GUIDELINES
FOR CLEANING AND HARMONIZATION OF GENERATIONS AND GENDER SURVEY DATA

Andrej Kveder
Alexandra Galico

Table of contents

1 Introduction

The Generations and Gender Programme (GGP) aims to improve cross-national comparative research on demographic and social developments. In light of this, the survey organizations implementing the Generations and Gender Surveys (GGS) need to follow consistent practices of data collection and data processing.

This document guides national focal points in the countries implementing the Generations and Gender Survey (GGS) through all stages of processing micro-data from its collection to analysis, focusing specifically on data cleaning and harmonization procedures. It begins by describing the initial post-fieldwork processing of the questionnaires and survey data to minimizing errors before the editing stage. The second section discusses methods of data editing and cleaning, and gives examples of common editing practices. Finally, the last section discusses harmonization practices and the format of the harmonized file, which are crucial for ensuring the international comparability of GGS datasets.

2 Data Processing

Data processing refers to the practice that transforms raw data from field collection into a cleaned and corrected state so that it can be used for analysis. Some aspects of processing depend on the method of data collection. The GGS uses either a Paper and Pencil Interview (PAPI) or a Computer Assisted Interview (CAI). For PAPI interviews, an interviewer writes down the responses to the posed questions on the paper version of the questionnaire. With CAI the interviewer reads the questions from a computer screen and enters the responses directly into the computer. Because of these differences, the PAPI questionnaire requires further processing stages than CAI as effort is needed to transfer the paper responses into a computer readable format. These data processing steps are highlighted and discussed throughout this section.

Data processing can be considered as a two stage process. The first stage prepares the PAPI documents for data capture, where data from the PAPI questionnaire is converted to a computer-readable format, while the second stage identifies and amends errors and inconsistencies in the data file.

The first stage or pre-edits of processing survey data includes:

- Document Reception and Control,
- Document grooming (PAPI only) and
- Data Capture (PAPI only).

The main aim of this processing stage is to prepare the file for editing stages.

After data capture, the final stages of data processing should include the following steps:

- Editing,
- Coding and
- Final file preparation (Harmonization).

Pre-editing stages represent a vital and often overlooked part of data processing as poor implementation of these stages will impact data editing and thus compromising data quality. It is recommended that the following minimum stages be implemented within document processing warehouses, to minimize errors in the final data file.

2.1 Pre-editing

2.1.1 Document reception & control

Document reception and control system is a diary for tracking the selected sample units, distributed questionnaires and other fieldwork aspects of survey data collection. Most often it is implemented in the form of a computer file which is updated daily, based on the reported information from the field, such as completed questionnaires returned. This procedure is put in place to ensure a tight control of all documents distributed to the field. PAPI questionnaires often have many paper components which are easily lost or misplaced during either the fieldwork or the processing stages. Lost or missing questionnaire components can severely affect the quality of data and can present serious challenges to the survey.

One possible example of the document reception and control is implemented at Statistics Canada. Survey department at Statistics Canada developed the so called 'Master Assignment Control' (MAC) file. This file contains a record of each sample unit assigned to the field and provides details on documents expected for return. This file should be updated daily and contain the following minimum information:

- Sample ID number – the unique identifier code given to each respondent.
- Interviewer Number – the unique number assigned to each interviewer.
- Fields for each document required for return.
- Notes.

Sample ID	Interviewer No.	BQ Returned	Core Booklet Returned	Core Score Sheet returned	Main Task returned	Newspaper	NOTES
10530001	105	"Yes" or "√"or "NA"	"Yes" or "√"or "NA"	"Yes" or "√" or "NA"	"Yes" or "√" or "NA"	"Yes" or "√" or "NA"	
10540002	059	"Yes" or "√"or "NA"	"Yes" or "√"or "NA"	"Yes" or "√"or "NA"	"Yes" or "√" or "NA"	"Yes" or "√" or "NA"	
ETC.							

Note: Fields included to indicate whether a document has been received could have the following values "Yes" or "✔" or NA if the case was a non-response.

Figure 1

This file can be easily created within Microsoft excel and/or other software packages developed for spreadsheet processing or database management.

Documents which have not been returned from field collection should be flagged in the computer file and attempts should be made to find the missing questionnaire components.

2.1.2 Document grooming

Document grooming is a process where data controllers check the quality of the written responses for legibility. This is done to ensure that PAPI questionnaires are legible for data capture. PAPI questionnaires can be filled with illegible writing and ambiguous marks which may compromise the quality of data capture. The following examples are some grooming procedures which should be preformed during this processing step:

Verify sample and ID numbers – Sample IDs should be matched manually to a master list of ID's. Any discrepancies should be further investigated and resolved prior to data capture.

