4 Consistency checks and editing

4.1 Introduction

Data editing is understood to mean the later amendment of electronically recorded 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 auxiliary variables, 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 collected in 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, data entry errors may have occurred (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 second HFCS wave 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 observations recorded (section 4.5), provide a detailed account of ex post editing (section 4.6) and describe formatting 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 65,000 of the close to 1.3 million observations collected in the second HFCS wave were edited, i.e. 4.8% of all data points are amended (see table 2).

Table 2

Number and type of edits

	Total observations[1]	Number of edits	Share of edited observations in total observations
All	1,347,529	65,237	4.8%
Edits based on expert judgment and follow-up phone calls	1,347,529	9,319	0.7%
Edits based on other survey information (e.g. verbatim records)	1,347,529	44,715	3.3%
Deleted observations	1,347,529	11,203	0.8%

Source: HFCS Austria 2014, OeNB.

[1] Includes only observable information. Filter missings are excluded.

The rows below "All" indicate the different types of edits. Edits resulting in changes to the collected values, i.e. real changes, occurred in the case of only

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

around 9,300 observations (see row "Edits based on expert judgment and follow-up phone calls"), which corresponds to a change rate of 0.7%. These changes involved primarily inconsistent values that were corrected as a result of subsequent queries on the phone and/or other information, or were deleted and replaced through imputation. More than two-thirds of all amendments (see row "Edits based on other survey information (e.g. verbatim records)", i.e. about 44,700 observations, could be derived from the verbatim records and the use of respondent-friendly questionnaire design (e.g. questions about life insurance policies or total annual net income). All in all, some 3.3% of all observations were amended through this type of editing. This rate indicates how important it is to allow for verbatim records on a large scale. Questionnaires as detailed as the one used for the HFCS in Austria must be user-friendly to ensure the participation of respondents and high quality standards. Various data – e.g. data on the occupation (ISCO code) of employed individuals – are only collected as verbatim responses to minimize the effort required from respondents. In around 11,200 cases (i.e. around 0.8% of observations), observations were set to filter missing (".")[2] mostly in the process of data cleanup (see section 4.6.2.1). Moreover, some items had been entered at a wrong position in the questionnaire. When transferring such information to the right position, the original entry must be deleted. In addition, there are cases (see below) in which a complete observation is set to filter missing (".") among other things because the corresponding head variable has been edited.

> A case in point[3] would be the duplicate recording of income from pensions, first under "Received employee income" and then under "Received income from public pensions." Here, the head variable "Received employee income" (PG0100) was changed to "No" and the value recorded for this variable was deleted because the respective income figure had been adequately recorded under the pension income variable (PG0300 and PG0310).

4.3 Consistency checks during interviews

The HFCS is based on computer-assisted personal interviewing (CAPI). CAPI has a number of advantages over the use of paper-based questionnaires or phone-based interviews. The interviewers use a laptop on which the survey software is installed and are guided through the questionnaire on the screen. The information collected is checked for integrity and consistency as it is being entered. Any questions of clarification that the respondents may have raised can be resolved immediately either by the interviewer or with the aid of the supporting documentation, and thus errors can be prevented during data entry.

However, consistency checks during an interview are subject to limitations in terms of scope. An excessive number of consistency checks during an interview would make it exceedingly long and thus wear out the respondents and in turn decrease the standard of the data collected. Interviews might even have to be broken off in individual cases.

[2] *The cleanup statistics do not reflect irrelevant variables cleaned up following the skipping of certain questions in a loop (see also sections 2.6.1 and 4.6.2.4).*

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

Moreover, restrictions arise from the fact that all information which should be used for the consistency checks must already be available. 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 immediately. However, the information necessary for more complex consistency checks often does not become available until answers are received in the latter stages of the interview.

The digital version of the questionnaire used for the HFCS provided for close to 250 consistency checks,[4] typically in the form of "soft" checks. Whenever a test criterion was violated, a warning message popped up.

If a household with a disposable net monthly income of EUR 1,000 (enough to cover the relevant household's average consumption) 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 message popped up:

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

The initial figures may in fact be confirmed in the cross-check, possible reasons being that the figures reported referred to different time periods, that the remittances were financed by the sale of assets, or that the household's income had since dropped as a result of one or more members losing their job. At any rate, these exemplary inconsistencies would prompt the respondents to confirm or correct the total household income, remittances and consumption expenditure.

Other consistency checks programmed into the digital version of the HFCS questionnaire in Austria would allow the survey to proceed only once an answer identified as incorrect or inconsistent had been amended. However, these so-called "hard" checks were only used in cases where a particular answer could definitely be ruled out.

If individuals stated, for instance, that they had lived in Austria for 40 years but gave their age as 30, the following 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 changing both variables).

