan. If it was lower, it was assumed that the amount given in the answer to the summary question was not the sum total of the two loans outstanding, but rather the amount outstanding of the second such loan. In both instances, the summary question was subsequently edited to filter missing. In the case of possibility (ii), the amount given in answer to the summary question was edited to filter missing. If only the sum total of the two loans outstanding was given – possibility (iii) – it was used as the upper bound of both the first and the second such loan for the imputation model. If no amounts were given at all, neither under the loop questions for each of the two loans nor in answer to the summary question – possibility (iv) – the summary question was edited to filter missing.

The editing procedure followed in cases of three loans and skipped loop questions prior to the recording of the individual amounts outstanding was similar to that used for two loans when loop questions were skipped. All edits were again properly flagged.

[10] *Where the amounts given were not identical, that given under the loop questions on the first loan was deemed to be more relevant than that given under the summary question. The reasoning behind this procedure is that the loop questions relating to the first loan contained a question asking expressly for the amount outstanding on an unsecured loan, so the amount given there is regarded as more trustworthy.*

[11] *This occurred in the case of only two households.*

Summary Questions

Every loop of questions ended with summary questions (chart 2 in chapter 2). As a rule, the variables on summary questions contained in the dataset reflect only the residuals beyond any three items of a household. As indicated in chart 2, the summary questions were ultimately also put to all respondents who had refused to indicate the numbers of a given item in the household. In such cases of non-response, the information provided here was used for multiple imputations (chapter 5) and deleted from the dataset ex post.

4.6.2.5 Personal Variables at the Household Level

Various variables recorded information on the individuals belonging to the household, but they are stored in the household file. Examples of such information are details on the real estate ownership patterns within the household, on which member of the household took out which of the different loans or on that member of the household who works in the business owned by the household.

In order to be able to cover even unusually large households, variables were created for up to 18 individuals per household. However, the largest household interviewed successfully in Austria had only 9 members, so that all such variables in excess of that number were deleted from the dataset. Each such variable was reviewed individually, and those that had not been filled were deleted. If, say, individual No. 6 was the last member of the household in the list of borrowers to take out a loan, all variables for individual No. 8 and beyond were deleted. The variable for individual No. 7, while not containing a "true" entry in the sense that all households are encoded "No further individuals listed," was retained in the dataset simply to reflect the latter.

4.6.2.6 Current Account Balances and Overdrafts

A number of households misreported a negative balance on their household current account as a current account balance (HD1110). There were also occasional duplicate entries, as well as misplaced entries, in this area that subsequently had to be edited.

4.6.2.7 Rent Variables

The HFCS questionnaire included questions on the amount of housing rent paid both excluding and including the running costs. In the case of some households, the rent excluding running costs was higher than, or equal to, that including such costs, which is simply not possible because housing cannot be "run" free of charge. Some of these households had entered only the running costs under the item "Rent including running costs." In the course of editing, these were added to the amount entered under "Rent excluding running costs" to arrive at the "Rent including running costs." In the case of other households, the "Rent including running costs" was edited to read missing and released for imputation, with the "Rent excluding running costs" serving as the lower bound to the "Rent including running costs."

In addition, the item "Rent including running costs" was edited to become the upper bound for the variable "Rent excluding running costs" and used for imputations whenever the answer to the latter was not an amount (i.e. read "Don't know" or "No answer").

4.6.2.8 Agricultural Businesses

As defined in the HFCS, farmers are owners of an agricultural business. Some farmers, however, did not regard themselves as persons running a business. Such cases thus had to be reviewed separately and the observations had to be edited accordingly. The edits were based on the definition used to classify a household that owned an agricultural business. The classification was undertaken with the aid of the employment variables of all household members. In cases where at least one member of the household indicated that he/she worked (on a self-employed basis) as a farmer, the number of investments in self-employment businesses was increased by one only if no investment in a self-employment business had previously been recorded in the agricultural sector. If the investment in a self-employment business had already been recorded earlier, the entry was not edited. The NACE code for this business was set to that for agricultural businesses, and at least that member of the household who had stated that he/she worked as a farmer was deemed to be employed in the agricultural business. The legal form of the business was edited to read sole proprietorship. Both the household's ownership share and the value of the agricultural business were released for imputation.

In some cases, the value of the agricultural business was wrongly entered under value of the main residence. Account was taken of such information in imputations through delimitations and/or the use of a proxy variable, in which the value of the main residence was recorded together with the value of the investment in a self-employment business (section 5.3.5).

The category of agricultural businesses was subjected to special individual reviews by experts. Particularly complex cases were covered by follow-up investigations and logical amendments were made where necessary.