Verify all variables and written-in responses – The clarity of written-in responses should be checked question by question to minimize the number of data capture errors. Below is a list of common problems likely within the PAPI version of the GGS:

Numerical responses - such as income or expenditure information, should be written legibly and clearly, with no ambiguity on the amount written.

Dates - Ensure that dates are provided in the format allowed by the questionnaire, i.e. in two-digit format for months and four digit format for years. Only 1 – 12 are allowed for month codes with the exception of 21, 22, 23, 24, and 25 which are codes for general seasons. Recorded responses should not exceed these values.

Births – Birthdates should be legible and valid. Checks on the consistency of answers should be made at this stage. For example, if a birth date is given as 21 December 1984 and age listed as 25 there is an error.

Single answer questions – verify that not more than one answer is provided. Ensure that the mark is clear and recognizable without ambiguity (Figure 2).

409. Couples deal with serious disagreements in various ways. When you have a serious disagreement with your partner/spouse, how often do you [keep your opinion to yourself]? Please choose your answer from the card.

Show Card 408: Frequency.

	never	seldom	sometimes	frequently	very frequently
a. keep your opinion to yourself	1	2	3	4	5
b. discuss your disagreement calmly	1	2	3	4	5
c. argue heatedly or shout	1	2	3	4	5
d. end up becoming violent	1	2	3	4	5

Figure 2

Written responses – For example, questions which require further explanation or 'other' fields. Text should be legible and in block capitals.

312. a. Was your partner/spouse born in [the UK]?

1 – yes → *go to 313* 2 – no ↓

b. In which country was he/she born?

Country of birth _____

Figure 3

Allowable Values or the acceptable range of values for the question. Only responses within these ranges, for example 1-10 for the below question should be allowed.

315. Which of the items on the card best describes what he/she is mainly doing at present?

Show Card 108: Activity

1 – employed or self-employed
2 – helping family member in a family business or a farm
3 – unemployed
4 – student, in school, in vocational training
5 – retired
6 – on maternity leave, parental leave or childcare leave
7 – ill or disabled for a long time or permanently
8 – looking after the home or family
9 – military service or social service
10 – other

Figure 4

Once these preliminary stages have been implemented, the PAPI questionnaire is ready for data capture.

2.1.3 Data capture

Data capture is the phase of a survey where hand-written responses recorded on the PAPI questionnaire are converted to a computer-readable format. This information can be captured manually by keying (i.e. individual data entry), or automatically by using scanning or optical character recognition (OCR) devices. Larger surveys are more likely to use automatic methods of data capture due to the large volume of data that needs to be processed. Data capture is not necessary for CAI questionnaires as the data is already in a computer readable format.

Data capture can be prone to errors unless it is properly controlled. To reduce keying or entry errors, verification of the entered data is necessary. Independent re-key verification is the primary method of quality control. It involves independent re-keying of all or a sample of the work of certain data-entry staff and comparing the newly keyed entries with the original ones. Discrepancies between these keyings can be flagged and checked by a computer. Any unresolved inconsistencies are checked manually.

Data capture represents an optimal opportunity to begin implementing pre-edits of the dataset. Most data-entry software programmes (e.g. SPSS Data Entry™) allow for the creation of certain rules or controls for each variable which can check for ranges of certain variables, logical consistency and routing of the questionnaire. For example, if a value exceeds the range of possible values or an answer field breaks the routing pattern of the questionnaire it will be brought to the attention of data entry staff during the entry process and will be flagged for further investigation.

After data capture, there is no difference between PAPI and CAI questionnaires and the file is now ready for editing.

2.1.4 Flagging

Flagging is an important component of data processing. It is simply a practice in which erroneous or suspected results are marked in a computer file for investigation. Flagged results are stored separately from the data in a 'flag file'.

These flag files can vary in sophistication, with the simplest containing a list of the original variables with a code, for example 0 or 1 denoting correct or suspected incorrect responses. More complex programmes will assign a specific code based on the type of error. For example different codes will be assigned for routing errors, duplicate errors and suspected anomalies.

Flagged results should be manually checked to determine their true nature. A file containing a list of all flagged results should accompany the released data file.

