Handbook on Population and Housing Census Editing Revision 2

United Nations Statistics Division
New York, 2020

ANNEXES

Acknowledgments

The current revision of the Handbook on Population and Housing Census Editing Revision was carried out by an expert group comprising census experts representing all regions of the world. The expert group for the revision of the handbook was convened in New York from 16 to 18 January 2019.

The United Nations Statistics Division expresses its appreciation to the members of the expert group for their contribution to the revision of the handbook: Mr. Danny Wall (Canada), Mr. Li Rui (China), Mr. Nashrul Wajdi (Indonesia), Ms. Rosa Maria Lipsi (Italy), Mr. Pedro Alain Lopez Condado (Mexico), Ms. Minerva Eloisa Esquivias (Philippines), Ms. Taeeon Kim (Republic of Korea), Mr. Papa Ibrahima Sylman Sene (Senegal), Mr. Elias Motshoene and Ms. Angela Kaba Ngyende (South Africa), Mr. Sergio Nelson (Suriname), Ms. Fern Leather (United Kingdom), Mr. Thomas Ondra (United States), Mr. Ismail Lubbad (ESCWA), Mr. David Thorogood (Eurostat), Mr. Friedrich Huebler (UNESCO Institute of Statistics), Mr. Yadigar Coskun (Unicef), Mr. Tapiwa Jhamba (UNFPA), Mr. Zurab Sajaia (World Bank), Mr. Frank Swiaczny, Mr. Francois Pelletier and Ms. Clare Menozzi (UN Population Division).

Special gratitude goes to the Statistics Canada and Italian National Institute Statistics and as well as Mr. Danny Wall and Ms. Rosa Maria Lipsi for contributing written input towards the revision of the Handbook. The contribution of Michael J. Levin, consultant to UNSD, to the revision of the Handbook is gratefully acknowledged. A noteworthy acknowledgment is given to UNSD staff Srdjan Mrkic, Meryem Demirci and Seiffe Tadesse for their coordination, compilation of inputs and review of the Handbook.

I. INTRODUCTION

A. PURPOSE OF THE HANDBOOK

1. A well-designed census or survey, with minimal errors in the final product, is an invaluable resource for a nation. To obtain accurate census or survey results data must be free, to the greatest extent possible, from errors and inconsistencies, especially after the data processing stage. The procedure for detecting errors in and between data records, during and after data collection and capture, and on adjusting individual items is known as population and housing census editing. This handbook looks primarily at censuses.

2. A census is a full count. A survey usually enumerates a smaller proportion of the total population. No census or survey data are ever perfect. Countries have long recognized that data from censuses and surveys have problems, so have adopted various approaches for dealing with data gaps and inconsistent responses. However, because of the long interval between censuses, the procedures that were used to edit the data are often not properly documented. Hence, countries must reinvent the process used in earlier data collection activities for a new census or survey.

3. The *Handbook on Population and Housing Census Editing* is designed to bridge this gap in census and survey data editing methodology and to provide information for concerned officials on the use of various approaches to census editing. The handbook is also intended to encourage countries to retain a history of their editing experiences, enhance communication between subject-matter and data processing specialists, and document the activities carried out during a current census or survey to avoid duplication of effort in the future.

4. Subject-matter specialists include demographers, social scientists, economists and others who are working in population, housing and other related fields. The *Handbook* is a reference for both subject-matter and data processing specialists as they work as teams to develop editing specifications and programs for censuses and surveys. The handbook follows a "cookbook" approach, which permits countries to adopt the edits most appropriate for their own country's current statistical situation. The present publication is also designed to promote better communication between these specialists as they develop and implement their editing programme.

5. With the advent of main frame computers in the 1950s and 1960s, countries started to edit their data electronically. Microcomputers came into play in the 1980s, but the methods remained about the same – the data were collected and then captured and then edits applied. In the early years of this century, most countries moved from keying their census data to scanning.

6. In recent years, however, an increasing number of countries are using electronic data collection technologies to collect their census data. The adoption of handheld electronic devices, Internet or telephone interviewing introduces significant changes in the process of data editing. These electronic data collection technologies allow data editing as they are collected. And, because these technologies can be programmed to edit data on entry, much more editing can be done during the interview.

7. Some countries are also implementing multi-mode data collection methods utilizing two or more data collection modes in one census. Multi-mode approaches can be implemented combining paper-based data collection (face-to-face interview or self-enumeration) and/or electronic data collection technologies (handheld electronic devices, Internet or telephone interview). For example, data can be collected first with the Internet followed by paper questionnaire based self-interview, and/or by face-to-face interview method with paper or electronic questionnaires on handheld devices.

8. The first version of the UN Editing Handbook for the 2000 round focused on paper recording and keying of data. Scanning was only beginning to be used for census data collection and capture. Enumerators recorded information on questionnaires which were structured to maximize data capture through keying.

9. The second version of the UN Editing Handbook focused on scanning for capture. Fewer and fewer countries keyed their data, and more countries used higher-level machine capture – OCR or OMR. The procedure required the questionnaire to be prepared in such a way that the machine could read the marks or written entries and convert them to a microdata set that could be edited and tabulated. However, the kind of editing based on specifications seen in the previous version of the handbook did not change very much because the enumeration was basically the same – the enumerators recorded what the respondents told them and did not edit the data before capture.

10. For the 2020 round, however, many countries are moving from a few cases of keying in very small countries and scanning in larger countries to use of electronic data collection technologies. Because electronic data collection technologies allow for editing the data as they are collected, this handbook will consider both traditional editing and the new "edit on entry". Discussions of changes because of electronic data collection will occur throughout the volume and be summarized at the end (see Annex 1).

B. THE CENSUS PROCESS

11. A population and/or housing census is the total process of collecting, compiling, evaluating, analysing and releasing demographic and/or housing, economic and social data pertaining to all persons and their living quarters (United Nations, 2017). Traditionally, censuses are conducted at specified times in an entire country or a well-delimited part of it, providing a snapshot of the population and housing. Most countries still conduct the census with the traditional method comprising of a field operation of actively collecting information from individuals and households on a range of topics at a specified time. However, increasingly, countries are exploring the use of alternative methodologies, for example, the use of registers in combination with field enumeration and/or other data sources and the application of a continuous survey methodology for producing detailed small area statistics on population and housing. For more information on alternative census methodologies, see the Principles and Recommendations for Population and Housing Censuses Revision 3, Part I, Chapter IV.

12. The fundamental purpose of a census is to provide information on the size, distribution and characteristics of a country's population by small geographical areas and for small population groups. Countries use the census data for policy-making, planning and administration, as well as in management and evaluation of programs in education, labour force, family planning, housing, health, transportation and urban/rural development. A basic administrative use is in the demarcation of constituencies and allocation of representation to governing bodies. The census is also an invaluable resource for research, providing data for scientific analysis of the composition and distribution of the population and for statistical models to forecast its future growth. The census provides business and industry with the basic data they need to appraise the demand for housing, schools, furnishings, food, clothing, recreational facilities, medical supplies and other goods and services.

13. All censuses and surveys share similar major processes that include (a) planning and preparatory activities, (b) questionnaire development, (c) mapping and geospatial data, (d) enumeration, (e) data processing, including data capture through keying or scanning (where electronic data collection technologies are used data capture and some editing is done during the enumeration process), editing and imputation, (f) analysis and validation of data, (g) dissemination of the results, and (e) evaluation of the census operation. In addition, the census operation also includes several over-arching processes that are applied throughout the census. Among these, the processes of quality management and metadata management are the most important ones. In the context of the census, quality management refers mainly to the quality assurance programme which is developed

for monitoring and assessing the quality of operational activities and census data. The over-arching process of metadata management refers to the creation, use and archiving of operational information and statistical metadata throughout the census process.

14. The census process given above is a general framework identifying the main phases of the census. Although the processes are described sequentially and the outcomes of one phase will serve as input to the following phase, it should be noted that these phases are all interrelated. The selection of the data collection mode(s) is one of the critical elements in designing census phases, as it could affect the order in which the census processes have to occur.

15. Data can be collected through the following modes:(a) Paper questionnaire with face-to-face interview (PAPI); (b) Computer-assisted face-to-face interview (CAPI); (c) Paper questionnaire with self-interview with (PASI); (d) Computer-assisted self-interview (CASI); and, (e) Computer-assisted telephone interview (CATI). The use of any type of electronic data collection technologies (handheld devices, Internet or telephone) will change the census process as data is captured and edited during data collection instead of the post-collection phase of data processing.

16. The below is an overview of the main phases of the census operation.

17. The **preparatory work** includes many elements such as determining the legal basis for the census; budgeting; developing the calendar; administrative organization; communication and publicity; developing of the tabulation program; and developing plans and training staff for enumeration, data processing, and dissemination.

18. The **development of census questionnaire(s)** refers to a process of deciding census topics and designing and testing census questionnaires. This process should be undertaken with the involvement of all stakeholders and users of census data. There are many elements that have to be considered during the questionnaire development. For example, census topics can be decided considering the need of users and national priorities as well as the burden on respondents. A decision on the use of paper questionnaire or electronic questionnaire will affect the way the questionnaire is designed. If a paper questionnaire is used, the design will be undertaken according to the technology that will be used for data capture (scanning or manual data capture). If electronic questionnaire is used, the data collection application will be developed considering the type of mode selected for data collection (CAPI, CASI or CATI). For more information, see the United Nations Guidelines on the use of electronic data collection technologies. Enumeration through CAPI, CASI or CATI is assisted by an electronic questionnaire which is a computer program running online or on handheld electronic devices. The program contains a set of built-in editing rules, which are used to assess whether the response is allowed by the survey criteria or should be discarded, that is whether an edit is satisfied or violated.

19. **Census mapping** plays a critical role in all processes from preparation to dissemination of census results. Census maps serve several purposes in the census process, including to ensure that every household and person in the country is covered, to support data collection and supervision of census activities during field enumeration and to make it easier to present, analyse and disseminate census results.

20. The **enumeration process** depends on the method of enumeration and technology selected, the timing and length of the enumeration period, the level of supervision and whether and how a sample is used.

21. After the data are collected either with paper questionnaire or electronic questionnaire, the **data processing phase** is implemented. In the traditional way of conducting **the census with paper questionnaire,** this process captures, codes, edits and validates data as well as generates the master database. When **electronic data collection is used**, data capture and some editing is done during data collection with the assistance of built-in edits and drop down list allowing to select single item from the list. Data processing produces both micro-

databases and macro-databases. National census/statistical offices use these databases for tabulations, time series analysis, graphing, developing online dissemination tools, and along with Geographic Information Systems (GIS), for thematic mapping and other dissemination techniques.

22. The phase of **analysis and validation of data** includes the preparation of outputs for validation and analysis before dissemination. Validation of the quality of the outputs is performed in accordance with the general quality framework and indicators adopted for this purpose. Validation activities can include: checking the population coverage and response rates: comparing the statistics with results from previous censuses and relevant surveys and administrative registers; validating the statistics against expectations, domain intelligence, and application of disclosure control for ensuring that the data to be disseminated do not breach the appropriate rules on confidentiality. The results are analyzed for both content and coverage using a variety of methods, including demographic analysis and post-enumeration surveys.

23. The **dissemination** phase manages the release of the statistical products to users. It includes all activities associated with assembling and releasing a range of static and dynamic products via a range of channels. These activities support users to access and use the outputs released by the statistical agency. The products for dissemination of census results can take many forms including interactive graphics, tables, public-use micro-data sets and downloadable files. This phase also concerns the active promotion of the statistical products produced, to reach the widest possible users.

24. **Evaluation** of the census process, as the last phase, manages the evaluation of each phase using operational information and metadata collected throughout the census process and feedback from system metrics and users and staff suggestions. The evaluation report prepared by the evaluating team should note any quality issues specific to any phase or activity of the process, and should make recommendations for changes, if appropriate, for the next census.

C. USE OF ELECTRONIC DATA COLLECTION TECHNOLOGIES IN CENSUSES

25. The basic census process has been in place since at least the 1960s when countries started using Main Frame computers. Even when countries moved to Micro-computers, the process essentially remained the same. Data collection was on paper; and then items like places, languages, industry and occupation, coded; then the data were captured, first by keying and later by scanning; then the data were edited, tabulated and disseminated. When paper is used, the data are collected during enumeration, but actual processing does not begin until the questionnaire is returned either to the field office or to the central data center.

26. However, the adoption of electronic data collection technologies changes this process in a fundamental way. From an operational point of view, electronic data collection means the integration of interviewing and the data entry process including data capture, coding and consistency checks. Electronic data collection with handheld devices, Internet or telephone allows the capture of information with relevant codes (there might be some exceptional variables that may require a coding in the office, such as occupation and industry). It also allows the identification of potential errors during the interview with pre-programmed consistency checks. Because the consistency checks are performed during the interview in real-time, errors and inconsistencies can be resolved, and corrective action can be taken by the respondent or the enumerator. However, introducing editing rules into the data collection application has to be carefully examined so as not to affect its performance significantly in the field, not to create a bias on data and not to affect the quality of the interview for questions that may not be answered properly, especially in case of collecting information from proxy respondents.

27. The use of electronic data collection makes it necessary to develop editing rules during the phase of the questionnaire development in order to integrate these rules with the data collection application. This means that identification of census topics, formulation of census questions, questionnaire design and developing editing

rules should be performed in an integrated manner. Testing procedures should also be designed to test data capture, coding and data editing. Moreover, when multi-mode data collection (the combination of electronic data collection technologies with/without paper-based data collection) is used, testing procedures have to be designed for testing efficiency of different types of data collection and processing procedures and method of integrating different databases to ensure comparability among modes of data collection.

28. The editing team should be involved during the process of developing the data collection application to ensure the proper incorporation of editing rules into the application. Certain decisions for consistency checks have to be made during the questionnaire development process. For example, what type of real-time edit checks will be applied to each field in the questionnaire, what variables can be left blank by a respondent or enumerator.

29. The editing team should carefully weigh the benefits of performing real-time consistency checks during the interview. Care must be taken so that the application of complicated consistency rules does not affect the performance of the application and does not increase the duration, and interrupt the fluency, of the interview. Therefore, a careful consideration needs to be given while designing the real-time consistency checks applied either with CAPI or CASI.

30. The use of electronic data collection technologies will require more time for designing, developing and testing the collection application (for capturing data and running editing rules) when compared to the time required for designing and testing of paper questionnaire. However, less time will be needed for data processing as there will be no activity for data capture and a significant part of coding and editing will have been completed during the data collection phase[1].

31. Considering the time needed for the development of the data collection application, the census work plan should allocate enough time for the questionnaire development process. This requires early planning and decision on whether electronic data collection technologies will be used in the census or not.

32. No matter how much edit on entry occurs, the computer editing described in this handbook is still necessary. Some people will not know some information, even their age, so the computer edit is needed to obtain the best estimate through imputation. While edit on entry will shorten the time needed to get to the computer edit and so get the results out to the users, it nonetheless, does not take the place of the office edit. A full edit program is still required. The resulting tabulations should be the same whether the original capture was on paper or electronic. The remaining activities will also be the same.

33. The United Nations editing handbook preserves the previous suggestions for editing census data. The previous handbooks assumed paper collection. The current handbook will make reference throughout to the use of electronic data collection as well.

D. Errors in the census process

34. Census data suffer from many sources of error that may be classified, generally, as coverage errors and content errors.

1. Coverage errors

35. Coverage errors arise from omissions or duplications of persons or housing units in the census enumeration. The sources of coverage error include incomplete or inaccurate maps or lists of enumeration areas or living quarters, failure by enumerators to canvass all the units in their assignment areas, duplicate counting,

[1] UN Guidelines on the use of Electronic Data Collection Technologies in Censuses.

omission of persons who are not willing to be enumerated, erroneous treatment of certain categories of persons such as visitors or non-resident aliens and loss or destruction of census records after enumeration. Coverage errors should be resolved, to the greatest extent possible, in the field. With recent development in the use of georeferenced data and digital maps for supporting enumeration, census coverage can be successfully improved. Another important tool for improving census coverage is to monitor the field enumeration in real time. The adoption of electronic data collection technologies makes it possible to transfer data during the enumeration, therefore enumerated population and enumeration status of living quarters could be easily monitored to ensure full coverage of population and housing units.

36. The electronic data collection edit-on-entry and the office editing process eliminate duplicate records. However, care must be taken to determine whether these are duplicate persons or households. Twins, for example, may have identical information, except for name and sequence number. Hence, the editing rules applied during this process determine when to accept and when to reject seemingly duplicate information, and when to make changes through imputation.

37. Structure edits, described in Chapter IV, check households for the correct number of person records, correct sequencing, and the existence of duplicate persons.

2. Content errors

38. Content errors arise from the non-reporting or incorrect reporting or recording of the characteristics of persons, households and housing units. Content errors may be caused by poorly designed questions or data collection application or poor sequencing of the questions, or by poor communication between respondent and enumerator. Errors also occur by mistakes in coding and data entry, errors in manual and computer editing, and erroneous tabulations of results. Edit trails (also known as audit trails) must be properly developed and stored at each stage of the process to ensure no loss of data. The following sections explain each of these errors.

(a) Errors in questionnaire design

39. Poorly phrased questions or instructions are one source of content errors. The type of questionnaire, its format and the exact wording and skip pattern of the questionnaire items merit the most careful consideration, since the faults of a poorly designed questionnaire cannot be overcome during or after enumeration. Pretesting should be used to minimize errors that may arise due to poorly designed questionnaires. If, for example, skip patterns are not clear or are not placed appropriately or invalid, the enumerator may erroneously skip sections of the questionnaire and fail to collect all the relevant information. When using electronic questionnaire, proper skip patterns and real-time consistency checks during enumeration will clearly improve the quality of the data.

40. As countries move to electronic data collection, skip patterns become even more important, because once a skip is made during enumeration, it is often very difficult to go back to correct an erroneous skip. In fact, skip patterns are particularly important with electronic data collection because when a skip is made, most enumerators or respondents will not backtrack to correct the erroneous skip, which can lead to further problems later. Testing of data collection application should be undertaken carefully to analyse for what cases a skip pattern will be implemented automatically and what cases a skip instruction will be given.

(b) Enumerator errors

41. Enumerators and respondents interact, unless the census is conducted using a self-administered questionnaire. The enumerator can err when asking the questions, either by abridging or changing the wording of the questions or by not fully explaining the meaning of the questions to the respondent. The enumerator may also add errors in recording the responses. The amount and quality of enumerator training are crucial factors in the overall quality of data collected.

42. Enumerators must be properly trained in all aspects of census procedures. They should be made to understand why their role in the census process is important and how the enumeration fits in with the other stages of the census. Moreover, since enumerators come from many different backgrounds and have varying levels of education, training must make certain that enumerators know how to ask the questions to obtain appropriate responses.

43. The use of handhold electronic devices makes the enumerator's job much easier. Enumerators do not have to record responses on paper, thus reducing reporting and recording errors. In many cases, entering the first few letters of an industry, occupation, place, ethnicity or language will produce a digitized code that can either be entered as numbers or by pressing the "enter" key on the handheld device. Supervisors must take care in training to ensure that the enumerators follow the skip patterns. Enumerators must also be trained to know when to leave a response blank for later office editing.

44. One of the more important issues as electronic data collection is introduced is how much editing should be done on entry and whether every entry must be filled. Ideally, there might be a desire that every item be filled during the interview. This procedure is likely to introduce errors into the data set for questions that the respondent does not know correct information as a proxy respondent and when the enumerators guesses. However, it is important to obtain information for key items, like sex and age, in the field since so much of the analysis and program and policy formation depend on these two variables. Also, relationship to the reference person or head of household should be collected in the field as an important information for checking the list of household members and visitors and ensuring all people are covered in the census. Other variables can also be obtained and checked in the field, like mis-matches between walls and roof, but in many cases, it is better to leave the full edit to the office computers rather than the many enumerators who make varying decisions about proper responses. See the Annex I on use of electronic data collection.

(c) Respondent errors

45. Errors may be introduced into the data when respondents misunderstand an item or make an error in recording the responses if it is self-administrated questionnaire. Errors may also occur from deliberate misreporting, or proxy responses (when someone other than the person to whom the information pertains provides the responses to the questionnaire). The quality of individual responses can be improved through publicity for the census as well as by training enumerators to explain the purpose of the census and the reasons for the various questions. Some countries use self-administered questionnaires, so no enumerator-respondent interaction exists. For self-administered forms, errors occur when the respondents misunderstand the questions or instructions or make mistakes in filling in the responses. Therefore, it is important to give appropriate instructions on the paper or online questionnaire and guidance document to help respondents in recording correct information.

46. Respondent and enumerator errors are best addressed at the enumeration stage while the forms, the respondents and the enumerators are still available. Electronic questionnaires via the Internet or handheld devices are particularly useful in this regard because they can receive an immediate message when inappropriate responses are entered. For example, a message "You just keyed a 79-year-old female with a 3-year-old child, did you mean that?" will signal the enumerator or the respondent in case of self-administrated questionnaire to do follow up before completing the questionnaire.

47. Supervisors must train enumerators appropriately. The supervisors must also check data collected by enumerators regularly during enumeration to ensure that enumerators do not introduce systematic bias into the data. Supervisors should deal with enumerator and respondent errors in the field before the questionnaires are sent to the regional or central offices.

48. It is crucial to train enumerators and supervisors for the questionnaire whatever paper or electronic questionnaire is used. However, the training programme will be more intense for electronic questionnaire as there is a need for a special training programme for the use of electronic data collection application in addition to the training of concepts and definitions. On the other hand, the quality of the work of enumerators will be difficult to control by supervisors when paper questionnaire is used considering that enumerators can introduce different types of errors during the interview. In case of using paper questionnaire, the supervisor can control data with eye movement to make sure that the enumerators understand the items and the skip patterns. However, discovering such errors with eye-tracing requiring full concentration is extremely difficult and even if errors are discovered, it is hard to reach the same respondent for correcting these errors after the interview completed. The use of electronic questionnaire can provide more systematic and effective control system if quality checks and monitoring procedures are implemented automatically using additional applications developed for this purpose.

49. For electronic data collection the data need to be transferred, and then the supervisor can check the quality of data based on standard reports produced by the software application used. Furthermore, subject-specialists can also control the data quality at headquarter using a standard tool developed for this purpose. If the supervisors are separated from the enumerators by space and time, the data can be transferred electronically for review; the supervisor can then analyze the keying and discuss any problems with the enumerators while they are still in the field.

(d) Coding errors

50. Most items are coded by the enumerators in the field, particularly the major items like sex or gender (usually 1 for male and 2 for female directly). Similarly, date of birth and age are coded as entered. Other items require simple coding which often appears on the questionnaire itself – like relationship to head and marital status. But others may require the enumerator to enter a specific geographical area, industry or occupation. Sometimes, they have a look up list, either on paper or electronically, that they can use to get a code to be entered directly. But, often they collect written entries which must be coded later in the office.

51. The introduction of electronic data collection also introduces the possibility of doing all coding by the enumerator or the respondent in the field. Some items remain self-coded, like age and year of birth. Others are easily changed to numbers for processing, like male and female. Still others will have short lists of possible entries, like relationship, marital status and the completed level of education, can be coded with drop down menus. And, others, like ethnicities, religions, languages and places (probably for geographic codes) could be coded using a look-up table for assigning appropriate code for each item. This can be done in two ways: the first approach is using a search tree or menu allowing enumerators or respondents to navigate through the look-up table by means of a two-level or three-level search tree and select the most relevant match. And second approach is to use semantic matching allowing respondents or enumerators to identify their occupation by typing text whereby matches with words in the look-up table that are instantly shown. Sometimes the enumerator or respondent will enter the alphabetic response and the machine will automatically convert to a number. But, when the electronic data collection is not fully automated and for items (such as industry) that may require a special expertise for coding, office coding will be needed for those items which are not already numeric.

52. If the enumerator or respondent makes an error in entry using electronic questionnaire, that error will remain forever unless it is invalid or inconsistent. This issue is one drawback to edit on entry. The enumerator could enter a valid response, and the responses could be compatible, but they could be wrong if more detailed editing rules are applied. In this case, the more sophisticated office edit should be implemented for ensuring full consistency in census database.

53. Errors may arise in office during coding since the coder may miscode information. Mis-keyings may introduce errors into the data during data entry. In general, lack of supervision and verification at this stage delays the release of data, as error detection and correction become more difficult later.

(e) Data entry errors

54. Range checks and certain basic consistency checks can be built into data entry software to prevent invalid entries. These checks can be implemented in the field in the case electronic data collection technologies are used or in the office after field keying. An intelligent data entry system ensures that the value for each field or data item is within the permissible range of values for that item. This system increases the chance that an enumerator or respondent using electronic questionnaire or data entry operator keying entries will key reasonable data. This procedure relieves some of the burden of data editing at later stages of the data processing.

55. These checks may, however, slow down the speed of data entry, especially with electronic questionnaires in the field. Therefore, the amount of consistency checking during data entry must be carefully weighed against the need to maintain a reasonable speed of data entry. A balance needs to be established beforehand, so that the enumerators/respondents or data entry clerks do not spend too much time on these efforts. In case of the use of paper questionnaire, procedures to verify keying inevitably improve the quality of the data. Keyed forms may be verified by rekeying the same information, often on a sample basis.

(f) Errors in computer editing process

56. The editing process refers to a range of procedures used for detecting errors and correcting invalid and inconsistent data by imputing non-responses or inconsistent information with plausible data. This process is a crucial step in census data processing regardless of whether electronic data collection technologies are used or not. The use of electronic data collection improves the data quality with real-time consistency checks but does not obviate the need for computer editing after data collection since unknowns, invalids, and inconsistencies will still occur. Data collection with paper questionnaire is exposed to these types of errors depending on the quality of the work of the field staff. Detecting errors in data and handling these errors with appropriate procedures are traditionally one of the most challenging and time-consuming parts of the census process. Paradoxically, any of these editing operations can introduce new errors. This type of editing errors can occur probably more in the case where paper questionnaires are used because all types of errors have to be dealt with only after the enumeration.

(g) Errors in Master Files

57. When data editing is in progress, new files consisting of clean data records for each person are produced; these can be assembled so as to build a master file for later tabulations (often called the micro-data file).

58. This master file, like the raw data files, can have a simple rectangular sequential format. There is usually no need for (but neither should it be discouraged) having the master file organized with a database structure with index files. However, the master file should usually be maintained in geographical order, starting with the lowest geographical entity, sorted by housing unit, household or family (United Nations, 2017, para. 3.200).

59. One of the most common and problematic errors in census files is that different enumeration areas can carry the same identification codes. Another common error can happen when different persons or households have the same code within a household or within an enumeration area. These types of errors can be observed even more dramatically for multi-mode data collection method, if case management procedures are not designed well for identifying duplication which may occur at enumeration area, household or individual level during enumeration or data processing. Upon sorting the file, these enumeration areas or households have been merged, generating households with abnormal characteristics such as two heads of households, twice the usual number of members with the same identification number. To avoid this kind of problem, the enumeration area geocodes and identification codes given to a housing unit, household, family and individual in a hierarchal format should

be checked carefully prior to the editing phase. This can be done by keeping a check file of all expected code combinations and marking a code as used. A module of this functionality can be part of editing programme.

(h) Errors in tabulation

60. Errors can occur at the tabulation stage owing to data processing errors or the use of information that is "unknown" (not supplied). Errors at this stage are difficult to correct without being sure not to introduce new errors. These types of errors can be determined through inter-tabulation checking for ensuring consistency between tabulated data. If such errors occur, rather than trying to correct the tables themselves, it is essential to maintain the processing system so that additional editing is done when inconsistencies in the tables appear.

61. If errors are carried through all stages of the process to publication, they will be apparent, and the results will be of questionable value. If “corrections” are made at the tabulations stage, say a few miscellaneous unknowns are found and placed in the “totals” but not in the distribution, the tables cannot be replicated by other analysts, and have less overall value. It is essential to see census processing as a feedback system so that changes to the data set are made during editing and not during the tabulation phase.

62. Before the release of tables, it is essential to conduct a thorough check to ensure that all planned tabulations are prepared for all intended geographical units. While range checks and consistency checks introduced at the editing stage can reduce most of the errors, an aggregate check after tabulation – sometimes called a “macro-edit” – is essential. Trained and experienced persons should go through the different tables to check whether the reported numbers in different cells are consistent with the known local situation observed from administrative registers and other sources.

63. Calculation of selected ratios and growth rates and comparison with previous census figures or other figures published by sample surveys can also be useful. However, comparison with other survey-based figures should be attempted only if the concepts used are comparable and compatible. If errors are found in the final tabulations, corrections should be made first on the data set.

64. Traditionally, minor errors appearing in tables were corrected by the programmers “on the run.” That is, the microdata were not changed, so invalids or inconsistencies remained. It is very important for data processors to avoid changing tabulation programs to correct problems in the data set. These changes will not appear in the microdata and will therefore not be replicable when other programs are developed and run. The team should make all changes to the full microdata set, partly to permit other data processors in the national census/statistical office to make comparable tables.

65. In addition, since national census/statistical offices sometimes release parts of the microdata files to researchers and other users in the public and private sectors, tables need to be replicable. In fact, most countries now release anonymized microdata – usually 5 or 10 percent of census data.

E. STRUCTURE OF THE HANDBOOK

66. Chapter II looks at the role of editing in censuses and surveys. The other chapters cover specific topics. Chapter III presents practical applications for editing and imputation. Chapter IV presents structure edits, edits that look at both housing and population items at the same time, and certain procedures to assist in the rest of the edits, such as determining whether one and only one head of household is present. Chapter V reviews population edits, and Chapter VI covers housing edits. Annex 1 discusses special issues with use of electronic devises. Finally, a series of annexes examine specific issues related to the editing and imputation of population and housing censuses.

II. EDITING IN CENSUSES AND SURVEYS

A. EDITING IN HISTORICAL REVIEW

67. Before the advent of computers, most census operations hired large numbers of semi-skilled clerks to edit individual forms. However, because of the complexity of the relationships between even a small number of items, simple checks could not begin to cover all the likely inconsistencies in the data. Different clerks would interpret the rules in different ways, and even the same clerk could be inconsistent.

68. Census editing changed markedly with the introduction of computers. Computers detected many more inconsistencies than manual editing. Editing specifications became increasingly sophisticated and complicated. Automated imputation became possible, with concomitant rules for the process (Fellegi and Holt, 1976). At the same time, the process allowed for more and more contact with respondents, or at least with the completed questionnaires of these respondents.

69. Many editing teams began to feel that the more editing the better; and, the more the sophisticated the edit, the more accurate the results. Programs produced thousands of error messages, requiring manual examination of the original forms or, for some surveys, re-interviews of the respondents.

70. With computers it became increasingly easy to make changes in the data set. Sometimes these changes corrected records or items. Many records passed through the computer multiple times, with errors and inconsistencies reviewed by different persons each time (Boucher, 1991; Granquist, 1997). The use of electronic data collection technologies adds another level to the ability to make changes to the data as they are entered in the field.

71. Several generalized census-editing packages came out of this whole process, and some of them are still in use today and others are being introduced for the 2020 censuses, particularly for electronic data collection. Initially the packages were developed for mainframe computers: some were later modified for use on personal computers. During this period, Fellegi and Holt (1976) developed a new method for generalized editing and imputation, which was not immediately put into practice, but which is increasingly being adopted today as national census/statistical offices become more sophisticated in their editing.

72. A major advance in census editing came in the 1980s when national census/statistical offices began to use personal computers to enter, edit and tabulate their data. Suddenly, data processors could perform edits at the data entry stage or soon after. The goal of automatic editing is to accurately detect and treat errors and missing values in a data file in a fully automated manner, without human intervention. In practice, the data file is checked record by record. If a record fails one or more edit rules, the method produces a list of fields that can be imputed so that all rules are satisfied[2]. For surveys and small country censuses, staff could develop programs to catch errors during collection or while entering data directly into the machine. Computer edits allowed more, continuous contact with respondents to resolve problems encountered in the editing process (Pierzchala, 1995).

73. In the early years, the process of making increasingly sophisticated and thorough checks on census and survey data seemed to be very successful. Editing teams created ever more complicated editing specifications, and data processing specialists spent months developing flow charts or decision charts and program code. Analysts seldom evaluated the packages or the results of the coding. It seemed that editing could correct any problems arising from earlier phases of data collection, coding, and keying. Nevertheless, it also became apparent to many analysts that in some cases, all this extra editing harmed the data, or at least delayed the results or caused bias in the results. Sometimes the program made so many passes through the data, correcting first one item, and then another item, that the results were far removed from the initial, unedited data.

[2] EU, Handbook on Methodology of Modern Business Statistics, Automatic Editing

74. During the 1990s, household surveys underwent a massive change in data collection methods. The use of Computer Assisted Personal Interview (CAPI), the Computer Assisted Self Interview (CASI) and Computer-assisted telephone interviewing (CATI) almost wholly replaced paper-and-pencil (PAPI) household interview surveys. Population and Housing Censuses received such change recently since the 2010 round of censuses.

75. In CAPI, CASI and CATI interviews, the data collection application (e.g. standard screens and standard editing rules on data entry) usually performs data entry, coding and editing functions, thus limiting the risk of introducing capture error. The edits performed during data collection usually comprise preliminary data checks and basic edits of summary information (such as total number of enumerated people and breakdown by sex). In most cases, editing during data collection will eliminate the going back to the actual questionnaires since the data edited on entry will already be set. Integrating edit rules in the collection system will decrease the time needed for data processing after data collection which will enable NSOs to produce reliable and timely data.

76. As national statistical organizations continue developing censuses and surveys, extensive computer editing is possible and even likely. Consequently, the issue that each national census/statistical office must face is what level of computer editing is appropriate for its purpose.

B. THE EDITING PROCESS

77. Editing is the process of reviewing the data for missing, invalid or inconsistent items and correction of errors in order to improve the quality and accuracy of the data and make it suitable for the purpose for which it was collected. Editing is an iterative and interactive process that includes procedures for detecting and correcting errors in the data. In this process, data is first edited, in order word, checked for missing, invalid or inconsistent values and then imputed for resolving problems and finally validated for ensuring the quality of imputation. Data editing must be repeated after the data are imputed, and again after if the data are altered during the process for protecting individual information.

78. The editing process refers to a range of procedures and processes used for detecting and handling errors in data. These procedures and processes are designed depending on the method of data capture: manual data capture, optical data capture or electronic data capture. The first two methods are applied to the paper questionnaire, while the last method is applied to electronic data collection technologies (CAPI, CASI and CATI).

1. Editing Process Steps

79. The overall editing process can be broken down into a number of sub-processes as shown in Figure 1. An explanation of these different sub-processes follows below.

80. **Step 1**. **Planning and designing**: All data processing stages (including data capture, editing, coding and imputation) should be planned in an integrated way within the overall census program to respect the quality goals, timeliness objectives and technical infrastructure of the program. Care taken in planning may help prevent defects in census results and avoid costly inefficiencies or problem resolutions during census operations. The plan for editing should be based on an understanding of the collection modes to be used, the operational sequence of events, and therefore which edits may be applied at any given part of the operational process.

81. During the planning phase, the editing team must work carefully and in close cooperation with teams that design electronic and/or paper questionnaires, that are implementing other portions of the processing work

flow, or that are working on information technology implementations within the census. Program data quality objectives and the timeframes available for completing work must also be taken into consideration in the edit plan.

82. If multiple collection modes are being used, then the edit approach within each mode needs to be taken into consideration. This is important as certain collection methods may allow for edits that prevent errors either by the respondent or the data entry operator. For instance, electronic questionnaires may enforce complicated skip patterns for respondents, or flag erroneous entries for correction through validity edits; for capture of responses from paper, validity edits may be used to flag possible data capture errors. A solid understanding of the collection mode will help with the implementation of edits within the operational processes and help configure methods to identify inconsistent or missing values in Step 4, or inform imputation approaches in Step 5.