4.6.2.9 Life Assurance Policies

Information on assets held in the form of life assurance policies was recorded through questions ensuring that the answers were both as precise as possible and prone to only few errors. In particular, there was no direct question on the value of the assets in life assurance policies, but rather a series of questions on the start of payments, the frequency of payments (monthly or yearly) and the amount of the current payments for every single life assurance policy in the household.[12] The value of the assets held in the form of life assurance policies was calculated as the sum total of all payments. In cases where one or several details were not given, the remaining observations were used for the bounds of the value to be imputed.

4.6.2.10 Income Variables

The following categories of personal income were recorded separately for every member of the household who was 16 years old or older:

- Employee income (PG0110)
- Income from self-employment (PG0210)
- Income from public pensions (PG0310)
- Income from private and occupational pension plans (PG0410)
- Income from unemployment benefits (PG0510)

[12] *Possible lump-sum payments at the start of a life assurance policy could be identified as such in the verbatim records.*

This information was supplemented by the following income categories that were recorded per household:

- Income from public assistance or welfare payments (HG0110)
- Income from private transfers (HG0210)
- Rental income from real estate assets (HG0310)
- Income from financial investment (HG0410)
- Income from investments in self-employment businesses or partnerships (HG0510)
- Income from other sources (HG0610)

In the case of the first four personal income categories, respondents could indicate their net income if their gross annual income was not known (chapter 2).

Where only a net amount was entered for an income category, the gross income was calculated with the aid of the Austrian Federal Ministry of Finance's converter of gross to net income,[13] with information on the type of income, the structure of the household (with reference to the tax credits for single parents and single earners), the employment status and age of children (if any), the Federal State and the employment status (white or blue-collar workers, pensioners)[14] being used as a basis.

Wherever both parents were gainfully employed, the single earner's tax credit was assigned to the main earner, i.e. the parent with the higher income (to the extent that the legal requirements were fulfilled and the partner did not earn more than EUR 6,000 per annum).

Given the far greater scope that the self-employed have for tax deductions, a precise conversion was dispensed with in the case of income from self-employment. A precise conversion of the net into the gross amount is given only in the case of annual incomes of less than EUR 11,000, which are regarded as tax-free, so that the gross is equal to the net. For all other values (in the case of some 25 individuals), the net income was converted (on the basis of the employment status of white-collar workers), with EUR 10,000 subsequently being added to, or deducted from, the amount obtained in order to arrive at an interval for the imputation of the exact amount. This reflected the uncertainty that such a conversion entails, without losing the important information on the actual range within which the value is to be found.

If the net amount had moreover also been recorded only as a range, the upper and lower bounds were converted into gross values that were subsequently used in the imputations. All converted values were assigned flag 3051.

Income from private pension plans (PG0410) was not converted, but simply taken as gross amounts since the amounts involved were insignificant.[15]

Using flags as a basis, table 3 gives an indication of the number of edits relating to employee income. The table also illustrates the use of flag variables (see also section 4.5). The question on the amount of employee income received (variable PG0110) was put to a total of 2,166 individuals. 1,263 of the respondents (58.3%)

[13] *See www.bmf.gv.at/service/anwend/steuerberech/bruttonetto/_start.htm (accessed on January 22, 2013, in German).*

[14] *"Apprentices" were categorized as "blue-collar workers" in the conversion, while "civil servants" were seen as "white-collar workers" on grounds of their more favorable tax treatment.*

[15] *Seven individuals in all gave their income from private pension plans in net terms. For six of them, the annual income was below EUR 2,900, while one indicated a net income of EUR 18,500 per annum from private pension plans. Given their low value, all amounts were taken as gross amounts, so that there was no conversion.*

Table 3

Number and Share of Edits of Gross Employee Income Based on Flags

	Number of persons	Share in %
Number of persons receiving employee income	2,166	100
Answer recorded, observation complete (flag 1)	1,263	58.3
Not imputed, originally "Don't know" (flag 1050)	49	2.3
Not imputed, originally "No answer" (flag 1051)	76	3.5
Not imputed, originally not collected due to missing answer to a previous question (flag 1052)	14	0.7
Not imputed, originally collected from a range or from brackets (flag 1053)	248	11.5
Not imputed, collected value deleted or value not collected due to a CAPI or interviewer error (flag 1054)	5	0.2
Not imputed, edited to "missing" due to incorrect answer to a higher-order question (flag 1055)	207	9.6
Edited, set to modified value as considered incorrect or unreliable (flag 3050)	28	1.3
Edited, adjusted on the basis of other information obtained in the (national) survey (flag 3051)	274	12.7
Edited, adjusted on the basis of the verbatim records (flag 3052)	2	0.1