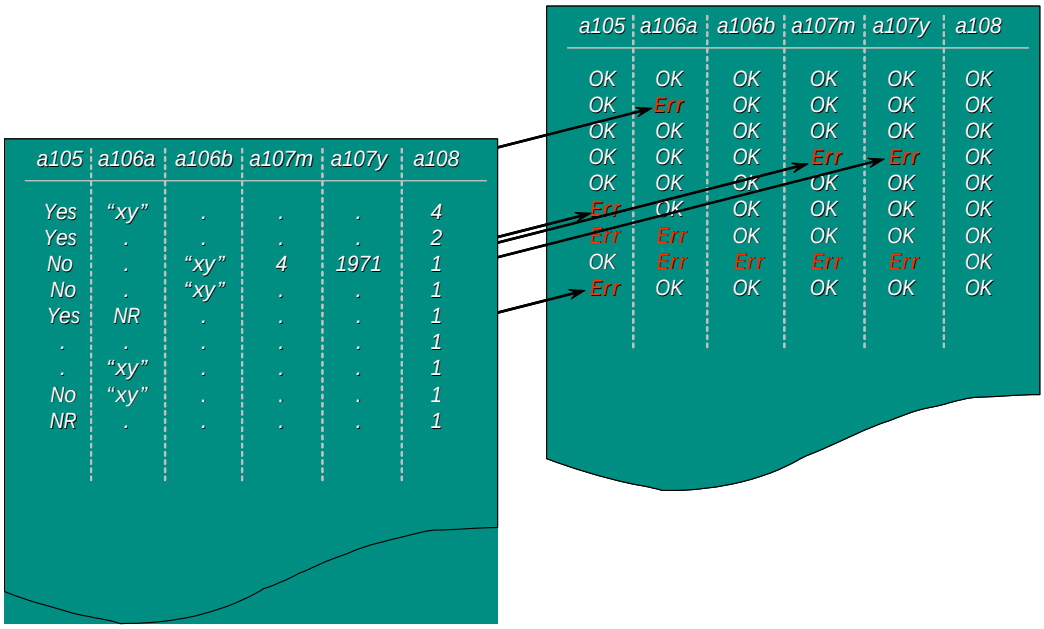


Figure 5 – Flag file example

2.2 Editing

Once all of the data has been captured into a computer readable format, editing determines the plausibility of data captured responses. Editing is an important stage of data processing as it seeks to eliminate redundancies and inconsistencies within the dataset. The editing process can be divided into two phases; a screening phase where errors are identified and a treatment phase where erroneous or missing values are corrected and/or verified.

In the screening phase, the data should be checked for the following:

- Duplicate or redundant records.
- Identifying valid response cases
- Identifying and finding missing records
- Inconsistencies

2.2.1 Duplicate or redundant records

The dataset should be purged of all duplicate or redundant information. Commonly, there are two types of duplicate records; full duplicates and ID duplicates.

Full duplicates are two or more identical records with exactly the same information in all fields. These are common in both PAPI and CAI environments and are often identified as having identical Sample ID numbers. The rule of thumb is the record with the least information is deleted, if they are identical then one of the records is arbitrarily retained.

ID duplicates are two records with the same Sample ID that contain different data. These are especially common in PAPI environments. Due to the nature of these errors, special attention is required and it is highly recommended that reference is made to the original questionnaire to determine the correct record.

2.2.2 Non-response

Non-response within surveys is common and ranges from; unit, item and partial non-response. Unit non-response refers to a situation where there is no data available on the

sample unit. For example, the household is selected for sampling but the interviewers are unable to establish contact or the target individual refuses to participate. These should be excluded from the dataset before processing begins. Item non-response on the other hand, occurs when a respondent fails to provide data for one or more items in the questionnaire. This should be included within the dataset and indicated by appropriate non-response codes. Partial responses, i.e. the result of a break in the interview by the respondent before completion, can often remain hidden until the editing phase. Therefore it is important to identify the broken responses before continuing with the rest of editing tasks. The partial response cases should be flagged in the main data file as such and not deleted.

2.2.3 Missing records

Since some countries implementing the GGS have many paper components of the questionnaire, it is possible that some records may be lost or misplaced. As such, it is recommended that a good document control strategy is put in place to eliminate these occurrences (see section 2.1.1).

Additionally, it may be worthwhile to compare the original sample file (the file before data capture) to the survey data file (the file at data capture) to identify any missing records. Missing records should be identified by Sample ID numbers and should be flagged for further investigation. Follow-ups with the interviewer should be made to locate the missing records.

2.2.4 Inconsistent results

Identifying inconsistencies should be based on knowledge of expected ranges of normal results and an understanding of common error types in order to screen for them. Screening this data is based on the principle that data points which do not fit 'pre determined rules' will require further investigation, correction or explanation. These rules are based upon the values that individual data items can take on, how these should relate to each other and how the data set 'should look'. Some common errors within the GGS relate, though not exhaustively to the below list:

- Logical consistency,
- Questionnaire routing – skip patterns,
- Life course events,
- Missing values,
- Birthdates / numerical values and
- Data non-conformity with expectations (anomalies).

2.2.4.1 Logical consistency

Logical consistency refers to the overall consistency of answers given by a respondent, in that there are no contradicting responses within the answer set. Editing for consistency requires examining the uniformity of responses between different variables. For example, Statistics Canada applies consistency rules based on individual variables. This can be done for a variety of different variables including; income, sex misspecification, birth history, education and employment.