83. Administrative data may also be used within the census or survey programme. Such data may be subject to edits before it is used. For instance, a data source may be subject to edits which prevent certain records from being used, triggering the need for some collection activity to be performed some other way.

84. **Step 2**. **Developing and testing** : The overall plan for edits will drive the selection of methods and tools (including software programs or applications) that will be used to implement the operational environment for processing steps generally and the editing/imputation steps specifically. Depending on choices made, sufficient time must be allocated for the development or configuration of software applications, and then for the testing of procedures, workflows and the interactions between different applications or systems within the program. Testing will ensure that process and production solutions perform as expected. (Methods of testing editing processes are discussed below). It should be noted that if multiple collection modes are used, then testing of edits within each mode may need to be performed at different times.

85. **Step 3**. **Correction of critical errors**: Systematic errors can occur at various steps during data collection. A respondent may misunderstand a question, an enumerator may make a mistake taking information, or a data capture operator may make a mistake in capturing data . Certain critical errors which have the potential of blocking further processing should be detected at an early stage in data processing. In figure 1, different flows are shown for paper and electronic questionnaires. A validity edit may flag a mistake by a data capture operator within the paper questionnaire flow, or a range edit may alert a respondent to a potential error when completing an electronic questionnaire.

86. These types of edits are normally performed during collection or processing operational activity and are aimed at correcting errors as early in the process as possible. These edits may span all of the micro edits explained in paragraph 94 and be applied differently in different collection modes. It is for this reason that the sequence of events during operational activity must be defined during the planning and design phase, and certain actions need to be taken before others can be performed. For instance, a final step to detect and correct duplicates may have to wait until returns are collected and converted into a common format, normally a database, if multiple modes of responses are used. This may be done even though steps detect or prevent duplicate returns within collection modes earlier in the process.

87. **Step 4**. **Identifying inconsistent or missing values**: Once the critical errors are defined and corrected, it is straightforward to check whether the values are inconsistent in the sense that some of the edits are violated. Basic principle is the data in each record and between records should be made to satisfy all edits with minimum changes possible in items. This principle may lead to many solutions, therefore correction stage of editing process should be carefully analyzed and evaluated to ensure this process does not introduce new errors.

88. **Step 5**. **Imputation**: Imputation is the process of resolving problems concerning missing, invalid or inconsistent responses identified during editing. Imputation works by changing one or more of the responses or

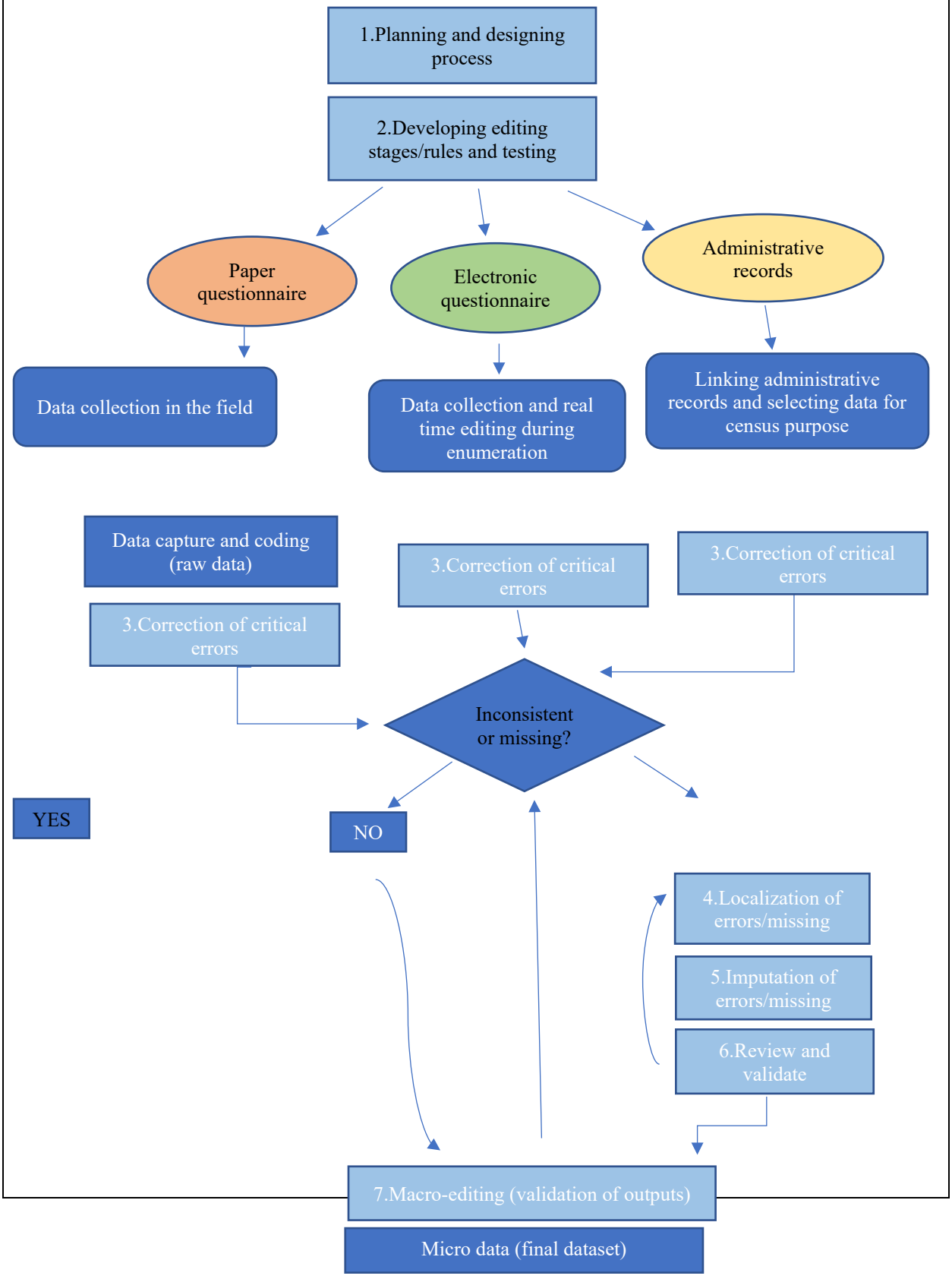
missing values in a record or several records being edited to ensure that plausible, internally coherent records result. Contact with the respondent or manual study of the questionnaire eliminates some problems earlier in the process. However, it is generally impossible to resolve all problems at these early stages owing to concerns with response burden, cost and timeliness. Imputation then handles the remaining edit failures, since it is desirable to produce a complete and consistent file containing imputed data. Imputation procedures are of crucial importance to make sure that census results are of high quality and internally consistent. When paper questionnaires are used, it is generally accepted that there will be more missing cases (item non-response) compared to censuses conducted with electronic questionnaire. However, in any case, there is a need to implement imputation procedures for achieving high quality of data.

89. **Step 6**. **Review and validate**: This sub-process examines data to identify potential problems, errors and discrepancies in editing and imputation. It can be run iteratively for validating data using quality indicators. This step is concerned with detection of errors occurred during previous steps of editing. This step includes also analysis of the impacts of imputation (see below section B.5 on quality assurance and assessing the quality of editing and imputation).

90. **Step 7**. **Macro editing**: The steps 1 to 6 use the data of a single record and related auxiliary information to check and correct it. Micro editing which corrects the data at the record level can be conducted from the start of data collection step until the step providing fully consistent data. After that, macro-editing stage, which analyses aggregate data (totals), starts for the purpose of verifying whether results to be released seem reliable. Macro-editing methods are selective editing techniques that help to discover possibly erroneous values.

91. Macro-editing can be accomplished by comparing aggregated data with the results of previous censuses and other relevant data sources such as household surveys and administrative registers. If methods used for validation of outputs give unusual or unexpected results, then it is possible to go back to the micro-editing procedures. In this stage, micro-editing procedures can use graphical or model-based methods to identify possible anomalous and suspect values in the micro data (see below section B.2 on Types of edits for explanations on types of micro-editing and macro-editing).

Figure 1: Editing Process Steps



2. Type of edits

92. During the editing process, whatever data collection mode is used, the data must be checked for the following errors:

i. Credibility, based on range checks to determine if all responses fall within a prespecified reasonable range;
ii. Consistency based on checks across variables within individual records for noncontradictory responses (i.e., no logical inconsistencies);
iii. Incorrect flow through prescribed skip patterns;
iv. Missing data (item non-response);
v. The omission and/or duplication of records;
vi. Inconsistency between enumeration unit, such as individuals, households and housing units.

93. There are two levels of data edits, whether data is collected via paper questionnaire or electronic data collection technologies: a) micro-editing and b) macro-editing (or validation of outputs).

a. Micro editing

94. Micro-editing corrects the data at the record level. This process detects errors in data through checks of the individual data records. The intent at this point is to determine the consistency of the data and correct the individual data records.

95. For micro editing, four types of edits can be implemented. They include[3]:

i. **Validity edits** look at one question field or cell at a time. They check to ensure the record identifiers, invalid characters, and values have been accounted for; (i) essential fields have been completed (e.g., no field is left blank for key variables, such as sex, age, relationship to the head of household and marital status, and no quantity field is left blank where a number is required, such as CEB and date of the last live birth); (ii) specified units of measure have been properly used, such as number of children, and (iii) the reporting time is within the specified limits, such as month, day, year.
ii. **Range edits** are similar to validity edits in that they look at one field at a time. The purpose of this type of edit is to ensure that the values, ratios and calculations fall within the pre-established limits. For example, minimum and maximum age for giving a birth.
iii. **Consistency edits** compare different answers from the same record to ensure that they are coherent with one another. For example, if a person is declared to be in the 0 to 14 age group, but also claims that he or she is retired, there is a consistency problem between the two answers. Inter-field edits are another form of a consistency edit. These edits verify that if a figure is reported in one section, a corresponding figure is reported in another. For example, number of household members should be equal to the number of individual records.
iv. **Duplication edits** examine one full record at a time. These types of edits check for duplicated records, making certain that a respondent or a household has only been recorded once.

Macro editing

96. Macro-editing detects errors in data but does this through the analysis of aggregate data (totals). The data are compared with data from the surveys, administrative files, or earlier versions of the same data. This process determines the compatibility of data.

[3] Statistics Canada, Data Editing https://www.statcan.gc.ca/edu/power-pouvoir/ch3/editing-edition/5214781-eng.htm

97. The outcome of micro editing is a set of records that are internally consistent and in which person records relate logically to other person records within the same household. This process does not, however, provide the full range of assurance necessary to accept data set as the best possible. A range of conditions could cause errors that cause the data to be consistent wrong: for example, perhaps a condition in the editing suite itself is set incorrectly; proportions in an imputation program may be mis-calibrated; or enumerators (or respondents in case of self-enumeration) may complete a census questionnaire incorrectly. To identify such consistent errors, it is necessary to critically review some key aggregate tables to isolate outlier aggregates and identify the cause of the unusual values. These key tables may be a subset of those intended for output or may be tablets specifically designed for this purpose (P&R Rev 3).

98. For this type of editing, the following approaches can be used:

i. **Historical edits** are used to compare census results in current and previous censuses. For example, any dramatic changes since the last census will be flagged. The ratios and calculations are also compared, and any percentage that falls outside the expected level or pattern will be noted and questioned.
ii. **Statistical edits** look at the entire set of data. This type of edit is performed only after all other edits have been applied and the data have been corrected. The data are compiled, and all extreme values, suspicious data and outliers are flagged for detailed analysis and making a decision if there is a need to continue micro editing.

3. Editing during data collection

99. The use of electronic data collection technologies in censuses allow data editing interactively during data collection. The goal of editing at the time of data collection is to take advantage of the collection application to improve the quality of the data and reduce the costs of the post-collection process.

100. Population and housing censuses which adopt CAPI, CASI or CATI offer the opportunity for new editing strategies. Moving editing closer to the respondents can significantly contribute to improving editing effectiveness. The possibility of using built-in edits allows respondents or enumerators to avoid errors as they are made. The elimination of data capture, either manual or with scanning technology, at the statistical agency directly gets rid of a common source of error. Hence, some of the traditional editing tasks could be reduced. However, not all post-collection editing processes can be moved to data collection editing. Firstly, some corrections can only be made based on an overview of all the collected data; secondly, complicated correction rules may be hard for respondents to understand and finally, they may be difficult to implement in the data collection application tool.

101. Editing is incorporated into data collection application using electronic questionnaire with CAPI, CASI or CATI. The interview is assisted by an electronic questionnaire, which can be considered as a complete software system containing a list of features that have to be met for effective and reliable data collection[4].

102. Handheld devices extend the field of editing to CAPI. The interviewer conducts a face-to-face interview using an interactive computer program with embedded edit checks. CASI questionnaires also adopt editing rules, in which the editing process is performed by the respondent. The prevalent self-administered data collection mode and the use of electronic questionnaires with incorporated edits enable the editing process at the respondent level. This solution could result in many benefits including: (a) decreases in costs and time needed for data processing, (b) improvement in data quality and response rates, (c) lowering of the perceived response burden, and (d) allowing timely dissemination of census results5.

[4] *For more information, see the Guidelines on the Use of Data Collection in Population and Housing Censuses.*
[5] *Eurostat, Editing During Data Collection, the Memobust Handbook on Methodology of Modern Business Statistics https://ec.europa.eu/eurostat/cros/content/editing-during-data-collection-theme_en*

103. Editing at the time of data collection should define all the skips and branching paths among variables, including the followings:

- Define the range of each variables;
- Define which type of control (soft or hard discussed below) has to be activated and when;
- Define the messages to be shown in relation to control activation;
- Specify the rules for text fields.

104. Data entered into the electronic questionnaire are checked for their correctness through the conditions that must be met to assume the response is accurate. The response item has a built-in edit rule to inform the user about an error in case the rule is not satisfied. This leads to the definition what is meant by assuming data are erroneous or data are supposed to be suspicious. Typically, validation in data collection application is notified in two ways: (a) an error message which means the situation is unacceptable and must be changed in order to continue or (b) a warning which notifies the possibility of incorrectness or to draw attention to a certain aspect of working being the consequence of earlier choices.

105. Considering these two options, there are two types of built-in edits:

i. **Hard edits** - must be unconditionally satisfied which prevent users (enumerators or respondents) from going further questions. In this case, users have to enter data in the field which satisfy with survey criteria;
ii. **Soft edits** – notify users that an item should be assessed for its accuracy. These kinds of edits do not prevent the user from moving to next question or submitting the questionnaire. In this case, a failure can be resolved either through changing the response or postponing to later stages by taking no action.

4. Imputation

106. Imputation is the process of resolving problems concerning missing, invalid or inconsistent responses identified during editing. Imputation works by changing one or more of the responses or missing values in a record or several records being edited to ensure that plausible, internally coherent records result. Contact with the respondent or manual study of the questionnaire eliminates some problems earlier in the process. However, it is generally impossible to resolve all problems at these early stages owing to concerns with response burden, cost and timeliness. Imputation then handles the remaining edit failures, since it is desirable to produce a complete and consistent file containing imputed data. The members of the team with full access to the microdata and in possession of good auxiliary information do the best imputation. The following are some key considerations:

(a) The imputed record should closely resemble the failed edit record. Imputing a minimum number of variables is usually best, thereby preserving as much respondent data as possible. The underlying assumption (which is not always true in practice) is that a respondent is more likely to make only one or two errors rather than several;
(b) The imputed record should satisfy all edits;
(c) Editing teams should flag imputed values, and the methods and sources of imputation should be clearly identified.
(d) The editing team should retain the unimputed and imputed values of the record's fields to evaluate the degree and effects of imputation.

107. The application of imputation procedures to missing data would require a careful review of the proportion of missing cases by each variable and reasons for such errors. Some NSOs do not impute missing data and release data with missing cases. This decision should be given by NSOs considering the cost and added value of imputation on data quality. On the contrary, some NSOs impute missing cases due to mainly three

reasons: missing data can introduce a substantial amount of bias; make the handling and analysis of the data more arduous; and, create reductions in efficiency. Because missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls involved with listwise deletion of cases that have missing values. Imputation preserves all cases by replacing missing data with an estimated value based on other available information. Once all missing values have been imputed, the data set must be analyzed carefully to ensure that original data is not changed significantly.

108. Imputations generally fall into one of four categories:

i. **Deterministic imputation** - where only one correct value exists, for example, fertility questions are asked only to females. A value is thus determined from other values on the same questionnaire.

ii. **Model based imputation** - use of averages, medians, regression equations, etc. to impute a value.

iii. Deck imputation - A donor is used to supply the missing value. The "nearest neighbour" search technique is often used to expedite the search for a donor record. In this search technique, the deck of donors comes from the same dataset and shows similarities to the receiving record, where similarity is based on other records that correlates to the data being donated. See Chapter III for the explanation of deck **imputation.**

iv. **Mixed imputation** - In most systems there usually is a mixture of categories used in some fixed rank fashion for all items. For example, first a deterministic approach is used. If it is not successful, then a deck approach is tried. (Ref: UN, 2000, Glossary of Terms on Statistical Data Editing).

5. Quality assurance and assessment of the quality of the process

109. Quality assurance is important in all census operations. Processing of census data is a complex exercise that usually involves many different processes. While each of these processes can be regarded as a separate entity, each one relies on the quality of the output from the preceding process. Consequently, formal quality assurance mechanisms should certainly be in place to monitor the progress of the census processes including the editing process (UN, Handbook on the Management of Population and Housing Censuses, Revision 2, 2018).

110. The editing process is a complex and iterative exercises which may introduce new errors. The quality of the editing and imputation processes should be systematically monitored and evaluated by a specifically dedicated team for this purpose. It is crucial to establish a management information system that will serve as the backbone for quality assurance in the processing system. It is suggested that the original value, edited value, type of error and data on failed checks should be monitored for analysis of editing changes. For every cycle of editing and imputation, change analysis should be undertaken based on the following types of information: i) frequency and type of edit flags; ii) magnitude of change; iii) extent of redundancy in consistency checks. This analysis should be done by item and by geographical areas.

111. The objectives of quality assurance include the following:

- Ensure that all edits are internally consistent (i.e., not self-contradictory)
- Ensure that no additional errors are introduced, considering the possibility of creating errors during the process
- Reapply edits to units to which corrections were made to ensure that no further errors were introduced directly or indirectly by the correction process.
- Make sure that the editing process does not change significantly the distribution of observed values. If there is significant changes in the distribution of observed values, subject-matter specialists should be able to explain reasons for creating implicit models imposed by the edits.
- Monitor the frequency of edit rejects, the number and type of corrections applied by stratum, collection mode, processing type, data item and language of the collection. This will help in evaluating the quality of the data and the efficiency of the editing function.

112. For monitoring the application of editing rules and imputation procedures, a number of indicators can be used. The following indicators are suggested:

a. **Edit failure rate:** It is the number of detected errors divided by total errors (detected and undetected errors) in the input file. This indicator varies between 1 and 0. High value means the editing rule is properly defined as a criterion for error detection. If it is low, then the rule fails and must be improved.
b. **Adjustment rate:** The number of households or people created under absent household imputation- if used- divided by total the number of households or people.
c. **Imputation rate:** It is the percentage of the imputed records in total records. This indicator should be calculated for each variable. Imputation rates are key quality indicators for the census as a whole, and for data users. Unit imputation must follow a statistically robust, transparent process, and be well documented including having an audit trail
d. **Dissimilarity index:** Degree of change of two distributions (observed and total including imputed values) at the variable level. This index (shown below) ranges from 0 to 100, which implies maximum dissimilarity.

$$ID = \frac{1}{2}\sum_{k=1}^{K}\left|f_{y_k} - f^{*}_{y_k}\right|$$

Where:
k: categories of the variable
f: percentage distribution of the variable before imputation
f*: percentage distribution of the variable after imputation

113. The performance indicators in editing process can also be a valuable tool in assessing the quality of the data. By indicating potential causes of problems, editing can also be an effective way of improving the census design.

114. In most cases the distributions of the attributes for an unedited data should be in the same proportions as the edited data (final data including imputed data). And, the distributions of the imputations themselves should also be in this same distribution. The mechanics to put the results of the imputation in place are relatively simple: three short programs will provide the distribution for each variable.

- The first run will make a distribution of the attributes of each variable for the unedited data, that is running on the unedited data set. For the variable on gender, for example, the results will be Male, Female, and anything else – three responses minimum, but some programs divide the "unknown" categories into unknowns, missing, or other problems.

- The second run will be similar, but on the edited data. The results should be about the same, but only male and female would be shown. If the imputation was done properly – that is the syntax worked properly, there should be no unknowns. The fact that there are no unknowns, of course, does not mean that the logic of the edits and imputation was correct. That needs to be checked with the summary lines and looks at sample households before and after the edits and imputations.

- The third run would be of the imputations. Some packages provide the distributions for the imputed values automatically. But if they don't, the programmer needs to have flagged the records with imputations for that item, and then a program can tally the flags. The tally would find the flag for the item and then tally the actual response for that flag.

115. Then, the three resulting series can be put into a single Excel sheet to check to see if the distributions are similar. For example, the distribution of the main material of walls in the housing unit, that is, the percentage

distribution, should be about the same for metal and concrete walls (say 60 percent metal and 40 percent concrete) should be the same for the unedited and edited data. Because imputation is used, the absolute number for each item should increase since invalid or inconstant values will be changed. Then, the percentage of imputation should also be the same distribution although clearly the values will be smaller. So, the edit should expect about 60 percent of the results from the unknowns and inconsistent values to be metal walls and about 40 percent to be concrete walls. But the edit will also check the combinations of materials for roof and walls. A special situation occurs in combinations. If a questionnaire has a concrete roof but thatch walls, one or the other may be changed, resulting in additional imputed values for some of them, and therefore perhaps throwing off the absolute percent distributions. The editing team will determine whether the results are appropriate.

116. Similarly, we would expect an edit to provide 50 percent males and 50 percent females during imputation (although with electronic data collection the results should have been hard edited and so no imputation should be needed for sex). But, in many cases, the distribution of imputed values will not be exactly half and half. If the edit specifications require a check of presence of fertility, the value for female would be assigned through deterministic imputation. An office might skew another part of the edit to account for this surplus of females being imputed.

117. The analysis of the aggregated imputed values should not be done in isolation. As discussed in Chapter III, a good edit program will produce three kinds of error reports: (1) summary statistics of each invalid or inconsistent set of variables, (2) use of a method to print out the housing unit before the edit, what edits messages are generated by the program, followed by a print out of the unit after the edit, and (3) the imputations by item to allow comparisons of unedited items, edited items, and the imputations for that item.

118. These should be generated at the same time, as noted in the discussion in Chapter III. The listing of the imputations is important because when the distribution does not look "right" the programmer can go back and assess what is happening in each edit for valid and consistencies for a particular item. The listings also can be checked by software application to allow analysts to see exactly what is happening to each variable and combinations of variables in the actual edit stream. Sometimes, the order of the edits causes skewing of particular items or sets of items.

C. CONSIDERATIONS FOR MULTI-MODE DATA COLLECTION

119. An increasing number of countries are offering multiple modes of data collection in their censuses, offering a combination of face-to-face interviews and self-enumeration with paper and/or electronic questionnaires. The main motivations for using multi-mode data collection are to improve coverage and reduce fieldwork costs. Another important factor is the possibility to compensate the weaknesses of one mode of data collection with the strengths of another mode. In principle, using a mix of modes allows the data collection manager to minimize both the costs and impacts on quality (due to coverage, non-response and measurement issues) associated with using any given single-mode approach. For example, an additional mode can help provide access to a group of respondents that would otherwise be hard or impossible to contact in the principal collection mode. Furthermore, combining modes sequentially, whereby one starts data collection with the most economical mode, and follows up non-respondents with increasingly expensive modes offers advantages with respect to both minimizing overall costs and increasing participation. For more information on types and methods of multi-mode data collection, see the UN Guidelines on the Use of Electronic Data Collection Technologies.

120. While the use of a mix of modes may offer solutions to problems of coverage and non-response—and may even help to reduce fieldwork costs—mixing modes of data collection has implications for the quality of the collected data, particularly for data comparability. One disadvantage of using a mixed-mode collection is that mode effects may occur. 'Mode effect' is the bias caused by the mode of the data collection. Mode effects are described as the delivery of different results as a consequence of using different means of collection. That is,

differences observed in the data can be attributed to how the data have been collected rather than to real differences in the population. Mode effect varies depending on the motivations for mixing modes and the type of mixed mode system adopted.

1. Main factors contributing to mode effect

121. The primary factors that contribute to mode-effect in multi-mode data collection include:

a. **Coverage**: It arises when not all members of the target population are enumerated. Designing the enumeration with multi-mode approach requires well-established system for case management, ensuring that all individual covered in the census and covered only ones. Automated process control should be established for following-up nonresponse based on whether multi-modes are used sequentially or concurrently (see the UN Guidelines on Use of Electronic Data Collection Technologies in Censuses). Quality assurance procedures across multiple modes have to be integrated in data collection process to ensure the enumeration is executed as designed. This type of data collection approach can be error-prone due to the lack of automated process control, workflow integrity and comprehensive tracking and monitoring system.

b. **Item non-response**: This is the failure to collect data for all items covered in the questionnaire. Potentially, self-enumeration modes have higher rates of item non-response compared to the interview-based data collection modes. On the other hand, it is recognized that electronic data collection has lower item non-response that paper methods. In this context, it might be concluded that PASI may give the highest item nonresponse compared to the PAPI and these methods may have higher item non-response compared to CASI, CAPI and CATI.

c. **Measurement differences**: This is defined as a bias in recorded responses, arising from design of the questionnaire and effects of interviews on respondents` answers. There is no enough scientific researches to discuss how each mode creates a bias in responses in censuses. There are some researches which were done for analyzing mode-effects in surveys. Some found there is some inconsistency between modes, especially between face-to-face interview and self-administration and some found differences may only affect specific items (especially sensitive questions) or subgroups. Measurement error in the census should be examined during pre-tests and pilot censuses and considered while assessing the impacts of editing and imputation.

d. **Processing error**: These are errors that can occur in data capture, coding, editing and imputation. When multi-mode approaches are used especially the combination of paper questionnaire and electronic questionnaire is used, one would expect that tendency for making mistakes in manual data capture is higher than electronic data capture. However, there is no enough research for types of errors that are introduced by enumerators or respondents when they are entering data through CASI or CAPI and when data is captured collected through PAPI.

122. One way for dealing with potential biases is to design questionnaires for multi-mode data collection. In principle, all questions and response categories and the questionnaire structure should be the same across modes. For example, the design should be consistent regarding explicitly offering or not offering a “do not know”. However, it should be kept in mind that if a specific mode is used as main mode for data collection -meaning that this mode is used for enumerating the majority of population and offered to whole population, it is preferable to use the main mode with its maximum potential. On the other hand, if each mode is equally important, one should use a generalized or universal mode design.

123. It should be noted that mode effects and other forms of measurement error may have an impact of the design of editing process and imputation procedures. In any case, it is recommended that validation of editing and imputation should be done separately for each mode of data collection and evaluated for understanding if there is any bias from any mode on data, especially for item non-response. This is particularly important if a specific mode is used for counting a special population group or counting people living in a specific geographical area.

124. For the operational point of view, it is important to include following procedures in editing process, when multi-mode approach is used:

a. Integrated database should be carefully checked for double counting of a person. household and housing unit,
b. Missing cases and inconsistency in data should be analyzed by a mode of data collection before imputation,
c. Analysis of edited and unedited data should be done by a mode of data collection

2. Considerations in designing paper and electronic questionnaires

125. When electronic and paper questionnaires are used in collection efforts, a disciplined and careful approach will help prevent mode effect biases. These questionnaires may be different in terms of presentation to the respondent or enumerator, and how the data from responses are subsequently captured or stored, but the risk of error due to these differences may be much reduced during planning, design and testing.

126. A good approach is to design electronic and paper questionnaires in a coordinated fashion. To the extent possible, the same team should manage both processes, although specialized knowledge may be required for the creation of a paper or electronic form. For instance, paper form creation will require people with knowledge of print processes and how to create specifications for them; electronic questionnaire creation will require people with expertise in web-based technology. Regardless, if directions regarding content and design for the specialized workers are managed by a single team, it will be easier to arrive at forms that are consistent across modes

127. When possible, questionnaire content in both the electronic and paper forms should be identical, as this will allow comparisons of responses obtained during testing and corrections to be made, therefore preventing mode effect bias during actual collection. This would apply to the phrasing and sequence of questions, and response types (e.g. numeric or character). In Canada, the main questionnaires (short and long) for private dwellings used in Census of Population have both electronic and paper versions with identical content. For some specific populations only a paper form is used, for instance in the enumeration of collective dwellings.

128. If electronic questionnaire is used as main mode of data collection. The initial determination of content, including the phrasing of questions, can be made during qualitative testing using the electronic questionnaire. Content determined through this process can then be verified using an equivalent paper version. Feedback from test groups can be used to modify content as required before the questionnaires are field tested.

129. Testing of proposed forms for either electronic or paper format is a critical part of the design and refinement process. For electronic forms, testing needs to include how respondents enter their answers, and how those responses are stored in a data base. For paper forms, testing needs to cover how respondents (for self-enumeration) or enumerators (face-to-face interview) enter the answers, how information from the paper form is captured, and how those responses are stored within a database. Errors within the entirety of each of these flows may lead to mode effect biases. Qualitative testing normally concerns itself only with the first part of these data flows, the actual responses. Capture methods for this portion of test activity may be considerably less robust than for field test, and of course actual production.

130. Within the context of reducing risk of mode effect bias, all types of collection mode for each questionnaire should be part of a field test. Field tests are larger in scope and may be conducted as part of a ‘pilot census’ effort. Because the number of returns for a field test will higher than for qualitative testing, it may be possible to detect mode effect bias that did not get identified in earlier rounds of testing. For example,

a field test will typically exercise the entire response, capture and data storage work flow for each responses type and allow response biases that exist for example, as a result of an error in a data formatting or transfer step in one of the collection modes.

D. THE BASICS OF EDITING

131. The raw data files in a census contain errors of many kinds which are usually introduced during data collection and data processing process. Data processing categorizes the errors into two types: those that may block further processing (critical errors) and those that produce invalid or inconsistent results without interrupting the logical flow of subsequent processing operations (non-critical errors). As noted in *Principles and Recommendations for Population and Housing Censuses, Revision 3* (UN, 2017, para.3.188), all errors of the first kind must be corrected and as many as possible of the second. The basic purpose of census editing at the processing stage, therefore, is to identify as many errors as possible and make changes to the data set so that data items are valid and consistent. Nevertheless, processing cannot correct all census errors, including questionnaire responses that are internally consistent but are in fact instances of misreporting on the part of respondents or misrecording on the part of enumerators.

132. Edits tend to fall into two categories: (1) **critical edits**, which identify errors with certainty, and (2) **non-critical edits**, which point to suspicious data items (Granquist, L. and Kovar, 1997: 420). Critical edits identify data items that are certainly in error, while non-critical edits point to data that are likely to be invalid or inconsistent. To maintain confidence in the census, particularly when the national census/statistical office decides to disseminate microdata, the editing process must detect and handle fatal edits. Non-critical edits are more difficult to correct, have fewer benefits than the detection and resolution of fatal edits, and add more to the cost of the total process.

133. Since all items in a census are included specifically because planners and policy makers need them, relatively more of the query edits must be resolved during census editing and imputation than for surveys. Nonetheless, in determining the final edits for a census, subject-matter personnel should investigate the edits developed and tested before the enumeration to make sure that individual edits have the expected cost of benefits. These investigations need to be part of the census evaluation. As Granquist and Kovar (1997, p. 422) note, data "on hit rates, that is the share of the number of flags that result in changes to the original data, are rarely reported in evaluations or studies of editing processes".

134. Another set of techniques and terminology relates to micro-editing and macro-editing (see paragraphs 94-98). As noted, census and survey editing detect errors in and between data records. This *Handbook* describes micro-editing, which concerns the ways to ensure the validity and consistency of individual data records and relationships between records in a household.

135. Another method, macro-editing, checks aggregated data to make sure that they are also reasonable. In this method, tables are run on the edited data and checked against predicted frequencies and tolerances to identify various problems with the data; if "errors" are found, the macro-edit can make a global change to the aggregated data, send a unit record back for reprocessing, or add new micro-edits to rectify the problem. For example, a country may have a very large percentage of persons without a reported age. After imputing for age to obtain a complete data set, checks at the macro, or aggregate level could make sure that selective under-reporting by older persons does not skew imputed values. The editing team could choose to take measures to alleviate the risk of potential skewing, depending on the results of the analysis. Both macro-editing and micro-editing require thorough testing before they are used.

136. As noted, editing should preserve the original data as much as possible. The editing team needs to have high quality, clean data, but also needs to preserve what the organization collects in the field. The original data need to be maintained at all stages of computer processing in case the editing team decides it needs to re-examine

the editing process. Sometimes the original data are revisited when the team discovers that a systematic error has occurred in the editing process. Sometimes a review occurs because part of the data set is found to be either missing or duplicated, and the data set must be re-formed and re-edited.

137. Sometimes the source of error is outside the processing office. Banister (1980, p. 2) notes that if "we know that a high proportion of some subgroup did not answer a particular census question, it means that they did not understand the question, that they resisted answering it or that they were apathetic about cooperating with the census". Hence, she argues that non-response rates for subgroups should be included on census archiving media. National census/statistical offices are now more likely to preserve these data for researchers and their own analysis to be able to improve response rates for such groups.

138. More and more evidence exists that no amount of computer editing can take the place of high quality census data collection. National census/statistical offices know that at some point computer editing is not only limited but becomes counter-productive: the edit adds more errors to the data set than it corrects. Changing a census item is not the same as correcting it. Hence, the editing team must work together to determine the beginning, the middle, and the end of the editing process.

139. Editing and imputation may or may not improve the quality of the data, but a clean dataset greatly facilitates analysis and use. The process begins with the design of the census questionnaire. Demographers and other subject-matter specialists usually determine its content, often in consultation with user groups. But, ultimately, census data are not produced "primarily for demographic purists but for a much broader audience of scholars, policymakers, and lay people" (Banister, 1980, p. 17). However, obtaining a census without invalid and inconsistent entries is essential when the credibility of the census and the national census/statistical office is at stake. As Banister notes, "Census organizations can cite instances of journalists writing humorous articles or citizens indignantly writing census officials about published tables showing three-year-old grandfathers and commuters riding non-existent trains".

140. The problem is determining how far to go to obtain a good quality dataset. As noted earlier, the advent of computers, first mainframe computers and then microcomputers and now electronic data collection technologies, has allowed for virtually complete automation of the editing process. In many national census/statistical offices subject-matter specialists have in fact, become editing enthusiasts. Hence, offices now perform many consistency tests that were difficult in the past, particularly those involving inter-record checking and inter-household checks. Unfortunately, this feature of microcomputers has also led to many problems, and the greatest of these is over-editing.

1. How over-editing is harmful

141. A data record that has been altered as a result of edits should be closer to the truth after those alterations than before. Edits are designed to detect and correct inconsistencies, and not to generate bias by imposing implicit models by the edits. When further editing has a negligible impact on the final results, it is called over-editing, and should be avoided. Over-editing has a negative impact on the editing process in several ways, including timeliness, cost, and the distortion of true values. It also gives a false sense of security regarding data quality. These concerns are reviewed below.

(a) Timeliness

142. The more editing a national census/statistical office does, the longer the total process will take. The major issue is to determine how much the added time adds to the value of the census product. Each editing team must test editing process and procedures and evaluate, both on an on-going basis and after the fact, the net benefits of the added time and resources for the overall census product. Often, the returns are so small in terms

of the time invested that it is better to have small "glitches" in the data rather than deprive prime users of receiving the information on a timely basis.