4.4 Postinterview consistency checks

4.4.1 Expert data analysis

During the field phase of the second HFCS wave in Austria, the data of households deemed to be final by the survey company were forwarded to the OeNB in 15 batches. This means that the OeNB received household data roughly every three weeks during fieldwork. All batches of data were subjected promptly to

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

expert data analysis.[5] 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.

The datasets for households actually interviewed and those for households that refused to participate were 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 probably have created a bias toward certain households (e.g. housewife or pensioner bias) and distorted the data accordingly. The interviewers knew that the list of addresses was limited to the 6,308 households of the gross sample (see also chapter 6). This ensured that interviewers would not select the less difficult households and then move on to a new set of addresses. The incentive for interviewers to use the strictly limited address material as efficiently as possible was supported with a performance-related payment 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 they had made sufficient effort to survey the households they were assigned at the time. The decision to exclude subsequent draws (substitute households) is among the key criteria for a successful survey, and is moreover essential for ensuring the representativeness of the sample (see e.g. Vehovar, 1999).

Initial analysis of the information on individual households during fieldwork covered the data 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 remarks made by interviewers, as well as the date, time and duration of the interviews. This information enabled a quick initial assessment of the interview's quality. The microdata on every single household were checked for consistency regarding their content and reviewed by at least two analysts from the HFCS team. 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 address errors or misunderstandings. The shortcomings identified in this process were often minor in their nature, but four interviewers whose results were not up to the required standards (e.g. regarding nonresponse) were excluded.

4.4.2 Follow-up queries

If individual data analysis did not reveal the type of problem or how it could be corrected, households were contacted again by the survey company to clarify uncertainties and ensure that data were recorded correctly. Given the timely submission of interview results to the OeNB (around every three weeks) and subsequent check by the HFCS team, the survey company was able to address any queries to the surveyed households promptly. A typical case of a data problem that

[5] *The data evaluation was conducted with the aid of the results of the first HFCS wave as well as external data sources such as the EU-SILC (conducted by Statistics Austria).*

was easy to spot and did not require queries was rewriting a negative sight account balance as a (positive) liability (overdrawn account) while setting the value of sight accounts to zero (see also section 4.6). This was simply a matter of adhering to the recording conventions as to where such liabilities should be recorded. Decisions on follow-up telephone queries were always guided by the principle that any ex post data editing and the burden on participating households should be kept to a minimum. Many unusual results (e.g. particularly high asset values) were confirmed or else corrected in the course of queries. All in all, follow-up queries (by phone) were necessary to confirm specific details of some 400 households, which is a smaller percentage of households than in the first wave. This decrease is, above all, attributable to the substantial increase in the use of comment fields (as a result of the experience gained in the first wave).

4.4.3 Investigation of outliers

The checks on a case-by-case basis were aimed in particular at recognizing and processing outliers (exceptionally high or low values). These outliers were recorded above all for wealth variables, the size the household income or the size of the dwelling. Any outliers that were not removed from the dataset were generally not the result of interview errors but largely confirmed by the follow-up queries. Our recommendation for future studies based on HFCS data is therefore 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 filtering and consistency

During the field phase, the consistency checks programmed into the digital version of the questionnaire and the rounds of expert data analysis were complemented with detailed automated consistency checks.

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. The technical review also covered the questionnaire's complete set of filters to prevent programming errors leading to extensive and costly follow-up queries. Comprehensive tests of the questionnaire's programming prior to the start of fieldwork as well as a pilot survey of 55 households made it possible to largely exclude programming errors from the outset. Minor difficulties, e.g. incomplete filtering with regard to the question on additional borrowing (HB150$x) (see section 2.5.2.3) were identified and corrected in a timely manner.[6] These filter checks also ensured that the coding of variables was consistent throughout the questionnaire.[7]

4.5 Flags

All edits (and imputations – see chapter 5) were documented with flag variables, which indicate how the individual HFCS observations were established (see table 3 for a list of the flags used to classify the observations). To comply with interna-

[6] *This problem resulted from an update of the questionnaire during the field phase and was relatively difficult to identify as previous interviews showed correct filtering techniques before the update.*

[7] *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).*

tional requirements, some flags were aggregated for the international datasets (section 4.7). The flags used can be divided into five groups.

Group I

The flags allocated to group I were used to identify recorded information. Specifically, all observations recorded during the interview were flagged "1" while all filter missing values (".") were flagged "0." Information recorded in loops (see section 4.6.2.4) was paired with a flag of 2 if it had to moved in the iteration of a loop. In other words, flag 2 observations were retained in the dataset exactly as they were recorded, but assigned a new iteration number. The yes/no question on silent partnerships (HD1000) was encoded with "yes" for those households that held investments in a self-employment business, but had no active role in running the business and were not self-employed in this business. Those (few) observations in this variable were given a flag of 12.