Table 1

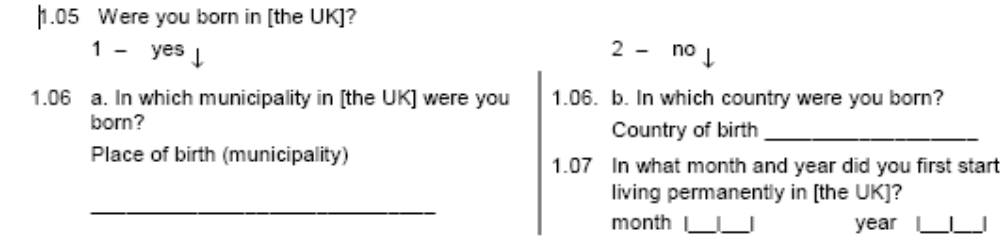
Edit #	Question	Rule
E1	Age	Age at interview is 16 or over
E2	Year of immigration	Year of immigration cannot be less than year of birth
E3	# of years of formal education	Verify # years of education and level attained are valid. Countries will have to define tolerance levels for these questions (i.e. upper and lower boundaries for number of years required to complete a programme)
E4	Age when took schooling towards…	Verify that Age entered is not greater than the Age of respondent
E5	Age when complete highest level of education	Verify that Age entered is not greater than the Age of respondent

An example of a common error which should be picked up by consistency editing is the detection of biologically impossible results, such as respondents who have children who are listed as older than them.

These consistency rules can be programmed into data-entry software which in turn automatically flags for inconsistencies within the dataset. These flagged results should be investigated and attempts should be made to determine the true nature of the error. If the true nature of these results cannot be ascertained, they should be flagged and included with the submitted file.

2.2.4.2 Questionnaire routing

The flow pattern of the survey questionnaire involves many 'skips' based on answers provided by the respondent. For example (Figure 6), if a respondent answers 'no' to question 1.05 then question 1.06b & 1.07 should be answered. If a respondent answers 'yes' to question 1.05 then the next question they should answer is 1.06a. This represents a valid skip in the questionnaire and should be identifiable within the data. Editing practices need to follow and replicate this flow pattern of the questionnaire. This is mainly a concern with PAPI questionnaires as CAI technology is programmed to automatically follow the flow of the questionnaire based on the respondent's answers.



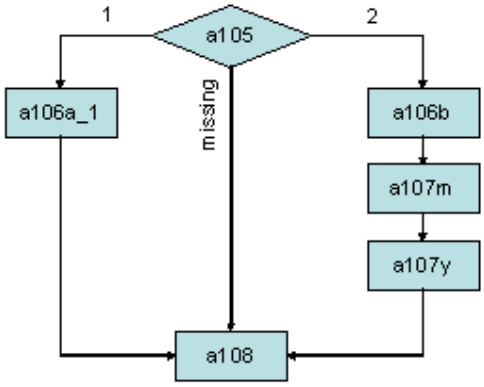


Figure 6

Questions which represent valid skips due to the questionnaire flow are converted to blanks or a system missing values. However, fields skipped due to non-response have to be set to one of the non-response codes i.e. 97, 98, 99 or an alternative code with the label 'omitted'. Often the information indicated by the filter variable and the subsequently recorded values do not match – e.g. the filter indicates the skip, however, some subsequent question that should have been skipped have recorded values. In the case of such a mismatch of recorded information there is a dilemma of which information to trust. Is the information provided with the filter variable more credible or is it the subsequently recorded information – e.g. if the respondent answers question a105 as 'yes' but also answers questions a106 and a107 which can be assumed as the correct response? From the operational point of view of data cleaning, trusting the information provided in the filter variable (e.g. a1.05), is more efficient and easier to handle. In addition a single question (filter) presents relatively low response burden, is more exact in its formulation and easier to understand, hence the answer provided could be considered more credible than the information collected from the subsequent questions.

Data editing related to the questionnaire routing therefore needs to follow the principles depicted in Figure 7.