(b) Finances

143. Similarly, the costs of the census process increase as the time increases. Each national census/statistical office must decide, as it increases the amount and complexity of its edits, whether the increases in costs are worth the added effort and whether it can afford these additional costs.

(c) A false sense of security

144. Over-editing gives national census/statistical office staff and other users a false sense of security, especially when offices do not implement and document quality assurance measures. Furthermore, odd results will appear in census tabulations no matter how much editing the team does, so it is important to warn users that small errors may occur. This is especially true now that many countries release sample microdata. National census/statistical offices would not want to release data detrimental to the planning process, so great care must be taken to assure that all crucial variables are edited properly and can be used for planning. For example, no national census/statistical office would want to release microdata or tabulations with unknowns for sex or age. On the other hand, variables such as disability or literacy work as well with less editing. While some inconsistencies in the cross-tabulations may appear because national census/statistical offices cannot edit all pairs of variables, editing teams should check the most important combinations. When editing teams find inconsistencies, correction procedures should be available.

2. Distortion of true values

145. Although the intention of the editing process is to have a positive impact on the quality of the data, increases in the number and complexity of the edits may also have a negative impact. Sometimes, editing teams change items erroneously for a variety of reasons: mis-communication between subject-matter and data processing specialists; mistakes in a very complicated, sophisticated program; or handling a census item many times in an edit. National census/statistical offices want to avoid this type of problem whenever necessary. Granquist and Kovar (1997) point out, for example, that imputing the age of a husband and wife using a set age difference between them can be useful but may artificially skew the data when many such cases exist.

146. For example, a set of rules may require that the child of a head of household should be at least 15 years younger than the head. However, a child of the head may be a social, rather than biological child: He or she might be the biological child of the spouse, but not the head. Hence, the difference in age might be less than 15 years. Since planners in most countries do not plan separately for children and stepchildren, if, under the above circumstances, the editing rules change the age of the child, inconsistencies in educational attainment, work force participation and other areas may develop. Hence, this rule should be tested to see the results before being fully implemented.

3. Treatment of unknowns

147. The editing team must decide early in census planning how to handle "not stated" or unknown cases. As noted earlier, columns or rows of unknowns in tables are neither informative, nor useful, so planners in most countries prefer these data imputed. Without treatment of unknowns, many users distribute the unknowns in the resulting tables in the same proportions as the known data, thus imputing the unknowns after the fact. The editing team needs to decide how to deal with the unknowns systematically.

148. It should be noted that if an electronic questionnaire is used, before making a decision on unknown cases which is a part of data processing, the editing team has to decide what fields will be checked with hard-edits and which fields with soft-edits. This means that fields that are checked with hard-edits must be filled during data collection, therefore there will not be any unknown cases for these fields. Other fields that are checked with soft-edits may have unknowns. These cases should be worked out during data processing post data collection.

4. Determining tolerances

149. The editing team must develop "tolerance levels" for each item, and sometimes for combinations of items. Tolerance levels indicate the number of invalid and inconsistent responses allowed before editing teams take remedial action. For most items in a census, for example, some small percentage of the respondents will not give "acceptable" responses, for whatever reason. For some items, like age and sex, which are used in combination with so many other items for planning, the tolerance level might be quite low. When the percentage of missing or inconsistent responses is low (less than one or 2 percent), any reasonable editing rules are not likely to affect the use of the data. When the percentage is high (5 to 10 per cent, or more, depending on the situation), simple, or even complex, imputation may distort the census results.

150. To reduce missing responses to a minimum, the national census/statistical offices should ensure that census workers make every effort to obtain the information in the field. Electronic data collection clearly assists in this task. If a given country decides that it does not need as much accuracy for some items, such as literacy or disability, the tolerance level for those items might be much higher. Sometimes editing teams can correct items that have too many errors, by returning enumerators to the field, by conducting telephone re-interviews, or by applying their knowledge of an area. Often, though, it is too costly to return to the field or do other follow-up operations, and the national census/statistical office may decide either not to use the item or to use it only with cautionary notes attached.

151. The question arises as to who should determine the tolerance level for an item. The editing team, including both subject-matter and data processing specialists, may have to decide on tolerance levels. The subject-matter personnel must use the items over time and therefore have a professional stake in making sure they obtain the highest quality data. The data processing specialists, however, may find that they cannot develop appropriate editing programs to reduce the tolerance to acceptable levels or that the data themselves may not permit any program to be successfully within tolerance.

5. Learning from the editing process

152. As the data are edited, detailed analyses of positive and negative feedback need to be recorded to improve the quality of both current census or survey and future censuses and surveys. The editing team must work constantly to determine what is working properly and what is not working. They must also determine whether those aspects of the process that are working properly can be improved and streamlined, so that the data can get to users even sooner. The earlier in the census process national census/statistical offices detect errors, the more likely they will be to correct them.

6. Costs of editing

153. This *Handbook* can assist countries in reducing the high costs involved in both time and resources to complete the edit and imputation of census or survey data. Built-in edits during electronic data collection can help reduce the costs. For most countries, editing activities take a disproportionate amount of time and funding, so each country must determine the return on its investment.

154. Excessive editing can delay census results. While national census/survey staff may have only anecdotal evidence for such experience with censuses, a study by Pullum, Harpham, and Ozsever (1986) found that machine editing of the World Fertility Survey contributed to a delay in the publication of the results by about one year. National census/statistical offices might better spend their funding on obtaining a higher quality census or survey enumeration in the first place.

7. A sample of census data

155. In recent years, more and more countries are choosing to release samples of their data for researchers and students to use as needed in their work. The researchers benefit, of course, because they do not need to request special runs from the National Statistical Office. The Statistical office also benefits because they do not need to take the time to make and check the tables. Also, the outside user's serve both to determine the level and flow of census information, and so assist in making decisions about future labor force (and other) surveys as well as the next census.

156. The size of the sample depends on the size of the country and the census characteristics. Usually, the sample is "systematic" or a certain percent. A ten percent sample might start with a random draw from 1 to 10 and then pull every 10^{th} housing unit after the first. Five or 10 percent samples are the most frequent when the country is largest enough to provide an appropriate sample.

157. Sometimes, the microdata sample created is weighted, with the weight attached to each housing unit or population or housing or agriculture record. When this happens, different sequences will be used and a run without the weights will give anomalous results.

158. When countries develop samples for distribution, they sometimes need to do additional editing for ensuring data confidentiality and quality.For example, most offices do not usually worry too much about the relationship item because it is usually mostly used to check to make everyone in the housing unit was reported. But if a country does not edit the data, a user-developed table might show age by relationship with very young grandparents or very old grandchildren. While this would not affect planning, it should be considered when developing editing strategies.

159. The other issue is developing confidentiality in sample data sets. One of the easiest methods, in most cases, is moving small numbers of attributes for a variable to an "all other" category. Certain variables like religions, languages, ethnicity, may have certain attributes with very small numbers; these would be edited into an "all others" group before distribution. Of course, the statistical office would maintain the original data for its own purposes.

160. In other cases, like number of rooms in housing units, an edit might top off a variable. So, if only a few housing units have 9 or more rooms, anything above 9 would be put into a 9 or more category.

161. Also, in developing sample data sets, the organization might choose to use switching, that is, switching information between individuals and households to preserve confidentiality. This method is used most often when confidentiality might not be maintained for a certain area. For example, if an area has intense migration from many other countries or ethnicities, an individual might be identified from the composition of the record or housing unit. So, sometimes organizations choose to switch some people or units with another unit some distance away. That way, no user can be certain that the house they think they are looking at is actually that house. But, even with switching, problems could still remain, even when switching units or individuals. In this case, whole units might be substituted.

8. Archiving

162. Part of the quality assurance process of the census or survey is to document all processes and then to archive that documentation. National census/statistical offices need to preserve both the edited and unedited data files for later analysis. Some procedures, such as many forms of scanning, automatically keep the original image. Similarly, immediately after keying batches, the data should be concatenated and preserved for potential analysis. But, with either procedure, it is important to archive original copies of the non-edited files. In fact, copies of the unedited data should be kept in several places within the Statistics Office, as well in other parts of the country, and outside the country as well.

163. The documentation should be complete enough for census or survey planners to be able to reconstruct the same processes later to assure compatibility with the census or survey under consideration. The processes and the results must be replicable. Finally, the unedited data as well as the edited data must be stored in several places, with appropriate measures to ensure their continued availability over time.

164. As noted elsewhere, part of the documentation involves the two types of edit reports. The first report provides the summary statistics giving numbers and percentages of errors (based on appropriate denominators, like total housing units, total population, working age population, adult females, etc.). The second report contains at least a sample of the "case" structure, with the unedited household or housing record, the listing of errors and their resolutions for the housing unit or individuals in the unit, and the edited housing unit or household.

165. The two types of reports should be provided at logical geographic levels, certainly for the major civil divisions, but providing error listings at lower levels of geographic levels could assist in targeting problems in enumerator training, quality control, or other issues connected with the enumeration. For more information on archiving, see Part 3 Chapter XII of the Principles and Recommendations for Population and Housing Census Revision 3.

E. TESTING OF EDITING PROCESS

166. Testing of the editing strategy is critical to the process of developing the most effective rules for collecting reliable data. Testing is an iterative process consisting of testing, making changes and corrections and re-testing, until the editing rules and imputation procedures are as much as possible error-free.

167. Once the editing process has been designed, it needs to be tested in order to assess its suitability for the observed data and to find the best set of methods and steps. The editing process contains a number of steps, such as for example the detection of critical errors, the detection of inconsistencies errors and the imputation of data. Consequently, the testing of the process should be split into the testing of the different steps/methods used. The testing aims at evaluating the performance of the methods in terms of efficiency. Based on the results from testing, some of the design decisions and/or parameters could be revisited to optimize the performance. (Eurostat, Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys).

168. Testing of editing procedures is a broad term that incorporates many different methods or combinations of methods. Each method has specific strengths and weaknesses that make them valuable for different types of problems. Consequently, they are useful at different stages of developing the editing process. In order to identify problematic issues and to suggest adequate improvements, the use of a combination of testing methods is indispensable.

169. The team for testing should be formed by staff with different background (census experts, programmers, subject-matter experts, researchers), since different typologies of errors are more easily identified by different experts.

170. Testing of editing process are explained in two major categories: pre-field and field methods. This distinction is an analytical one, but nevertheless reflects some principal and operational differences.

171. **Pre-field methods** are normally (although not necessarily) applied under laboratory conditions. This means that the interviews are not carried out in exactly the same way as later on in the field. "Laboratory conditions" refer to an observational environment which may totally or partially differ from the actual field conditions. Testing of the editing process undergoes rigorous tests before conducting field tests in order to verify the process works the way designers planned and validate the editing rules developed for all topics. Editing rules might be tested with a sample of data, collected from small number of people and/or using the dataset of existing household surveys. Pre-field methods are particularly suitable to verify the rules for identifying inconsistent data and missing information. The focus would be on single topic or multiple related topics rather than the whole topics. They should include expert group reviews, especially for ensuring to include extreme cases or maximum and minimum values in the editing process such as maximum number of children ever-born, year of birth, number of working hours per week, so on.

172. **Field methods** are those used to evaluate the editing process under field conditions. This means that the interview is carried out in a way very similar to the subsequent fieldwork (regarding setting, lengths, choice and order of questions, etc.), and the majority of the conditions mirror the real census situations. The test might be conducted as a part of whole census process, e.g. in the context of a pilot study, in conjunction with the actual data collection, or in parallel to ongoing or recurring surveys. Therefore, field testing often includes bigger sample sizes and allows quantitative analyses. The focus is more on the complete editing process including editing during data collection and after data collection, census questionnaire (s) and data collection application.

173. When CAPI or CASI is used, the adoption of electronic questionnaire has prompted redesign efforts to explore new prospects in data collection. This has added a new level of complexity to questionnaire testing. One new dimension is usability testing, which is intended to assess if the collection tool is user friendly and whether the interaction with the computer is intuitive and simple for the users of the electronic questionnaire -either respondents or enumerators.

174. Testing of built-in editing process aims to:

- Ensure edit checks are understood by respondents/enumerators; errors are clearly described and appropriate actions are explained to whether correct or keep the response as it is and instructions are well designed for reducing the risk of respondent error;
- Consider a number of edits to ensure that non-response isn't introduced;
- Consider inclusion of hard checks for key variables and soft edit checks for others;
- Consider how edits should be presented to users.

175. This type of test will be conducted in different perspectives. Respondents/enumerators will be observed to see how they react to the various edits i.e. are they noticed? and do the respondents/enumerators take right action? Users' feedback is also sought on how the edit design could be improved. The results of this type of tests will lead to make a decision on what variables should be edited with hard checks and how many edits to include.

176. While testing editing steps and procedures for electronic data collection, the following types of tests should be considered[6]:

177. **Usability testing** – Generally, usability testing results suggest the need for a good visual questionnaire design that uses fonts of different size and color for questions and answers, can facilitate the answering process and reduces the completion time (Hansen and Couper, 2004). Though usability tests have their limitations, as they are conducted on a small number of users and try to test an entire questionnaire, not only the editing aspect, they can be a source for best practices for designing edit rules.

178. Usability principles advocate the basic rule: user needs should be at the centre of the design. All tasks to be performed should be under the user's control. Throughout the response process, during the data entry stage, edit messages can appear several times and in various forms. The user needs to be able to choose the right moment to deal with them and to ensure the action taken is effective, which requires inter-connectivity between edit messages and the item back and forth as desired. The policy on how data with unresolved items will be submitted should be included in the instruction manual. It should be clear whether data marked as erroneous can be submitted. In other words, the question is whether strict conformity to edit rules should be required or rather whether users should be allowed more freedom in this matter, which will make them more likely to provide data, thus reducing the rate of non-response.

179. **Analyses of collected data** – Data collection for population and housing censuses will be implemented by many and different types of users. This makes necessary to test the data collection application in different environments and for different population groups. A way to evaluate the set of built-in editing rules can be the number of non-response items. An issue when users tried to fit values to the upper or lower bound of a range edit when it exceeded the range can be an example for tracking too rigorous edit rules in questionnaires (Anderson et al., 2005).

180. **The burden** – Incorporating edit rules into the questionnaire does not necessarily increase response burden (Anderson et al., 2005). Usability studies showed that some automatic checking of data entries are awaited by users to be performed by a computer. If the goal of edit checks is clearly understood by respondents the tolerance and acceptance for them can be easier gained. The limit for the scope of edits can be drawn from usability testing. The aim of edit rules is to improve data quality and not to encourage non-response.

F. THE EDITING TEAM

181. As national statistical offices prepare for a census, they need to consider a variety of potential improvements to the quality of their work. One of these is the creation of an editing team. The editing process should be the responsibility of an editing team which usually consists of a number of task teams and a group of supervisors and managers. Task teams should be established for different stages of census editing such as a team for checking census coverage and for census topics such as a team for editing data on demographic characteristics and a team for editing data on economic characteristics of population. Supervision of and coordination between these task teams should be under the responsibility supervisors and managers. These teams should be consisting of census managers, subject-matter specialists and computer programmers.

182. Managers and supervisors of editing process should be set up as soon as preparations for the census begin, preferably during the drafting of the questionnaire. The editing team is important from the beginning and remains so throughout the editing process including evaluation of the quality of census outputs (see the editing process steps). Care in putting together the team and in developing and implementing the editing and imputation rules assures a census that is faster and more efficient.

[6] Eurostat, 2014. "Editing During Data Collection". This module is part of the Memobust Handbook on Methodology of Modern Business Statistics

183. Meetings between census officials and the user community concerning tabulations and other data products can provide insight into the edits that need to be performed. Frequently, users request a table or type of tables that requires extra editing to eliminate potential inconsistencies. The editing team should plan to implement these tables during the initial editing period rather than implementing them as special tables after census processing.

184. Subject-matter and data processing specialists should work together to develop the editing and imputation rules. The editing team elaborates an error scrutiny and editing plan early in the census preparations. The census or survey editing team creates written sets of consistency rules and corrections. In addition to developing the editing and imputation rules, the subject-matter and data processing specialists must work together at all stages of the census or survey, including during the analysis. The risk of doing too much editing is as great as the risk of doing too little editing and having unedited or spurious information in the dataset. Hence, both groups must take responsibility to maintain their meta-databases properly. The editing team must also use available administrative sources and survey registers efficiently to improve subsequent census or survey operations.

185. It is necessary also to build an effective collaboration between the editing team and the evaluation team which is mainly responsible for the planning and implementation of evaluation of the data quality and overall evaluation of census operations. Collaboration between two teams should allow good knowledge about each other`s work. Particularly, the evaluation team should know the editing process and procedures and other issues that may affect the work of evaluation team such as why item non-response is relatively higher among some variables compared to others. This collaboration should also involve a shared vision regarding the data quality and in depth understanding of the roles of each team with the goal of achieving a high quality.

186. Developing the editing rules and the computer programs during a pretest or dress rehearsal makes it possible to test the programs themselves and leads to faster turn-around times for various parts of the editing and imputation process. The editing team then ascertains the impact of these various processes and takes remedial action if necessary.

187. When electronic data collection technologies are used, these procedures may be adapted. Since some editing will be done on entry and other editing will be done in the office, the editing team should work together to obtain the best distribution of the edits which can be implemented during data collection or data processing in the office. That is, the amount of editing done in the field may be too much or too little relative to the amount of editing done in the office. The proper balance will provide the best data set in the shortest possible time.

H. EVALUATION OF THE EDITING PROCESS

188. Evaluation of the overall census operation is vital for identifying strengths and weaknesses of census phases, including planning, enumeration, data processing and dissemination, and also for the purpose of analyzing the quality of census statistics, which are the major output of these processes. Census evaluation with all dimensions[7] of quality requires a comprehensive evaluation programme for assessing and documenting the outcomes of each process using appropriate and customized methodologies. Methodologies for evaluation should be planned well in advance, in the planning phase of the census. It should be noted that this is a continuous process implemented from the planning to the end of census operations. It is also appropriate to consider it as

[7] UN, Principles and Recommendations for Population and Housing Censuses Revision 3, Part Two, Chapter XIV. Quality assurance

being the first step in the subsequent census cycle. Similarly, evaluation of one process within a census cycle could be the first stage in the next process of the same census cycle.

189. Plans for the evaluation of the editing process should be an integral part of the overall census plan and must be planned from the start of census activities. The editing process like other census processes consists of a series of interrelated and repetitive activities and sometimes can be very expensive statistical activity if the process is not designed well. Therefore, this process should be evaluated to provide lessons learned from one census to the next. For this reason, evaluation is generally undertaken as a part of overall evaluation of the census to be able to evaluate its impact on the quality of other census processes[8,9]. Evaluation of the editing process can assess the effectiveness of the process design and procedures and also efficiency of systems and applications used in the process and their impact on data quality.

190. In general, evaluation of the editing process can be conducted for assessing two major issues: (a) Whether editing methods help to achieve pre-determined levels of data quality and (b) Whether the quality of the editing process achieves to produce a complete and consistent database. Evaluation should be focused on the efficiency of editing processes and raise the question whether resources spent on editing are justified in terms of quality improvements. In some cases, the question can be answered by studying the impact of editing changes on the estimates. More specifically, the following questions can be considered in the evaluation of the editing process:

a. Is the use of resources spent on editing justified?
b. Can more effective editing strategies be applied?
c. What are main achievements and difficulties in use of new technologies and methodologies, and identification of possible improvements for the next census?
d. Is the planned calendar for editing process achieved? In the case of changes to the calendar, what are the reasons and consequences?
e. Can the quality perhaps be improved by allocating some of the editing resources to other processes to prevent errors, such as questionnaire design and data capture?.

191. The evaluation of the editing process should be undertaken by census managers, subject-matter specialist, programmers and data processors according to the agreed goals and methodologies covering all possible dimensions of quality7. To obtain measures on how editing affects quality, it would be necessary to conduct re-interview studies, record check studies or simulation studies[10].

[8] UN, Principles and Recommendations for Population and Housing Censuses Revision 3, Part Three, Chapter XIII. Overall evaluation of the census
[9] UN, Handbook on Management of Population and Housing Censuses Revision 2, Chapter VII. Evaluation
[10] UN, Technical Report on Post Enumeration Survey, Chapter 1. Overview of Census Evaluation and Selected Methods

III. EDITING APPLICATIONS

192. This chapter provides a general overview of the applications for the editing and imputation process. It provides a framework for the general flow of the census or survey edit, from raw scanned or keyed data, through structure editing and content editing, to provide an edited data set. It gives selected examples to illustrate which kinds of problems unedited data may present for users and why edited data are more useful. It considers issues of keying and coding as part of the preliminary editing process. The chapter also presents general issues in computer editing along with guidelines on topics such as checking for validity and consistency. The two generic types of computer editing, static imputation (cold deck) and dynamic imputation (hot deck) techniques, are reviewed in detail.

193. When this handbook was originally written for the 2000 round censuses, almost all countries keyed their data. In the 2010 round censuses, most countries scanned, sometimes with keyed follow-up. For the 2020 round censuses, many countries use electronic data collection technologies, primarily handheld electronic devices and/or Internet for data collection during the field enumeration. The current handbook attempts to take keying and scanning (entry from paper) and electronic data collection "edit on entry" into account for the structure and content edits. Just as technologically developing countries had problems with scanning in the early 2000s, many countries are also facing several challenges in the adoption of electronic data collection technologies in censuses

194. Whether census data is collected with paper questionnaires (scanned or keyed at the office), a certain general flow pertains. The census edit team starts with the unedited data. In most cases, all data have been precoded by the enumerator or by office staff, so the data set is ready for the structure edit. In some cases, an operation is needed to convert the scanned data into another machine-readable form for the editing process, depending on the editing package to be used. Also, in some cases, however, the scanned data require a second automated coding operation to fill in items like birthplace, industry, and occupation.

195. In either case, the unedited data should appear in a form allowing the computer programs to develop the **structure edits** (as described in Chapter IV in more detail). The structure edit checks to make sure that all of the major civil divisions are presented in geographic or numerical order, and within each major civil division, each minor civil division occurs, and in geographic or numerical order. Then, within each minor civil division, each locality must appear, and with geographic or numerical order. This procedure continues down to the lowest geographic level. As described in the next chapter, appropriate procedures must be taken to make sure that each housing unit appears once and only one in the data set.

196. The structure edit must also make sure that all record types are present when appropriate, and that no record types are repeated when they should not be. So, for a population and housing census, either the population or housing records must come first, and then that convention must be repeated throughout the whole data set. In most cases, only housing record should be present, so that surplus records must be dealt with, and programmers must supply housing records to households without housing records. Similarly, population records must be present for occupied housing units (usually defined as such on the housing record) and must be absent for vacant units.

197. After the structure is set, it is not really final. It is important to note that, inevitably, the structure edit will be re-visited during the content edit, and often beyond, as glitches appear during the various census processes; this is normal census procedure, should be expected, and time, personnel, and equipment requirements should be built into the total system.

198. Then, the **content edit** begins. Each population and housing item must be considered alone and usually in combination to determine the validity of each item, and the best fit among the items. Chapters V and VI cover the various population and housing items in the U.N. Principles and Recommendations for Population and Housing Censuses, Revision 3.

199. When the content editing is done, a completely edited data set should be established. The unedited data should be stored in several secure places, and the important unedited items (or all the unedited items) should also appear at the ends of the various types of records. Again, it is important to note that as the tables are developed, the content edits may have to be re-visited as well to take care of any specific problems resulting from particular cross-tabulations.

200. The purpose of editing censuses and surveys is to discover omissions and inconsistencies in the data records; imputation is used to correct them. Editing establishes specific procedures to deal with omissions and various types of unacceptable entries. Imputation changes invalid entries and resolves inconsistencies found in the dataset. The product is an edited microdata file for tabulation, containing acceptable and consistent entries for all applicable data items for each housing unit and person enumerated.

201. It is important to note, again, that no amount of editing can replace high quality enumeration. The editing process works well when imputations are used to deal with random omissions and inconsistencies. However, if systematic errors occur during data collection, editing cannot improve the quality of the data no matter how sophisticated the procedures. The choice of topics to be investigated is of central importance to the quality of the data obtained. When interviewed, respondents must be willing and able to provide adequate and appropriate information. Thus, it may be necessary to avoid topics that are likely to arouse fear, local prejudices or superstitions, as well as questions that are too complicated and difficult for the average respondent to answer easily in the context of a population census. The exact phrasing for each question that is needed in order to obtain the most reliable response will of necessity depend on national circumstances and should be well tested prior to the census. It is therefore of the utmost importance that national census/statistical offices should allocate sufficient resources to obtain the highest quality census data.

202. To implement the office computer editing phase of the process, the editing team prepares written editing instructions or specifications, decision tables, flow charts and pseudocode. Pseudocode is a set of written editing instructions or specifications for the programmers to follow as they develop the edits.

203. Flow charts can help the subject-matter specialists understand the various linkages among the variables and make it easy to write editing instructions. Sample flow charts are given in Annex V. The subject-matter specialists write the editing instructions in collaboration with the computer specialists, describing the action for each data item. The editing instructions should be clear, concise and unambiguous since they serve as the basis for the editing program package.

204. The whole census editing team, both subject matter specialists and data processors, should have extensive exposure to demographic data processing and analysis. Unqualified personnel may unintentionally introduce additional errors and bias into the census.

A. CODING CONSIDERATIONS

205. During much of the second half of the 20th century, as noted earlier, countries keyed their data. Most countries now use electronic data collection technologies or scan their censuses. For any kind of data capture method, paper questionnaire or electronic questionnaire, certain variables still need to be translated from words to numbers. The process of making machine-readable numbers and alpha-numerics is called **coding.**

206. Codes that are completely alphabetic characters or alphabetic characters combined with numbers (called alphanumerics) should be avoided whenever possible. When forms are scanned, alpha-numerics are not a great problem, but many computer packages require considerable manipulation or at least consideration in their use. In

many cases, editing programs require that alpha characters be placed between quotation marks, or in some other manner, in order to process them.

207. When developing a coding scheme, census and survey staff must consider the returns of each investment of time, energy and funds. Coding considerations are reasonably insignificant for small countries or small surveys since the amount of processing is much less than for a census. Also, data that are scanned don't suffer as much from additional columns of information.

208. When a census or survey uses two columns for the item *relationship*, for example, rather than one, scanning will introduce errors that would not be present when a single column of information occurs. That is, if you have codes 1 through 9, the scanner may pick up an alpha character, or a blank, or a stray mark converted to some readable character, but these issues are readily handled in the edit, as described later in the text. However, when a census uses two columns, say codes 1 to 10, then a whole new realm of errors can be introduced. Instead of legal values 1 to 9, there are other values coming in that could range anywhere from 0 to 99, as well as the aforementioned alpha characters, blanks, and stray marks. When the editors receive a value of 13, they must start making strategic decisions about what to do with this value. Was it meant to be 3, and the 1 is erroneous? Was it meant to be 10, and the 3 is wrong? In most cases, the subject specialists provide the edit specifications for the item, but these values automatically increase the time and complexity of the edit and could decrease the quality of the final data set.

209. One of the most common problems, and one discussed later on specifically, has to do with items in the fertility series. Many countries now collect information on children living in the household, children elsewhere, and children dead, and sometimes collect the sum of these children, and by sex of child. So, countries could have up to 12 items of information. The issue here is how many digits each of those items should be. When two columns are used, the boys in the house could be anywhere from 0 to 99; when only one column is used the numbers can only range from 0 to 9. However, since it is extremely unlikely that a female would have more than 9 male children in the household, having two digits introduces high probability of picking up stray marks or scanning misreads – reading 9 for a 0, for example, so 91 children instead of 01. `

210. So, for boys and girls present in the house, currently elsewhere, or dead, single columns would probably be most appropriate. However, for total children in the house, total children elsewhere, total children dead, and total children, two columns might be more appropriate. Much depends on the fertility levels in the country. Occasionally, an unusual household will actual have more than 9 people in a particular category, but, as always in census work, the statistical office will have to decide on the relative balance between errors and good data.

211. For ordinal variables, consider the following series of codes for relationship:

0. Household reference person or head of household
1. Spouse
2. Child
3. Spouse of child
4. Adopted or step-child
5. Sibling
6. Parent
7. Grandchild
8. Other relative
9. Nonrelative

212. This set of standard codes covers the majority of relationships for most countries. Some countries may start with "1" if the classification is not detailed and less than 10 categories.

213. Even these codes can be used to obtain household composition as shown in Annex III on derived variables. However, many countries, particularly those living in the extended household (see the UN P&R revision para.4.146)

or experiencing the HIV/AIDS epidemic need much more detailed information than cannot be provided by these codes. These countries may need specific information on children-in-law, parents-in-law, grandparents, nieces and nephews, and so forth. In this situation, additional codes are required for the statistical office to carry out its mission, and so two-digit codes are required.

214. When a country decides to use multiple columns, it also needs to decide on how to use those columns. In the example above, the assumption is that the codes for relationship will be sequential. However, once the decision is made to use two columns, the subject matter specialists for this item may choose to use the columns to have significance. For example:

Sample of coding scheme using generation and relationship to the reference person or head of household

10	Head of household
11	Spouse
12	Sibling
13	Sibling's spouse
21	Child
22	Adopted child
23	Step child
24	Niece/nephew
31	Parent
32	Parent-in-law
33	Uncle/aunt
41	Grandchild
77	Other relative
88	Non-relative
90	Institutional population

215. This scheme uses generation in the first column – 1 for head's generation, 2 for one generation down, 3 for one generation up, 4 for two generations down, etc., and then numbers the types of relatives within each of the categories. These values could be useful in family reconstruction, but, of course, could be more cumbersome for the office staff and certain, general users.

216. This type of coding, though, should be considered for certain social and economic variables. For ethnicity, for example, the major tribal or ethnic grouping would be in the first of two columns and the minor tribal or ethnic grouping (like a sect) would be in the second digit. When more than 10 minor groups appear, two numbers in the first column would obviously need to be used.

217. Similarly, for three or four digits, like occupation or industry, the first digit would be for the major occupation or industry, the second digit for the minor occupation or industry, and the third digit for specific occupation or industry. Most international coding schemes by the United Nations agencies, and others, already have the levels imbedded in the codes, so the statistical office does not have to do any additional work.

218. As national census/statistical offices develop lists of codes for the editing programs and for subsequent tabulations, they may wish to establish common codes for some items. For example, in many countries, place codes (birthplace, parental birthplace, previous residence and work place), are very similar. A common coding scheme for "place" might be developed considering administrative levels in a country. At the global level, country codes can be developed as three-digit codes with the first digit representing the continent, the second the region, and the third the specific country. National census/statistical offices can use the *"Standard Country or Area Codes for Statistical Use"*

(originally published as Series M, No. 49 and now commonly referred to as the M49 standard) prepared by the United Nations Statistics Division (see the link below for the list of countries or areas contains the names of countries or areas in alphabetical order, and their three-digit numerical codes[11]). A set of common codes for closely related variables can reduce coding errors and assist the data processors during the edit. Common codes also allow data processors, where appropriate, to use an entry from one item to determine another.

219. The structure of coding can facilitate the coding process as well as later processing during editing, tabulation and analysis. For large countries with many immigrants or ethnic groups, codes based on continent, region and country, with different codes or digits assigned to each, would be preferable to a simple listing.

220. Figure 2 provides examples of common codes for such items as birthplace, citizenship, language and ethnicity. For the Philippines, the codes for speakers of Ilokano and Tagalog are different from the general code for the languages of the Philippines. Depending on the specific country situation, these codes could be different from each other as well. While the English language has a single code, it is spoken by more than one ethnic group. Therefore, the codes for birthplace, citizenship and ethnicity in Canada and the United States are slightly different. For persons born in France, having French citizenship, speaking French and having French ethnicity the same code is used. Hence, if one of these items is missing and if the editing team decides this solution is appropriate, a data processor can move the code from one of the other entries.

221. If a group of items on a questionnaire is not independent of each other, national census/survey staff probably should not ask all of them. The editing team must decide, on a case-by-case basis, when to use other items directly for assignment, and when to use other available variables.

Figure 2. Examples of common codes for selected items

Group	*Birthplace*	*Citizenship*	*Language*	*Ethnicity*
France/French	10	10	10	10
Spain/Spanish	20	20	20	20
Latin America	25	25	20	25
Philippines/Filipino	30	30	30	
Ilokano			32	
Tagalog			32	
England/English	40	40	40	40
Canada	50	50	40	50
USA	52	52	40	52

222. Another problem occurs when definitions differ between censuses (or between a census and a survey) for variables such as work or ethnicity. The national census/statistical office must decide how to take these changes into account, both for currently edited data and for datasets from the prior census, in order to show trends. If the original, unedited data are available, data processors can make changes to the appropriate edits and rerun all of them.

223. For example, a European country may use a single code for country of origin for all of the South Asian countries when only a few cases are identified. Because of changing migration patterns, however, the next survey or census may require separate codes for India, Bangladesh, Pakistan, Sri Lanka, and other South Asian countries all the way through the processing.

[11] Standard country or area codes for statistical use (M49) https://unstats.un.org/unsd/methodology/m49/

B. MANUAL VERSUS AUTOMATIC EDITING

224. Data editing can be performed manually, with the assistance of computer programming, or a combination of both techniques. Very small countries use only manual editing in the census. Most countries use a combination of both techniques that is, limiting the manual editing to influential errors, particular for census coverage, and by automating the editing process as much as possible. In designing an editing process, it is important decide which tasks can be done automatically and for which tasks (additional) manual editing is required. Such decisions can be made in advance based on the specific nature of the task. For example, manual editing might be needed for geography edits and for cases that are failed when automatic edits are used.

225. Manual editing of a census or survey is likely to take months or years, presenting many possibilities for human error. Manual editing is a weak alternative to computer editing, partly because it is impossible to create or reconstruct an edit trail for the manual correction process. Computer, or automated, editing reduces the time required and decreases the introduction of human error. Both computer and manual editing check the validity of an entry by looking for an acceptable value, but computer programs also check the value of the entry against related entries for consistency. Finally, and most importantly, automated editing allows for the creation of an edit trail and is therefore reproducible, while manual editing is not.

226. In the early years of computer entry, no editing on entry was possible. That is, all correction had to be either manual as part of the coding and checking office operations, or as part of the computer operations, but after the data were keyed. Newer packages have built in edit functions so that invalid entries cannot be entered, unless forced by the keyers, and inconsistencies can be flagged, to be corrected by the keyers, manually, or by a programmer. As scanning has become more prevalent, this sequence has been repeated; in the early years of scanning, no edit during entry was possible, but recently validity edits and data conversions and recodes can also be built into the scanning systems.

227. When censuses and surveys collect large volumes of data, staff cannot always refer to the original documents to correct errors. Even if the original questionnaires are available, the data recorded on them may sometimes be wrong or inconsistent. A computer editing and imputation system corrects or changes erroneous data immediately and generates reports for all errors found and all changes made. Computer edits should be carefully planned to save staff time for other data processing activities. While running large quantities of data through a computer system can be time-consuming, it is never as time-consuming as manual correction.

228. Manual correction takes several forms. Consider a simple example of an error in the sex response: a supervisor checks an enumerator's work and finds an obvious error, such as assigning "male" to someone named "Mary". In changing the sex to "female," the supervisor performs a manual edit. If the supervisor does not correct the questionnaire, but instead sends it to the field office, the office workers there may observe the problem and manually correct it. At the central office, during coding, coders might see the mismatch between the name and the sex and make the manual correction then. Or, the coders might not observe the problem, but when the keyers are entering the data for the questionnaire, they may notice the mismatch between the name and the sex and make the manual correction before keying.