Table 3

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
	12	Recorded as found in other source, not collected in survey
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 higher-order question
	1053	Not imputed, originally collected from a range
	1054	Not imputed, collected value deleted
	1055	Not imputed, value not collected due to a CAPI error
	1056	Not imputed, set to missing due to incorrect answer to a higher-order question
	1057	Not imputed, collected value deleted but range information available
	1058	Not imputed, set to missing due to red button
	1075	Not imputed, specific answer code
Group III	2050	Missing, set to missing for anonymization purposes
	2051	Missing, set to 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 to missing (".")
	3075	Edited, set on the basis of follow-up with household
	3076	Edited, set on the basis of follow-up with interviewer
Group V	4050	Imputed, originally "Don't know"
	4051	Imputed, originally "No answer"
	4052	Imputed, originally not collected due to missing answer to a higher-order question
	4053	Imputed, originally collected from a range
	4054	Imputed, collected value deleted
	4055	Imputed, value not collected due to a CAPI error
	4056	Imputed, originally value not recorded due to incorrect answer to a higher-order question
	4057	Imputed, collected value deleted but range information available
	4058	Imputed, set to missing due to red button

Source: HFCS Austria 2014, 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 a specific 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 1056). If the edited observation was available as a range, it was assigned a flag of 1057. If an observation was not available due to a CAPI error, it was given a flag of 1055. Observations that were not available because questions in a loop were skipped were flagged "1058" and special missing values were flagged "1075." In these cases, alternative information was collected.

For example, if gross income was unknown, but information on net income was provided, the variable for gross income was flagged "1075."

Observations with group II flags were not imputed (see chapter 5).

Group III

Group III flags identify observations and/or variables that were not recorded or that were recorded but later deleted from the datasets on account of anonymization requirements.

Group IV

Flags from group IV indicate an ex post edit of an observed value. The following types of ex post edits can be distinguished: edits as a result of logical inconsistencies (flag 3050); calculations that were adjusted using other information obtained in the survey, for instance with regard to life insurance contracts (see section 4.6.2.9 for details; flag 3051); coding that was subsequently adjusted on the basis of verbatim records (see section 4.6.2.3; flag 3052); edits made to delete a value and set the observation to missing, as in the case of duplicate entries (flag 3053); and information from telephone queries put to households (flag 3075) and interviewers (flag 3076).

Group V

Flags from group V mirror those from 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, which was subsequently imputed, rather than a specific figure, this observation was flagged with "4053" after multiple imputations. This ensures that all information can be tracked even after the imputations.

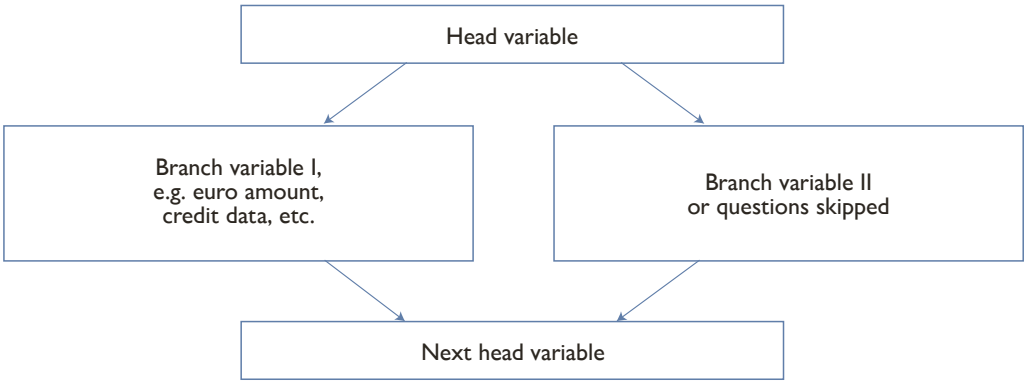
Chart 3 indicates how questions were typically structured in the HFCS questionnaire. Let us take employee income to give an example for the structure of question blocks[8] and the use of flags.

The head variable for recording employee income serves to ascertain whether or not a household has an income of this kind. If this yes/no question was answered with "Yes," the amount was recorded in the next question 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 (i.e. responded with "Don't 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 "1" or "0." If the

[8] *See chapter 2 for details of the structure of the whole questionnaire.*

Chart 3

Sequence of questions



Source: HFCS Austria 2014, OeNB.

response to a subsequent question (e.g. on employment) revealed ex post that a "No" given to the question on employee income was in fact incorrect, the initial response was corrected and flagged "3050" (Edited, set to modified value as considered incorrect or unreliable) and the corresponding variable for the value of the income was flagged for imputation. Following imputation, the value was then reflagged "4056" (Imputed, originally value not recorded due to incorrect answer to a higher-order question).

Or, if the question on a household member's highest education qualification (variable (A)PA0200) was answered by selecting the category "Other qualification" and if that answer was subsequently found to match one of the predefined categories, the observation was flagged "3052" (Edited, adjusted on the basis of the verbatim records) in the flag variable of the individual dataset.