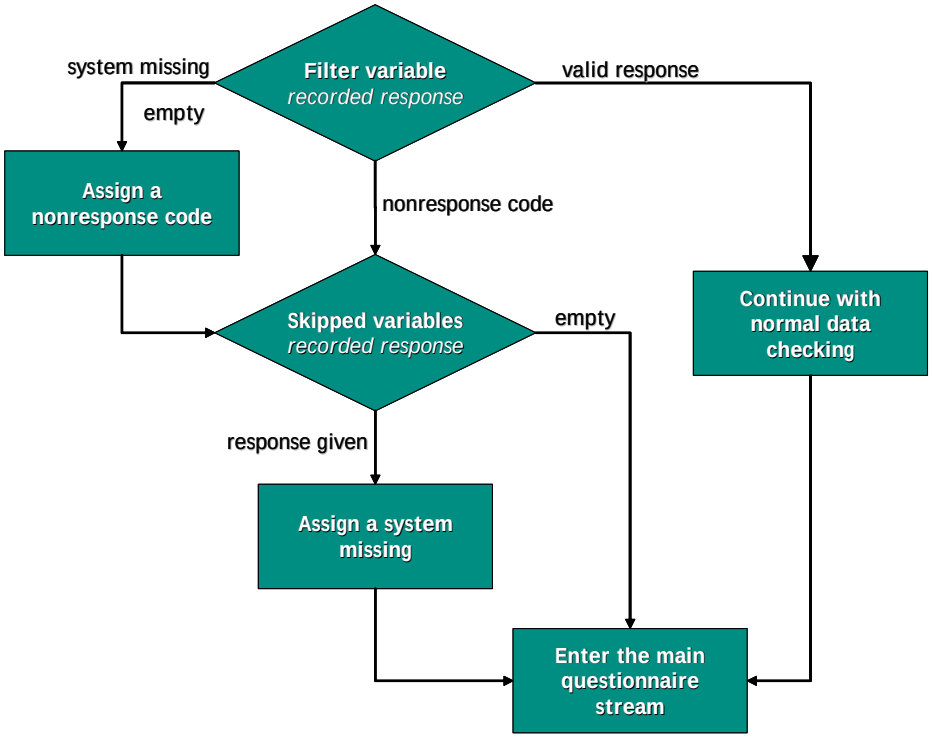


Figure 7

Any data edits following the routing checking procedures are not modifications of data values. The only possible editing actions are:

- Deletion of the erroneously entered value (i.e. substitution of a value with an empty cell);
- Substitution of an erroneous value with a nonresponse code;
- Substitution of an empty cell with a nonresponse code.

2.2.4.3 Missing values

Identification and verification of missing values is an important component of data editing. Some values are supposed to be missing and reflect the routing i.e. valid skips in the questionnaire. These should have system missing codes (or empty cells) and be easily identifiable within the dataset. Additionally, item non-response, i.e. missing information due to refusal or other types of errors, should be indicated by non-response codes:

- 7, 97, 997, 9997, etc. – for "don't know",
- 8, 98, 998, 9998, etc. – for "refusal" and
- 9, 99, 999, 9999, etc. – for "not applicable" or "other missing".

Especially routing errors can include large sections of missing data. Ultimately, some of this missing information will be identified as more important and will be considered a priority. In this case, attempts should be made to locate the original questionnaire to rule out any data-capture error. If this information cannot be located, then it is recommended that the respondent is re-contacted.

2.2.4.4 Life course events

The cleaning of the life course events should be minimal and should include solely editing to the extent that the provided information is valid. However, the recorded information is not necessarily logically correct. No imputation of dates should be attempted. With respondent recall of life course events, it is typical that respondents remember the year of

the event but not necessarily the month. Interviewers should be trained to propose seasonal codes highlighted in the questionnaire if this occurs (i.e. 21 -25 seasonal codes). Years should be coded in a four digit format.

2.2.4.5 Anomalies

Anomalies or values outside of the normal range of possibilities are common within survey data. Screening methods differ between categorical and continuous variables. Similar to other editing processes, rules can be input into data processing software and applied to the dataset to screen for anomalous values.

Categorical Variables:

Screening categorical variables for anomalies includes inputting an acceptable set of values or possible outcomes for the variable. For example, variable "respondent's sex" should have only two valid codes 0 and 1 denoting male and female, or 1 and 2 depending on how they are coded. Any value which does not fit into this pre-specified set of valid values is considered to be erroneous and should be flagged and further investigated with reference to the original questionnaire. If this does not resolve the issue, then the value should be changed to a non-response code.

Continuous Variables:

Screening of continuous variables is less straightforward as it is difficult to ascertain the true nature of suspected anomalies. Many surveys implement screening cut-off points for each variable with set range checks. Further, soft and hard cut-off points are often set out for each variable to separate impossible results from suspect results. However, this can be problematic as suspected errors may fall in between the soft and hard cut-offs and diagnosis will be less straightforward.

Software can scan the dataset for these range, consistency and routing checks and flag suspect variables. Printouts of variables not passing these checks should be further investigated; if the true values cannot be determined then efforts should be made to treat these anomalies. Treatment of these anomalies is discussed in the next section.

2.2.5 Treatment phase

Once errors and missing values have been identified within the dataset they can be treated with the following options:

- correcting,
- deleting or
- leaving the values unchanged.