Source: HFCS Austria 2010, OeNB.

expressed their annual income in gross terms, as had been requested. A further 49 respondents (2.3%) answered "Don't know" and 76 individuals (3.5%) opted for "No answer". Yet another 14 individuals answered the yes/no question on whether they had any employee income with either "Don't know" or "No answer," so that the flag (1052) precluded that they were asked about the amount. 248 respondents (around 11.5%) gave the amount of their income in the form of a range. The responses of 212 individuals (9.8%) were edited and flagged as missing and "to be imputed"; by far most of the respective edits (those of 207 individuals) were due to an incorrect head variable (flag 1055). 274 of the respondents (12.7%) were only able to provide their net income, which was then converted with the aid of the Federal Ministry of Finance's converter of gross to net income. The responses of the remaining 30 individuals (1.4%) were edited in value on the basis of analyzes by experts, with most edits due to follow-up investigations or to the aforementioned problem of additional or missing "zeros."

4.6.2.11 ISCO Classification

In accordance with the requirements of the ECB questionnaire, the main occupations of respondents were recorded (in variable PE0300) on the basis of the categories set out in the *International Standard Classification of Occupations* (ISCO88). Making individual members of each household classify their respective occupation themselves, however, would have been extremely difficult for respondents who do not have any advance knowledge of the ISCO categories, so that the observations received would have been inadequate. On those grounds, the questions on the job title and/or the tasks of relevance to the main occupation were asked orally in Austria, with the answers being recorded verbatim. That information was then transposed ex post into the corresponding ISCO code, as taken from Statistics Austria's list of Austrian ISCO designations ("Alphabetikum zur OEISCO"). As required by the ECB, classification was based on the ISCO aggregation of occupational groups (two-character codes). To this end, the verbatim record of the job title and related main tasks was supplemented with personal data of relevance for

the ISCO classification (in particular, the respondent's professional qualifications and the main activities of the company in which the respondent worked).

In order to ensure the international comparability of this data across all euro area countries, the ISCO codes used for classification were those of the old ISCO88, as required by the ECB. In the flags, the variable PE0300 to be submitted to the ECB was first flagged with 3051 – which means "Edited, adjusted on the basis of other information obtained in the (national) survey" – and aggregated in a next step (section 4.7).

4.6.2.12 Exclusion of Successful Interviews

For various reasons, the final data do not include those on the following three groups of, all in all, 28 households that were interviewed successfully.

- Households not belonging to the target population: The target population for the HFCS in Austria comprises all households that do not live in institutions (e.g. boarding schools, homes for elderly people, prisons, etc.). A number of the households that had been interviewed successfully were excluded from the survey because the respondents were living a home for elderly people. The four households in this group were edited out and flagged as "Not belonging to the gross sample." Two households that were not interviewed successfully were likewise eliminated from the gross sample and given the same status.
- Households with a high proportion of nonresponse items: This group, which comprised 22 households, had to be deleted from the dataset because the respondents refused to answer too many questions.
- CAPI errors: In the case of two households, a programming error in the digitalized questionnaire made it impossible to access the personal questions for all members of the household.

The observations for the households in the two latter groups were edited to "Nonparticipation on other grounds" and assigned a nonresponse weight of zero (section 7.2.3).

4.6.2.13 CAPI Errors Encountered with the Questionnaires

Category "Other" in Answer of the Question on the Country of Birth (RA0400)

A verbatim record was not possible for this category, which meant that "Other countries" could not be specified ex post and had to remain "Other" in the dataset. All in all, 11 individuals chose this category, and these answers were retained as such in the dataset.

Educational Attainment of Parents (APA02$0[16])

This question was not to be put to children of the household's financially knowledgeable person because the information on educational attainment had already been obtained from the latter. However, the filter did not work properly in all cases. Four individuals – parents of households' financially knowledgeable representatives – were not asked this question at all. The variables for these individuals was set to missing and released for imputation. In the case of a further 18 individuals who were all children of households' financially knowledgeable representatives, this question was asked by mistake, so that the data were deleted ex post.

[16] *This is a noncore variable specific to Austria that is not enclosed for in the international HFCS dataset.*

Uncapped Range in the Case of Investments in Self-Employment Businesses (HD0801)
At the start of the field phase, the filter programming of the euro loop for the value of the first investment in a self-employment business was faulty. In cases where a range without an upper bound was entered, the filter did not direct to the recording variable the currency and the required confirmation, but rather to the variable for predefined ranges. This occurred only once, and the entry involved was obviously a euro amount and was also recorded as such.