This flag system allows the origin of every single observation in the HFCS to be tracked. To allow for the merging of datasets, no flags were used to encode the variables for identifying households and individuals, nor were the country codes and the imputation's iteration number flagged. The flags described here provide for a more detailed breakdown by category than those incorporated into the international HFCS dataset that can be obtained from the ECB. For reasons of international consistency, the flags were aggregated prior to being submitted to the ECB (see 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 queries and ex post editing. Specifically, respondents' answers were checked for plausibility against known benchmarks, including descriptive statistics (e.g. on average income) compiled on the basis of completed HFCS interviews and external sources of data. Moreover, the review process heavily relied on auxiliary variables that recorded information in aggregated form and/or in a variety of other ways.

Both interviewers that produced nonstandard results (see chapter 3) and follow-up queries made by the survey company were reviewed in particular detail. Expert judgment was generally used to resolve the following issues through ex post edits:

- Double entries: Cases where an inheritance, for instance, was recorded under both "Household main residence inherited" and in the "Inheritances received" 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.
- Implausible values: Values that remained implausible after follow-up queries had to be set to missing and were subsequently imputed.
- Often, information could be gained from the many additional comments made by respondents. If the additional comments made it necessary to change the collected information, the changes were made.
- Data entry errors by interviewers: In one instance, the contact month for an interview conducted in 2014 had been entered as January 2014 ("1"). This was changed to October 2014 ("10"), because the data on the relevant household were submitted in November and the preceding and following contact attempts had taken place in October 2014.
- Also, all data obtained through follow-up queries were used in this step to correct individual observations in the dataset where necessary.

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 queries and/or the use of auxiliary variables (such as verbatim records) failed to provide further information – inconsistent observations were set to missing and flagged for imputation. Inconsistent or implausible observations were processed with great care and only deleted if there was absolutely no doubt about the inconsistency.

4.6.2 Structural editing

4.6.2.1 Data cleanup

When answering the HFCS questions, respondents occasionally gave inaccurate answers but subsequently corrected those answers when they proceeded backward 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 "wrong" initial path through the questionnaire, however, remained in place for transparency reasons and had to be cleaned up (i. e. observations needed to be deleted ex post).

4.6.2.2 Currency conversion

Respondents could specify any amount in any currency (see chapter 2). The edits set out below relate both to specific amounts and ranges indicated by the respondents (predefined ranges had to be specified 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.[9] Some amounts were also given in Deutsche mark (DEM). These amounts were also converted at the irrevocably fixed conversion rate, namely EUR 1 = DEM 1.95583.[9]

In a few cases – in particular for 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 2014 exchange rate, while the total value 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[10] being used as exchange rates.

Individual cases[11] involving other currencies were converted with great care using the annual average of the respective currency's exchange rate to the euro. Two inheritance cases from before the introduction of the euro were first converted from Deutsche mark into Austrian schillings on the basis of the applicable exchange rate at the time and then from Austrian schillings into euro according to the fixed ATS/EUR exchange rate.[12] Likewise early amounts in terms of the reference year in Canadian dollars were first converted into Austrian schillings and then into euro.[13]

4.6.2.3 Verbatim records

For many questions, respondents were given the option of choosing the category "Other" and providing a verbatim response, mainly with a view to making the questionnaire as user-friendly as possible. Thus, a verbatim description could be recorded 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, which proved to be possible in the majority of cases. Wherever this could not be done, the initial categorization of the observation as "Other" was retained. Some data, such as data on the occupation (ISCO coding in the variable PE0300) of an employed individual or the main activity (NACE coding in the variable PE0400) of the company where the individual is employed, were collected entirely in verbatim form and coded ex post. All observations subjected to ex post edits on the basis of verbatim records were flagged "3052" (see section 4.5 for details on the flags).

4.6.2.4 Navigation of loops

As outlined in detail in section 2.6.1, some pieces of information were recorded in loops, which required interviewers to run through an identical set of questions for each individual item from a group of items owned by the household. Information on the following items was collected using loops:

[9] *See https://www.oenb.at/isaweb/report.do;jsessionid=31767F3B9E6FA661A8A4CD5CB700B5A7?report=2.12 (accessed on December 9, 2016).*

[10] *See www.oenb.at/en/Statistics/Standardized-Tables/interest-rates-and-exchange-rates/Exchange-Rates.html (accessed on December 9, 2016).*

[11] *Some observations of variables specific to Austria that are not contained in the internationally available core dataset required the use of historical OeNB exchange rates. However, these are not included in this documentation, as they are not part of the available data.*

[12] *The relevant exchange rate was taken from the OeNB's Statistisches Monatsheft of December 1998 (OeNB, 1998).*

[13] *The relevant exchange rate was taken from the Mitteilungen des Direktoriums der Oesterreichischen Nationalbank 1979 (OeNB, 1979).*

- mortgages on the main residence
- real estate assets apart from the main residence
- mortgages secured against these other properties
- unsecured loans from family and friends
- other unsecured loans
- businesses owned by the household
- inheritances and gifts

Below 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 secured against the main 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 were prompted to record the inheritances received in descending order of relevance for the household's current wealth situation. They were, however, instructed to indicate amounts as transferred rather than current amounts. After all, certain inheritances could have gained (or lost) more in value than others since the inheritance date; or inherited residential property might since have been passed on to children, causing it to be irrelevant for the household's wealth situation at the time of the interview.