In general treatment of errors should not include the imputation of values. Substitution of missing information with an estimate should be avoided, especially if the imputation is simple in nature (i.e. substitution of missing information with sample mean or similar point estimate). If countries do decide to impute missing or erroneous data, the imputation should be well documented and a copy of the new, imputed variable, as well as the original one should be included within the data file.

Editing is a necessary step within any data processing activity. However, extensive editing that does not lead to quality improvements or 'over-editing' can actually introduce additional bias into the dataset. This section provides some basic rules which should followed within the treatment stage of data editing.

- Any imputation is discouraged. Erroneous and suspected results should be flagged and submitted with the data file.
- Impossible values should never be left unchanged but should be corrected if a correct value can be found. Otherwise they should be deleted and changed to a non-response code.

- Statistical techniques can help with identifying anomalies and determining the influence of anomalous data points on analysis results. This can help with the decision on whether or not to leave erroneous values unchanged. If the value is changed, the results should be flagged in the data file and the original not changed variable should be included.

Data editing is not required by the survey and it is done at the discretion of the participating country. Identifying and flagging known and suspected errors is the most important step in data cleaning. Values which have been edited or altered should be flagged and reported with the data file submission.

2.3 Coding

Coding is a classification process in which responses are assigned to specific categories. These categorical variables have a unique meaning without any ambiguity and should be based upon the codebook supplied with the survey questionnaire. Any alteration to the codes supplied with the questionnaire should be noted and submitted with the dataset.

The majority of the coding, especially the coding of closed sets of answer categories, is part of the preparations for data entry. However the coding of open-ended responses need to be performed once the data have been entered into a digital form. This coding can be conducted manually by an operator or automatically by specially designed coding software. Coding can be most susceptible to errors when proper coding rules are not applied as it is a highly subjective activity. In addition to often subjective actions of the coders, the "written in" answers can also be a source of error. As the responses may not always be adequate to assign a code number unambiguously, coders need to use their judgment and "read between the lines". This is often the case when the codebook does not contain examples of all possible open-ended responses and there is disagreement amongst coders over the proper code to assign.

Coding of open-ended responses is the responsibility of the country implementing the survey. The codes of occupation and education variables should be based on international coding standards described in the ISCO 'International Standard Classification for Occupation' and ISCED 'International Standard Classification of Education'.

2.3.1 Coding of multiple response questions

Multiple response questions need to be coded in a standardized manner. The most versatile form of standardized coding is by transforming the questions into a sequence of dichotomous variables representing all items within the question. All the indicator values should have a unified response pattern, i.e. same number of respondents in valid categories and have same coding of non response. Value '1' indicates that the respondent has selected the item and value '0' that the respondent has not selected it, but has in other sense provided a valid response. The pattern of nonresponse codes need to be standardized across all the indicator variables – if a respondent refused to provide the answer to a given question, the nonresponse code should be recorded in all the indicator variables for that particular case.

Another possible standardized way of recording the multiple response questions is to organize the provided information into response variables. Response variables represent separate responses; hence the number of response variables created should be equal to number of allowed responses or at most equal to the number of possible answers. Each response variable has a set of valid response values, which corresponds to the possible responses given in the questionnaire. The nonresponse codes need to be recorded in first response variable only – if a respondent refused to provide the answer to a given question, the nonresponse code should be recorded in first response variable for that particular case and should be empty (system missing) for all the other response variables available.

The latter method of recording multiple responses is recommended for the GGS data. Although the first approach is more general and easier to handle from the analyst's perspective, the second method offers better comparability across multiple datasets. The versatility of the second approach can be shown when adding a new country specific

response category to the international database. When the response variable method is used, the only necessary intervention to the international database is then the update of the value labels for the response variables in question. In the case that the indicator variable method would be used, then the intervention would require additional variables to be included into the international database in order to correctly represent the additional country specific response. The response variable method clearly requires less processing then the indicator variable approach and is therefore in the case of internationally comparative datasets superior.

6.09 Did you or your partner/spouse use or do any of the things listed on this card to prevent pregnancy at the time it occurred? Please name all of the things you used or did.

Show Card 6.09: Contraception

[Comment: Country-specific list that should cover the range of available methods with commonly understandable labels in a country]

1 – condom 2 – pills 3 – intra-uterine device (coil, loop) 4 – diaphragm/ cervical cap 5 – foam/ cream/ jelly/ suppository 6 – injectables (e.g. Depo-Provera) 7 – implants (e.g. Norplant) 8 – Persona 9 – hormonal emergency contraception afterwards ("morning-after pill") 10 – withdrawal 11 – safe period method (rhythm)	→ ***go to 6.30***
0 – did not use or do anything (not on the card)	→ continue with 6.10

Figure 8 – Example of multiple response question

Let's look at the following example of multiple response question (Figure 8). In case of question a6.09 eleven response variables should be constructed with valid responses from 1 to 11. The first response variable would have an extended valid response set to include value 0 for non-use of contraception as well as 97, 98 and 99 for valid nonresponse codes.