229. However, if the error is not noticed, and the keyer enters the code for "male", a number of different procedures may be followed at this point. For gender-related items such as the fertility block, the editing program might flag the fact that this is a male with fertility information and produce a message to that effect while the keyer is entering the data. The keyer could then look at the questionnaire, find that indeed this is a female and make the correction manually. Alternatively, if the national census/statistical office uses an editing program independent of the keying, the computer program might flag this person as a male with fertility information. Then, by using the geographical information, office workers can find the original questionnaire in the bins, pull it and determine that the respondent, named "Mary", was erroneously reported as "male" instead. At this point, the office staff can take this information back to the keyer, who can pull up the record and make the manual correction. It should be noted that, depending on

the types of edits on entry implemented, usually electronic data collection applications will not allow the entry of fertility information for males. When sex information is not correct, enumerators first have to change "male" with "female" then they will be able to enter fertility information. In order to ensure the quality of sex information, it is suggested that sex information should be validated during data collection by asking this question while listing household members and then re-asking (or getting a confirmation) in the individual questionnaire. Information on age should also be validated using the same approach to ensure the data are correct and will not create any error to collect all required information from all eligible persons.

230. This example of using "name" for validating sex information shows both the advantages and disadvantages of manual editing. At any of the steps outlined above, a census worker could note the error—the mismatch between the name and the sex—and make the correction. However, national census/statistical offices that use manual editing probably have staff checking for this relationship at every stage. An enormous amount of time and energy is expended in this activity, and the results are probably little different, particularly in the aggregate, than if the staff were instructed to do no manual editing.

231. Originally, the only way to make corrections in a dataset was to make this change manually. Some countries still do not feel comfortable using automatic correction, so they use manual correction at one of the stages described above. If the dataset is small, timing is not crucial or the work force is labour-intensive, then manual correction will work in many cases. The advantage is that if the information is both complete and accurate on the questionnaire, and the inconsistency can actually be resolved by looking at the form, the quality of the survey will probably improve marginally (the editing team has to assume, for example, that "Mary" is not "Gary"; that if fertility appears, it was actually supposed to be collected for this person – that it was not collected erroneously). In fact, editing and imputation procedures rarely improve the quality of the data collection. They only change certain elements.

232. Sometimes, looking up a questionnaire for manual correction is fruitless. The information is not there, for whatever reason. Sometimes a person does not want to provide his or her age, so the item is blank on the questionnaire. In this case, examining the questionnaire will not resolve the issue. Then, the editing team must make a decision about how to handle the situation. For manual correction, the national census/statistical office must either assign "unknown" or use some set of values to assign the age item.

233. Manual correction inevitably lowers quality and consistency unless the respondent is contacted. It takes more time, and it costs more. Computers do not tire and are faster; they do not have personal problems that might interfere with maintaining quality or consistency; and, in most cases, they make processing cheaper. Most countries now use some kind of automatic correction.

234. Missing and inconsistent responses reduce the quality of data and make it difficult to present easily understood census tables. Some users prefer to tabulate missing and inconsistent responses as a "not reported" category, while others prefer to distribute these cases proportionately among the reported consistent entries. Still others recommend rules for imputing "likely" answers for missing or inconsistent responses. The use of computers makes it feasible and efficient to impute responses based on other information in the questionnaire or on reported information for a person or housing unit with similar characteristics.

235. Since the computer can look at many characteristics, the editing process should take advantage of this feature. Thus, editing procedures involving many related characteristics may result in imputing more reasonable responses than a simple edit could produce. On the other hand, poorly designed editing may lead to the production of poor census data. The editing team should be composed of experienced subject-matter specialists from different relevant disciplines as well as data processors. The members of the editing team should carefully select the variables to examine in the tests for consistency in order to determine the editing and imputation specifications. The program outputs should include the percentage of responses that were changed or imputed. Analysts will then be in a better position to judge the quality of the data; for example, a high percentage of imputations would be a warning to use the data with caution.

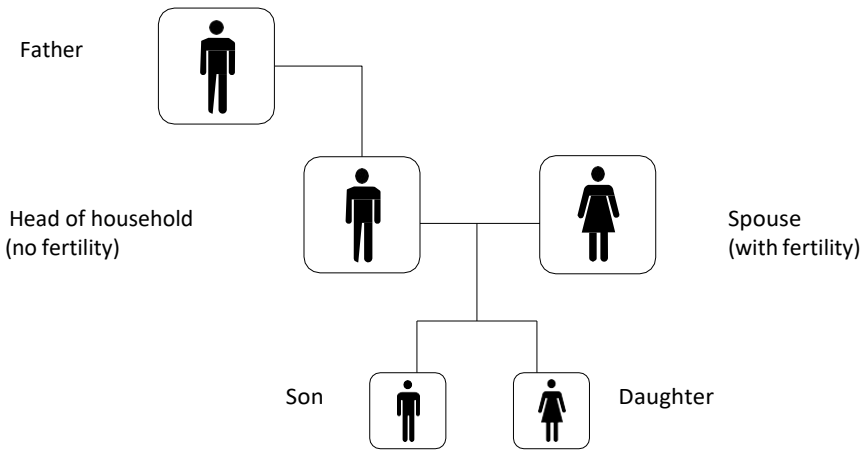
236. The edit, or audit, trail shows the changes made to each variable. The trail is used to trace the history of the responses from the receipt of the data through the editing and imputation process.

C. CONSIDERATIONS FOR CORRECTING ERRORS

237. Whether performed manually or automatically, editing should make the data as nearly representative of the real-life situation as possible by eliminating omissions and invalid entries and by changing inconsistent entries.

238. Consider the following diagram (figure 3) for a particular household. The diagram shows a household with consistent relationships and sex entries. The head of household is male and has no fertility information; the spouse is female and has appropriate fertility information.

Figure 3. A typical hypothetical household including relationships, sex and fertility of the members



239. In many instances, however, information is inconsistent. The following questions then arise: what should the editing process be for a household with inconsistent entries? How should the editing team perform the edit, if the head of household and spouse are both reported as male, as in figure 4. In the past, the typical editing rule would have assumed that the first person in a couple is male, particularly if that person is the head of household, and that the second person, or the spouse, is female.

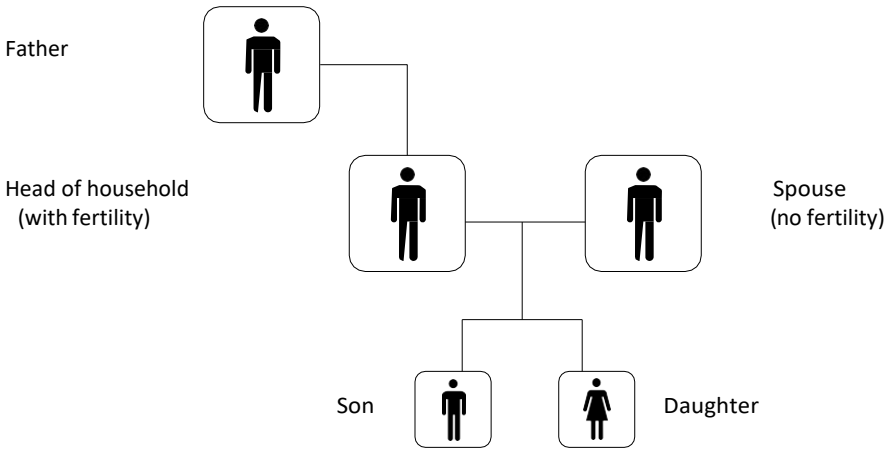
240. If the head of household in this case happens to be the wife rather than the husband, then the editing rule adopted would be wrong and the national census/statistical office would end up with four errors:

(a) The head of household's sex would be wrong;
(b) The spouse's sex would be wrong;
(c) The head of household would lose her fertility information;
(d) The male spouse would erroneously be assigned fertility.

This is clearly not good editing procedure.

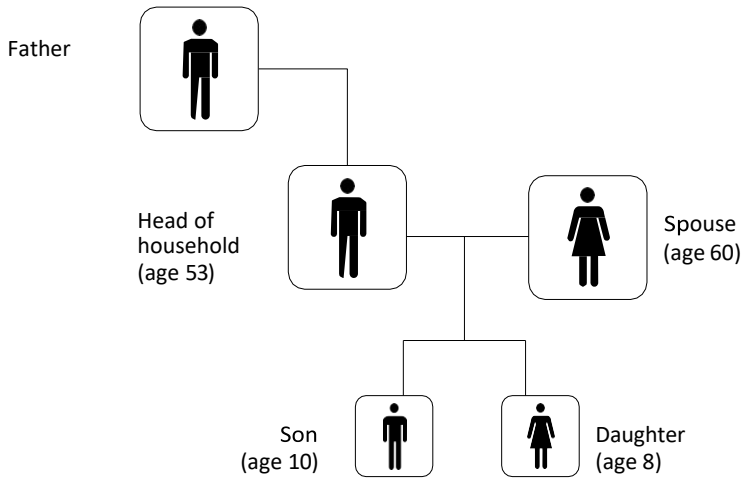
241. In contrast, when a good editing procedure finds that the head and spouse have the same sex, it then checks both persons for fertility. Since only the head has fertility, the head becomes the female. The editing rules for these items are then satisfied.

Figure 4. Example of household with head and spouse of the same sex



242. Another example, in figure 5, also illustrates the point. Most countries consider the age for child-bearing to be between 15 and 49 years old. Suppose a woman reports having a child at age 52, based on direct evidence through line number indicated for the child's mother or the computed age difference (the age difference between mother and a biological child should probably not exceed 50; adopted children could have larger age differences). The editing team must decide whether the age difference is acceptable or whether it must change, with the edit replacing one or the other of the ages. If the edit increases the acceptable age range for having children, and other women report having children at older ages, more anomalies may enter the data set if the age itself is misreported. Again, the editing team must decide the appropriateness of reported ages for particular variables.

Figure 5. Example of household with ages of some household members

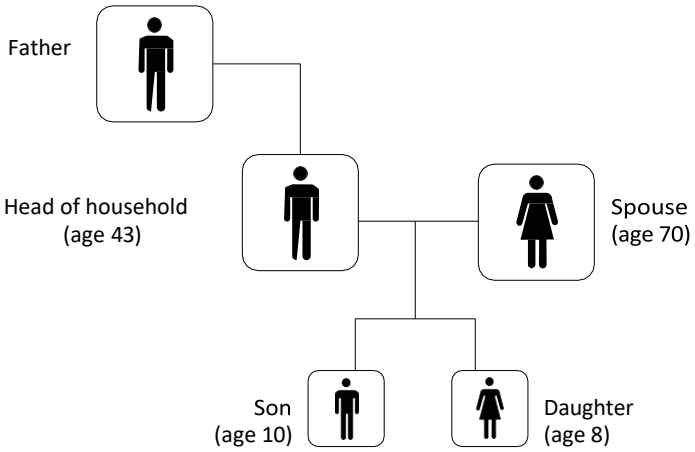


243. Figure 6 offers another possible scenario. Suppose the edit finds a 70-year-old female with children aged 10 and 8 as in figure 6. This situation is possible because the husband might have had the children with a previous wife. Under these circumstances, the children are related to the head of household, not the spouse per se, even though it may be more likely that keyers made an error by keying "7" when they meant to key "4" for 40. For whatever reason, suppose the subject-matter specialists require the data processor to change the age of either the mother or the child when more than 50 years separate the mother and children. This requirement leads to another, more complicated edit. Since the woman is 70 and the first child is 10, the editing team must decide whose age to change. The editing team

could decide to change the first child's age to 20, and that would resolve the problem for that first child, or it could change the spouse's age. A problem still remains with the second child's age, which also requires editing.

244. When considering only the ages of the mother and one child, an imputation would randomly assign age and would be right about half the time. However, when the edit also looks at the husband's age, the editing team would be more likely to change the spouse's age, based on this additional information. That one change would make the ages of the whole family more compatible.

Figure 6. Example of household with potential inconsistencies in age reporting



D. VALIDITY AND CONSISTENCY CHECKS

245. One of the major requirements in editing is that no item may contain invalid values. Additionally, responses for all related items within and between records must be consistent. Invalid entries are those that are unacceptable for technical or aesthetic reasons. For example, only codes for male and female are allowed for gender. (Introduction of transgender and other self-identified other sexes may change this.) Any other value would be unacceptable and would need to be changed to "unknown" or one of the two acceptable sexes; since most countries do planning and policy formation on the basis of sex for many other variables, having unknowns in the data set would complicate obtaining single values for the work. Similarly, tabulations with inconsistencies like "thatch walls and concrete roof", "females 13 years old with 20 children", a "3-year-old with a PhD" would make the statistical office appear inept, even if the few cases of inconsistency would not affect actual planning for a country.

246. Imputation should take into account all the information about related variables at the same time, to the greatest extent possible, and not necessarily sequentially with respect to related variables. In some cases, however, the edit may make a consistency check before determining the validity of an entry. If the imputation assigns a value based on the consistency check, it must compare the value to the original entry to ascertain whether it is an actual change. If it is not a change, the original entry remains as is.

247. For example, during the edit for marital status, relationship is checked first to see if the entry is "spouse"; if it is, and the spouse is not reported as married, "married" is assigned to marital status. Before the assignment of the

code for married, the program checks to see what the original response was. If the code for married is already present, the program does not change the entry and no error has occurred.

1. Top-down editing approach

248. The top-down editing approach starts with the first item to be edited (the "top"), which is usually the first variable on the questionnaire, and then moves through the items in sequence, until completing the edit of all items. The usual approach is to first take into consideration the response rates and the relative importance of the various items. Because of their importance, particularly in dynamic imputation, the edits usually start with sex and age. While the top-down approach does not completely preserve the relationships among the data items, it does provide an adequate framework to complete the edit.

249. During the editing process, some edits change the value for an item more than once. This procedure can introduce one or more errors into the dataset. An imputed value may be inconsistent with other data. Even when variables are dealt with sequentially, a particular variable should be edited against all other variables concurrently, if possible. For example, a child's age, imputed on the basis of the mother's age, may be inconsistent with the child's reported years of school or years lived in the district. In this instance, the age will be re-imputed until it is consistent. An imputed age is an intermediate variable until final assignment. In creating the edits, imputed intermediate variables should not be recorded as changes until the final assignment.

250. Although for a few items and conditions, the editing program might accept a blank or "not reported" entry, related information can supply entries for most items left blank or having erroneous entries. Entries supplied in this manner may or may not be correct on an individual basis. However, the extensive capabilities and speed of the computer for comparing different stored values permit the determination of replacement entries that reasonably describe the situation. The resulting tabulations in most cases will be sometimes more consistent than those from unedited records or records in which imputation converts all unacceptable entries into "not reported".

251. The editing program must also perform structural checks (see Chapter IV). The edit should then check population items (see Chapter V) and housing items (see Chapter VI). In addition, the editing procedures should probably create one or several recoded variables (derived variables) on the individual record required for the tabulation, as noted in annex II.

252. It is extremely important to avoid circular editing—making changes to an item or several items, and then, at some later point, changing them back to the way they were. Elsewhere this *Handbook* notes that staff must make several runs to make sure they completely edit all items. It is possible to create editing criteria that change the data during a first run, but that, when applied to the changed data during a second run, change it back to the original configuration. This procedure can continue through multiple runs. The editing team should avoid introducing such criteria into the editing process.

2. Multiple-variable editing approach

253. The "top-down" approach to census and survey editing which is the procedure that was introduced in Section 1 above, may not always give the best results—those that come closest to the real distribution of the variables. As indicated, the top-down approach, if applied without proper precautions, frequently causes problems in the edit.

254. Another approach is multiple-variable editing, which is based on the Fellegi-Holt system. This approach requires more computing expertise and computer power but probably obtains results that are closer to "reality". Different kinds of multiple-variable editing appear in annex VI, "Imputation methods". In the multiple-variable editing system it is necessary to determine a set of positive statements to test the relationship between the variables. Then, the edit tests each statement against the data in the household to see whether all statements are true. For any false statement, the edit will keep track, on an item-by-item basis, of invalid entries or inconsistencies. After all tests,

the editing and imputation system must assess how best to change the record so that it will pass all edits. Editing teams usually use a minimum-change approach and change the smallest possible number of variables to obtain an acceptable record.

255. The 11 declarative statements in figure 7 provide an example of rules that could be applied in a multiple-variable edit of selected population characteristics. In this example, the head of household must be 15 years of age or older. For generalized edits, it would be better to use "X" years where X is the determined minimum for the country. The statements in the example, such as relationship, sex, age, marital status, and fertility, focus on other important primary variables. The variables are closely related, hence editing teams should look at them together for the most efficient way of editing the data. It should be noted here that while all variables are important, some variables are more crucial for data presentation than others.

256. Figure 7 shows a simple case where, for some reason, both the head and spouse have the same sex – as it turns out, both are male, and one of them is a male with fertility. It is pretty clear that the sex is wrong (as indicated by the summary at the bottom) and that the male with fertility should be changed to female.

Figure 7. Example of rules for a multiple-variable edit of selected population characteristics

No.	*Rule*	*Relation*	*Sex*	*Age*	*Marital status*	*Fertility*
1	Head of household should be 15 years or older					
2	Spouse should be 15 years or older					
3	A spouse should be married					
4	If spouse present, head of household should be married					
5	If spouse present, head of household and spouse should be opposite sex	1	1			
6	Person less than 15 years old should be never married					
7	Male should have no fertility		1			1
8	Female less than 15 years old should have no fertility					
9	For female 15 years or older fertility entry should not be blank					
10	A child should be younger than head of household					
11	A parent should be older than head of household					
	Totals	1	2			1

Note: the "1s" show when two or more items are inconsistent. For example, in item 5, the head and spouse are the same sex, so the edit fails for relationship and sex, and the 1s appear in these cells.

257. In the example in figure 8, both spouses are from the same population as those in figure 6. Both are reported as male. Here the editing procedure is simple and straightforward. The variable with the greatest number of errors tallied is the one that will be edited first. In figure 7, the editing program implements the imputation procedure for "sex" since, based on the data in figure 6, that variable is most in error with respect to (1) relationship and sex, and (2) fertility and sex. When the editing program checks fertility and finds that the head of household has fertility data but the spouse does not, imputation assigns "female" to the head of household. Finally, when the editing team rechecks the series of tallies, and all positive statements are true, no further editing is required.

Figure 8. Example with head and spouse of same sex in an unedited data set and its resolution

Person	*Relationship*	*Sex*	*Children ever born*
Unedited data			
1	Head of household	Male	03
2	Spouse	Male	BLANK
Data after editing for sex			
1	Head of household	Female	03
2	Spouse	Male	BLANK

258. The editing specifications for this edit can be written as shown in figure 9. If fertility is complete for both, the edit will work. However, the edit is clearly not complete since it only takes care of the case in which fertility is complete and accurate for both the head of household and the spouse.

Figure 9. Sample editing specifications to correct sex variable, in pseudocode

```
If SEX of the HEAD OF HOUSEHOLD = SEX of the SPOUSE
   If FERTILITY of the HEAD OF HOUSEHOLD is not blank
      If FERTILITY of the SPOUSE is blank
        (if the SEX of the head of household is not already female) Make the SEX = female
endif
        (if the SEX of the spouse is not already male) Make the SEX = male endif
      else       Do something else because they have same sex and both have fertility !!!
          [The "something" could be using the sex of the previous head, or alternating the
sex of the
            Head, or using ratios of sexes of all heads for an appropriate response, etc.]
      endif
    Endif
 Else        This is the case where the head of household's fertility is blank
    If FERTILITY of the SPOUSE is not blank
       (if the SEX of the head of household is not already male) Make the SEX = male endif
       (if the SEX of the spouse is not already female) Make the SEX = female  endif
    else        Do something else because BOTH have no fertility!!!
          [The "something" could be using the sex of the previous head, or alternating the
sex of the
            Head, or using ratios of sexes of all heads for an appropriate response, etc.]
    endif
  Endif
Endif
```

259. The figure below (figure 10) is an example in which an editing procedure considers a female head of household 13 years old who is widowed but with three children, according to the keyed information. When the program runs through the editing rules, the following results:

Figure 10. Example of multiple-variable edit analysis for very young widow with 3 children

Number	*Rule*	*Relation*	*Sex*	*Age*	*Marital status*	*Fertility*
1	Head of household should be 15 years or older	1		1		
2	Spouse should be 15 years or older					
3	A "spouse" should be married					
4	If spouse present, head of household should be married					
5	If spouse present, head of household and spouse should be opposite sex					
6	Person less than 15 years old should be never married			1	1	
7	Male should have no fertility					
8	Female less than 15 years old should have no fertility		1	1		
9	For female 15 years or older fertility entry should not be blank					
10	A "child" should be younger than head of household					
11	A "parent" should be older than head of household					
	Totals	1	1	3	1	0

260. Again, we are considering a 13 year old widow head of household with three children. The first edit described in rule 1 – a head of household 15 years or older – fails because the head is less than 15 years old. She is 13 years old, so the boxes for "relationship" and "age" are marked, since the inconsistency is between these two variables. She is not a spouse, so neither rule 2 nor 3 is triggered. Also, rules 4 and 5 are not triggered for the same reason, that they apply only to the spouse. However, in rule 6, a person less than 15 years old (in this case 13 years old) should be never married. But our 13 year old widow is "widowed" so the rule is violated. Rule 7 is for males, so is not triggered. And, for rule 8, females less than 15 should have no fertility, but this person does have fertility. And rules 9, 10 and 11 do not apply to this person.

261. Based on the series of positive statements, the variable for age is most in error, and that is the one to change first. When we change the value for age, the tests are re-run and the edit will be finished if the change resolves all inconsistencies. Otherwise, the program edits the variable with the next highest number of inconsistencies.

E. EDITING APPLICATIONS FOR ELECTRONIC QUESTIONNAIRES

262. The introduction of electronic data collection in conducting population and housing censuses changes the way edits are done. This is because electronic data collection can allow editing of a variable as soon as it is entered. As discussed in Chapter II, *Editing during data collection* looks at the individual data items within a questionnaire or case and examines the validity of each item and the consistency of each item with respect to other related items, as it is entered. Data entered into the electronic questionnaire are checked for their correctness through the conditions that must be met to assume the response as accurate. A built-in edit rule informs the user about an error in case the rule is not satisfied when a response item is entered. This leads to the definition of what is meant by assuming data are erroneous or data are supposed to be suspicious. Typically, the validation outcomes in the data collection application are notified in two ways: (a) an error message which means the situation is unacceptable and must be changed in order to continue, or (b) a warning which notifies the possibility of incorrectness or to draw attention to verify certain response items.

1. Hard and soft edits

263. The goal of editing at the time of data collection is to take advantage of the measurement instrument to improve quality of the data and reduce the costs of the post-collection process. Data typed into the questionnaire are checked for their correctness. This requires defining the conditions that must be met to assume the response is accurate. The response item has a built-in edit rule to inform the user about an error in case the rule is not satisfied. This leads to the definition what is meant by assuming data are erroneous or data are supposed to be suspicious. Typically, validation in software technology, when a reaction is expected from a user or a user should be informed about something, is notified in a dual way[12]:

(i) First kind of edit rules called "hard edits" are the rules that must be satisfied unconditionally. It is like an error marked in red color which means the situation is unacceptable and must be changed in order to continue and a warning which notifies the possibility of incorrectness or to draw attention to a certain aspect of working being the consequence of earlier choices;

(ii) A second kind of edit rules can be called warnings or "soft edits". These kinds of edits only notify users that an item should be assessed for its adequacy. In this case two types of resolution of that kind of failures can pointed out: correction or no action. However, no action should be confirmed by enumerators or respondents as a user of the software as their selection.

264. Hard checks are applied to ensure editing rules are met unconditionally, while soft checks are applied to make a closer investigation of the inconsistent data items and resolve suspicious data during the interview, if possible. In other words, hard edits cause data containing errors to be rejected by the data collection application, while soft edits may allow data in which problems have been identified to be, nonetheless, accepted by the data collection application. However, sometimes the enumerator or respondent can get stuck with a hard check and introduce new errors in data, especially for those cases that might have extreme values which are not recognized by data collection application. Therefore, the use of hard checks require an extensive test and discussion to ensure it will work for all.

265. So, a hard edit at entry would not allow the enumerator to key fertility information for male. A soft edit would allow the enumerator to key in the gender but would then check back when fertility is entered. If the enumerator initially keyed in "male" for a person but then started keying fertility information, the edit can be programmed to provide a message "You are keying in fertility for a male" and then the enumerator can correct the variable in error.

266. Among other issues that might come up in this procedure is that the program might skip fertility altogether if the enumerator keys in "male" and the respondent was actually female. So, the enumerator would need to go backwards, or have been programmed with a key stroke to go back to enter fertility.

2. Decision for hard and soft edits

267. Making decision for the type of edits that will be applied to each field may depend on many factors. To achieve fully consistent data, it might seem ideal to use hard edits. However, the experience in several countries shows that excessive use of hard edits could be counter-productive.

268. Using many hard edits can create a burden on respondents or enumerators. Respondents or enumerators are generally receptive to edit checks. However, in situations with numerous edit checks, respondents could be bombarded with edit failures if they happen to trigger each edit rule. If it is likely that a large number of edit checks will be triggered, the edit rules may be too strict or there is a problem in the question phrasing or response field placement causing respondents' answers to trigger multiple edit checks. It should be kept in mind that the purpose of edit checks is to help the users of electronic questionnaire submit accurate, internally consistent data. Edit checks that do not clearly foster this result may be annoying to users.

[12] Eurostat, Memobust Handbook on Methodology of Modern Business Statistics Theme: Editing During Data Collection, 2017.

269. While considering types of edits, it is also important to keep in mind that errors can be introduced in the interview due to different factors. For example, enumerators can enter the wrong information, respondents may not remember the right information, especially proxy respondents, and some extreme but real cases may not conform with editing requirements.

270. For variables for which soft edits are applied, if respondents or enumerators have not changed data in a way to satisfy the edit failure (assuming that respondents usually provide accurate data and enumerators enter the right information), it should be assumed their response is correct or reliance has to be made on post-collection editing.

271. When making a decision on the types of edits to implement, the following two principles might be considered:

i. Let the user be in control to the extent possible (a usability principle); and,
ii. Obtaining some data, even edit-failing data, is better than obtaining no data at all.

Since the first principle is respondent centered and the second is from the operational perspective, editing strategies should be developed through balancing these aspects.

272. In the context of population and housing censuses, it is recommended to apply hard check to key census variables especially age, sex and relationship to the reference person. Several other variables are age and sex dependent (for example, fertility, educational attainment, labor force participation, female only occupations); therefore, it is paramount to hard-check these two key variables. This means that there will be no missing case for key variables and information will be consistent with other variables. If soft edits are applied to other variables, the users will be in control of correcting errors or leaving responses as it is without any correction.

273. When hard checks are applied to sex and age, data on these topics should be collected first in the part which lists all persons living in a household. In the individual questionnaire, age and sex should be asked again or be confirmed to ensure that data is correct. If these two attempts are not consistent, then the data collection application should not allow continuation of data collection before these two attempts are made consistent.

274. Other variables might need hard checks, especially for the reference person or household head. Some countries might need immediately explicit information on the household head's ethnicity, birthplace, citizenship, or religion. Usually, if the information is collected by hard edit for the head, the others in the housing unit could be soft edited, receiving the head's information in most cases.

275. Certain housing variables should also be considered for hard checks. For example, type of living quarters, type of housing unit and occupancy status might need hard checks as these variables are key variables for validating census coverage and distinguishing institutional population from household population.

276. In parallel with the application of hard and soft checks, two types of approaches can be used for skipping questions. Automatic skips depend on information already keyed to direct the cursor movement. For example, in a census, if a sex is female, the cursor can be directed to skip over the information on children-ever born with the application of automatic skip. It is clear that such skips depend entirely on the accuracy of the data keyed prior to the skip; if a "Yes" response is mistakenly entered as a "No," the cursor will be misdirected. This type of skip can introduce errors in data and therefore, it should be used when the probability of creating such errors is negligible.

3. Designing editing rules

277. Editing procedures ensure that the information provided is accurate, complete and consistent. Depending on the structure and content of the questionnaire, built-in edits are designed to check invalid characteristics, missing fields and inconsistencies between different answers from the same record or answer from different records, such

as persons who are unenumerated in the same housing unit. Also, built-in edits can be developed for checking duplication to ensure that census units are enumerated only once.

278. In general, built-in edits can be designed in three different ways;
   a. Single-variable rules
   b. Cross-variable rules
   c. Multi-case rules

279. The first two types are applied at the individual level or at the housing unit level. The third type of rules are applied to a group of cases and therefore this approach is more complicated. Applicability of this type of built-in edits has to be evaluated thoroughly to ensure this will not create additional burden on users and on the performance of data collection application.

a. Single-variable rules:

280. Single variable rules are used for checking internal inconsistencies, such as invalid values. Normally, checks for out-of-range or invalid values and missing values are included in this category. For example, a value for age should be entered as a number ranging from 00 to 99 or the highest age with three digits. Assume the highest age is 120 in this case, number above 120 will be invalid values and the data collection application should give an error message for this inconsistency.

281. There are two ways of dealing with this problem: (a) if hard-edits are applied to this field, a valid value should be entered by users (enumerators or respondents), (b) if soft-edits are applied to this field, users can leave this number or leave the field empty (missing value). This type of cases will be checked again in the data processing phase and a decision on whether impute or leave as it is will be made in editing and imputation stage of data processing phase.

b. Cross-variable rules:

282. This contains rules for checking inconsistencies in a variable through the values of other variable in the same case, such as checking inconsistencies between the questions in the individual questionnaire or the questions in the housing unit.

283. During data collection, this type of inconsistencies should be identified by editing rules and relevant error message should be displayed to warn users about what variables are inconsistent. For example, if the respondent declares that a child at age 10 completed secondary education which assume can be reported for people age 15 and over, an error message will warn users for this inconsistency. There might be several approaches to solve this problem. Application of hard-edits will make necessary to correct this inconsistency-either with changing age or changing data on level of education. On the other hand, if soft-check is decided for this kind of error, enumerators or respondents will check the data entered and change the data or leave the value as it is or leave the field on the level of education empty if there is no information. This type of errors has to be resolved during data processing phase.

c. Multi-case rules:

284. These rules can be applied to a single variable or a combination of variables in a group of cases (individuals or housing units or households). One of applications of multi-case rules is checking whether there are duplicates in the dataset, such as cases that have been entered more than once for a single household, or a household that has two reference persons/heads in the same household.

285. This type of editing rules should be carefully decided for ensuring its value in improving data quality and coverage. It might be worth to apply such editing rules if there is any risk for copying cases or households by enumerators for any kind of reason. Also, this can be used for checking consistency between the data entered for members of households, such as age difference between fathers/mother and their own children. However, identifying this type of inconsistencies may take a significant time during the interview, so that it might be better approach to apply these edits at the end of the interview.

286. If there is a high probability of triggering numerous edit checks, the number of edits incorporated into the electronic questionnaire should be reduced not to affect the performance of the data collection software significantly. According to the usability principles as much as possible should be left under the control of users of data collection application (enumerators or respondents); it is therefore recommended that users be allowed to submit data with unresolved edit rules to prevent non-response and respondent's perspective to provide most accurate data they have.

287. Questions arise as to what type of edits can be incorporated into data collection application balancing the need of high quality data and the necessity of a smooth interview. Some crucial questions arise: What kind of edits should be implemented on the electronic questionnaires? How many? What kind of edits should be mandatory?

288. If it is likely that a large number of edit checks will be activated, perhaps the edit rules are too strict or there is a problem in the question phrasing or response field placement causes respondents` answer to activate multiple edit checks. Edit failures may occur also due to poor design of electronic questionnaire. The purpose of the edit checks during data collection is to help the users submit accurate and internally consistent data. Edit checks that do not clearly foster this result may have negative impact on the performance of the data collection application and increase the burden.

289. No matter how much edit on entry occurs, the editing after data collection described in this handbook is still necessary. Some people will not know some information, especially information that is provided for people who are absent during data collection, so the editing during data processing is needed to obtain the best estimate through imputation. But edit on entry will certainly shorten the time needed to get to the post-enumeration edit and so get the results out to n of good auxiliary information.

4. Structure and timing of editing messages

290. An error message is needed for every kind of checks, hard or soft checks. The structure and timing of the editing messages is very important.

291. Basically three types of validation messages are possible[13]: i) Non-response messages appear when respondents have not answered a question, (ii) Invalid response message appear for numerical responses, when users enter a number outside the range established for a question and (iii) inconsistent message appear when a response for two variables are not consistent.

292. A further dilemma is how to present messages about data that do not satisfy edit rules contained in the application. The possible solutions can be to present the message immediately after the field has been left or after the page was filled or at the end of the interview. Immediate edits allow the respondent to correct the error straight away and can prevent similar mistakes later on. From the other side, edits involving more than one variable raise the issue of waiting with edit execution for last variable completion. Another question is when such messages should be presented to the user: immediately after a value was typed in or after the entire portion of data was entered.

[13] Technical report on Designing Interactive Edits for US Electronic Economic surveys and Censuses: Issues and Guidelines, January 2005

293. For example, if a date of birth is not indicated, then the message "date of birth not indicated" will appear. If it is invalid, then message "you entered invalid number, please enter correct age" will appear.

294. The structure and timing of the message can make the difference between a smooth enumeration with the respondent to a stumbling, difficult one. Some edit messages should show up immediately and should have a very simple structure. For example, "You just keyed in a 3 year old child with a PhD. Did you mean that?

295. In other cases, multi-cases rules can be applied for checking consistency between people enumerated in one household. For example, after the age of a spouse and a child, a message might be "You just keyed a spouse aged 79 with a 3-year old child. Did you mean that?" In this case, one of the two ages is wrong or a spouse is not father of that child. The timing of error message should be carefully evaluated. Like this example, some of the inconsistent data especially involving multiple cases can be made to appear at the end of the interview.

296. Edit on entry can also look between record types. So, if the questionnaires asks for the numbers of males and females in the housing unit as a check on the reported members, the edit can check to see if the sums of the genders in the population records equals the number recorded for the housing total. If there's a discrepancy, a message "The housing record says 5 males, but the population records only report 4 males." This message can only be generated after all the population records are collected as well as the housing record.

297. It is important to remember that while all kinds of edit checks can be done, with appropriate structure and messages, too many edits can create extra burden on users and would be very complicated to deal with during data collection. Therefore, it necessary to conduct a series of tests for developing the most suitable editing for self-enumeration with the Internet or face-to-face interview with the use of handheld devices.

298. Testing editing procedures is critical for ensuring to meet the set quality targets in collecting data. Problems with skip instructions may result in missing data and frustration of the enumerators and/or respondents. Poor visual design of electronic questionnaires can easily lead to confusion and inappropriate measurements. Testing of editing rules is discussed below.

5. Testing and evaluation of editing applications

299. When new procedures are adopted, such as the move from paper to electronic devices, considerable testing should be done before the actual fieldwork begins. After the dictionary and screens are determined, staff should test all editing procedures, using both pre-field and field methods (see Chapter II). Testing editing rules under laboratory conditions are particularly important for testing the skip patterns and the messages. This kind of test is usually done in the statistical office, with a small number of people and using the dataset of existing household surveys or censuses. This procedure will show if the questionnaire is too long or too complicated or has too many checks and too many messages.

300. The office should then do an initial analysis of the burden on respondents or on enumerators and impact on the performance of data collection applications. The experimental test in the office should also estimate how long it will take to administer the questionnaire. The questionnaire and its skip patterns and messages should then be changed to reflect the solutions to the issues that come up during the pre-field tests. This type of tests should be performed to make sure the length and complexity of the questionnaire is now appropriate.

301. Then, the office should do a pretest and pilot in a field situation. The best time to do this is exactly a year before the actual census date so that the climatic and other conditions will be about the same as during the census.

302. Several pretests may be needed to determine the best set of skip patterns and messages. The objective to do the minimum amount of messaging rather than the maximum. As noted, some items must be hard edited, like sex and age, but most of the others can be checked with soft edits during the interview and ascertained in the statistical office during computer editing.