Every variable within a loop that was replaced with observations recorded for the same variable in another iteration was flagged "2" (see section 4.5). Wherever a variable set to filter missing in one iteration was replaced with the same variable set to filter missing in another iteration, it was flagged "0" ("Not applicable (skipped due to routing)").

Skipping 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 to 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 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 from 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 and private businesses, the relevant summary responses had to be edited accordingly. For ease of reference, examples of these edits are described below on the basis of the section of the questionnaire dealing with other unsecured loans (see section 2.5).

In the 18 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 (1) indicated the outstanding amount only

in response to the summary question; or (2) both when going through the first loop of questions and in answering the summary question; or (3) neither during the first loop of questions nor in answer to the summary question. If the respondent had indicated the outstanding amount only in their answer to the summary question (variant 1), this amount was entered as the answer to the appropriate question (in the first loop) and the entry under the summary question was set to missing. If the respondent had indicated identical amounts in answering both the loop and the summary question (variant 2), the latter was set to missing since it was a duplicate entry.[14] Where no amount was given at all, neither within the loop nor in the summary (variant 3), only the summary question was set to missing.

In cases where a household had taken out two unsecured loans and had skipped questions within a loop,[15] the type of edit depended on whether the respondent had (1) specified the value of the highest outstanding loan and indicated an aggregate amount in response to the summary question; or (2) indicated outstanding amounts in response to both question loops and the summary question; or (3) specified an amount only in the answer to the summary question; or (4) given no amounts at all, neither in the answers to the loop item questions nor in the answer to the summary question.

If variant 1 was the case, the amount outstanding for 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 in the first loop. This, however, was only done if the sum total of the two outstanding loans exceeded the amount outstanding from 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 outstanding loans, but rather the amount outstanding for the second loan. In both instances, the summary question was subsequently set to missing. If variant 2 was the case, the amount given in response to the summary question was set to missing. If only the sum total of the two outstanding loans was given (variant 3), it was used as the upper bound for both the first and the second loan for the imputation model. This was the case for one household that held two other unsecured loans and had skipped some loop questions. If no amounts were given at all, neither in response to the loop questions for each of the two outstanding loans, nor in answer to the summary question (variant 4), the summary question was set to missing.

The editing procedure followed in cases with 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. This case did not occur in the example given above.

All edits were again flagged correspondingly.

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

[15] *In the second HFCS wave in Austria, only one household opted for this route.*

Summary questions

Every loop of questions ended with summary questions (see chart 2 in chapter 2). The variables for these questions exclusively contained information on any additional items above three per household. As indicated in chart 2, the summary questions were ultimately also put to all respondents who had refused to indicate the number of a given item in the household. In such cases of nonresponse, the information provided here was used for multiple imputations (chapter 5) and deleted from the dataset ex post.

4.6.2.5 Sight account balances and overdrafts

A few households misreported a negative balance on their household sight account as a negative value of sight accounts (HD1110). For this, however, a separate variable was available. In this area there were also occasional duplicate entries, as well as misplaced entries that subsequently had to be edited.

4.6.2.6 Rent variables

The HFCS questionnaire included questions on the amount of housing rent paid both excluding and including utilities. In the case of some households, the rent excluding utilities was higher than, or equal to, rent including such costs, which is logically impossible as housing cannot be "run" free of charge. Some of these households entered just the utility costs under the item "Rent including utilities." In the course of editing, these were added to the amount entered under "Rent excluding utilities" to obtain the "Rent including utilities." In the case of other households, the "Rent including utilities" was set to missing and flagged for imputation, with the "Rent excluding utilities" serving as the lower bound to the "Rent including utilities."

In addition, the item "Rent including utilities" was set as the upper bound for the variable "Rent excluding utilities" and used for imputations whenever the answer to the latter was not an amount (i.e. read "Don't know," "No answer" or "Rent excluding utilities unknown") (see also section 5.4.6 on the use of bounds in the imputations).

4.6.2.7 Agricultural businesses

As defined in the HFCS, farmers are owners of an agricultural business. Separating the asset components of households that own an agricultural business sometimes posed a problem to respondents, in particular with regard to their main residence and the investments in their business. Such cases, therefore, had to be analyzed separately. In this context, the extra questions and guidance added to the questionnaire for the second wave (see also section 2.6.3) proved very helpful during the various steps of data processing.