3 Harmonization

Harmonization aims at achieving a clear and comparable format of the GGS micro-data files. While most of this work is supposed to be carried out centrally, countries are strongly recommended to implement the following minimum steps to improve comparability of their data.

3.1 Variable naming

Variable names are unique identifiers given to the variables so they are recognizable within the dataset. Variable naming should be consistent between all countries and is based on the question numbers supplied in the GGS Wave 1 Full Questionnaire. A full list of question numbers, variable IDs (Var-IDs) and variable descriptions can be found in the Data Availability Report (http://www.unece.org/pau/ggp/materials.htm).

Naming variables entails the following:

- Var-ID begins with the letter "*a*" followed by the question number in the full GGS questionnaire. For example, question, 1.05 as it appears in the questionnaire should be named as *a105* in the file.
- For questions which include dates, years and months, the variable names end with *y* and *m*, respectively. For example, *a107m* and *a107y*.
- Variables relating to event history information are ended by an underscore followed by the sequential number of the event. For example, *a212_1* is the sex of the first mentioned non-residential child.

- Frequency information is recorded in two variables; the frequency of an event and unit. The variable name of the unit variable is terminated with letter "u".

  3.25 How often do you see him/her?
  _____ times per: W M Y

  For example, question 3.25 should be recorded by two variables *a325* and *a325u*, where *a325* holds the information about the frequency of visit and *a325u* the information about the unit (i.e. 1'per Week', 2'per Month', 3'per Year').

3.2 Variable and value labels

Variable and value labels for all the variables should be included in the data file based on the specified values in the Harmonized Data File (HDF). The variable label is a short description of the variable based on the text of the question as presented in the questionnaire. Value labels are short descriptions of the responses given and are attributed to the underlying numerical values.

If different answer categories were used during the national implementation of the survey, the data should be recoded so that categories have the same values as specified in the HDF. This may not always be possible. If this is the case, the following example should be followed.

Consider that a country uses four categories in variable *a122*, two of which are different from the standard list and the other two values are in a different order. To illustrate:

1.22 Does your household own or rent this accommodation or does it come rent-free?

1 – owner .....→ continue with 1.23
2 – tenant or subtenant, paying rent .....→ continue with 1.23
3 – accommodation is provided rent-free .....→ ***go to 1.30***
4 – other.....→ ***go to 1.30***

Country specific values include the following:

1: 'tenant or subtenant, paying rent'
2: 'owner'
3: 'staying in a hotel'
4: 'homeless'

It is recommended that the values are recoded so that values 1 and 2 follow HDF specification and 3 and 4 are recoded so it is made clear that they differ. Value 1 'tenant or subtenant, paying rent', should be recoded to 2 and 2 'owner' should be coded to value 1, so the variables match the original question. Values 3 and 4 should be recoded to 333 and 444 (or similarly obviously impossible values) so it is immediately clear that this variable does not comply with the HDF specification.

When recoding to the new values it is important to also take the missing values defined in the HDF into account. Country-specific values should not overlap with any value defined in the HDF for the respective variable. However, any country-specific values that were included in the national questionnaire should not be dropped from HDF. That is, all answer categories for a certain question should be included, not just those mentioned in the HDF.

3.3 Coding standards

Coding of variables needs to comply with the guidelines discussed in the coding section of this document. Often, countries will use their own national codes which need to be converted to the international equivalent to ensure international comparability of all participating countries. The examples of such international standards are International Standard Classification of Education (ISCED) and International Standard Classification of Occupation (ISCO).

3.4 Organization of life course

Data from event histories should be organized in the same sequence as the questionnaire, i.e. by columns or events. For event histories, no imputation of missing data should take place.

In the data recorded in the household grid, household members should be sorted according to the following rules:

- Household members should be first sorted according to their relationship to the respondent, defined in variable *agh3_**. For example, partner or spouse first, biological children of current partner second and non-relatives last.
- Within type of relationship, household grid data should be sorted according to age, in the order of oldest first.
- When sorting the household grid, care needs to be taken that other variables which make use of it (e.g. *a206 & a207*) are referencing the correct person in the household grid. This means that when a206_1= 3, this refers to the household grid questions *ahg*_3* for this child. When *a207*=5 this refers to questions *ahg*_5*, and so forth.

Event history data should be provided in a wide data structure and not in a long or relational structure.

3.5 Consolidating scattered information

The paper version of the questionnaire requires repeating some questions. For example, section 5 on parental home has many such repetitions. Where possible, this duplicated information will be merged into a single variable. Consolidating will be performed at the central level and should not be attempted at time of the preparation of the data file for submission. In order for the consolidation to work, respecting the routing as defined in the core questionnaire is essential. Although some variables may seem to duplicate information already available, it is essential to respect the questionnaire routing and HDF coding to the highest degree possible.