F. METHODS OF CORRECTING AND IMPUTING DATA IN THE STATISTICAL OFFICE

303. As mentioned above, blanks in data records from "not reported", "unknown" or otherwise missing information occur in all censuses and surveys. Invalid entries also occur from respondent, enumerator or data entry mistakes. Methods of making corrections vary depending upon the item. In most instances, data items can be assigned valid codes with reasonable assurance that they are correct by using responses from other data items within the person or household record or from the records of other households or persons.

304. This *Handbook* presents two computer techniques to correct faulty data. One is the static imputation or "cold deck" method, which is used mainly for missing or unknown items. The other, more current method is the dynamic imputation or "hot deck" method, which may be used for missing data as well as for inconsistent or invalid items. Different computer packages and different programs within those packages, using various methodologies, employ cold deck and hot deck in different ways, as illustrated in the annexes. We use "hot deck" and dynamic imputation interchangeably.

1. Static imputation (or "cold deck" technique)

305. In static imputation, which is also known as cold deck imputation, the editing program assigns a particular response for a missing item from a predetermined set, or the response is imputed on a proportional basis from a distribution of valid responses. In the cold deck method, the program does not update the original set of variables. The values do not change from those in the initial static matrix after processing records for the first, second, tenth or any other persons. The original values provide imputations for any missing data.

306. Static imputation is a stochastic method, as is dynamic imputation, but the values do not change over time. This approach is described in Annex VI.

307. Sometimes static imputation uses a ratio method, assigning responses based on predetermined proportions. As an example of the proportional distribution of responses, suppose a tabulation of valid data, that is, data from completed as opposed to missing items, on time worked per week by males 33 years old who were employed in agriculture showed that 25 per cent worked 50 hours a week; 40 per cent worked 60 hours a week; and 35 per cent worked 70 hours a week. Missing or invalid responses for time worked for males 33 years old employed in agriculture would be replaced 25 per cent of the time by 50 hours, 40 per cent of the time by 60 hours, and 35 per cent of the time by 70 hours. However, unless reliable data are available from previous censuses, surveys or other sources, this technique requires pre-tabulation of valid responses from the current census, which may not be economically or operationally feasible.

2. Dynamic imputation (or "Hot Deck" technique)

308. The primary current method of ridding the data of unknowns is dynamic imputation (or hot deck technique), which allocates values for unavailable, unknown, incorrect or inconsistent entries. The United States Census Bureau originally developed the method, but other agencies have since added refinements. Dynamic imputation uses one or more variables to estimate the likely response when an unknown (or, in some circumstances, several unknowns) appears in the dataset. Dynamic imputation has become increasingly popular for census edits because it is easy and

produces clean, replicable results. In addition, by eliminating unknowns, trends between censuses and surveys are easier to obtain since the analyst does not have to deal with the unknowns on a case-by-case basis.

309. For dynamic imputation, known data about individuals with similar characteristics determine the most appropriate information to be used when some piece (or pieces) of information for another individual is unknown. These characteristics, for example, include sex, age, relationship to head of household, economic status, and education. The imputation matrix itself is a set of values, similar to the cards in a deck. These matrices store, and then provide, information used when encountering unknowns. The deck constantly changes by updating and/or by logically "shuffling the deck", so that response imputations change during data processing: hence the term "dynamic".

310. The values stored in the deck represent information about the "nearest neighbors" with similar information. Note that the nearest neighbor is usually the nearest *previous* neighbor because, especially in the top-down approach described elsewhere, housing units and people in those units are only considered once, and then the program moves on. So, within a village for example, when a person's maternal orphanhood is unknown, for example, the deck will contain information about the most recent person encountered with the same sex and age and valid maternal orphanhood. This approach is particularly important in countries having relatively large migration movements or HIV/AIDS. Similarly, housing characteristics are more likely to be similar within a village or set of villages than to other parts of the country.

311. As a simple illustration, a single value can be stored as the deck. For example, if a person's sex is invalid for some reason, the deck is assigned an initial value (male or female) arbitrarily, thus determining an initial value. The seed value becomes the sex of the first individual encountered with unknown sex. If the first person's sex is valid, however, the sex of the first person replaces the seed value. If the second person's sex is unknown, then the imputation matrix assigns the stored sex. In this case, the imputed sex is the sex of the first person. In essence, when the edit finds an acceptable value for an item, it puts it into the imputation matrix. When it finds an unacceptable one, imputation replaces it with the valid value from the imputation matrix.

312. One of the problems with the dynamic imputation method described here is that if two different items have unknown values, the same "donor" individual may not be used to assign valid responses. Each value may come from a "real" person, but these may be different persons. A better method would be to assign both variables at the same time, from the same person. Programming these complicated matrices however, may present some difficulties.

313. The data below (figure 11) illustrate a household for a set of ten individuals. The blanks, as illustrated by brackets [ ], show where missing data occur. Often, the numbers 9 and 99 are used to show missing information, in this case for sex (a 9 for a single digit) and age (99 for two digits), indicating missing information. But sometimes value 9 needs to stand for another, real value, for example in a limited number of relationship codes, so these values should be used very sparingly; and, if another value, like "." or ".." can be used, it probably should be. Note that although other variables are available for use in imputation, such as education and occupation, they have not been included in this short example.

Figure 11. Sample household as example of input for dynamic imputation

ID number	*Relationship*	*Sex*	*Age*
1	1	1	39
2	2	2	35
3	3	1	13
4	3	[ ]	10
5	4	2	40
6	4	1	[ ]
7	4	2	13
8	5	[ ]	[ ]
9	5	1	44
10	5	2	36

NOTE: [ ], [ ] = missing information

314. If the initial value for the imputation matrix called SEXARRAY is male (code=1), the imputation matrix will look something like this: SEX = 1

315. After person 1 is processed, the value will remain 1. The value will change to 2, however, after processing the second person, since that person is female. The variable will now look like this: SEXARRAY = 2

316. For each valid entry for the sex of a processed individual, the code for the sex of that person replaces the imputation matrix value. When the third person is processed, imputation changes the value to 1, or male, again.

317. For the fourth person the sex is unknown, so the edit looks at the imputation matrix value, which in this case is male, and replaces the unknown value with the imputation matrix value. Person 5 is female, so it replaces the previous value in the imputation matrix from person 3 (male). This process continues until person 8.

318. The edit uses imputation again, and person 8 becomes female since the imputation matrix value obtained from person 7 is female. The edit used the imputation matrix to obtain values twice: once to obtain a male and once to obtain a female. Since the sexes appear in approximately equal frequencies, over the long run the imputation uses each sex approximately half the time. After processing all ten individuals, the variable will look like this: SEXARRAY = 2

319. Although an imputation matrix assigns sex in this way, other, more complicated ways of using the procedure exist. For instance, the editing can use the relationship to head of household and the sex to aid in determining the age for an individual. Consider the following partial list of relationship codes:

1 = Head of household
2 = Spouse
3 = Child
4 = Other relative
5 = Non-relative

320. The data processor can create initial age values that might approach the real situation for the relationships by sex. These values are not very important since the editing process will almost certainly replace them before using them. Also, the edit rule calls for imputation of many values, so few initial values affect the final tabulations. These values might be as shown in figure 12.

Figure 12. Initial static matrix for age based on sex and relationships

	Relationships				
	Head of household (1)	*Spouse* (2)	*Son/daughter* (3)	*Other relative* (4)	*Non-relative* (5)
Male (1)	35	35	12	40	40
Female (2)	32	32	12	37	37

321. Consider again the 10 individuals introduced in figure 11. Since the first person in our sample is listed as head of household (code=1) and he is male (code=1), his age (39) replaces the first element (coordinates 1,1) during the imputation. The deck (dynamic imputation matrix) then contains the values displayed in figure 13.

Figure 13. Example of a dynamic imputation matrix after one change

	Relationships				
	Head of household (1)	*Spouse* (2)	*Son/daughter* (3)	*Other relative* (4)	*Non-relative* (5)
Male (1)	39*	35	12	40	40
Female (2)	32	32	12	37	37

322. The second person is spouse (code=2) and female (code=2), so her age (35) replaces the value in the second row of the second column, changing the deck to these values. The ages of other individuals in the household similarly replace imputation matrix values, through the fifth person.

323. Note that the previous sex imputation procedure assigned sex 1 to person 4. Because the edit requires imputation of a value for sex, the edit does not update the array with that person's age. The edit will update only with values from records where sex and relationship are both initially correct. When the edit gets to person 6, however, it finds that the age is unknown. The person is male and he is an "other relative" of the head of household. Therefore, the edit uses the imputation matrix element for males whose relationship group is "other relative" (the fourth column in the first row) and assigns the value of age for that category ("male other relative" – in this case, 40).

324. The eighth person has neither sex nor age reported. The edit imputes sex as female and then allocates the age based on this allocated sex and the relationship code (5). In this case, the age is 37.

325. Although the edit imputed the value for age from the known relationship, it used a previously allocated value for sex for the other variables. Here, the use of allocated values for further imputation is an example of poor editing procedure (see section 3(d) below). It would be better to look for other known data items, such as marital status, for use in the imputation.

326. After the tenth person, the imputation matrix values are given in figure 14. In this example, both imputations used the initial static matrix. Usually only a small number, if any, of initial values will be used in imputation. The majority of cases will use values assigned from the enumerated population.

Figure 14. Example of a dynamic imputation matrix after multiple changes

	Relationships				
	Head of household (1)	*Spouse* (2)	*Son/daughter* (3)	*Other relative* (4)	*Non-relative* (5)

Male (1)	39	35	13	40	44
Female (2)	32	35	12	13	36

3. Dynamic imputation (hot deck) issues

(a) Geographical considerations

327. If the editing program uses dynamic imputation to impute missing values, it should attempt to use data sorted by the smallest geographically defined area. This procedure should increase the probability of obtaining a correct answer, since people living in the same small geographical area are usually somewhat homogeneous with respect to their demographic, housing, and other characteristics. Where the population is not homogeneous, no correlation will exist, so the editing team must look at variables on a case-by-case basis. Also, as will be discussed later, some areas should never have certain variables – like central heating in very warm places – and the edit should take this into account.

(b) Use of related items

328. Before using dynamic imputation to obtain missing values, an effort should be made to use related items to assign a value that is likely to be correct. For instance, if the marital status of a person is missing, the editing program will determine whether the person has a spouse in the household. If so, the program will assign the code for married without using an imputation matrix. However, when no such evidence is present, the program may have to rely on an imputation matrix value.

(c) How the order of the variables affects the matrices

329. National census/statistical offices that use imputation matrices should consider which variables they need as they develop the order of their edits. For population items, the offices will want to edit sex and age at the beginning, so they can use these in the other imputation matrices. The overall edit should not use unedited (not checked with editing rules) variables in imputation matrices, although most computer packages will accept "unknown" rows or columns. Response rates and distribution of attributes within variables will assist in determining the best variables, and the most useful attributes within those variables, to assist in developing the dynamic imputation matrix. Subsequent imputation matrices can use the data items after editing. However, whenever possible, statistical offices should consider excluding imputed data from the imputation matrix.

330. For example, if the edit imputes age based on sex and relationship, cells in the array for this imputation matrix (sex by relationship), should not be updated if either the sex or the relationship was imputed. As a rule, only when age, sex and relationship are all valid and consistent should the editing package enter age in the cell for the appropriate sex and relationship. However, sometimes the use of imputed data is unavoidable because of other factors. It is important to note that most countries ignore this suggestion, and impute from previously imputed values.

(d) Complexity of the imputation matrices

331. The national census/statistical office increases the probability of obtaining a consistent, "correct" imputation matrix value by making the imputation matrix more detailed. For example, the program could impute marital status using relationship alone. However, the likelihood of widowhood or divorce increases with age. Therefore, it makes sense to impute marital status by age and relationship. Using the age and relationship of the current person, the editing program takes the value for marital status from a person with the same characteristics in the immediately preceding valid record stored in the imputation matrix.

332. Nonetheless, the procedure described above can create new problems. The national census/statistical office usually edits questionnaire items in a fixed sequence, with age edited after marital status in a top-down approach. If this is the case, when both marital status and age are missing from a record, it is impossible to take the value for marital status from the immediately preceding record with the same age and relationship values.

333. The best editing practice is not to use edited values in hot decks. Sometimes this practice is difficult to follow, either because of timing for results or difficulty in the computer programming. In these cases, one of several variables would be imputed, its value placed in the appropriate hot decks, and then used to impute subsequent variables

334. As a result, the program may not be able to determine the age category for this record. Another solution would be for the imputation array to have a row or column for "not reported" items. This procedure would allow the program to assign a value for marital status using the marital status category from the immediately preceding record with the same relationship and age "not reported". Two factors, however, argue against this approach. One is that "not reported" cases in the same combination are so few that it would be difficult to update the imputation array for the missing item. Secondly, it is essentially impossible to obtain proper cold deck, that is, initial values for these combinations of "unknown" values for a hot deck since they do not exist in the "real" world.

335. The solution to the problem described above creates more work for the data processor but results in a cleaner product. The editing program first tests to determine whether the items have valid codes. If the record for the current person does not have a valid code for the item, the imputation matrix does not use the item for this record. Data processors can facilitate the process by creating a simpler imputation array. To continue the earlier example, if the program must impute marital status because the value is missing, the imputation array will ordinarily have two-dimensions: age and relationship. If, after testing, the program finds no valid code for age, it will impute marital status by relationship alone. Because the edit for relationship comes before marital status, the relationship code will be valid. The program uses these same principles for all dynamic imputation procedures.

(e) Imputation matrix development

336. The subject-matter staff, in collaboration with the data processors, should prepare the appropriate imputation matrices. (Some editing teams use multiple imputation matrices). Only valid responses update the imputation matrices; editing teams do not use allocated or imputed values. Both subject-matter specialists and data processors must check editing specifications and hot decks for consistency and completeness.

337. Considerable time and thought should go into the development of an imputation matrix, including research into the use of administrative records and the results of previous censuses or surveys, particularly for cold deck values. Even after research and development, editors should not apply imputation matrices randomly. When imputation matrices are not internally consistent, considerable effort is required to reconcile them. When imputation matrices do not use standard conventions, staff must consider each one separately.

338. Although for the examples in this *Handbook*, each cell in the imputation matrices has one value, some editing teams keep more than one possibility for each cell. You can imagine this as a two-dimensional matrix, with a third dimension, like going back into a blackboard. These cells provide an extra dimension. To illustrate, if the ages of all the children in a family are unknown, as for example, in a family with four male children, the computer will not assign the same value four times, creating quadruplets. Instead, four different ages will be assigned. However, even here the same value may be assigned more than once, depending on what is stored in the matrices.

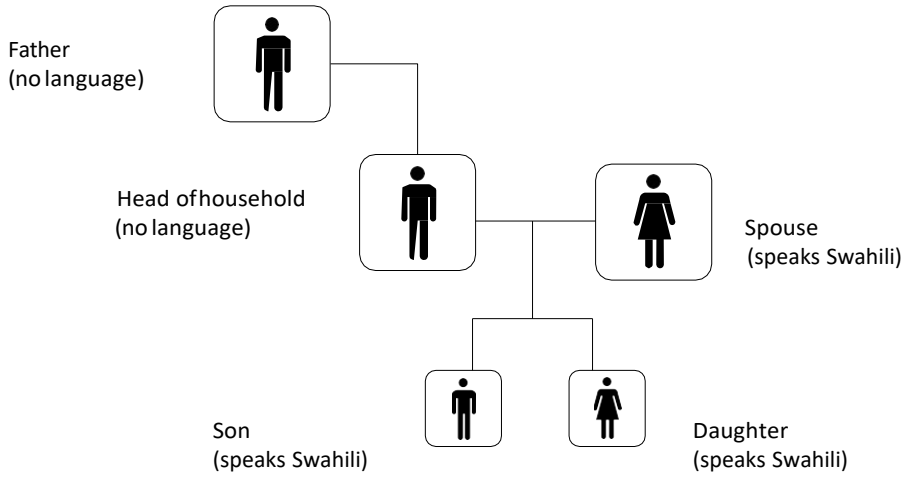
(f) Standardized imputation matrices

339. Standardized imputation matrices can streamline the editing process. Imputation matrices with standard dimensions for various social and economic variables, such as age groups and sex, can be tested and applied quickly.

340. For example, the national census/statistical office may want to develop an imputation matrix to determine a code for language when none is given. The first place for the editing program to look will almost certainly be within the household for another person reported as speaking a given language. Failing that, the program can select the language of a previous person of the same sex and age group (having updated the imputation matrix when all three items were valid). This procedure will give a likely language, since persons speaking the same or similar languages are usually located geographically close to each other.

341. In figure 15 the variable "language" contains no information for, the head of household. For whatever reason, the scanner or the keyer may not have picked up the language entry or code, or something else may have gone wrong. However, since the spouse and children all speak Swahili, that language can be assigned to the head of household and to the father of the head of household, whose language entry is also missing. Note that the household head in figure 14 is female.

Figure 15. Example of head of household and head's father without assigned language



342. When no language is reported for anyone in the household, the editing program must do something else. First, the edit looks for other variables to give an indirect estimate of the language used. Sometimes race, ethnicity or birthplace gives an indication of the appropriate language to impute. If such an identifier is available, then the editing team might choose to use that to determine the language for the head of household. If not, the edit can use age and sex for imputation. The imputation matrix might look something like figure 16.

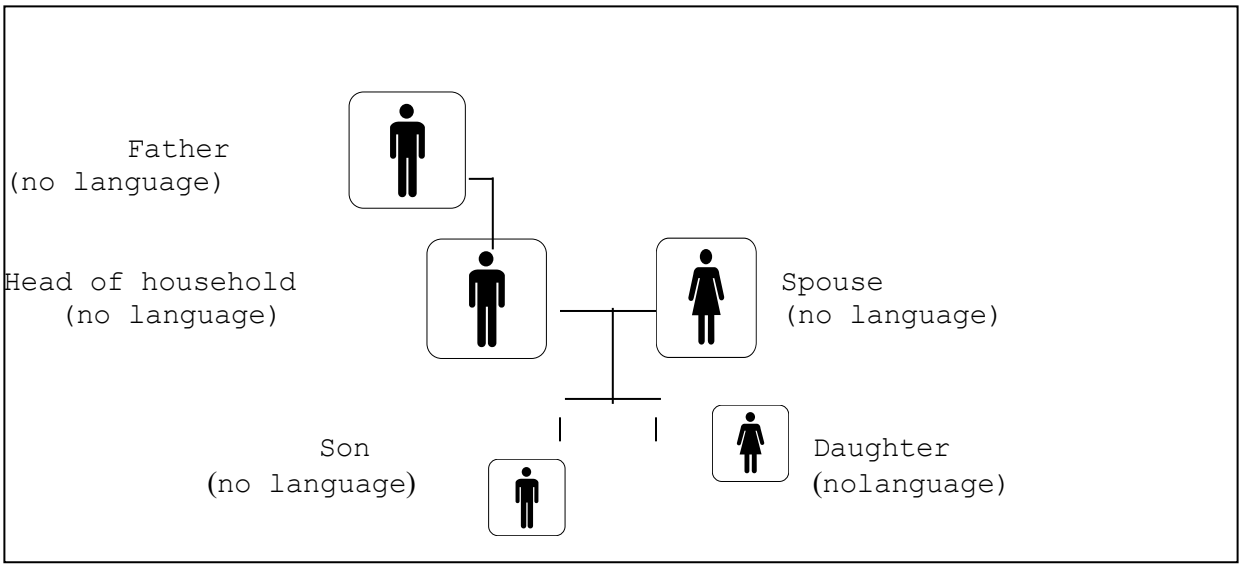
Figure 16. Initial values for a dynamic imputation matrix for language

Sex	*Age*					
	Less than 15 years	*15-29 years*	*30- 44 years*	*45-59 years*	*60-74 years*	*75 years and over*
Male	Language 1	Language1	Language 1	Language 1	Language 1	Language 2
Female	Language 1	Language 1	Language 1	Language 1	Language 1	Language 2

343. If it is decided to impute, the program assigns the head of household a language based on age group and sex. In this case, the entries in the imputation matrix will be for previous heads of household only, since all other persons in a given household receive the same language code as the head of household.

344. At this point, if the household still has no one who reports speaking a defined language, the editing program uses the imputation matrix to assign a language to the head of household based on the head of household's age and sex. The language assigned is the most recent one in the data file spoken by another head of household of the same age and sex. Since the imputation matrix is "updated" continuously as acceptable cases are encountered, the assigned language is likely to be a language spoken in the general community (see figure 17).

Figure 17. Example of members of a household without an assigned language



345. Exceptions to the editing rules will occur at the very beginning of an edit run. Staff must be careful to take note of language changes that may occur when they move from one geographical area to another. Some countries must also be concerned with localized mixtures of language speakers. However, even in this case, unless selective under-reporting for certain languages exists, the percentage of allocated and unallocated values resulting from the imputation should be about the same.

346. Another edit might look at religion. Again, the responses for religion may be imputed by age and sex. The editing program will continue updating when all information is available and will pull responses from the imputation matrix for "unknown" information. This imputation matrix will look like the one for language, but with religion in the cells instead of language.

347. This explanation assumes a top-down, sequential approach. Editing teams using sophisticated methods such as Fellegi-Holt and the New Imputation Method (NIM) (see Annex VI) apply all related edits concurrently. The present procedure also assumes the existence of an appropriate order for the edits.

348. Many of the economic characteristics, such as labour force participation, time worked last week, or weeks and time worked last year, can be imputed using similar characteristics. By using similar imputation matrices, the editing program can quickly check the value for the characteristics of the variables, and the editing process should proceed faster overall.

349. It is sometimes difficult to obtain appropriately edited characteristics for the first imputation matrices in a series. Usually a statistical office does not want to include unedited items as dimensions for an imputation matrix; the edit would not use either sex or age as imputation matrix dimensions before they have been edited. Hence, the first few imputation matrices will use different variables that need no editing or those that cannot change in value. For the very first imputation matrix for population items, the edit might use the number of persons in the housing unit including a zero for vacant units.

350. For housing edits in general, the first imputation matrix might also use the number of persons in housing units as the initial dimension, but the editing team might modify actions for housing items to account for vacant units. For example, if the first housing edit is for "construction material of outer walls" or "type of walls", the initial values might be based on the number of persons in the housing unit, including a value for when the unit is vacant.

351. When the unit is vacant but "type of walls" is valid, the edit updates the first cell with the type of outer walls. When the type of walls is known, for an occupied unit the edit updates the cell corresponding to the number of persons in the unit. When the construction material for the outer walls is unknown, however, the imputation matrix will supply a value for the construction material of the outer walls, based on the number of persons in the unit.

352. After the initial use of this imputation matrix, the editing team might then want to switch to some other housing characteristics, such as "type of roof" or "tenure". Whatever is selected must distinguish clearly between units and provide enough diversity that the same attribute will not be selected repeatedly. Recurring selection of the same attribute can give quasi-cold-deck rather than dynamic imputation (hot deck) values. Using dynamic imputation, for instance, in an army barracks "group quarters" might cause the same value to be used repeatedly if the only characteristics selected are age and sex. In this case, all of the residents would probably be male, and most would be within a limited age range. Hence, that particular matrix might not give the best results. If "tenure" has sufficient diversity, with sufficient percentages of owners and renters, this variable could work. Otherwise, the country could use different types of roof.

353. In general, many editing teams find that by using comparable dimensions for imputation matrices, they do less checking, get their results more quickly and probably get them more accurately.

(g) When dynamic imputation is not used

354. If the editing team chooses not to use dynamic imputation at all, the sequence of the edits is still important. For example, age is related to many items, including relationship to head of household, level of schooling, employment and fertility (for females). Consider the household members identified in figure 3.18:

Figure 3.18. Example of head of household and child with child's age missing

Person	*Relation*	*Age*	*Grade*	*Working*	*Occupation*	*Children ever born*
1	1	40	12	1	33	BLANK
3	3	X	7	BLANK	BLANK	BLANK

NOTE: X = Missing age
BLANK = Does not apply

355. The record for person 3 has relationship 3 (child), but no reported age. To find the age, the editing program can use the difference in age between the head of household and child (either a cold deck value or a value obtained from a previous unit by imputation). If that difference is 25, for example, the child's age becomes 15 (the head of household's age of 40 minus the age difference of 25).

356. The number of years of schooling is also known, which in this case is 7 years. Age 15 may well correspond to this grade level. Since the range of appropriate years of schooling for a particular age is smaller than the range of ages for the difference in age between the head of household and the child, it is better to check first whether the level of schooling is appropriate. If the level is reported, an age difference determined by either static (cold deck) or dynamic (hot deck) imputation can be used to provide an appropriate age. If the level is not known, then the age difference between head of household and child can be used to assign the age.

357. However, even age difference information may be missing. In fact, in most countries, it is more likely that the level of education is missing than age. The following example illustrates the steps the editing team may take if both age and grade are missing.

Figure 3.19. Example of head of household and child with child's age and grade missing

Person	*Relation*	*Age*	*Grade*	*Working*	*Occupation*	*Children ever born*
1	1	40	12	1	33	BLANK
3	3	X	X	BLANK	BLANK	BLANK

358. In figure 3.19 neither age nor grade is present, but other information exists. Person 3 is not old enough to be employed, and is too young to have had children (or is male). Using the employment information, a set of cold deck values can obtain an age, but it will be an age lower than the lowest acceptable age for working. Alternatively, if the editing team uses dynamic imputation, an imputation matrix value gives a value for age. The selected age probably should use the head of household's age as one of the variables to maintain consistency. For example, if the head of household's age is 20 rather than 40 it would obviously be inappropriate to assign age 14 to person 3. When the age is imputed, then the grade can also be determined, and the latter should thereby be consistent with both age and working status.

359. If the editing team decides to impute all or most of its items, it should develop a strategy for building the edit in a logical way. For population items, the edit should begin by considering all items potentially having unknowns. Editing teams should use information from surveys and administrative records, earlier censuses, the pilot for the census under consideration, and other information available to help determine each item's inclusion in the first, and subsequent, imputation matrices. While development of the details of imputation matrices is very country-specific, all national census/statistical offices are likely to have some information available for this purpose. Testing of various sets of variables in the hot decks will assist in getting the most appropriate set for the particular country.

360. Many editing software packages keep track of the number of persons in the housing unit as they go along. An imputation matrix for unknown sex, for example, could allow for assignment of male or female depending on the number of occupants in the housing unit. Hence, the initial value to be selected for a person of unknown or invalid sex for a one-person house might be male. For a two-person house, the initial value might be female. For a three-person house the value would be male and so on. The matrix would be used only as a last resort after all consistency edits, such as the sex of the head of household and the spouse and the presence of fertility information, had been tested and resolved.

(h) How big should the imputation matrices be?

361. Most computer packages can accept multidimensional imputation matrices. The following points should be taken into consideration before setting up the imputation matrices.

(i) Problems that arise when the imputation matrix is too big

362. One of the biggest problems that some national census/statistical offices have as the team of subject-matter and data processing specialists work together is that of over-eager editors. It is easy to get carried away in developing the editing packages so that the programming takes much longer than necessary and slows the census or survey processing. The editing team may decide, for example, that in order to determine age, in addition to "sex", "educational attainment" and "labor force participation", "number of children ever born" must also be included for females. The addition of "number of children" ever born may provide a slightly better age estimate, but the increased complexity of the programming may not justify it. Editing teams have to decide how many imputation matrix dimensions will give the best results, in terms of both accuracy and efficiency. Imputation matrices that are too big (with too many cells) cannot be updated thoroughly, and cold deck values may inappropriately be used instead.

(ii) Understanding what the imputation matrix is doing

363. In addition to imputation matrices that are too big, paths may be confusing. It is important to make sure that the subject-matter personnel as well as the data processors are able to follow all the paths. Together, they must make sure that the imputation matrix is performing its intended task. Again, the subject-matter persons and data processors must work together to verify that each variable or dimension of the imputation matrix is implemented properly. Moreover, they must ensure that all of the combinations are working properly.

(iii) Problems that arise when the imputation matrix is too small

364. The imputation matrix is too small if it has too few dimensions or if, because of groupings (such as too few age groups or educational levels), the same imputation matrix value is used repeatedly before being updated. For example, without a dimension for sex in an age array, all children in a family are more likely to receive the same age when age is unknown. Subject-matter personnel should work with the data processors to test the imputation matrices for all of the different combinations and should ensure that none occur too frequently.

(iv) Items that are difficult for imputation matrices

365. Some items, such as "occupation" and "industry" have proven notoriously difficult to edit. While separate imputation matrices for occupation and industry may produce inconsistent results, an effort to crosscheck all pairs of occupation and industry entries can be costly and difficult. For example, if barbers or hairdressers are found working in fish processing plants, some other type of edit is needed. In addition, the large number of occupations and industry categories can make dynamic imputation very difficult. For some items the editing team may decide that editing is counter-productive and, instead, opt to use "not stated" or "not reported." Otherwise, use of a static imputation (cold deck) approach may suffice.

4. Checking imputation matrices

366. The basic structure of the imputation matrix in an editing software package should look something like the display in Figure 20. Editing specifications must identify the arrays used for the imputation and use cold deck values for the initial set of values.

(a) Setting up the initial static matrix

367. The procedure outlined below updates the imputation matrix each time it finds a person with valid values in all three items—in this case, "relationship", "sex" and "age". However, when the editing program finds an invalid (or blank) sex, the imputation matrix selects a value based on valid relationship and sex codes (variables that have already been edited).

Figure 20. Sample set of values for a cold deck array and sample imputation code

```
.
.
.
   A01-AGE-FM-SEXRL (2,6)
```

Head of household	Spouse	Child	Other relative	Parent	Not reported	Sex
40	40	10	20	65	20	Male
40	40	10	20	65	20	Female

```
.
.
.
 if AGE = 0.98
          let A01-AGE-FM-SEXRL (SEX, RELATIONSHIP) = AGE
else
          message 'Age is unknown, so imputed' AGE
          write 'Age is unknown, so imputed, Age =' AGE
          impute AGE = A01-AGE-FM-SEXRL (SEX, RELATIONSHIP)
          message 'AGE is now known' AGE
end-if
.
.
.
```

(b) Messages for errors

368. Editing packages should provide several methods to make certain that they implement edits and imputations properly. Two of these features, message commands and write commands, are reviewed below.

369. One source of information is the display of a message, as seen above in figure 3.20. This command generates specific messages and summary counts (the total number of times the message occurs) for levels of geography (e.g., enumeration area, minor civil division, major civil division) as well as for each questionnaire. For all questionnaires, a summary report might look something like figure 21:

Figure 21. Example of a summary report for number of imputations per error

Count	*Error number*	*Message*	*Line number*
-	14-1	Too many children per woman	2629
-	14-2	Too many children per woman	2645
2	14-3	Boys present not stated	2669
2	14-4	Girls present not stated	2678
33	14-5	Month last birth not stated	2723
7	15-6	No children ever born; age difference between mother & child OK	2892

NOTE: Here "14" simply refers to item 14 in a given series; errors are numbered sequentially.

370. A report organized by questionnaire (figure 22) might give the questionnaire number, including all of the specified geographical codes. The report could then list the errors found in the program, by item (in this case age), and by line number in the software program, seen below on the right. In this example, the age was blank, but the imputation matrix provided the age of 48, based on the relationship and sex of this person. For this case, the specific age was unknown, but the message command could also write that information, also, if desired.

371. Of course, while it makes sense to list all individual errors on sample tests or small, selected data sets, the amount of output in production runs would be so large and cumbersome (and leading to meaninglessness after a while), that a trigger should be set to turn off all or parts of the individual questionnaire problems for the complete census. The summary statistics would remain, of course.

Figure 22. Sample report for errors in a questionnaire

Questionnaire ID: 01 01 017		*Line number*
AGE (1) =	Age is unknown, so imputed	#46
AGE (1) = 48	Age is now known	

(c) Custom-made error listings

372. The software might also provide another command, allowing for a more detailed analysis of the editing specifications and edit flow. The command may be used to show the information before a change is made, and then all of the changes made. Finally, it shows the record or records again, with the changes made. In this way, the analyst can make certain that the edit follows all paths properly. The results may be as shown in figure 23. The first line of the output gives the variables (e.g., province, relationship, sex, age). Then, the incoming data are shown, followed by the error (in this case, no age), and then the data after the change was made.

Figure 23. Example of supplementary error listing by questionnaire including multiple variables

	Province	*District*	*Head of household*	*Relation*	*Sex*	*Age*
Incoming data	01	01	17	1	1	
Error	Age is unknown, so imputed age = BLANK					
Edited data	01	01	17	1	1	48

373. This procedure assists the editing team in determining whether the edit is taking the proper paths.

374. Testing is an important part of census and survey editing. The following method represents one possible way of testing editing procedures. The process might begin by having specialists perform the analysis systematically by creating a "perfect" household. A perfect household is one that is a complete household—head of household, spouse, children, other relatives and non-relatives—with all their characteristics. The perfect household must pass all of the edits without any errors. Then, the unit is duplicated repeatedly in a single file. The procedure continues as outlined below:

a) The data processors introduce a single error into each household, in sequence, to correspond to the sequence of the editing specifications and the editing program;

b) The analyst then checks all of the paths early in the editing process;
c) Once the edit follows all paths properly, data processors run a sample of the whole data set, looking for idiosyncrasies in the actual data set and making modifications as necessary;
d) Finally, the data processors run the whole dataset.

375. When satisfied that the messages are working properly and the appropriate modifications have been made, the data processor may decide to turn them off for lower levels (like for each questionnaire). If large countries were to run their whole data sets with message statements left in for each questionnaire, the resulting quantity of lines and paper would be prohibitive. However, the summary report for these messages should continue because it gives useful information for the various levels of geography. The output will look something like that in figure 21.

376. Computer edits usually include a safeguard procedure. The edit trail shows all data changes and tallies for cases of changes and substituted values. Reference to the edit trail will determine whether the number of changes is sufficiently low for the group of records to be accepted.

377. If a particular item has too many errors, the item may not have been adequately pretested, either on its own, or in relation to other items, indicating that enumerators or respondents did not understand the item. Sometimes enumerators get confused, for example, and collect fertility information only from male adults and not from females. If this type of data collection is systematic, the editing team might have the programmers move the fertility data from the males to the females in a married couple. Otherwise, the editing team can do little at this stage to correct the error.

378. Usually the editing program needs to look at several different files to cover all situations. In addition, the data processors will need to make changes because of faulty syntax or logic. Even the most experienced data processing specialists occasionally key a "greater than" sign in place of a "less than" sign, and the error is found only after several runs are made since the particular problem may not be immediately apparent. Similarly, small flaws in logic may not be apparent at first. Again, the subject-matter and data processing specialists need to work together to resolve these issues early in the editing process, if possible.

(d) How many times to run the edit?

379. As noted, as soon as the questionnaire is set, development and testing of edit specifications and programs should begin. Individual items should be developed separately when a top-down approach is used, but even when several variables are to be edited at the same time, edits for individual items will need to be tested on small parts of the whole data set. The edit specifications should be developed by the subject matter specialists, and then individual edit programs implemented by the programmers. The total edit can then be built, and run on larger and large parts of the data set, refined along the way.

380. In general, for both the parts of the program and the whole program, it is a good idea to run an editing program three times, as explained below:

381. The first edit run supplies the imputation matrices with real values rather than the values created in the initial static matrix. Some countries use data from other sources—either a previous census or survey or administrative records —to supply cold deck values for an array. The data processor runs the complete dataset, or a large part of it, to supply values for the imputation matrix. Cold deck values from the actual dataset are more likely to be accurate and current. The edits use only about two percent of this initial static matrix: the rest are dynamic imputation values.

382. The second edit run performs the actual editing. The second edit run consists of several repeated runs in order to cover all situations. At this time, the data processors will need to make changes in order to correct errors resulting from faulty syntax or logic. In addition, even the most experienced data processing specialists may make mistakes and, since the particular problem may not be immediately apparent, the error may be found only after a few runs. Similarly, small flaws in logic may not be apparent at first.