Some farmers did not report their agricultural business as an investment in self-employment businesses. For these households, data on investments in a self-employment business had to be imputed. The NACE code for such businesses was set to that for "agricultural businesses," and at least the individual who stated that he/she worked as a farmer was deemed to be employed in this agricultural business. The legal form of the respective business was edited to read "sole proprietorship." Use of the additional guidance to the respondents during the interview made it possible to reduce the number of such cases considerably compared with the first wave.

For all farmers, additional auxiliary variables were created for the combined value of the main residence and the agricultural business (business assets) as well as for the main residence's share in this amount. For households that were not able to separate their assets and specify the share themselves, information on the total value and on the main residence's share was used. For households that had specified both the value of their main residence and that of their private business, as required, the combined value and the share of the main residence was calculated. If information was partially missing, it was flagged for imputation (see section 5.3).

The category of agricultural businesses was subjected to case-by-case reviews. Particularly complex cases were clarified through follow-up queries and corrected where necessary.

4.6.2.8 Individual variables for investments in self-employment businesses

The variables for household members employed in a business owned by the household were edited as follows:

To be able to cover even unusually large households, variables were created for up to 18 individuals per household for the CAPI version of the questionnaire. The largest household successfully interviewed in Austria had only 8 members, however, so all variables in excess of that number were deleted from the dataset. Moreover, the coding was changed from yes/no questions for each household member (the type of coding used in Austria) to the list of individual IDs that were required for the internationally available dataset (which only contains six variables for individuals).

At the same time, all NACE codes for household members employed in the business were checked against the information contained in the P-file and corrected where necessary.

4.6.2.9 Life insurance policies

Information on assets held in life insurance contracts was recorded through questions ensuring that the answers were both as precise as possible and not very error-prone. In particular, there was no direct question on the total value of such assets, but rather a series of questions on the start of payments, the frequency of payments (monthly, yearly or single payment), the type of life insurance (benefits to be provided at the death of the policy holder or at a given date, or a hybrid form) and the amount of the current payments for every single life insurance contract in the household. For all life insurance policies with a set payout date and/or all hybrid policies, the value of the assets held in life insurance contracts was calculated as the cumulative sum of all payments. In cases where one or several details were not given, the remaining observations were used as bounds for the value to be imputed. Insurance policies (term-life insurance) which do not pay out capital if the insured lives beyond the term period do not constitute wealth; they were therefore excluded from this calculation.

4.6.2.10 Income variables

The following categories (variable name in parentheses) of personal income were recorded separately for every member of the household who was 16 years old or older:

- employee income (PG0100 and PG0110)
- income from self-employment (PG0200 and PG0210)
- income from public pensions (PG0300 and PG0310)
- income from private and occupational pension plans (PG0400 and PG0410)
- income from unemployment benefits (PG0500 and PG0510)

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

- income from public social transfers (HG0100 and HG0110)
- income from private transfers (HG0200 and HG0210)
- income from real estate assets (HG0300 and HG0310)
- income from financial investments (HG0400 and HG0410)
- income from private businesses or partnerships (HG0500 and HG0510)
- income from other sources (HG0600 and HG0610)

In the case of the first four personal income categories, respondents could indicate their net income if they did not recall their gross annual income (see chapter 2). Likewise respondents could indicate their net income from financial investments if they did not know their gross income in this category.

Where only a net amount was entered for individual incomes, the gross income was calculated with the aid of the Austrian finance ministry's gross-to-net calculator,[16] based on 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 any children, the province and the respondent's employment status (employed as a blue-collar or white-collar worker or retired).[17] 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 (as long as the legal requirements were fulfilled and the partner did not earn more than EUR 6,000 per annum).

Given the far greater scope for tax deductions for self-employed people, the gross-to-net conversion of income from self-employment was not generally based on the precise figures. Precise conversions were recorded only for annual incomes less than EUR 11,000, which are classified as tax-free, so that the gross amount is equal to the net. For all other values (for some 45 individuals), a range was created for imputing specific amounts by adding EUR 10,000 to and by subtracting EUR 10,000 from the amount converted subject to the conditions for white-collar workers. This range reflected the uncertainty that such a conversion entails, without losing the important information of the actual range within which the value is placed. In all conversions, all types of income were always included together. If a respondent indicated both employee and self-employed income, the sum of both incomes was used and converted. The employment status of the respondents was dependent on their main source of income. The total gross value was then split in proportion to the net income proportions indicated. To calculate the gross income from financial investments, 25% withholding tax (capital income tax) was added to amounts given for net income.