3.6 Standard file format

Datasets should be standardized so that the same number of variables appears in the same order for all datasets, which would allow both horizontal matching with a unique identifier and vertical matching by appending. Since the implementation of optional sub-modules varies by country, placeholders will be inserted for all the possible variables within the sub-modules that were not implemented. In the same manner, variables for the histories of partnerships and children have to be filled up to the maximum number of episodes in these histories, that is, for example, the maximum number of children and relationships included in any one file.

4 Pre-harmonization

Preparation of data for harmonization, also called pre-harmonisation, entails checking of the questionnaire routing and converting the data structure to HDF format.

4.1 Conversion of data structure

The second pre-harmonization activity is the conversion of the data structure to a standardized format described in the Harmonized Data File Description[1]. All the variables

[1] ibid.

should be renamed to standard variable names and standard variable labels applied to them. Special attention needs to be brought to the following two aspects of data re-formatting:

- All values, including missing values, need to be recoded to the values as described in the Harmonized Data File (HDF) description.
- All values in the dataset, except years of birth, ages and other amounts and quantities, need to be labelled in English according to the HDF. The country-specific values should not overlap with HDF values for a given variable.

In the conversion process, special attention to detail needs to be paid as any modification or restructuring of the data can lead to logical and structural inconsistencies if not performed correctly. This holds in particular for the construction of grids. The grids containing information on the household, non-resident children, step-children, and income are complex to construct. When looking at individual variables, only the most obvious errors can be identified. However, attention needs to be paid to the multivariate structure of the grids. All information on a certain event or on a person mentioned in a grid should be in the same line of the grid, namely in the variable holding the same *_x* subscript.

Much of the necessary data transformation is straightforward. However, the great amount of data and complexity of the underlying questionnaire can cause unforeseen mistakes. Therefore, checks for value range and logical consistency need to be implemented in all stages of the pre-harmonization process. The checks should not focus only on separate variables, but should also incorporate broader checks across a grid, topic or the entire questionnaire.

4.2 Routing check

At the beginning of pre-harmonization, the routing check of the nationally implemented questionnaire needs to be performed. Ideally, routing should correspond to the Full GGS Wave 1 Questionnaire. The Data Availability Report Template[2] provide a convenient overview of routing by each variable.

Routing patterns are transparent for most parts of the questionnaire. However, certain parts, such as Fertility (section 6) and Parents and Parental Home (section 5), are more complex and need great attention when implementing the checks. Additional complexity may arise from deviations of the applied routing from that specified in the GGS Wave 1 questionnaire. To provide a comparable dataset that contains as much information as possible, routing used in the harmonized datasets should be identical.

Certain sections of the GGS Wave 1 questionnaire have identical information scattered over multiple questions in order to facilitate the interviewing process. The most pronounced example is section 5 Parents and Parental Home. As mentioned in section 3.5 of these guidelines, this information is consolidated into new variables. However, in order for this consolidation to be successful, the pre-harmonized data must also follow the same routing as in the GGS Wave 1 questionnaire.

For successful and timely harmonization of the data, it is essential that the pre-harmonization and especially the checking of the data routing complies with the GGS Wave 1 Questionnaire as much as possible. Any deviations from the standard routing still present in the pre-harmonized dataset need to be thoroughly documented in the Data Availability Report[3].

---

[2] The document can be found on GGP web page at the following address:

http://www.unece.org/pau/ggp/materials.htm

[3] ibid.

4.3 Reporting

The submission of the national dataset for harmonization and international dissemination needs to be accompanied with extensive documentation on the national questionnaire and survey process in order to ensure a smooth harmonization process. First and foremost, the Data Availability Report provides a per-variable overview of all variables available in the dataset. It also indicates any potential deviations from the GGS Wave 1 questionnaire routing or deviations from the default values as specified in the HDF. The computer programme scripts (syntax) used to pre-harmonize the original dataset should be included as part of the submitted documentation.

4.4 Check-list

The following documents are a necessary in order to be able to perform data harmonization.

- Pre-harmonized dataset that was checked for logical inconsistencies at least by inspecting frequency distributions and value labels;
- Syntax used to construct the pre-harmonized dataset;
- Completed Data Availability Report that has details on all variables in the dataset.

The following documents would greatly reduce the time spent on clarifying questions with those responsible for pre-harmonization:

- Original questionnaire in English and national language;
- Technical report on sampling, survey design, fieldwork and data entry;
- Original dataset;
- Full sample with the final disposition codes, sampling variables (indication of strata, clusters, etc.), basic demographic information from the sampling frame (where available) and ID for linking to the original dataset for each of the sampled units.