Status of the Last Main Job and Duration of Employment (PE0900 and PE1000)
Faulty CAPI filter programming resulted in these variables being missing in the case of some individuals, while they were recorded unnecessarily in others. Accordingly, variable PE0900 entries for individuals who had been covered by mistake (i.e. for all individuals who were gainfully employed) were deleted in the course of the editing process in order to ensure conformity with international requirements. However, the information involved is easy to reinstate via the data recorded in answer to the next variable in the questionnaire (PE1000). In addition, the observations relating to 14 individuals who had mistakenly not been asked the question on the duration of their employment were released for imputation.

Expected Age of Retirement (PE1100)
The CAPI filter programming for the HFCS in Austria did not coincide with that of the ECB questionnaire in that the filter for the question on the age of retirement did not relate (as required by the ECB) to only the main job (PE0100a), but also to other jobs (HFCS conducted in Austria). For this reason, the question was not asked in the case of nine individuals. In order to ensure conformity with international requirements, personal observations under this variable were edited to missing and released for imputation.

Regular Payments of Occupational Pension Plan Benefits (PF0800)
The filter programming for the question on the payment of regular occupational pension plan benefits caused some individuals in Austria who had a claim to future payments of benefits under an occupational pension plan not be asked this question. The information on individuals who had mistakenly not been asked this question was imputed.

4.7 Formating and Editing after Multiple Imputations

Any information collected at a greater degree of granularity in Austria than in other countries was processed further upon imputation so as to bring the level of aggregation into line with the international requirements. The most important aggregations can be summarized as follows:

- Marital status: The categories "Married and living together with spouse" and "Married, but separated" were aggregated as "Married."
- Education: Categories specific to Austria were assigned categories under the International Standard Classification of Education (ISCED). As allocating domestic education paths to the respective ISCED is not straightforward, we recommend using the classification established for Austria.
- Labor/employment status: More detailed subcategories were aggregated.

- Main residence – ownership: More detailed subcategories were aggregated.
- Loan installments: The installments for repaying (secured and unsecured) bullet loans were set to "0" as such loans are repaid with a single lump sum upon maturity. Assets accumulated for repayment can be analyzed on the basis of variables that are specific for Austria.
- Number of other vehicles: The vehicle categories "Vans" and "Mobile homes and caravans" were aggregated as "Vans."
- Purpose of a loan: The category "To finance a deposit for the housing association" was recorded in the category "Other."
- Legal form of the business: More detailed subcategories were aggregated.
- Balances on savings plans with savings and loan banks and life assurance contracts: Data recorded on those two investment methods are aggregated into savings (HD1200 and HD1210).
- Investment behavior – willingness to take risks: The optional answer "No uniform allocation possible" was coded as "Don't know."
- Kind of assets received (survey questions on inheritances and gifts): The sequence based on asset values was dissolved.
- Provider of assets (survey questions on inheritances and gifts): More detailed subcategories were aggregated.
- Purpose of saving: The sequence by relevance was dissolved.
- Paradata: The variables HR1100 and HR1200 were recoded from a single-response to a multiple-choice answer and vice versa, while the nonrecorded variable HR14001 was edited to "missing."
- Flags: The more detailed flags specific to Austria were aggregated to conform to international standards, i.e.
  - flag 1055 was recoded as 1054;
  - flags 3051, 3052 and 2 were recoded as 1;
  - flag 3053 was recoded as 0;
  - flag 4055 was recoded as 4054.

The additional data over and above those in the ECB's HFCS datasets, which are collected at the national level and contain all the variables specified by the ECB, will probably be available from the OeNB as of spring 2013. The additional information includes additional variables, as well as a more detailed breakdown of certain variables. Datasets may be merged on the basis of both the identification numbers and imputation numbers.

4.8 Concluding Remarks and Online Appendix

The underlying rationale of editing was to edit only those observations that had most probably not been recorded correctly. In cases of ambiguity, the possibility of conducting ex post investigations on the phone was always considered first. Recourse to this option allowed many observations either to be corrected or to be confirmed as correct.

Knowledge of the steps undertaken to check the consistency of the data is essential both for any analysis of the data and for understanding how the observations have come about. In addition, the use of flags makes it possible for users to develop an imputation model of their own, to dispense with imputations, or to resolve the problem of item nonresponse in some other way.

The online appendix which supplements the information provided here on the edits and consistency checks applied in the HFCS in Austria contains a list of the consistency checks programmed into the digitalized version of the questionnaire.[17]

[17] *All documents included in the online appendix are available at www.hfcs.at.*