[16] *See www.bmf.gv.at/service/anwend/steuerberech/bruttonetto/_start.htm (German only) (accessed on December 9, 2016).*

[17] *"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.*

Table 4

Number and share of edits of gross employee income based on flags

	Number of persons	Share in %
Number of persons receiving employee income	2,638	100
Answer recorded, complete observation (flag 1)	1,067	40.4
Not imputed, originally "Don't know" (flag 1050)	32	1.2
Not imputed, originally "No answer" (flag 1051)	105	4.0
Not imputed, originally collected from a range (flag 1053)	326	12.4
Not imputed, collected value deleted (flag 1054)	7	0.3
Not imputed, set to missing due to incorrect answer to a higher-order question (flag 1056)	36	1.4
Not imputed, collected value deleted but range information available (flag 1057)	66	2.5
Edited, set to modified value as considered incorrect or unreliable (flag 3050)	1	0.0
Edited, adjusted on the basis of other collected (national) variables (flag 3051)	965	36.6
Edited, adjusted on the basis of the verbatim records (flag 3052)	2	0.1
Edited, adjusted on the basis of follow-up with household (flag 3075)	30	1.1
Edited, adjusted on the basis of follow-up with interviewer (flag 3076)	1	0.0

Source: HFCS Austria 2014, OeNB.

If the net amount was only recorded as a range, the upper and lower bounds were converted into gross values that were subsequently used in the imputations. All converted values were flagged "3051."

Using flags as a basis, table 4 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 2,638 individuals. 1,067 respondents (40.4%) expressed their annual income in gross terms. A further 32 respondents (1.2%) answered "Don't know" and 105 individuals (4.0%) opted for "No answer." 326 respondents (12.4%) specified their income amount using a range. For 66 individuals, a range could be calculated from other information given. The responses of 43 individuals (around 1.6%)[18] were edited, set to missing and flagged for imputation; the vast majority of these edits (for 36 individuals) were due to an incorrect head variable (flag 1056). 36.6% of the respondents (965) provided their net income, which was then converted with the aid of the finance ministry's gross-to-net converter. Expert judgment was used to edit the income of one individual. The responses of the remaining 33 individuals (around 1.2%) were corrected on the basis of follow-up queries and verbatim records.

4.6.2.11 ISCO and NACE classification

As required by the euro area blueprint questionnaire, the main occupation of respondents was recorded (in variable PE0300) using the occupation codes and titles set out in the *International Standard Classification of Occupations* (ISCO-08). Making individual members of each household classify their jobs themselves, however, would have been extremely difficult for respondents without any advance

[18] *Different from table 4 on account of rounding differences.*

knowledge of the ISCO codes, possibly giving rise to misclassifications. Therefore verbatim answers were recorded for the Austrian HFCS question on job titles and/or main job tasks. That information was later paired with the corresponding ISCO codes, using the German version of Statistics Austria's information on ISCO-08.[19] As required by the ECB, classification was based on the two-digit ISCO codes (major subgroups). To this end, the verbatim record of the job title and related main tasks was supplemented with individual data relevant for the ISCO classification (in particular, the respondent's level of education and the main activity of the company where the respondent worked). The variable PE0300 to be submitted to the ECB was first flagged "3052" (Edited, adjusted on the basis of the verbatim records and aggregated in a next step (see section 4.7).

Also, the main activity of the company (PE0400) where the respondent worked was first recorded verbatim and then assigned a single-digit NACE rev. 2 code.[20]

4.6.2.12 Highest education qualification

To account for latest developments in the *International Standard Classification of Education* (ISCED), the highest education qualification of all household members was recorded in substantially more detail than during the first HFCS wave. Respondents were not asked about ISCED categories, however, but were prompted to indicate their qualification based on Austria's education system. Additionally, the bachelor, master (including Magister and diploma degrees) and doctorate were recorded. The bachelor degree may have been inadequately chosen by some respondents. As this is a fairly new degree in Austria, respondents aged 40 or above are rather unlikely to have graduated with this degree. Therefore, these individuals were assumed to have a master degree (including Magister and diploma degrees); the corresponding variable was flagged "3051." As the national degree hierarchy was aligned with the ISCED codes (see also section 4.7), this aspect does not play a role for the international dataset, though.

4.6.2.13 Exclusion of successful interviews

For various reasons, the final data do not include those from altogether 42 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. children's homes, retirement homes, prisons). A number of households that had been interviewed successfully were excluded from the survey because the respondents were living in student housing. The four households in this group were edited out and flagged as "Not belonging to the gross sample." 33 other households that were not interviewed successfully were likewise eliminated from the gross sample. In addition to households living in institutions, these households also comprised addresses whose occupants had died and addresses that were used commercially.