383. The third edit run makes certain (1) that no errors remain in the data set, and (2) that the editing program did not introduce new errors. When the processors run the edit this last time, no errors should appear in the error listings. If errors remain, the logic of the edit is probably faulty, so the data processor needs to modify it. In addition, this run usually tells the data processor if the edit accidentally introduced new errors by the logic of the edit.

5. Imputation flags

384. Imputation flags are one method used to retain information about unedited data. As mentioned previously, many editing teams are concerned about the loss of potential information when unedited responses are changed. In cases where a value is changed because of an inconsistency, the editing teams may wish to save the original value or values in order to carry out further demographic or error analysis after the census. Both subject-matter specialists and programmers will want to analyse various aspects of the missing, invalid or inconsistent data. Members of the editing team need to make sure that the imputed and unimputed distributions are consistent, to see if any systematic error appears in the editing and imputation plan.

385. For example, sometimes data processing specialists accidentally use only cold deck values because the program neglects to update the imputation matrix. If the country conducted a census pretest, the editing team may need to investigate the relationships between some of the variables after the pretest in order to finalize the questionnaire. In prior censuses, before microcomputers with large hard disks were common, many statistical offices did not have the space on their tapes or other storage media to maintain extra data; however, these days, for most countries, keeping information about unedited data is no longer a problem.

386. Some countries choose to maintain a simple, binary accounting variable as a flag for each item. This method is simple and takes up a single byte for each variable. For example, the United States Census Bureau places imputation flags for each variable at the end of each record, for both housing and population records. For each housing variable, for example, the variable for the flag was initially "0", but was changed to "1" if the original item is changed in any way. The program does not retain the original value, although offices sometimes compile these, either for each record or in the aggregate.

387. Other methods are available to save unedited responses. In the example in figure 24, the national census/statistical office has changed a spouse's age from 70 to 40 using an imputation matrix. The national census/statistical office can easily put the pre-imputation value, in this case 70, in the area reserved for imputation flags and reserve the variable used for published tabulations for the allocated value, in this case 40. In order to examine changes in the data set, the statistical office can make frequency distributions or cross-tabulations of the allocated and the unallocated values. If, following this analysis of the effects of the edits on the data set, the tabulations based on the edit appear suspicious or anomalous, the editing teams might want to consider changing the edit or part of the edit flow. And, because hard disk capacities have increased so much in recent years, all initial values can be stored on the records for later use. Offices will probably want to maintain at least two files since a file of all edited data is likely to run slightly faster.

Figure 24. Sample population records with flags for imputed values

Person	*Sex*	*Age*	*Children ever born (CEB)*	*Sex flag*	*Age flag*	*CEB flag*
1	1	40	BLANK			1
2	2	40	7		70	

388. Figure 25 illustrates the case of a female 13 years of age who is recorded as having borne a child (children ever born is 1). However, the editing team has decided that the minimum age at first birth will be 14, and that births to females younger than 14 are more likely to be errors than fact. As always, this raises the question of whether this case represents noise in the data set versus a real value.

389. Under the editing rules, imputation "blanks" information for children ever born. Note that the CEB flag is a little more complicated since it must account for a *BLANK* that was imputed, as well as for numerical entries. Suppose the subject-matter personnel want to study the numbers and characteristics of persons 13 years old reported as having had a child. The data processors can place the original information in an area of the record set aside for flags, usually at the end of the record. Then, the set of published tables will exclude the children ever born information for this female, but, because of the flag if it is used or the original item at the end of the record, the information will still be available for later research. At some later time, particularly when planning a follow-up survey or the next census, the editing teams can use the information about children born to 13-year-old females to decide whether they need to lower the age for inclusion.

390. One problem in the use of imputation flags is that the procedure just described takes up considerable space in the computer. When the flags repeat each variable, the edited data set will be approximately twice as large as the unedited data set. For many countries, this would be unacceptable for long-term storage. However, the original data and the edits could be stored for later reconstruction.

391. Although large populations will need more storage than small populations, no country should be restricted any longer in keeping all the original, collected data, and a series of flags, and the edited data on the records. The use of flags can assist in (1) evaluating the current census, (2) doing intercensal analysis, particularly of demographic items, and (3) preparing wording and editing for the next census or intercensal survey.

392. Countries with very large populations might prefer to use imputation flags on a sample basis for research purposes. For example, depending on the size of the country, it might want to create a data set with every 10^{th} or 100^{th} housing unit. Then the edit would run with imputation flags on this smaller set, helping to evaluate how the edit affects the quality of the data and determine what differences exist between the unedited and edited data.

Figure 25. Example of a flag for a young female with fertility blanked and flag added

Person	*Sex*	*Age*	*Children ever born (CEB)*	*Sex flag*	*Age flag*	*CEB flag*
Fertility blanked						
4	2	13	1			
Fertility blanked and flag added						
4	2	13	BLANK			1

G. OTHER EDITING SYSTEMS

393. Most of this *Handbook* describes the use of top-down methods for census and survey computer editing. A few countries implement another, more complicated, procedure for computer editing, known as multiple-variable editing (see above section D.2).

394. Fellegi and Holt (1976) were the first to develop these procedures, which are usually applied to the most important variables in a census or survey: age, sex, relationship and marital status. However, they can be applied to any group of variables, or all of the variables on a census or survey questionnaire. In the method, the edit program looks at responses to these items simultaneously for one person or for all of the persons in a household in order to identify missing or inconsistent responses. When unknown (blank), invalid, or inconsistent entries are found, a series of tests determine which of the selected items is most in error, and that one is changed first. Then, the tests are repeated to determine that no invalids and inconsistencies remain; if they do, an edit changes the item with the most remaining problems. The procedures are repeated until no errors remain.

395. Statistics Canada developed the Fellegi-Holt approach and used it for Canadian censuses from 1976 to 1991. For the 1996 Canada Census, this approach was refined and called the New Imputation Methodology (NIM). It permitted for the first time, "minimum-change imputation of numeric and qualitative variables simultaneously for large [editing and imputation] problems" (Bankier, Houle and Luc, n.d.). NIM was first implemented in CANEDIT then replaced by the Canadian Census Edit and Imputation System (CANCEIS) (Bankier 2005, Chen 2007). CANCEIS[14] has been in use at Statistics Canada since 2001, and is used to process data after certain edits during collection and data capture have been applied such as validity edits in electronic questionnaires and checks for record duplication. CANCEIS now processes all Census variables for the Canadian Census of population (from both short and long form populations). NIM finds potential donors first and then decides on the minimum number of variables to be imputed based on each potential donor. The reversal of the imputation steps fulfills the data-driven approach and gives NIM significant computational advantages, while still meeting the objectives of imputing the fewest variables possible and preserving subpopulation distributions as much as possible.

396. If the editing process is carried out using traditional dynamic imputation or hot deck method, the imputation information for a series of questionnaire items may come from many different individuals, depending on the information used to update the imputation matrix. For example, if person A's sex, relationship and marital status are correct, these values will update the appropriate imputation matrices. If A's age is missing or invalid, it will, of course, not be used to update imputation matrices. In fact, other items will update that value. So, if the next person has an inconsistent sex and "sex" is imputed, person A will donate the sex. If the age is also unknown, the editing program will use some other person's age.

[14] See Annex VIII for more information about CANCEIS.

397. The NIM uses donors for items, with the hope that all missing or inconsistent information can come from a single donor or a few donors. To obtain all or most of the information from a single donor, whole data records must be stored in the computer's memory. Then, when both age and sex are unknown or invalid, the same, stored variable provides values for both items.

398. The objectives of an automated dynamic imputation methodology should be as follows:

(a) The imputed household should closely resemble the failed edit household;
(b) The imputed data for a household should come from a single donor, if possible, rather than two or more donors. In addition, the imputed household should closely resemble that single donor;
(c) Equally good imputation actions, based on the available donors, should have a similar chance of being selected to avoid falsely inflating the size of small but important groups in the population (Bankier, Houle, and Luc, n.d).

399. Under the NIM these objectives are achieved by first identifying the passed edit households that are as similar as possible to the failed edit household. This means that the two households should match on as many of the qualitative variables as possible, with only small differences between the numeric variables. Households with these characteristics are called "nearest neighbours". The next step is to identify, for each nearest neighbour, the smallest subsets of the non-matching variables (both numeric and qualitative) that, if imputed, allow the household to pass the edits. One of these imputation actions that passes the edits and resembles both the failed edit household and the passed edit households is then randomly selected (Bankier, Houle, and Luc, n.d.).

IV. STRUCTURE EDITS

400. **Structure edits check coverage and determine how the various records fit together.** The structure edits will change somewhat because of (1) improved scanning and (2) the use of electronic data collection technologies; this is also true for the content edits as well. These structure edits must assure that (a) all households and collective quarters records within an enumeration area are present and are in their proper order; (b) all occupied housing units have person records, but vacant units have no person records; (c) households must have neither duplicate person records, nor missing person records; and (d) enumeration areas must have neither duplicate nor missing housing records. Hence, the structure edits check to make sure that the questionnaires in general are complete.

401. The specific structure edits used for one census or survey may need to change over time since the technology used for determining and correcting structure errors changes so rapidly. Specifically, with technological developments in geographic information and tools, it is possible to check the structure more efficiently as the data are collected. The statistical office would need to implement rigorous quality control if the structure of the data sets is to be set during entry. In most cases, it is also important to redo the structure edits during the office computer editing. This chapter examines the more general issue of item validity and the relationship of items between and within records. Chapters V and VI deal with specific individual population and housing items.

BOX 3. GUIDELINES FOR STRUCTURE EDITS

Structure edits should manage the following tasks:

(1) Make sure each enumeration area (EA) batch has the right geographic codes (province, district, EA, etc.), and that common practice is used to name the batches;
(2) Make sure that every housing unit (occupied and vacant) is included; and that all households in an EA are entered;
(3) Merge the households into their appropriate EAs, and merge the EAs into the appropriate higher level of geography;
(4) Assist in deciding between person pages and household pages within or outside questionnaire booklets based on the size of the population and the layout of the questionnaire;
(5) Assign each individual record to its valid record type;
(6) Handle group quarters or collective housing records separately from housing units;
(7) Make sure a correspondence exists between the various types of records: for example, vacant units contain no persons, occupied units contain at least one person. Make sure the number of person records for each household corresponds to the total household count on the housing record. Make sure the correct number of questionnaires are present when multiple documents are used for a single household, and that they are properly linked;
(8) Eliminate duplicate records both within households (duplicate persons) and between households (duplicate households, or parts of households) to avoid over-coverage;
(9) Handle blank records within a record type;
(10) Handle missing housing units.

A. GEOGRAPHY EDITS

1. Location of living quarters (locality)

402. A locality, according to *Principles and Recommendations for Population and Housing Censuses, Revision 3* (United Nations, 2017, paragraph 4.89) is defined as "a distinct population cluster… in which the inhabitants live in neighbouring or contiguous sets of living quarters and that has a name or a locally recognized status". Additional information relevant to the location of living quarters may be found under the definitions of "locality" and "urban and rural" in paragraphs 4.89-4.100 of *Principles and Recommendations*. It is essential for those concerned with carrying out housing censuses to study this information, as the geographical concepts used to describe the location of living quarters when carrying out a housing census are extremely important, both for the execution of the census and for the subsequent tabulation of the census results (United Nations, 2017, para. 4.463).

403. The coding of each housing unit to a small geographic area, often the enumeration area, or to a specific longitude and latitude, allows for flexible production of different geographic outputs and production of comparable area-based geography over time. This is the numerical code that provides the link between aggregated census data and the digital enumeration area boundary database stored in the case of using Geographical Information System (United Nations, 2017, paragraphs 3.70-3.72).

404. The use of electronic data collection technologies can help significantly in improving the quality of the location of living quarters by capturing geographic information during data collection. Handheld electronic devices with built-in global positioning systems (GPS) capability could be programmed to automatically capture geographic coordinates of each housing unit. This information can be used for checking the location of housing units with the data available in the GIS database created during mapping phase.

405. It is crucial to check geographic codes before field enumeration to ensure that all enumeration areas are included in the census frame with correct geographic codes. Validation of geographic codes in the field by local census committed is necessary to have any significant problem during and after enumeration. When editing for location the geographical codes must be completely accurate. Getting complete, accurate codes for the geographic hierarchy for data processing is one of the most difficult tasks of the whole census. If the geography is miscoded, data entry operators may assign the housing unit or units to some other part of the country. It is often very difficult to correct this kind of error, if geographic codes are not validated before enumeration and checked by supervisors during enumeration.

406. In most cases, the structure edit and the content edit are done independently, and by separate teams. Structure edit is glorified bookkeeping, but without it, the content edit cannot move forward. The structure must be set, and in most cases, proven experts must work with the data to make sure that the hierarchy is implemented and resulting in a fully structured file before handing off to the content editors. Even so, the content edit will inevitably require revisiting parts of the structure edit, just as the tabulations will require revisiting the content and structure edits.

2. Urban and rural residence

407. The traditional distinction between urban and rural areas within a country assumed that urban areas, no matter how they were defined, provided a different way of life and usually a higher standard of living than that are found in rural areas. In many industrialized countries, this distinction has become blurred, and the principal difference between urban and rural areas in terms of the circumstances of living tends to be a matter of the degree of concentration of population. Although the differences between urban and rural ways of life and standards of living remain significant in the developing countries, rapid urbanization in these countries has created a great need for information related to different sizes of urban areas (United Nations, 2017, para. 4.93).

408. Most countries determine which geographical areas are "urban" and which are "rural" before the census and make needed adjustments after census data are collected. If the country attributes codes for urban and rural residence (such as 1 for urban and 2 for rural), these codes can be entered during keying of paper questionnaire -or during enumeration with electronic questionnaire- or can be determined during the edit, based on the criteria the editing team prescribes. When the editing team provides a list of the geographical units that are urban and those that are rural, the data processors can easily assign the appropriate codes to the housing records.

409. Efforts should be made to ensure that population characteristics are generally consistent with the enumeration area. For example, in some countries, except for doctors, teachers and persons in similar occupations few professional people should be found in rural areas and few farm workers should be found in urban areas. The editing team should check to make sure that the geographical area has been classified correctly.

410. Sometimes urban and rural designations cannot be made merely from the major and minor civil divisions. Lower levels of geography may be needed to uniquely identify the urban and rural areas. At times these can be identified before the census through GPS or other geographic designations. But other times the definitions of urban and rural can only be obtained after the census when population totals permit cutoffs to be known and so urban and rural designations set.

411. Some countries use a third designation – semi-urban – which also requires geographic and population designations. The three geographic designations – urban, semi-urban, and rural – will be obtained similarly to the methods used for urban and rural alone.

B. COVERAGE CHECKS

1. Enumeration of present population and usual resident population

412. *Definition of usual residence.* In general, "usual residence" is defined for census purposes as the place at which the person lives at the time of the census, and has been there for some time or intends to stay there for some time. Generally, most individuals enumerated have not moved for some time and thus defining their place of usual residence is clear. For others, the application of the definition can lead to many interpretations, particularly if the person has moved often. It is recommended that countries apply a threshold of 12 months when considering place of usual residence according to one of the following two criteria: (*a*) The place at which the person has lived continuously for most of the last 12 months (that is, for at least six months and one day), not including temporary absences for holidays or work assignments, or intends to live for at least six months; (*b*) The place at which the person has lived continuously for at least the last 12 months, not including temporary absences for holidays or work assignments, or intends to live for at least 12 months (United Nations paras 2.49-2.50).

413. Given this definition of usual residence, National census/statistical offices tend to enumerate present population *o* (where persons are found on census night) or usual resident populatio*n*e (where they usually reside). The edit for checking the relationship between housing records, particularly the count of persons in the living quarters and the individual person records, must consider the type of census. Sometimes countries collect both information on present population and usual resident population. An item for each person can indicate whether he/she is (1) always resident, (2) temporarily visiting but with a usual home elsewhere or (3) usually resident in this household but temporarily absent. Tabulations on a *de facto* basis use only (1) and (2) if all three types are present; tabulations on a usual residence basis use only (1) and (3) if all three types are present.

414. National census/statistical organizations implementing these categories must be very careful in their use, not only during data collection and processing, but also to make tables of the entire data set are not run during later analysis. When these three categories are used, users must be aware of the selected population since analysing the

whole dataset will result in including some persons twice. If a present population is required, the tabulation must exclude category (3), persons temporarily away; if a *usual resident* population is required, the tabulation must exclude category (2), persons temporarily visiting. During initial tabulations, the tabulations for the printed reports and supplementary media, the editing team might choose to make a subset of the total data set for processing. For later tabulations, file documentation should state explicitly how to handle the various possible tabulations. Multiple files might be more appropriate.

415. The editing program should be designed to ensure that, when all three types of information are collected, a person should be classified under one group. If a person has few answers, this record indicates that there are enumeration problem requiring special editing procedures for identifying whether this persons belongs to present population or usual resident population.

416. Many countries now choose to collect data on both a present population and usual resent population. As noted, some individuals who are temporarily absent from their housing units will appear more than once. Usually, the best procedure is the develop separate present and usual resident population data sets after editing so that only one type of data will appear in the data set, and no persons will appear twice in the tables. Obviously, statistical office staff must be very careful in making subsequent tables that they use the appropriate data set.

2. Hierarchy of households and housing units

417. Chapter VI examines the relationships between households, housing units and buildings. Implementation of these concepts depends on the individual national census/statistical organization. However, before proceeding with the individual housing edits, the editing team must develop methods of checking to make certain that the hierarchy is respected during data collection and keying.

418. Electronic data collection applications can be programmed to check for this relationship on entry. Unless specifically programmed otherwise, the enumerator would normally be expected to enter households in numerical order. When the housing units are listed using the information available in digital census maps, and these units do not have to be collected in numerical order, a program can be written to test for completion after the fact. However, this hierarchy should be carefully checked during data collection in the office for ensuring that the hierarchy between households, housing units and buildings is properly constructed.

3. Fragments of questionnaires

419. Before editing item by item, the computer program must check for valid records, missing records and duplicate line numbers as part of the structure edit. It must also determine whether the records being edited are for persons living in group quarters. Data entry operators can make a mistake in entering a record, and on occasion, they will forget to delete fragmentary information (parts of records). One function of the preliminary edits should be to examine the file for fragmentary records, to delete them. The most common case will be a record that contains geographical codes but only some population or housing items.

420. When electronic data collection technologies are used for data collection, the edit on entry program should be designed for catching fragments and a message should go to the enumerator (if handheld devices are used) or the respondent (if the Internet is used for data collection) to provide information for missing records. The edit on entry can be designed to validate if all individuals listed in the roster are interviewed accordingly. When there is any inconsistency, missing, double or erroneously entered, the enumerator will then need to decide whether to delete the fragment, or maintain it because the person being interviewed either was not present in the house or refused to provide part of the information. Similarly, the respondent should complete the information for this person or delete the fragment if the person was listed erroneously.

421. In the office, this type of problem is most likely to occur when data are keyed, depending on the entry package. If the keyer starts to enter a person and then finds out that no person is to be entered, a fragment may occur. Alternatively, when scanning, stray marks or partial entries may need to be deleted. And, finally, on electronic data collection entry, the entry package would need to be set up with appropriate skip patterns, and to skip out altogether after the last person is entered.

4. Changing geography

422. Some countries change geography over time. When minor civil divisions become major civil divisions between censuses, staff may want to be able to make tables and trends using either or both levels of geography. When this happens, a second set of geography needs to be implemented to represent the previous census or survey. So, for a census taken in 2020, the basic tables will be compiled for the current geography. However, an additional set of items can be included for "2010" – that is, the major civil division for 2010 and the minor civil divisions. Then, tables can be made for either census, or for both to see trends. Similarly, the 2020 geography can be placed on the 2010 data set for retrospective analysis.

C. STRUCTURE OF HOUSING RECORDS

423. One of the topics that may be included in the collection of information through national housing censuses or surveys is the number of dwellings in a building. In this case, the unit of enumeration is a building and information is collected on the number of conventional and basic dwellings in it (see United Nations, 2017, para. 4.427).

424. The term "general edit" refers to the practice of ensuring that the number of housing units as parts of the building matches the total number of housing units in the housing record. In the case of a mismatch, the number of housing units entered as a characteristic of the building should be corrected to match the number of housing unit records. If the building in question is coded as having five housing units, but the actual count of individual housing unit records for that building is four, the editing team must decide which adjustment to make: (a) to change the first figure on the basis of the count of individual records (which in most cases would prove to be more acceptable); or (b) to introduce another record using information about existing records (which should be avoided).

D. CORRESPONDENCE BETWEEN HOUSING AND POPULATION RECORDS

425. If the census or survey includes both housing and population records, a structure edit is needed to make sure that the two record types agree.

1. Vacant and occupied housing

426. A vacant housing unit should have no population records, but an occupied housing unit must have population records. Where population records are present, but housing is listed as vacant, the vacancy status will be changed to occupied. For electronic data collection, this function should be automatic. Sometimes the record layout includes vacancy status and tenure together in the same item, so this information has to be taken into account as well in making the determination. Also, if a response is available for value of unit for owner-occupied units or "rent paid" for renter-occupied units, then the editing programs uses this information in the determination; otherwise, an imputation matrix may be needed.

427. For electronic data entry, the relationship between vacant and occupied units, within and without population, should be determined at the unit, and the appropriate correspondence set. A computer office edit should not be needed.

428. But if traditional enumeration takes place, with either keying or scanning, if no population records appear for what is supposed to be an occupied unit, then the editing team must decide whether to count it as a vacant unit or substitute persons from another unit. If the unit is vacant, imputation can easily change the variable for vacancy status. If the unit is occupied, however, then the editing team must decide whether and how to assign persons from another unit with the same number of persons, with similar characteristics, if possible. Since it is impossible to know the characteristics of missing persons, this method should be used, if at all, only when the editing team decides it has no other alternative. Three possible alternatives are outlined below:

(a) Choosing to leave a housing unit vacant

429. In this case, the editing team decides that unoccupied housing units coming in it from the field should be left as vacant, so no values are imputed. Housing edits for unoccupied units are described in chapter IV.D.1.

(b) Revisiting the housing unit several times to complete questionnaires

430. The national census/statistical office may choose to implement procedures requiring enumerators to keep returning to unoccupied units during enumeration (usually three times maximum) to understand if these units are vacant or are occupied and then the enumerators will collect data on these households at least minimal characteristics. In this case, the editing team should develop edits that check to see whether the unit is vacant or has enough characteristics to be considered "occupied" or may be not vacant, but no one was present during enumeration. Depending on what the editing team decides is "minimal" information for a personal record, the regular edit described in Chapter IV is applied, or data from donor records are supplied for "missing" persons, as described above.

(c) Substituting another housing unit for missing persons

431. Procedures for substituting whole households or individual missing persons are described elsewhere in this chapter. These procedures require an assumption that the missing persons have the same characteristics as the substituted persons, which is almost certainly not usually the case, and the procedures themselves are very difficult. Still, without these procedures, the counts of numbers of persons, and persons by characteristic, may decrease. In electronic data collection enumeration, substitution is not usually be used as missing persons can be identified during enumeration through checking especially household roster and personal information. If there is any missing person is observed, this procedure is best left to the office computer edit.

2. Duplicate households and housing units

432. For electronic data collection, the data collection application should check to make sure that units are enumerated once and only once. A second entry at the same housing unit should be caught immediately, so that the unit will not be recorded twice. There might be a variety of reasons for duplication of individuals within a household or duplication of housing units within an enumeration area. For example, the wrong geography can be used or the same IDs for different units can be used or some technical errors may arise during data transfer. It is also possible that more than one enumerator could either go to the same unit a second time. So, care must be taken to make sure that the units are enumerated only once.

433. In keying, sometimes an individual data entry operator will input the same housing unit twice. Sometimes different data entry operators will accidentally rekey the same housing units or even whole enumeration areas because of a lack of quality assurance in the national census/statistical office. Thirdly, an enumerator might record the geographical code for a housing unit improperly, creating duplicate information, by assigning it the same geographical identity as that of another housing unit.

434. If the office monitors keyed batches, duplicates will probably not occur. Nevertheless, an editing program should be developed that will make certain that duplicate households do not occur because data entry operators have

keyed the same household or households twice. Countries should not sort their data until the structure checks are finished and problems with duplicate records eliminated. Before sorting, staff can correct batches manually; after sorting, the staff may not be able to find the problem. When the data are sorted, an edit can check for duplicate households and use imputation to eliminate subsequent duplicate entries.

435. Special care is needed when a census or survey has two or more household identifiers being the same but the people in the house are completely different. It is important not to sort the files prematurely and so mixing two independent housing units together. Methods need to be developed to investigate and then manually or develop a program to assign a new identifier to one of the two units. It is also important to make sure that the new designation does not duplicate yet another housing identifier. Hence, it is a good idea to use a series out of range of the original series.

436. When multi-mode data collection is used, it is highly possible to receive duplications across collection modes. Detecting and correction of duplicate persons or housing units might be more difficult, considering that duplications across data collection codes can be detected only when individual records are integrated at the central level. For detecting such error, a household identifier which is usually assigned based on the address frame is extremely important. Using this identifier, duplicate returns can be checked within a specific mode or across multiple modes of data collection. An editing program should be designed to detect double enumeration of persons, households and housing units before finalizing structural edits. More detailed discussion is provided in the next section on Duplicate Records.

3. Missing households and housing units

437. A structure edit can be added to the edit on entry, as discussed here, but determining missing households or housing units may not be a simple edit. Sometimes units no longer exist, and so the housing unit will look missing, when it doesn't exist. One method of coping with this would be to have the enumerator enter a code that indicates that the unit is demolished or otherwise uninhabitable.

438. If more than one enumerator is working in an area, the structure may disjointed, collected in parts, and so a check on missing units probably should not be included in the edit on entry. In fact, in most cases, it would probably be a better edit when included in the office edit and not in the edit on entry.

439. Similarly, during office editing after sorting, missing households may become apparent. For example, the editing program anticipates a sequence of households within the lowest level of geography, such as 1,2,3,4, but receives only 1,2,4. Then a decision must be made either to renumber the units or to find some "acceptable" method of substituting another unit for unit 3. Several ways are available for adding missing households when it is clear they are, in fact, missing and need to be supplied. One method is to simply duplicate the previous household. But, if you know the number of people in the household, as you often do (even though you don't know their characteristics), you work backward and duplicate the previous unit with the same number of people. Similarly, if you know the age and sex of the household members, that information can be used to assist in obtaining a substitute house. It is not a good idea to try to use dynamic imputation to create information about household members since this method often produces variables inconsistent with each other. However, adding missing households in the census database will arbitrary change the census coverage, therefore, if missing households are added, it is necessary to consider them while measuring census coverage.

4. Correspondence between the number of occupants and the sum of the occupants

440. The number of occupants recorded on the housing record should be exactly equal to the sum of the persons in the household. The editing program sums the number of persons and then compares this value to the number of occupants on the housing record. If the sum differs from the value for number of occupants, either the value for number of occupants must be adjusted to equal the sum of persons, or the individual entries must be adjusted.

441. Electronic data collection application edit on entry should note any discrepancy between the number of occupants and the total number of household members and the problem should be fixed during data collection. Chapter VI elaborates on the housing edit for number of occupants.

(a) When the number of occupants is greater than the sum of the occupants

442. Electronic data collection applications should make sure the number of occupants and the sum of the individual population entries is the same; when not, the correction should be made immediately. For office edit, however, if the value for a specific variable for the "number of occupants" on the housing record is greater than the sum of the individual person records, the editing team has a real problem. No one can know the characteristics of missing persons. Hence, editing teams choosing to impute missing persons characteristic by characteristic or by substituting persons from similar households may face a dilemma. Missing persons should not be substituted. However, if the value of number of occupants is accepted, the alternative is to decrease the size of the enumerated population. The editing team must analyse the whole picture and then decide on an appropriate path.

443. Several ways exist for locating and substituting missing records, none of them completely satisfactory. Whole households can be saved with different, important characteristics. When a household with some, but not all individuals is found, the file can be searched for a household where all or most of the known characteristics match, and then missing persons can be adjusted based on the other persons in the donor household. However, the programming for this operation is very complicated, so national census/statistical offices using this approach should start planning long in advance for this operation.

444. A variation on this procedure is to flag all households with missing records and proceed with the rest of the edits. At the end of the editing process, after all individual entries have been corrected, the editing team can choose to have the data processing specialists go through the file making additions and changes using the fully edited dataset. By using this top-down approach, the editing team may find acceptable donors.

(b) Checking numbers of persons by sex

445. Sometimes the number of occupants is reported by sex on the housing record. In this case, the edit – whether edit on entry with electronic data collection or later in the office edit – must sum the number of persons for each sex separately. Again, if the sums differ from the numbers of occupants, one of the values must be adjusted in each case. Usually, totals on housing records are adjusted rather than adding "missing" records or deleting records having useful information because the enumerator is likely to have made a mistake on the dwelling form.

(c) Sequence numbering

446. Electronic data collection applications will normally force the person or sequence numbers in a housing unit to be entered with fixed numerical order. Even if a person is deleted part way through the enumeration in a housing unit, the edit on entry should adjust the numbering to account for this. Similarly, if a person, like a son-in-law has to be inserted into the sequence, the edit should be able to adjust for this as well. It is important to note that in countries where mother's and/or father's line numbers are collected, the edit program would also have to account for this too.

447. When the data come to the office, population records should be sequenced—numbered in order. These numbers should appear as a variable, such as a line number or sequence number on the questionnaire. Also, sequence numbers should appear in numerical order. Errors may occur: sometimes the questionnaires or person forms get out of order because enumerators assemble the information in the wrong order, or they may skip pages, unintentionally leaving blank pages in the dataset.

448. Although a lack of sequencing usually does not affect either edit or tabulation, many national census/statistical offices choose to re-sequence the persons in the proper order. Hence, the editing program must be able to locate out-of-order persons and re-sequence them. As re-sequencing will sometimes affect the relationship to head of household, it must be considered in the editing specifications. Re-sequencing will definitely affect such variables as mother's line number or husband's line number. Therefore, decision on re-sequencing should be taken after careful review of its effect on other variables, especially if a paper questionnaire is used.

449. If the persons in the housing unit are renumbered, and mother's person number is collected in the census or survey, a program would need to check to see whether renumbering the persons also affects the numbering for the mother's person number. It is likely that these would change as well. (Some surveys may also include father's person number or spouse's person number, and these would also have to be checked for change.)

5. Correspondence between occupants and type of building/household

450. The type of relationship between household members should be consistent with the type of housing unit. Sometimes household members appear in a house declared as collective living quarters or vice-versa. In those cases, the type of relationship or the type of housing unit must account for the size of the household and other variables. The edit on entry for electronic data collection should be programmed to check for this.

451. Some countries define a collective living quarters by the number of unrelated persons living together. Sometimes the enumerator collects the information assuming the unit is a housing unit, but the editing team determines that the unit should have been a collective because of the composition of those living there. Then the designation must be changed, usually on the housing record. The edit is a simple one, basically adding up the number of unrelated people living together and then administering the change when above the cutoff.

E. DUPLICATE RECORDS

452. Whether paper or electronic questionnaire is used, as mentioned before, records can be duplicated during enumeration or data capture. When electronic data collection methods are used, duplicate records should be caught immediately as the edit on entry should check the variables from one person to the next to see if they are the same. Sometimes young twins are enumerated, with the same sex and age and other characteristics, so care is needed. When paper questionnaire is used there should be certain procedures for checking double records.

453. Duplicate line numbers are not likely to appear in optically read or other scanned questionnaires. For forms that are to be keyed, the national census/statistical office may choose to check the correspondence between the household list and the line numbers for the household to be keyed manually. This manual check may improve the quality of the keyed data, particularly in comparing (1) the names of persons appearing on a page where all persons in the household are listed with (2) the data on the person columns, rows or pages. Two persons who initially seem to be duplicates may be twins when reference is made to their names.

454. Keyed forms should not have duplicate line numbers if data screens and skip patterns are properly set up. Most contemporary software packages create sequence numbers automatically as part of the data entry process. An error may be introduced when staff enter duplicate records for a person, or an erroneous line number may create a duplicate record. As each record is processed, the editing program compares it with the previous population records for the housing unit. The edit must ascertain that each line number has been captured correctly. Duplicate line numbers are errors and must be changed.

455. Countries may choose to develop their own keying schemes, rather than use an off-the-shelf package. Then, the editing team must decide on the acceptable level of errors. Many methods are available for making these decisions. One method might be to follow the guidelines below:

(a) If the line number for two different records is identical and the number of characteristics that differ is 2 or less, the edit will eliminate one of the records since it is a likely duplicate.

(b) If 3 or more characteristics are different, the line number will be changed. Traditionally, duplicate records were tracked down and corrected manually, but more and more, these are at least partially automated.

F. CONSIDERATIONS FOR DOUBLE RECORDS FOR MULTI-MODE DATA COLLECTION

456. The detection and correction of duplicate questionnaires is an important part of the editing process. Naturally, duplicate responses may be returned within a single collection mode, but the problem of detection and correction across collection modes may be more difficult. As countries consider utilizing more than a single collection mode type, this situation needs to be taken into account.

457. Further, if household records are obtained from administrative sources, the potential exists for obtaining, for example, an electronic return for a household which already has information from an administrative source. Editing rules will need to exist in order to deal with this situation. The remainder of this section will deal with detection of duplicates across electronic and paper returns. However, the use of administrative data may also need to be considered when considering the risk of duplicate returns, and approaches to this may depend very much on the precise use and content of such data.

458. Duplicate returns from a single household may arise for several reasons. Some examples are:

- Two members of a household may inadvertently submit separate responses for the entire household.
- Someone moving homes may take their self-response paper questionnaire or electronic access code delivered to them to another address and return it from there.
- Some issue with the return of a completed response may cause a follow up action to be triggered for an (apparently) non-responding household. Both responses may be received.

459. Within a single mode of return, as explained before, preventative controls (such as a dwelling register check) may be in place to help limit duplicate returns. When multiple mode responses exist, such measures may not in themselves prevent duplicate returns across modes, and detection of such duplicate responses need to be considered. It is also important to consider that the timeline for initialization of response to response return is much shorter for electronic than for paper questionnaires.

460. The detection of duplicate responses is based on the householder identifier used. This identifier may be assigned based on the address of a dwelling when it is known before the start of field operations, or designated during field operations if it is not. In Canada, all duplicate detection is based on the use of this identifier.

461. The detection of duplicate responses across collection modes can be done once the receipt of returns, or the record of these receipts, are available to determine that a household may have returned questionnaires in both. This action may be carried out using a monitoring process on a database inventory of returns. Because such records or signals may be received too late to take an action during field activity, remedial actions will normally have to be taken in post collection operations.

462. The correction of duplicate responses across collection modes can be done once returns are available for review within a common format. Correction may be carried out, for instance, within a processing step using an application developed for this reason running on a database of all responses. As this step must be after

detection, it too will be done as part of post collection operations. While the detection may be a largely automated process, the correction of duplicates will most likely require an interactive process.

463. There are of course legitimate reasons why multiple returns may be received for a single household. Using paper questionnaires, there may be too many people in the household to put on a single form. Such cases do not require the type of consideration discussed below.

464. When multiple returns are received, whether across response modes or not, an interactive process may be used to select the returns which are best suited for further processing, and flag returns which should not be processed any further. This may be a simple process if the responses involved are identical (for instance if full information for the household was provided by two different people), or may involve decisions based on pre-determined rules if there are differences between the forms.

465. A response may be associated to a household identifier to which it may not belong, for example because of an issue with the drop off of a paper questionnaire or a letter containing an activation code for an electronic self-response. And this may become apparent if a second, legitimate, return is received for the same address. In this case an interactive operator would normally select the form correctly associated with the household in question and flag the other form for re-assignment to the correct household identifier. This attempt at re-assignment will only be possible if address information or other field notation is available within the return itself. The attempt at re-assignment may require considerable effort but may be worthwhile as the form in question most likely should be associated to a household that is flagged as non-responding.