[19] *For further information, see www.statistik.at/web_de/klassifikationen/oeisco08_implementierung/informationen_zur_isco08/index.html (German only) (accessed on December 9, 2016).*

[20] *For explanations, see www.statistik.at/web_en/classifications/implementation_of_the_onace2008/index.html (accessed on December 9, 2016).*

- Households with an excessive proportion of nonresponse items: This group, which comprised 38 households, had to be deleted from the dataset because the respondents refused to answer too many questions. The observations for these households were edited to "Interview completed, but rejected after fieldwork" and assigned a nonresponse weight of zero (see section 7.2.3).

4.6.2.14 CAPI errors encountered with the questionnaires

Additional borrowing (HB150$x)

Because of a CAPI error, a total of 39 households were not asked the question on additional borrowing (HB150$x), although they should have been routed to this question. These observations were set to missing with a flag of 1055 and released for imputation. Following imputation, they were reflagged "4055."

The few other problems encountered with the programming of the questionnaire (section 2.5.2.3) did not require any further editing measures.

4.7 Formatting and editing after multiple imputations

Any information collected at a greater degree of granularity in Austria 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 paired with ISCED 1997 codes[21] and classified as ISCED level 0 ("Compulsory education not (yet) completed"); ISCED level 2 ("Compulsory education completed"); ISCED level 3 ("Apprenticeship (vocational school)" and "Other vocational middle school"); ISCED level 4 ("Nurses' training school", "Secondary academic school – senior classes" and "Vocational or technical school – college, high-school graduates' training course"); ISCED level 5 ("Master craftsman – works master training" as well as "University, academy, technical college – bachelor" and "University, academy, technical college – Magister/diploma degree/master") and finally ISCED level 6 ("University: doctorate").
- Employment status/relationship: More detailed categories were aggregated.
- Main residence – tenure status: More detailed categories were aggregated.
- Reasons for refinancing: With regard to collateralized loans, "For the conversion of a foreign currency loan" was available in Austria as an additional category for this variable. This category was added to "Other" in the international dataset.
- Loan repayments: 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 to Austria.
- Use of additional real estate property: In this variable, "Buy-to-let apartment" was available as an additional category in Austria; in the international core dataset, this category was added to "Other."

[21] *For explanations, see www.statistik.at/web_de/klassifikationen/klassifikationsdatenbank/weitere_klassifikationen/bildungsklassifikation/104092.html (German only) (accessed on December 9, 2016).*

- 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 allocated to "Other."
- Rejection of a loan application: Information of this type was recorded in three variables with multiple responses; it was then aggregated into two variables.
- Business' legal form: More detailed categories were aggregated.
- Silent partnerships: The yes/no question recording silent partnerships was not put to households that held investments in a self-employed business but were not actively involved in its management or working for the business on a self-employed basis. For reasons of consistency, the respective variable (HD1000) was encoded with "1 – yes" ex post and flagged "12."
- Investments in savings plans with building and loan associations and life insurance policies: Data recorded on these two investment methods were aggregated into savings (HD1200 and HD1210).
- Type of assets received (survey questions on inheritances and gifts, HH030$a-i): The sequence based on values was abandoned.
- Provider of assets (survey questions on inheritances and gifts): More detailed categories were aggregated.
- Purpose of saving: The sequence ordered by relevance was abandoned.
- Paradata: The variable on alarm systems and other security measures (ASC0700a-h)[22] was aggregated (the possibility of multiple responses was abandoned).

In Austria, more specific flags were used in some areas (see also section 4.5). To conform to international standards, these flags were, in general, aggregated as follows:

- flag 1057 was recoded as 1053
- flag 1058 was recoded as 1051
- flags 3051, 3075, 3076, 1075 and 2 were recoded as 1
- flag 3052 was recoded as 11
- flag 3053 was recoded as 0
- flag 4057 was recoded as 4053
- flag 4058 was recoded as 4051

The following variables are exceptions to this rule:

- Income variables[23] flagged "3051" after net-to-gross conversions were reflagged "13" (rather than "1") following aggregation in line with international requirements.
- The variable featuring rent excluding utilities (HB2300) was flagged "1075" to identify the special cases when only the rent including running costs was known to respondents. These costs were subsequently imputed, with the rent including utilities serving as bounds, and reflagged "4053."

In analogy to the first wave, some of the additional data over and above those of 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 2017. The additional information includes additional variables, as well as a

[22] *These variables are not contained in the international core dataset.*

[23] *This concerns the following variables: PG0110, PG0210, PG0310, PG0410, HG0410.*

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 clearly not been recorded correctly. In cases of ambiguity, the first possibility considered was to conduct follow-up inquiries by phone. This option allowed many observations either to be corrected or to be confirmed.

Knowledge of the steps undertaken to check the consistency of the data is essential both for the analysis of the data and for understanding how the observations originated. 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 another way.

The online appendix, which contains the information provided here on the edits and consistency checks applied in the HFCS in Austria, includes a list of the consistency checks programmed into the digital version of the questionnaire.[24]

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