G. SPECIAL POPULATIONS

1. Persons in collectives

466. Whether data is collected by electronic data collection technologies or not, the structure edit should treat persons living in collectives such as institutions, barracks or nursing homes differently from those living in regular housing units. Since collectives will not usually have a head of household, countries must determine how best to distinguish between the types of units. One method is to have a different record type for collectives. Another method is to assign a particular code for relationship, one that stands for “group” or “collective” quarters.

467. For collectives, the editing team will need to decide about whether the first person in a collective is designated as head, that is, receiving code 1, and if so how that would affect the edits for ethnicity, nationality, language, religion, etc. Tabulations will differ when institutions are or are not included; users need to be informed either directly in footnotes or elsewhere in a text.

(a) When collectives are a different record type

468. When the national census/statistical office chooses to use a separate record type (due to using a different form for people living in collective quarters or administrative records can be used for full or partly enumeration of people living in collective quarters) , the editing team will have no difficulty determining which records are collectives or collective records. Tabulations for collectives can be easily done by referring directly to these records only. Variables that are unique to the collective records, such as type of collective, can be edited and imputed separately. Variables that are excluded from the collective records can easily be checked to make sure they are actually blank. However, a bulkier file results, since these records are likely to be shorter than the regular population records, but will take up as much room as in a rectangular file. Also, during editing and imputation, some programs may have to check both population and collective records for some items.

(b) When a variable distinguishes collectives from other records

469. When using a separate variable, rather than a separate record type, the editing team may have more difficulty determining which records are collectives or collective records. Under the circumstances, tabulations for collectives can still be easily produced only done by referring to the variable itself, which notes which records are persons in collectives. Variables unique to the collectives, such as type of collective, can still be edited and imputed separately. Variables that are excluded from the collective records easily can be checked to make sure they are actually blank by referring to the code for collectives. A more compact file results, since the additional records for persons in collectives are not needed but are simply included as population records with a different code for the variable for household/collectives. During editing and imputation, the program will have to check only population records, and not both population and collective records, for some items.

(c) When the "type of collective" code is missing

470. The code indicating collectives may be missing or invalid, or a mismatch may occur between the collective code and the relationship codes. The suggested solution when the code for collectives is missing but the relationship codes indicate a collective is to change the collective code accordingly. If the collective code is present, but relationship is missing, the relationship code might be determined from the type of collective.

(d) When the collective code is present, but all of the persons are related

471. If a code for collectives is present, but all persons in the housing unit are related based on the relationship codes, then the code should be changed to indicate a housing unit. On the other hand, if the unit is coded as a household, but no two persons in the unit are related, it might be necessary to change it to group or collective quarters. A household could have 5 or 6 unrelated persons and still not be collective. As emphasized above, consultation among the members of the editing team may be necessary to resolve specific, unusual cases.

(e) Distinguishing various types of collectives

472. Most countries distinguish various types of collectives. They often break the information down further into specific types of collective quarters. This information can be either coded separately as a "type of collective quarters" item or included as multiple possibilities in the household relationship codes.

2. Groups Difficult to Enumerate

473. Edits for the following difficult-to-enumerate groups (United Nations, 2017, para. 4.48) should be planned carefully taking into consideration some of their characteristics.

(a) Seasonal migrants

474. In some countries with seasonal migration, the interviewer will need to know whether a unit is vacant or occupied because of the time of reference. So, even if the household has complete information, this household could also be counted (enumerated) in another place. Of course, the opposite is also true. A household that has two dwellings in different places (these residents are sometimes called *snowbirds* because they live in different, preferred areas in different parts of the year) could be missed altogether if care is not taken.

475. Sometimes, on a very regular basis, whole households live in one place for part of the year, and another place for the rest of the year. The national census/statistical office and the editing team must decide how to handle various types of situations. For example, some persons spend part of each year in another home, such as those who live in a colder part of a country in the warm parts of the year and in a warmer part of the country in the cold parts of the year.

Another case is that of nomads who travel for part of the year but are sedentary for a part of the year—perhaps the part of the year when the country chooses to do its census.

(b) Homeless persons

476. By definition, the record of a homeless person will not have housing information. Electronic data collection applications should account for this automatically, creating a dummy housing record. Creating a "dummy" record (a new record that initially includes blank values for some variables) will make structural checking easier and make the record consistent with the structure of the other housing units. The editing team will have to decide whether to create this dummy housing record to assist in the data processing and tabulation procedures.

477. Since many countries are collecting information on homeless, efforts might be made to obtain information on birthplace, language, education, and so forth. Regular hot decks might not be appropriate, so other methods of imputation might be needed, or the invalid or inconsistent might be made "unknown". Tables might need to exclude homeless if sometimes they have and others don't have ethnicity, language, etc.

(c) Nomads and persons living in areas to which access is difficult

478. Again, like for the homeless, a structure edit may be very difficult. Some countries will collect some "housing" information, so this information can be used to assist in editing the structure of the "unit". Hence, the housing edits would differ from those used in standard units. Population information should be collected as for persons living in standard housing units, and edited in the regular way, following the guidelines below.

(d) Civilian residents temporarily absent from the country

479. In *de jure* censuses, civilian residents temporarily absent from the country, but living in households who can report them, should be included in the standard population edits. For the *de jure* census, some indicators should show persons who are temporarily absent to allow for both usual resident population and present population to be determined. The housing edit will not differ because of the absentees. However, obviously, in a *de facto* census, these people will not be included, so will not appear in the population edits.

(e) Civilian foreigners, who do not cross a frontier daily and are in the country temporarily, including, undocumented persons, or transients on ships in harbour at the time of the census.

480. For a *de facto* census, everyone present in the country at the time of the census should be included, so these persons should be included as well. Individuals should be included in the place that they were present at the time of the census, and edited there, using standard edits for the population items. If housing is not collected, for a collective, or other non-standard housing unit, then that edit will not be done for these individuals either. If ships in harbors are considered housing units, then the housing characteristics should be described, and edited, using the information for other ships from the hot decks.

481. Foreign persons only in the country temporarily, presumably are not included in usual resident population (*de jure* censuses). Undocumented persons would be included, particularly in those countries not distinguishing documented and undocumented persons separately in the census (which would normally produce a better census result). Transients would not be included in the *de jure* census after editing, unless they are transient for the local area, but still usually reside in the country. If a ship is usually harbored in the country port, then presumably the persons on the ship would be included as usual residents and would be edited as such.

(f) Refugees, asylum seekers and internally displaced persons

482. Refugees, asylum seekers and internally displaced persons may be in temporary quarters and may require an indication on a particular variable, a separate record type or a dummy housing record to account for their condition. The editing team will need to develop and implement the appropriate procedures. Normally, the housing and population items will use the standard edit, with hot decks including "refugee housing" as an indicator.

(g) Military, naval and diplomatic personnel and their families located outside the country and foreign military, naval and diplomatic personnel and their families located in the country.

483. For a *de jure* census, military, naval, and diplomatic personnel and their families both inside and outside the country would normally be included. For many countries, information about the military is not obtained in a census, and the country's statistics office must deal with simple counts, or counts with minimal other information. Limited information will make use of hot decks difficult, and likely to introduce errors into the data set, so it is usually better not to include military households reported in this way in the census. Diplomatic personnel may have similar problems. However, enumeration within a country may produce good results when the standard questionnaires and procedures are used, so these housing units should be included in the regular edit, but with an indicator for the special status of the housing unit. Housing units outside the country may not be enumerated in the standard way, so care is needed in assessing whether or not to include these units in the edits; they could still be included in some of the tabulations.

484. For enumerating present population (*de facto* census), usually only the housing units inside the country would be included. Military, naval, and diplomatic households living outside the country would normally not be included. Housing for these personnel would normally be reported by those living in their units in the sending country; population of current residents would be included.

(h) Civilian foreigners who cross a frontier daily to work or study in the country.

485. Civilian foreigners who cross a frontier daily to work or study in the country would normally not be included in either the *de jure* or *de facto* censuses because they were not present in the country at the census reference moment, nor do they usually reside in the country. They would normally be reported in their sending country, in both *de jure* or *de facto* censuses.

(i) Civilian residents who cross a frontier daily to work in another country

486. Civilian residents who cross a frontier daily to work in another country are residents of the country doing the census and should be included in both the *de jure* and *de facto* counts. Both their housing and population items would be edited in the standard way.

(j) Merchant seamen and fishermen resident in the country but at sea at the time of the census (including those who have no place of residence other than their quarters aboard ship).

487. Merchant seamen would be enumerated in a pure *de jure* census, and also in a modified *de jure* census (a census adjusted to include people who have no other residence), but not in a *de facto* census. When included, housing edits need to include reference to the special type of place, but population items should be able to use standard edits when the country's regular questionnaire is used on the ships.

H. DETERMINING REFERENCE PERSON OR HEAD OF HOUSEHOLD AND SPOUSE

1. Editing the reference person or the head of household variable

488. In identifying the members of a household, it is traditional to identify first the head of household or reference person and then the remaining members of the household according to their relationship to the head or reference person. The head of the household is defined as that person in the household who is acknowledged as such by the other members. Countries may use the term they deem most appropriate to identify this person (head of household, household reference person, householder, among others) as long as solely the person so identified is used to determine the relationships between household members. It is recommended that each country present, in its published reports, its concepts and definitions (United Nations, 2017, para. 4.129).

(a) The order of the relationships

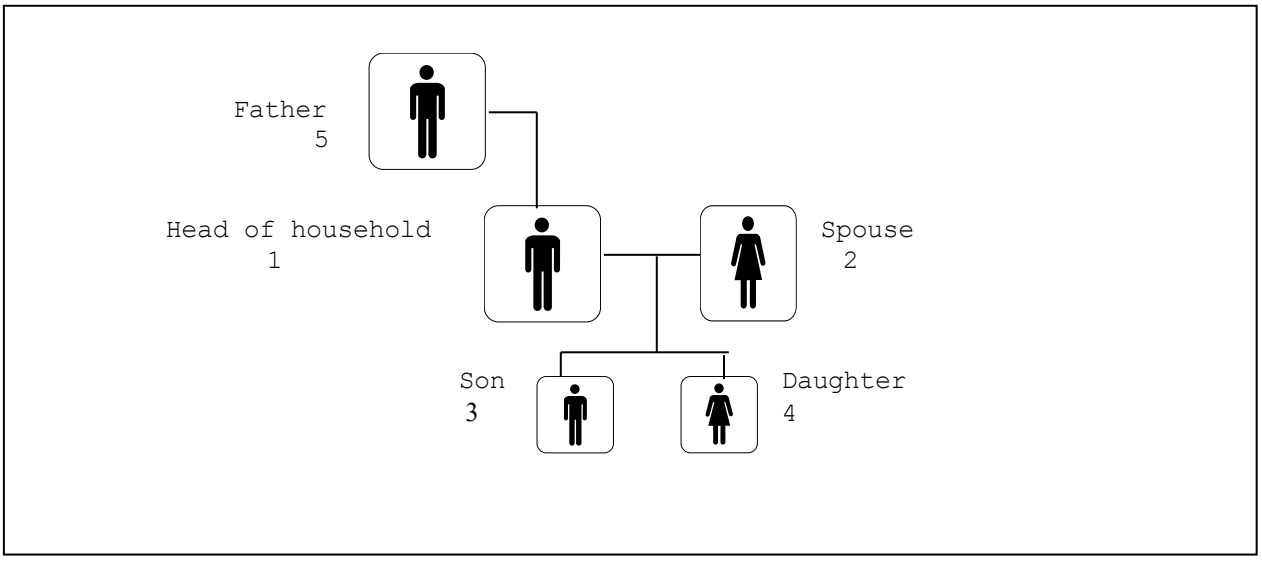
489. The order of the relationships in the unit effects the edits since many of the edits assume that the head of household or the reference person is the first person and his/her data will be edited first. For example, variables such as language, ethnicity, and religion are checked first in the edit for the head of household. If the head of household/reference person has valid information for any of these variables, that information is often imputed for any other person in the household where it is missing, miscoded or mis-keyed (refer to Chapter V). The head of household needs to be edited first since his or her characteristics are used to assign or impute values to other household members.

490. Without a head of household or a reference person, if a country does not use imputation, that is, uses "unknown" when invalids or inconsistencies exist, a different edit would be needed. Otherwise, if no one in the house has a religion, then all of the people will get religion "unknown" even if they are in a particular religious community

(b) When the reference person or the head is not the first person

491. Actions that the enumerators take in the field, based upon the different kinds of situations they encounter with respect to designation of the head of household, affect the editing process. To better understand the issue, consider first the household illustrated in figure 26.

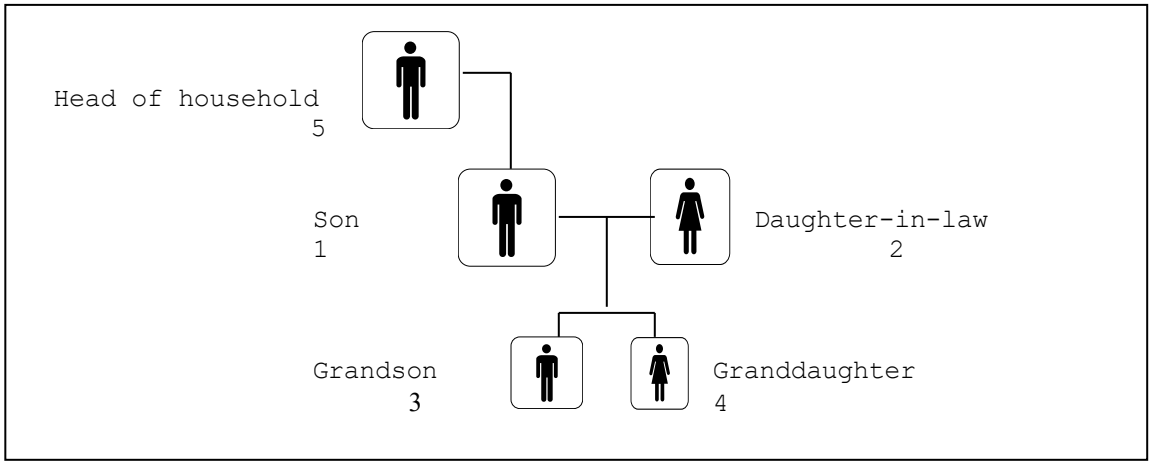
Figure 26. Example of household with head of household listed as first person



492. This household shows a typical situation encountered in the field: a head of household and spouse, their children and the head of household's father. If the enumerator collects the information in this manner, an edit based on the head of household being in the first position in the household will run smoothly.

493. However, if the enumeration is conducted in such a way that the grandfather is designated as the head of household, the relationships are reconfigured, as in the second depiction in figure 27. This situation would occur if an enumerator went into a house, found a nuclear family of husband and wife and two children, and, during the interview, the head of household's father entered the room and claimed that he was the head of household. Based on the agreement of the putative head of household, person 5 would become the head of household, with person 1 becoming the son, person 2 the daughter-in-law, and so forth.

Figure 27. Example of household with head of household listed as fifth person



494. Obviously as illustrated by these two households, the edit paths based on different designated heads of household would be different. Three different possibilities exist for determining the actual head of household for the rest of the edits and tabulations: (i) a pointer can be used to note which person is head, and the pointer can be used throughout the edits and tabulations; (ii) if the head is not listed as the first person, he or she can be moved to the first position, and the persons higher on the list can each be moved one position down; or, (iii) the relationship codes can be changed to have the first person as head, no matter what the other relationships.

495. When electronic data collection applications are used for entry, the edit on entry program should automatically reassign the appropriate line numbers for the head and other household members. The relationships can be programmed to be reassigned as well. And, if mother's person number is included, that can also be readjusted.

(i) Assigning a pointer for the head's record

496. In the editing procedures regarding the head of household or reference person, a pointer is used to determine the line number of the head of household in the unit. If the head remains in the position collected, a pointer can be set to that position, and the head can always be easily found whenever needed for a particular edit or tabulation. A variable "head-pointer" can be set to the line number of the head of household or reference person and used during the edit to assign or impute missing or invalid characteristics for other persons in the unit. If the head is the first person in the household, the value of the variable "head-pointer" is "1".

(ii) Making the first person the head

497. The editing team may choose to move the head to the first position in the household. The programming for this is somewhat more complex than that required for (i) above. The data processing specialist must develop a program that moves the head to the first position on the list, followed by the person who was previously in position 1, then the person who was in position 2, and so forth, until reaching the person just before the person who was the head. So, if the head is in position 5, the order of persons will change from 1,2,3,4,5 to 5,1,2,3,4. After this change is made, the head will be in position 1, which makes the rest of the edits easier since the head will always be in that position. Nonetheless, if this operation is carried out, some "damage" is done to the integrity of the data set. Since the order of persons has been shifted, analysts may have difficulty determining the actual order of persons collected from the field and the potential affect of this order on the interpretation of the results.

498. The programming for moving the head to position one is no longer complicated since most packages can easily handle it, many with a single command. But, mother's and spouse's person numbers may be affected by this. Relationship codes will also have to be changed, but this is also straightforward. Finally, if a country develops family and sub-family codes, those also might be affected.

(iii) Reassigning relationship codes to make the first person the head or the reference person

499. If the editing team decides that the first person listed is to be the head of household, then procedures (a) and (b) need to be followed in the edit:

(a) The first person is assigned the value for head of household;

(b) A routine is implemented that reassigns values to other persons in the household to adjust the household.

500. For example, in figure 26, the parent starts out as the head of household. When person 1 is made head of household, person 2 will need to be assigned "spouse", persons 3 and 4 will be assigned "child", and person 5 will be reassigned "parent" (as shown in figure 25). The subroutine will need to contain a matrix to hold the initial and changed values.

501. The integrity of the dataset is affected to an even greater extent with this procedure. The order of persons is not shifted as in the previous example, and analysts will not have difficulty determining the actual order of persons collected from the field. However, all of the relationships will change, and analysts will not know which person was initially selected as head of household. Also, if mother's person number, father's person number or spouse's person number is collected in the census or survey, these must be taken into account in any renumbering scheme. On the other hand, tabulations may be nominally easier with the head in the first position. Unlike the previous example, for this procedure programmers do not have to physically move the records around.

(c) More than one head

502. When more than one head of household or reference person is found, the editing team must determine who is to be designated as the head of household. If electronic data collection applications are used, edits on entry will not allow to enter second head of household or reference person. So, this kind of errors should not occur where electronic data collection is used. For paper questionnaire, the editing must be performed based on characteristics set by the subject-matter specialists and by edit flows. The editing program must then reassign the relationship of the other person(s) who were identified as heads of household.

503. A special case exists in those countries permitting "co-heads" either because of socio-economic conditions (like male heads frequently leaving for mining or other activities and leaving the spouse as head) or because respondents insist on "equality". Traditionally, for editing purposes, it is important to designate one and only one head of household, with original data maintained on the record in these cases. If a country chooses to include co-heads, these must be maintained in the edit; however, many of the suggested subsequent edits in this handbook would

also have to be modified; when the co-heads have different religions or tribes or other demographic and social characteristics, a single person can no longer be used in the imputation procedures.

504. Also, some countries now recognize same sex marriages. If both members of a couple – whether heterosexual or homosexual – insist on being enumerated as "head" or "reference person", the editing team may want to assign one as head/reference person for the edits and the tabulations, while maintaining the structure for other purposes.

(d) No head or reference person

505. Similarly, when using electronic data collection edits on entry are used, one member of the household has to be determined as a reference person or head of household, otherwise, the application will not allow to move on. When paper questionnaires are used, during office edit, if no head of household is found, the edit must determine who is to be designated as the head of household. In this case, it is likely that the relationships between other persons in the household will need to be adjusted through editing. In determining a head in this way, variables such as age, educational attainment, and economic activity should be taken into account to get the most likely head. A flow chart for a sample edit for head of household appears in Annex V.

2. Editing the spouse

(a) When exactly one spouse is found in monogamous societies

506. If exactly one spouse is found, the variable "spouse-pointer" keeps track of the line number of the spouse for later edits. These edits might include looking for opposite sex for head of household and spouse, for appropriate age differences, or for other relevant characteristics. (In countries with same-sex spouses, the edit would need to be adapted).

(b) When more than one spouse is found in monogamous societies

507. In a monogamous society, if more than one spouse is found in the dataset, then whether electronic data collection edit on entry or later in the office, an edit must determine who is the spouse, and reassign the relationships of the other persons who were identified as spouses. Again, subject-matter specialists must determine what the characteristics and flow of the edits should be.

(c) Spouses in polygamous societies

508. If more than one spouse is found in a polygamous society, the editing team may want to leave the information as it is, or do some consistency checking. For example, at a minimum, each of the polygamous spouses should have the opposite sex of the head. If same sex spouses are found, the earlier edit for spouses of the same sex should be applied.

(c) Other characteristics of head/reference person and spouses

509. Good editing practice is to impute other important items for head and spouse when they are identified in this part of the overall edit. These items include age of head and spouse and marital status, which may be needed later in imputation files and for other edit purposes. Also, it is also a good idea to get "social" items such as religion, ethnicity, and language of head at the beginning, particularly if the head is not listed as the first person; since most packages start with the first person and work down, having the head's information in place before editing the other people in the unit is important. A sample flowchart for edit for spouse appears in Annex V.

510. It is important to note that when the head/reference person is not the first person and/or the spouse is not the second person, additional editing may be required to make sure that the head and spouse are edited before anyone

else. For example, if the head or reference person is recorded as the last person in the housing unit, the other persons in the unit will be edited first, which will affect items such as religion or ethnicity. So, other people in the house may get the value 0 assigned to them if edited before the head.

I. AGE AND DATE OF BIRTH

1. When date of birth is present, but age is not

511. When electronic data collection application is used for data entry and age is checked with hard-edit, date of birth (or completed age) will be known for all persons. When the data is captured in the office, when the date of birth is collected, but age is not, the latter information can be obtained by subtracting the date of birth from the date of the census or survey. Some national census/statistical offices choose to obtain the age based on the year of the census and the year of birth only, giving a value with potential deviation. If year and month are used, the age will be more accurate, but using day, month and year will give the most accurate results.

2. When the age and date of birth disagree

512. When the census or survey obtains both completed age and date of birth, a "computed" age is obtained by subtracting the date of birth from the reference date. If this value is different from the reported age which may happen during the interview or data capture, the editing team might want to take remedial action. Normally, date of birth takes precedence over reported age, and the computed age is substituted for the reported age. However, before making a decision on a computer age, it is important to check whether date of birth is consistent with other information such as education, recent births and so on. Edits should consider the possibility of making a mistake in capturing date of birth and correctness of the reported age.

513. More and more countries can rely on date of birth, but some older people may not know their birthdates in a few remaining countries.

J. COUNTING INVALID ENTRIES

514. Some editing teams may choose to implement procedures for counting the number of invalid and inconsistent entries for the major variables (or all of the variables), such as age and sex, before starting on the actual editing. If the editing team prepares itself beforehand or conducts periodic surveys using these same items, they may have several different dynamic imputation arrays available to them. If the percentage of invalid or inconsistent entries is very small, the editing team may decide to use only a few variables for the imputation. If the percentage of errors is larger, the editing team may need to use more variables to account for the large number of imputations required.

515. Smaller imputation matrices are usually better because they are easier to check out as the edits and imputations are being developed, and they are easier to use during the actual editing. However, if values are used repeatedly, a larger, more varied imputation matrix will be needed.

V. EDITS FOR POPULATION CENSUS TOPICS

516. Chapter five covers edits for population items, including those related to demographic, migration, social and economic characteristics. The list of topics covered in this document is taken from the recent revision of the *Principles and Recommendations for Population and Housing Censuses, Revision 3*.

517. The specifications for these edits take into account the validity of individual items and consistency between population items as well as between population and housing items. Having some knowledge of the relationships among the items makes it possible to plan consistency edits to assure higher quality data for the tabulations. For example, population records should not have 15-year-old females with 10 children or 7-year-old children attending tertiary school.

518. When assigning values for population items, the editing team must decide whether to assign "not stated"; a static imputation (cold deck) value for an "unknown" or other value; or a dynamic imputation (hot deck) value based on the characteristics of other persons or housing units.

519. Most countries now use hot deck. If "unknowns" remain, they must be dealt with at the time of tabulation and analysis when no similar geographic, village, or family information is known. With the unedited data on the records, further analysis can be done later.

520. In most cases, dynamic imputation is preferred since it eliminates editing at the tabulation stage, when only the information in the tabulations themselves is available to make decisions about the unknowns. Imputation matrices supply entries for blanks, invalid entries or resolved inconsistencies when no other related items with valid responses exist. Some countries have some variety in population characteristics across the nation, but very little variation in most individual localities. Others may have considerable variation among localities, particularly concerning urban and rural residence. This variation must be considered when developing imputation matrices and, particularly, when establishing the initial cold deck values. The editing team should specify the circumstances in which entry should be supplied for a blank. This entry should come from a previous housing unit with similar characteristics.

521. Population record serial numbers assist in data processing. The structural edits described in Chapter III check for correspondence between the sequence number and the order of serial numbers. When the serial number does not correspond to the order within the housing unit, an edit must determine whether duplicates or missing persons are involved. If mother's person numbers are present in the file, care is needed to assure that in correcting missing or duplicate values, the line number changes are appropriate.

522. The editing team should edit each population record for applicable items only. The edited items may differ depending on urban/rural, climatic, and/or other conditions. It is desirable to edit selectively, depending on these conditions, but in practice few countries have the time or expertise to develop and implement multiple arrays to deal with missing or inconsistent data. Even fewer countries actually implement this added procedure.

523. Information collected on the questionnaire also sometimes applies only to selected population groups. For example, fertility is asked only of females, and economic activity is usually asked of adults.

524. Sometimes the editing team should allow a "not reported" entry for certain items. The editing team may lack a good basis for imputing responses for some characteristics. The decision to leave "not reported" responses must be balanced against the requirement to produce appropriate, tabular characteristics for planning and policy use. As long as the "not reported" cases have the same distribution as the reported cases, allocating the "not reported" cases when planners need selected information should pose no problem. If the "not reported" cases are somehow skewed,

however, the post-compilation imputation could be problematic, particularly for small areas or particular types of conditions. For example, if teenage female respondents refuse to reveal their fertility information, and no fertility is collected, the editing process will not be able to assist in obtaining this information.

525. Examples of appropriate "not reported" would be industry and occupation. In some cases, place of birth or residence in the past might be other examples. For most countries, hot decks for occupation and industry would be very difficult because people of all ages, sexes, and educational attainment could be reported in most occupations or industries. When the number of occupations is limited, however, that is, most of the population is subsistence farmers, then a hot deck might be appropriate. All hot decks should be thoroughly tested.

526. Population edits tend to be more complicated than housing edits because cross-tabulations are generally much more sophisticated. Most countries compile individual housing characteristics only by various levels of geography but may have many layers of cross-tabulations for the population items. As explained above, countries choosing not to use dynamic imputation should determine an identifier for "unknown" values for use when invalid or inconsistent responses occur.

527. For countries that use dynamic imputation, editing teams should develop simple imputation matrices with dimensions that differentiate population characteristics. For most countries, age group and sex are the best primary variables for dynamic imputation, so they should be edited first. National statistical/census offices using multiple-variable editing should edit age, sex, and other variables, such as relationship and marital status, simultaneously. Other items that may be helpful in dynamic imputation include level of educational attainment, economic activity, and employment status.

528. Editing teams must be very careful not to skew the data during imputation. Teams should not assume that the unimputed and imputed data will necessarily have the same distributions. Often, the unknown data are skewed themselves. For example, older people are less likely to report their age than younger people. Similarly, when sex is not known, if fertility is checked and the person is made female by its presence, and then all other unreported sexes are even divided then the imputations will add more females than males. But, in most cases, the distribution of unedited values, edited values, and imputed values should appear in the same proportions.

529. Use of electronic data collection technologies adds another layer of assistance in obtaining the highest quality data, but also can cause problems when over-editing occurs during entry. As noted previously, if the data collection application produces an error message You just keyed in a 79-year-old woman with a 3 year old child. Did you mean that?" The enumerator and respondent must then work out what the actual relationship or ages were supposed to be. However, the child could also be adopted, and so the message should be carefully discussed with respondents.

A. GEOGRAPHIC AND INTERNAL MIGRATION CHARACTERISTICS

1. *Place of usual residence (core topic)*

530. When countries collect de jure census data, the enumeration is by "usual residence", compared to de facto collection, where the enumeration takes place using the place where a person is present at the census reference moment. Hence, countries taking de jure censuses should not be asking a separate item on usual residence.

531. Countries doing de facto censuses, however, may include an additional item on "usual residence" to obtain de jure as well as de facto information. Edits for this item will vary depending on the particular country situation. For persons who have never moved, the usual residence will be the same as the place of enumeration, so missing information can be filled directly.

532. But, when the data show that a person is a visitor or not residing in the place of enumeration, the situation becomes more complicated. Usually, countries assume that when this item is left blank, the usual residence and the place of enumeration are the same, and the enumerator and/or respondent simply left the information out.

533. However, when the data show evidence, by relationship to the reference person or some other evidence of residence somewhere else, then the statistical staff may want to try to develop methods of obtaining best estimates for particular geographic areas or for the whole country. Although the specific edit will depend on the particular country's situation, a category for "unknown" should probably be used as a last resort.

534. If the enumerator is instructed to leave the entry blank if the usual residence is the same as the place of enumeration, the code for the place of enumeration should be placed in the item for usual residence during edit. Another variable should indicate that the editors have made this change. Having a complete set of codes will assist users of the public use sample in making complete tabulations of their data. The usual residence edit should probably be done in the office edit so that all of the migration variables are at hand can be readily used to assist in obtaining the best entry.

2. *Place where present at time of census (core topic)*

535. Normally the place where present at the time of the census should not need editing unless the enumerator recorded the detailed information improperly for any reason. If the entry is invalid, the households on either side – and if all levels of geography except the household number agree, then that information should be assigned to the housing unit. When the geography differs, the information should be looked up in the register. This information should not be imputed.

3. *Place of birth (core topic)*

536. The place of birth is, in the first instance, the country in which the person was born. It should be noted that the country of birth is not necessarily related to citizenship, which is a separate topic (see United Nations, 2017, para. 4.64 to 4.71 and the section on country of birth in this handbook). For persons born in the country where the census is taken (natives), the concept of place of birth also includes the specified type of geographical unit of the country in which the mother of the individual resided at the time of the person's birth. In some countries, however, the place of birth of natives is defined as the geographical unit in which the birth actually took place. Each country should explain which definition it has used in the census (United Nations, 2017, para. 4.65).

(a) Relationship of entries for place of birth and duration of residence in current place

537. The entries for place of birth and duration of residence can be checked for consistency since strong relationships exist between the two items. Also, relationships exist between the different members of a household, and assumptions can be made from other family members as to whether or not the person in question has migrated.

538. The duration of residence cannot be greater than the age. If they moved before they were born, a method should be used to change the entry. If they never moved, but the subject specialists insist on duration being filled, a different code should be used for "never moved" than moved within the first year of life.

(b) Assigning "unknown" for invalid entries for place of birth

539. If a country chooses not to use dynamic imputation, any invalid responses for place of birth should become "unknown." Usually a country should not edit inconsistent responses among family members or for geographical areas unless the coding is amiss.

(c) Using static imputation for place of birth

540. The entry for place of birth should be altered only if it is out of range. If the code for duration of residence is “always”, the code for the country or if a person is born in a country, the code of the place of residence should be assigned. If the entry is other than “always”, information for a previous person can be used. For example, if a previous person is the mother, the number of years the mother lived in the place could be compared with the person’s age. If the mother’s entry is greater than or equal to the person’s age, the code for “this country” should be assigned; otherwise, “mother’s country of birth” should be assigned. If country of birth cannot be assigned based on the mother’s entries, the entries of other related persons can be used in the same way. If an entry cannot be assigned after these tests, country of birth could be assigned as “unknown”.

541. Because countries now provide samples of their data for public use, it is important to provide a specific code during edit to those entries that are blank because they are skipped by enumerators following skip patterns. That is, often the questionnaire tells the enumerator to skip the question on place of birth of the person always lived in this place. During edit the code for the specific place should be assigned to assist users so they will not need to look in two places when making their own cross-tabulations later.

(d) Using dynamic imputation for place of birth

542. As before, the entry for country of birth should be altered only if it is out of range. If the entry for years in the place is always, the code for “this country” should be assigned country of birth. If the entry is other than “always”, information from other people in the household should be studied for clues to this person’s country of birth. The place of birth edit should be the same, whether edit on entry or when in the office. However, in the office, all of the migration information will be available – current residence, place of birth, residence at a point in a past, previous residence and citizenship, so combinations of these can be used to obtain the best entry; some of these will not be available in the top down approach to data collection.

(e) Assigning place of birth when a person’s mother is present

543. If the country of birth is blank or invalid, and the duration of residence is other than “always”, a search can be made for the person’s mother. If the mother is found in the household, the entry for the mother’s duration of residence is examined. If her entry for years lived in the place is “always”, the person’s country of birth can be assigned as “this country”. If the person’s mother did not always live in the place, but this person’s age is less than or equal to the number of years that the mother has lived in the place, the program can also assign “this country” to the country of birth. If this person’s age is greater than the number of years the mother has lived in the place, and the mother’s country of birth is valid, this person’s country of birth is assigned the same country of birth as the mother’s.

(f) Assigning place of birth for child of head or reference person

544. If the person’s mother is not in the household, but this person is a son or daughter of the head of household, then to obtain the birthplace several checks can be made using information from the head of household’s record. If the entry for the head of household’s years lived in the place is “always”, the program should assign “this country” as country of birth to the person’s record. If the head of household’s years in the place is not always, but this person’s age is less than or equal to the number of years that the head of household has spent in this place, the program should also assign “this country” as the person’s country of birth. However, if this person’s age is more than the number of years the head of household has spent in this place, the program should assign the head of household’s country of birth if it has a valid code for country of birth.

(g) Assigning place of birth for child, but not of head or reference person

545. Quite different imputations can be made depending on whether or not a person is above or below a given age (age X) set by the country's editing team. If a person is less than age X, country of birth should be imputed from the first previous record for a child under age X, by age and sex.

(h) Assigning place of birth for adult females with husband

546. If this person is age X or older and is female, the program should check to see if she has a husband in the household. If the woman has a husband, and he has a valid code for country of birth, the program should assign his country of birth code to her record. If the husband does not have a valid country of birth code, his entry for years in the place of residence should be looked at. If the husband's years in the place is coded "always", the woman's country of birth should be assigned "this country". If the husband's years in the place is not "always", then the woman's country of birth should be imputed by age and sex.

(i) Assigning birthplace for adult females with no husband

547. Although a woman over some minimum age set by the editing team does not have a husband in the household, she may be the mother of a child in the household. In this case, the program should search for her eldest child. If the child cannot be found, the program can impute country of birth by age and sex. If the child has a valid country of birth code and the mother's reported years in the place of residence are greater than the child's age, the program should impute country of birth by age and sex. But if the mother's years in the place are less than or equal to the child's age, the program should assign her the child's country of birth.

(j) Assigning place of birth for males

548. To obtain the place of birth for a male, several approaches can be applied. If a country has minor international migration, the editing team can decide to find country of birth of his wife. Another approach is if he is the head of household, the program should try to find his children. First, the program attempts to find the man's wife. If she is found, and his years in the place of residence are less than or equal to hers, the wife's country of birth can be assigned to the man's record. If the man's years in the place are greater than his wife's, the country of birth should be imputed by age and sex using an imputation matrix. When the man is the head of household of the family, has a son or daughter present in the household, and has been in the place for an amount of time equal to or less than the child's age, then the program should assign the same country of birth as his child's. If his time in the place is greater than his child's age, the program should impute by age.

4. *Duration of residence (core topic)*

549. The duration of residence is the interval of time up to the date of the census, expressed in complete years, during which each person has lived in (a) the locality that is his or her usual residence at the time of the census, and (b) the major or smaller civil division in which that locality is situated (United Nations, 2017, para. 4.72).

(a) Edit for duration of residence

550. Like country of birth, the duration of residence is important when compiling statistics on the mobility of the population. In some instances, a subgroup of the population may be far more mobile than the nation as a whole. The edit for this item takes into account the person's place of birth and the responses for other members of the households. "Duration of residence" should be edited with "place of previous residence" or "place of residence at a specified date in the past".

551. In some ways, the duration of residence item can be easily edited during data collection if electronic data collection technologies are used. An error message can tell enumerators immediately if they have entered a length of

residence which is greater than the age of the respondent or other residents in the housing unit. Efforts can then be made to correct the information immediately.

(b) De facto/de jure residence and duration

552. The edit may be affected by whether the census is a de facto or de jure census. Because the de jure census collects information at the usual residence, duration of residence may not elicit the same information as in a de facto census where persons are enumerated where they were present at the census reference moment. In addition, codes and edits must take into account persons who either "always" lived in the place or "never left." For these individuals, the editing program should skip consistency and other edits.

(c) Relation of age to duration of residence

553. The first part of the edit should check for consistency between age and place of birth and for a valid entry in years lived in the locality or civil division. The number of years a person has lived in a locality or civil division cannot be greater than the person's age. In addition, a person who was born outside of the country cannot have always lived in the locality or civil division. The program should assign "always" to years lived in locality or civil division, if years in locality or civil division is greater than age and country of birth is this country. If years in locality or civil division is greater but country of birth is not this country, the person's age should be assigned to years in locality or civil division. In that case, it is assumed that although born outside of this country the person moved into the locality or civil division when he/she was less than 1 year of age.

(d) Relation of birthplace to duration

554. In the case of out-of-range entries, the same tests as those for place of birth should be used. A search should be made for related previous persons (mother, head of household, husband, child). Imputation should be based on the information found. However, before a value is assigned it must be consistent with the age and place of birth of the person whose record is being edited. Edit on entry for electronic devises will assist in this edit, so it is a useful one to include, knowing that further office edit will be needed to check for the relationship of all migration variables.

(e) For persons who have always lived here

555. If the response for the number of years a person has lived in the locality or civil division is "always", but the country of birth is not "this country", the editing team might want to assign the person's age to the duration of residence in the locality or civil division. The specialists will assume that although born outside of this country the person moved into the locality or civil division when less than 1 year of age. The next part of the edit will check for a valid entry in years residing in locality or civil division. Since the length of time a person lived in the locality or civil division cannot be greater than the person's age, age will be assigned to the years in locality or civil division for this situation.

556. Sometimes it is a good idea to put in the age for the duration of residence for those who never moved to assist in cross-tabulations, instead of leaving the field empty. It is important that this be done in a way that "never moved" is uniquely identified. So, if the code for "never moved" is 98, then the maximum length of residence would be 97, and persons older than 97 would still have 97 as their duration of residence, while those who never moved would have code 98.

(f) Person's duration from mother's duration

557. If the category does not have a valid code, the program can perform an inter-record check by searching for the person's mother in the household. If found, the mothers record can provide information helpful in assigning

missing values. If the person's mother has always lived in the locality or civil division, and her country of birth is "this country" (as it should be), the program will assign "always" to this person's years in locality or civil division category. If the mother's country of birth is not "this country", even though the entry for her years in the locality or civil division is "always", this indicates that something is wrong with the mother's categories. The program will then ignore the mother's country of birth and assign age to duration of residence in locality or civil division. If the entry of the mother's years in the locality or civil division is not "always", but is a valid code, and the person's age is less than the number of years the mother has lived in the locality or civil division, the edit will go back and check the mother's country of birth. If the mother's country of birth is "this country," the program will assign this person's age to years in locality or civil division. However, if a person's age is equal to or greater than mother's years in locality or civil division, the program will assign "mother's years in locality or civil division" to this person's years in locality or civil division.

(g) Person's duration from child's duration

558. If the person in question is a child (son or daughter), the editing program should check the head of household's record for possible information to aid in assigning values for missing data on duration of residence. When the head of household was born in "this country" and has always lived in this locality or civil division, the program will assign "always" to the child's years in locality or civil division. When the head of household has always lived in the locality or civil division, but was not born in "this country," the child's age will be assigned to locality or civil division. When the head of household's entry for years in locality or civil division is not "always", but is a valid code, this information can be used if it is consistent with the age in the record of the child being edited.

559. If the child's age is equal to or greater than the number of years in the locality or civil division of the head of household, the program will use the head of household's years in locality or civil division as the years in the locality or civil division of the son or daughter. If the child's age is less than the head of household's years in locality or civil division, the program will assign a value depending on the country of birth of the head of household. This value will be "always" if the head of household was born in "this country"; if not, the program will assign the son's or daughter's age to years in locality or civil division.

(h) Person's duration when no other information available

560. When all of the above efforts fail to produce a valid value, the program can assign "not reported" or "unknown" to years in locality or civil division for this person. If the value is still invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics to obtain "known" information from similar persons in the geographical area.

(i) Relationship of Duration of Residence to Year of Arrival

561. It is important to note that some countries will concentrate on internal migration and include the item on duration of residence (often with previous residence). Other countries focusing on international migration will include the item on year of arrival (often with residence preceding the move). Most countries have either considerable internal migration and little international migration or considerable immigration and little internal migration. Some countries, however, will have both internal and international migration, and so will include both items.

562. When both items are included, statistical office staff must be very careful to develop edits that do not end up being internally inconsistent. That is, the variables for age, duration of residence, and year of arrival must be considered together to be certain that sum of the duration of residence and time since arrival is not greater than the age. Hence, programmers will need to consider all three variables at the same time when this occurs.

563. When dynamic imputation is used, the statistical staff may need to use a hot deck that includes multi-dimensional arrays to account for the various ages and years. Also, when duration of residence and year of entry are single years, the hot deck must also use single years, or the update for a 5-year group, for example, may cause a conflict during imputation.

564. Also, great care is needed when grouped data for duration of residence or year of entry or both are collected when doing this checking, and when developing and implementing hot decks. Grouped data cause problems of overlap. Countries may decide that supplying an "unknown" may be the best approach for this situation.

5. *Place of previous residence*[15] *(core topic)*

565. The place of previous residence is the major or smaller civil division, or the foreign country, in which the individual resided immediately prior to migrating to his or her present civil division of usual residence (United Nations, 2017, para 4.75).

(a) Previous residence edits

566. The item "place of previous residence" should be edited with "duration of residence". If the person was born in this place (country, locality or civil division, depending on the census item) and never moved, either this item should be left blank, or a specific code for "never left" should be assigned. However, blanks can cause problems during tabulation, so the editing team needs to decide on the best approach for their situation.

(b) Previous residence when boundaries have changed

567. Boundaries of countries change over time, so care should be taken to make sure that appropriate correspondences are reflected in the coding schemes. In addition, the codes should be set up in a way that allows for logical groupings. For example, as mentioned above, in a three-digit code, the first digit might represent the region of residence, the second digit the province within the region and the third digit the district within the province.

(c) When person has not moved since birth

568. Data processors make tabulations on certain individual items. So, specialists should make certain that a special code for "born here" is used in addition to the other place codes. In this way, the program can distinguish between persons born in a place and those who were born in one place but moved to another place within the same geographical area.

(d) Use of other persons in unit

569. When "place of previous residence" is invalid or inconsistent, edits similar to those performed for "duration of residence" usually apply. The editing program can examine the mother's previous residence if she is in the housing unit. The program can then look at the head of household's previous residence for both children, and adults in those countries where adults do not move often.

(e) No appropriate other person for previous residence

570. If none of the above produces a valid value, the program can assign "not reported" or "unknown" to years in previous residence for this person. If the value remains invalid, "unknown" should be assigned when dynamic

[15] Place of previous residence is the place that the individual resided immediately prior to migrating into the place of present usual residence.

imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics to obtain "known" information from similar persons in the geographical area.

6. *Place of residence at a specified date in the past (core topic)*

571. The place of residence at a specified date in the past is the major or smaller division, or the foreign country, in which the individual resided at a specified date preceding the census. The reference date chosen should be the one most useful for national purposes. In most cases, this has been deemed to be one year or five years preceding the census. The former reference date provides current statistics of migration during a single year; the latter may be more appropriate for collecting data for the analysis of international migration although perhaps less suitable for the analysis of current internal migration. Also, to be taken into account in selecting the reference date should be the probable ability of individuals to recall with accuracy their usual residence one year or five years earlier than the census date. For countries conducting quinquennial censuses, the date of five years earlier can be readily tied in, for most persons, with the time of the previous census. In other cases, one-year recall may be more accurate than five-year recall.

572. Some countries, however, may have to use a different time reference than either one year or five years preceding the census because these intervals may present recall difficulties. National circumstances may make it necessary for the time reference to be one that can be associated with the occurrence of an important event that most people will remember. In addition, information on year of arrival in the country may be useful for international migrants (United Nations, 2017, para. 4.80).

573. "Place of residence at a specified date in the past" is like the edit for previous residence. Usually, countries will ask either "duration of residence" and "place of previous residence" or simply "place of residence at a specified time." If the person was born in the place of enumeration (country, locality or civil division, depending on the census item) and never moved, this item might either be left blank, or a specific code for "never left" may be present. As mentioned before, blanks may cause problems during tabulation. Then, the same procedures for previous place of residence, described in the three preceding paragraphs, apply.

7. *Total Population (core topic)*

574. The total population in the housing unit is normally determined by a roster at the beginning of the housing unit enumeration to account for all persons living in the unit. After all persons in the unit are recorded, the program should sum then and compare the total with what is recorded on the housing record as the total population. When these values do not match, the enumerator or the responded should be warned immediately to determine which value is wrong, and either list the persons missed during the initial enumeration or change the value of the total population so they will match.

575. In the central office, during editing, the total population will be matched with the sum of the persons enumerated. When these do not match, the analyst will have to determine whether it is worth re-contacting the enumerator or respondents to rectify the situation. Otherwise, the total population on the housing record should be adjusted to be the sum of the population records.

8. *Locality (core topic)*

576. The locality should be recorded before the enumerator goes into the field. If the locality is not filled, the geography section should determine the code for the locality, and it should be entered on the records.

9. *Urban and rural (core topic)*

577. Most countries hard code urban and rural residence. The geography of the urban and rural areas (and sometimes semi-urban changes over time so care must be taken when doing trends analysis.

578. If the urban and rural areas are predefined, that is, before the enumeration, so guesses about the population size and structure are set, then the variable for urban and rural residence would already be on the records. Part of the edit might sum those living in a defined urban area to see if a minimum is actually reached (like 2,500 people); if the minimum is not reached, the analysts may decide to reclassify the area.

579. When the determination of urban and rural residence is not pre-determined, so on the records, the analysts can provide the detailed geography of what constitutes an urban area and what constitutes a rural area, and the program can assign the value in a variable at the ends of the records, usually on the housing record.

580. An alternate method would be to use a program that adds up the numbers of people in a set area and assign urban or rural depending on size of the area and perhaps certain housing assets, like air conditioning or freezer.

B. INTERNATIONAL MIGRATION CHARACTERISTICS

581. The demographics of a country change over time as a result of natural increase (fertility and mortality) and net migration. Migration can be lifetime migration (since birth) or current migration, measured by previous residence and duration or at a previous, specified point in time. Since these items are often interrelated, a joint edit similar to the one described for the basic demographic variables might be appropriate for some countries. If the top-down approach is used, the order of the edits becomes important since certain items must be edited before others.

582. Migration items often require more detailed codes than other items since smaller geographical units may be necessary for planning and policy use. Detailed information on small areas may be needed for staff planning for a new school or health clinic. Also, different coding schemes and different edits may be needed for places inside and outside the country.

583. Traditionally, most countries did not experience very much international immigration, so emphasis was on internal migration. Internal migration continues to be of primary concern. However, in an increasingly globalized world, more and more emphasis are placed on international migration as well.

584. For within country migration (internal migration), data on within country place of birth and years living in the district should be checked for consistency, since obviously relationships exist between the two items. Additionally, some reasonable relationships exist between responses for various members of the household. For example, if no response appears for the number of years living in the district for a child, it can be imputed from the response for the mother, and the editing program will check that the value imputed does not exceed the child's age.

585. For international migration, country of birth, citizenship and year of entry into the country are of concern.

1. Country of birth (core topic)

586. Countries having little immigration have an easier time since most people will have been born in the country. When the householder's country of birth is unknown, usually because it was not recorded, then the program should look through the housing unit for another person with country of birth, and that country should be assigned to the householder. If no one in the housing unit has a country of birth, then dynamic imputation might be used to obtain the country of birth for the head of household or reference person. All others in the unit would get his/her country of birth.

587. When only country of birth is to be reported, this procedure should work when there is little immigration to the country. However, if more detail is collected, like major civil division, this procedure may be more difficult to apply. People move during their lifetimes, internal migration may make it difficult to assign a value when the original value is unknown.

588. Countries with immigration have a more difficult edit. If the country experiences little migration or if certain receiving areas are targeted, the procedure above can be used. When immigrants move to any part of the country, a risk occurs that the country assigned by dynamic imputation may not be the best choice. It that case, the category "unknown" might be used.

589. As with other variables, like occupation and industry, where "unknown" values are accepted, it is possible that a skewing of the data will occur in analysis. If the analysts assume that the distribution of the unknowns is the same as the distribution of the knowns, a skewing might occur. So, dynamic imputation, even with its possible faults, might be the better way to go.

2. Citizenship (core topic)

590. Information on citizenship should be collected to permit the classification of the population into three categories: (a) citizens by birth; (b) citizens by naturalization, whether by declaration, option, marriage or other means; and (c) foreigners. In addition, information on the country of citizenship of foreigners should be collected.

591. It is important to record the country of citizenship as such and not use an adjective to indicate citizenship, since some of those adjectives are the same as the ones used to designate ethnic group. Hence, French-Canadian could have several interpretations, and so should be avoided if possible.

592. The coding of information on country of citizenship should be done in sufficient detail to allow for the individual identification of all countries of citizenship that are represented among the foreign population in the country. For purposes of coding, it is recommended that countries should use the numerical coding system presented in *the online version of the United Nations publication "Standard Country or Area Codes for Statistical Use" originally published as Series M, No. 49 and now commonly referred to as the M49 standard.* (https://unstats.un.org/unsd/methodology/m49/). The use of standard codes for classification of the foreign population by country of citizenship will enhance the usefulness of such data and permit an international exchange of information on the foreign population among countries. If a country decides to combine countries of citizenship into broad groups, adoption of the standard regional and sub regional classifications identified in the above-mentioned publication is recommended (United Nations, 2017, Para. 4.112).

(a) Citizenship edits

593. Citizenship depends on each country's definitions. In most countries, but not all, persons born in the country are automatically citizens by birth. Hence, an edit should look at the relationship between birthplace and citizenship and may need to assign "citizens by birth" to persons born in the country. When electronic data collection is used, it is advisable to use the list of standard country codes so that the enumerator or the respondent can select a country name from the list. This will improve the data quality, especially in cases of changing geography over time. However, it is still suggested that the citizenship edit as well as the other migration variables should be done during office edit when all of the migration variables and other related variables are available for consideration.

(b) Relationship of ethnicity/race to citizenship

594. Some countries also collect "ethnicity" or "race" which may give additional information to be used in determining citizenship, particularly when the collected response is invalid. For many countries, first and second-

generation migrants should have almost complete consistency between their ethnic origin and their citizenship. For countries with a long history of international immigration, this characteristic may be less valuable, but still might be considered with other variables.

(c) Relationship of naturalization to citizenship

595. In countries where naturalization occurs, the requirements for naturalization may or may not be covered by the census items. If, for example, a residence period is required, an item on "duration of residence" could be used to test for fulfillment of the naturalization period. Then, if a person is born abroad and has an invalid or inconsistent response for citizenship, the editing teams may choose to assign "naturalized" for citizenship. Other persons who do not fulfill the duration of residence requirements for naturalization would be assigned as "foreign," using the cold deck method of imputation.

(d) Relationship of duration of residence to citizenship

596. The item "duration of residence" may not appear on the questionnaire or may be ambiguous in determining citizenship, or the editing team may choose not to use it. Then, if the value for citizenship is invalid or inconsistent with birthplace, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics (and one should probably be birthplace) to obtain "known" information from similar persons in the geographical area.

3. Acquisition of citizenship

597. Some countries, especially those with high levels of immigration, the method of acquisition of citizen is collected. When the information is unknown, the program would check to make sure about the actual citizenship.

598. Then, if the analysts determine that everyone in the household is likely to have acquired or not acquired citizenship based on length of time in the country or other variables, the standard edit could be applied.

599. So, if the head of household or reference person does not have acquisition reported, the other household members would be checked, and if the acquisition variable was filled, that would be assigned to the head of household or reference person. If no one in the household reported acquisition, the analysts might decide that they have not acquired citizenship because they were not in the country legally, and so should be coded not acquired. Then, dynamic imputation could be used to assign acquisition to the head of household or reference person based on the acquisition of a previous householder with the same country of origin and other characteristics like age and sex. Others in the housing unit would get what the head of household has.

600. When only some of the people in the housing unit could have acquired citizenship, the acquisition of the head of household could be used, if available, that is, the head of household also had a foreign birth and so an acquisition. Otherwise, the program would need to either use dynamic imputation or assign a value or make the information "unknown."

4. Year or period of arrival (core topic)

601. The Principles and Recommendations divide migration variables into internal migration and international migration. As noted in the edit for duration of residence refers to the length of time living in the particular designated "Area". The year of arrival normally refers to the year of arrival from a place outside the country into the country. Therefore, year of arrival is an item usually asked with its complement, that is, place of residence before arrival in this country (United Nations, 2017, para.4.117 to 4.120).

(a) Relation of age to year of arrival

602. The first part of the edit should check for consistency between age and place of birth and for a valid entry in year of arrival in the locality or civil division. The number of years a person has lived since arrival in a locality or civil division cannot be greater than the person's age. In addition, a person who was born outside of the country cannot have always lived in the locality or civil division. The program should assign "always" to year of arrival in locality or civil division, if years in locality or civil division is greater than age and country of birth is this country. If years since arrival the in locality or civil division is greater but country of birth is not this country, one method would be to assign the person's age as years in locality or civil division. In that case, it is assumed that although born outside of this country the person moved into the locality or civil division when he/she was less than 1 year of age. Edit on entry with electronic data collection can provide an error message when respondents report arriving before they were born; the enumerator or respondent can then correct the inconsistency immediately. Otherwise, this edit is better performed during office edit because of the various relationships described in the next paragraphs.

603. To assist users of public use samples, statistical offices should provide codes for "less than one" and "always" in this item. The "always" code should normally actually be the place of current residence to assist in making tables directly. The "less than one" code will allow users to be certain that they have looked at everyone in the population in their cross-tabulation.

(b) Relation of birthplace to year of arrival

604. In the case of out-of-range entries, the same tests as those for place of birth should be used. A search should be made for related previous persons (mother, head of household, husband, child). Imputation should be based on the information found. However, before a value is assigned it must be consistent with the age and place of birth of the person whose record is being edited.

(c) For persons who have always lived here

605. If the response for the number of years since arrival for a person who lived in the locality or civil division indicates "always lived here", but the country of birth is not "this country", the editing team might want to use the person's age to assign the year of arrival in the locality or civil division. The specialists will assume that although born outside of this country the person moved into the locality or civil division when less than 1 year of age. The next part of the edit will check for a valid entry in year or arrival in locality or civil division. Since the length of time a person lived in the locality or civil division cannot be greater than the person's age, age will be assigned to the years in locality or civil division for this situation.

(d) Person's year of arrival from mother's year of arrival

606. If the category does not have a valid code, the program can perform an inter-record check by searching for the person's mother in the household. If found, the mothers record can provide information helpful in assigning missing values. If the person's mother has always lived in the locality or civil division, and her country of birth is "this country" (as it should be), the program will assign "always" to this person's years in locality or civil division category. If the mother's country of birth is not "this country", even though the entry for her years in the locality or civil division is "always", this indicates that something is wrong with the mother's categories. The program will then ignore the mother's country of birth and assign age based on year of arrival in locality or civil division. If the entry of the mother's arrival year in the locality or civil division is not "always", but is a valid code, and the person's age is less than the number of years since the mother arrived in the locality or civil division, the edit will go back and check the mother's country of birth. If the mother's country of birth is "this country," the program will assign this person's age to years in locality or civil division. However, if a person's age is equal to or greater than mother's years

since arrival in locality or civil division, the program will assign "mother's arrival year in locality or civil division" to this person's arrival year in locality or civil division.

(e) Child's year of arrival from head's year of arrival

607. If the person in question is a child (son or daughter), the editing program should check the head of household's record for possible information to aid in assigning values for missing data on year of arrival. When the head of household was born in "this country" and has always lived in this locality or civil division, the program will assign "always" to the child's years in locality or civil division. When the head of household has always lived in the locality or civil division, but was not born in "this country," the child's age will be assigned to locality or civil division. When the head of household's entry for year of arrival in the locality or civil division is not "always", but is a valid code, this information can be used if it is consistent with the age in the record of the child being edited. If the child's age is equal to or greater than that determined by the year of arrival in the locality or civil division of the head of household, the program will use the head of household's arrival year in locality or civil division as the year of arrival in the locality or civil division of the son or daughter. If the child's age is less than the head of household's arrival year in locality or civil division, the program will assign a value depending on the country of birth of the head of household. This value will be "always" if the head of household was born in "this country"; if not, the program will assign the son's or daughter's age to years in locality or civil division.

(h) Person's year of arrival when no other information available

608. When all of the above efforts fail to produce a valid value, the program can assign "not reported" or "unknown" to arrival year in the locality or civil division for this person. If the value is still invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use an appropriate number of characteristics to obtain "known" information from similar persons in the geographical area.

C. HOUSEHOLD AND FAMILY CHARACTERISTICS

1. Relationship (core topic)

609. The relationship item is used to assist in determining household and family structure. It appears near the beginning of most census and survey questionnaires and assists in making sure everyone in the housing unit is counted. The enumerator and the respondent use the information about the relationships among the household members to make sure no one is missed. The relationship item also assists in checking for consistency for sex and age among household members. Determination of one sole head of household or reference person and no more than one spouse (in non-polygamous societies) is covered in the structure edits.

(a) Relationship edits

610. Since statistics on relationship are becoming more important, some care should be taken in developing edits that allow for family and subfamily formation for various types of tabulations. Developing appropriate relationship codes in the first place will obviously assist in this endeavor (see Annex II - Derived Topics for "family type", and subfamily number and subfamily relationship recodes).

611. When relationship cannot be assigned, and dynamic imputation is not used, "unknown" must be assigned for invalid or inconsistent responses. With the use of dynamic imputation, relationship may be allocated from an imputation matrix by age and sex, or other appropriate characteristics. The imputation matrices should not impute relationships that would conflict with already established relationships within the household. For example, second

and third spouses should not be imputed, even in polygamous households, unless the editing group decides to implement such an edit.

612. In most countries, the overwhelming number of households will be of related persons only, and so "other relative" could be assigned when all else fails. "Not reported" have the same problems as for all other variables. If you are trying to get percentages of heads, spouses, children, etc., you must do something with those not reported – usually these are dropped altogether, and percentages based on the knowns are used, so imputing after the fact.

613. Detailed relationships are important in developing edits for derived variables showing family and household composition especially for household structure under stress situation. For example, describing missing generation households (grandparents and grandchildren only) could be used to assess the social and economic impact of the epidemic or refugees or other migrant situations, to assist governments in planning. See Annex II for suggested derived variables for these structures.

(b) When the head or the reference person must appear first

614. If the head or the reference person does not appear as the first person, the structure edits introduced in chapter III indicate that a pointer can be used to keep track of the head's position. If the editing team wants the head to be the first person, the head can be placed in the first position either by rearranging the order of the persons or by leaving the household in place but rearranging the relationships, as noted in the chapter on structure edits. The former method requires considerable programming expertise, the latter method may do damage to the dataset if extreme care is not taken. It should be noted that when electronic questionnaire is used, the reference person or head of household appears first in the household listing. Data collection application will now allow to enter more than one person as head or reference person.

615. As noted elsewhere, contemporary computer packages now allow the head to be easily switched into the first position. However, this does not mean that the other relationships to head will be unaffected. They might have been related to the original person in the first person, rather than the new head and so could be wrong. An edit will change the relationships appropriately.

616. However, in cases where a couple considers themselves co-heads, this determination may be more difficult. The risk is that the respondents will refuse to continue with the enumeration if they are forced to choose just one as head. In this case, it is probably best to accept what they say -- that the unit has two heads -- and go on to the other items, allowing to let the subsequent office computer edit select the person to be used as head of household. Because some tabulations require age and sex of the head/reference person, or ethnicity or language or religion, conventions expect one and only one head.

617. When the electronic data collection application finds no head of household/reference person after keying the individuals, a message, "There is no head" should appear, so that the enumerator and respondent can resolve the issue. The respondent should select one adult to be listed as head. The other persons should then be recorded in their relationship to that head, even if they had previous relationships recorded. Usually the first person will be the head, and that will be the default in most cases.

(c) When the relationships are coded upside down

618. Sometimes enumerators collect the relationships "upside down": rather than collecting the relationship of each person in the household to the head, they collect the relationship of the head to each person. Hence, the relationship of the third person as "parent" rather than "child". The household may end up with four parents instead of four children. When the editing team finds a systematic problem of this sort, it must develop a solution that does not do too much damage to the household.

619. One procedure for inverting the relationships involves using a small "look up" file containing the original relationships and the inverted relationships, taking sex into account.

620. When electronic data collection is used, a message should be provided by the data collection application for warning about mistakes in the relationships, looking at the age difference between the head/reference person and the grandfather. Edits on entry should make an attempt to check the age difference between father/mother and son/daughter when it is appropriate. For example, after the section of the list of members of the household containing usually data on age, sex, relationship and marital status.

(d) When polygamous spouses are present

621. The structure edits, if performed as indicated in chapter IV, will have already checked for "one and only one" head/reference person and "no more than one spouse" for monogamous households. For polygamous households, the editing team should decide when polygamous relationships are permitted and when they are not. Sometimes households that seem to have polygamous relationships are actually mistakes.

622. In most cases, polygamous relationships are rarely real – detailed analysis of some censuses shows that usually these are mis-captures or mis-codes. Usually, polygamous spouses head up or live in separate households. If the census questionnaire is set up to be able to link the polygamous unions and families, then these can be edited and analyzed.

623. For example, a household might have a head and spouse identified, but another couple reported as "spouses" to each other, making three spouses in all. The edit should check to make certain that the second couple is not actually father and mother, son and daughter-in-law, sister and brother-in-law or some other combination. Sometimes these relationships can be determined with some certainty, and sometimes they cannot. When the above detailed relationships are coded, the editing team should expect to see appropriate imputations. When the additional spouses are actual spouses, in polygamous households, the edit should check for sex and, perhaps, age.

624. For electronic data collection, the entry program should take into account whether polygamous marriages are possible. The entry program can be developed to allow polygamous marriages for some groups (for example, by religion or ethnicity) or allow for all polygamous unions or no polygamous unions. Depending on the decision, if the relationships are violated, a message "You have keyed in multiple spouses" should appear, and a resolution obtained, if possible.

625. However, if multiple spouses are not permitted, then either a program in the entry program should use information like age and sex and placement on the listing form to determine the most appropriate spouse. The other spouses would be made "other relative" if no further communication occurs between the enumerator and the respondent or can be adjusted when the enumerator and respondent determine the appropriate responses.

626. The electronic data collection application can be programmed to look for same sex marriages. More and more countries allow same sex marriage, so then entry program can take this into account. But, in the cases of multiple spouses, a resolution is still needed, and can be either programmed or the enumerator and respondents can work out the appropriate relationships.

(e) When multiple parents appear

627. Households should have no more than two "parents" reported, and the parents should be of opposite sex. When more than two parents appear, the additional parents should probably be made "other relative". Sometimes

censuses or surveys have a code for "parent" or "parent-in-law" which would allow for up to four "parents" rather than two, with no more than two parents of each sex.

628. With electronic data collection, the entry program can check for multiple parents, and, in fact, should be able to distinguish between parents of head and parents of spouse when separate codes are used for these. Hence, when more than two or two of the same sex are encountered, a message should appear for noting the problem to the enumerator or respondent to make necessary correction. If the problem cannot be resolved during the enumeration, it might be better to deal with this problem in the office edits. In countries with same sex marriage, both parents could be of the same sex, so this would need to be taken into account in the programming.

(f) When censuses collect sex-specific relationships

629. Some censuses or surveys collect sex-specific relationships: "husband" and "wife" separately, instead of "spouse"; "son" and "daughter", instead of "child"; and so forth. If these responses are not edited, tabulations may contain data with "male" daughters or "female" husbands. The editing team must decide on the priority of the edits—whether relationship or sex takes precedence. In some cases, such as husband and wife, the edit is more important than for others, such as a young child. Note that it may not be a good idea to use sex-specific relationships since redundancy does not clarify the relationships and require additional editing.

630. When electronic data collection is used, resolutions of these problems can be done onsite. For this edit, the entry program can probably be constructed to check for this as keyed in or at the end of the record. Then, the program can check to see whether fertility has been keyed, for example, and the correction can be made.

(g) When relationship and marital status do not match

631. Relationship and marital status should agree. Persons who report the relationship "spouse" should be "married" in the marital status item. The editing team makes choices about which variable to change when the items do not agree. Sometimes, relationships are ambiguous, so care must be taken in developing editing specifications. For example, in many countries, a brother-in-law could be either the brother of a spouse (and would not have to be married) as well as spouse of a sibling (and would have to be married).

632. Several other, more contemporary problems in relationship reporting currently appear. When two unmarried persons of the opposite sex live together outside of marriage, the relationship code might be "unmarried partner" or it might be "spouse". If the census or survey has a code for unmarried partner, then the appropriate marital status should not be "married" unless the person is married to someone other than the person with whom they live.

633. Similarly, persons of the same sex now live together either in romantic or non-romantic relationships. Persons in a non-romantic relationship might be coded as "roommate" or "nonrelative". For those in romantic relationships, the category "unmarried partner" might be appropriate for some countries. Then, the editing team must also decide on the appropriate corresponding marital status. Censuses cannot distinguish between platonic and romantic relationships.

634. As more countries sanction same sex marriages, census questionnaires may try to provide methods of determining this situation. Deleting same sex couples could do damage to the data sets; accepting all of them would also do damage.

635. For electronic data collection, after keying in the housing unit, the entry program can check to make sure that the relationships and martial statuses agree. Most countries do not use either of these variables in planning and policy development, so the resolution of inconsistencies can wait until the office computer edit.

2. Household and family composition (core topic)

636. Household and family composition is determined after all of the population items are edited. The relevant items for determination of households and families are relationship, sex and age. Families contain at least one other person related to the householder, but a household does not have to contain persons related to the head and can be a single person (United Nations, 2017, paras 4.121-4.128). A discussion of this item appears in the Derived Variables in Annex II.

3. Household and Family status (core topic)

637. Similarly, the household and family statuses are obtained without additional editing. Determining types of households and family status using the relationship to the reference person or head of household are discussed in the Principles and Recommendations for Population and Housing Censuses, Revision 3 (United Nations, 2017, paras 4.140-4.148) A discussion of these appears in Annex III.

D. DEMOGRAPHIC CHARACTERISTICS

638. Sex and age are considered to be the most basic of all demographic variables. Of all the topics included in population censuses, sex and age are more frequently cross-classified with other characteristics of the population than any other topics. Apart from the importance of the sex-age structure of the population in itself, accurate information on the two topics is essential for most census tabulations. A very important use of census data on the sex and age composition of the population is the evaluation of the data especially with respect to coverage. The variables are therefore very crucial, and it is important that this information be reported in respect of every person for whom census information is been collected. It is therefore recommended that where this information is incomplete it should be imputed for census purposes rather than being reported as 'not stated' (United Nations, 2017, para. 4.149).

639. Data on sex, age, relationship and marital status for each person are basic to any census and should probably be edited together.

640. The multiple-variable approach to editing population and housing data was introduced in chapter II of this Handbook. Since the demographic variables are integral to all census planning, this approach should be used if time and expertise permit. The quality of the overall dataset is almost certain to benefit from a priority edit looking at age and sex and other selected variables to determine errors or inconsistencies. The items most in error are edited first, followed by those items less in error or inconsistent.

1. Sex (core topic)

641. Sex is one of the easiest characteristics to collect but requires some thought in its editing. It is among the most important variables since most population characteristics are analyzed based on sex. Sex imputation requires some comparison with other variables. In some cases, sex should be based on differences between the sexes of related persons, usually the head of household and spouse, but also between parents and in-laws. Sex should probably not be left as "invalid" or "unknown" since it is such an important variable. Hence, some thought should be put into how best to obtain results comparable to a country's real situation. As example, Annex IV contains a flowchart for an edit for sex of head and spouse.

642. If a person is not the head of household or the spouse of the head, no other persons exist to refer to; therefore, other items within the person's record should be checked. If sufficient fertility items occur, the code for female should be assigned. However, if this person's sex is missing or invalid, for example, but a spouse exists, for whom sex is indicated, the edit can impute opposite sex to this person.

(a) When the sex code is valid, but the head and spouse are the same sex

643. In instances where contradictory evidence seems strong the code for sex should be changed even though a valid code exists. For example, the record shows that a second married couple is present when the household already has a head of household and spouse or married couple in a subfamily. If both persons in the second couple report the same sex (particularly in countries not sanctioning same sex marriage), information about fertility and other items can be used to determine which is the male and which the female. Then, the erroneous person record can be changed.

644. As noted previously, when same sex marriages are allowed, unless the entry program checks for other variables, like incompatible ages, usually the responses would be accepted, and the sexes remain the same. If one of the pairs clearly has fertility information, then that person should be made female. If both have fertility information, then both should be female when same sex marriages are accepted.

(b) When a male has fertility information, or an adult female does not

645. The edit may detect a male with fertility information and/or children in the house, an error that can be based on the mother's person number or a similar variable. If no spouse is present, the sex may be changed to female rather than deleting the fertility information. Similarly, an adult female with no fertility information and without accompanying children may be changed to male under certain circumstances determined by the editing team.

646. Some enumerators or supervisors in various countries become confused and collect fertility from the males and not the females. In this case, an edit must look for adjacent couples and move the fertility from the male to the female. Otherwise, fertility will be deleted from the male and imputed for the female from some other female.

647. In electronic data collection method, a message should be displayed for males with fertility or females without fertility as collected. In most cases, the skip patterns would be set up so that the enumerator could not enter fertility for males, and configurations would expect entry for fertility for females. Only when the sex is entered as male, and the fertility is skipped even though it is there, and the enumerator has to go backwards to enter female and then fertility, could there be a holdup.

(c) When the sex code is invalid, and a spouse is present

648. If the entry for sex is blank or invalid, the editing program should use the entries for relationship to head of household and sex of spouse, if the sex of the spouse is valid, to determine the correct code. If the relationship to head of household or reference person is "head of household/reference person", the program then checks to see whether a spouse is present (by checking for another person in the household whose relationship is spouse). By determining the sex code of the spouse, the opposite sex code is assigned to the head of household. With electronic data collection, this will be straightforward, and should be done automatically during data collection.

(d) When the sex code for spouse is invalid

649. If the relationship of the person to the head of household or reference person is "spouse", and the sex of the head of household/reference person is given, the program assigns to this person the sex opposite that of the head of household.

(e) When the sex code is invalid and female information is present

650. Numerous clues in the questionnaire indicate whether a respondent is female. If the program has not yet determined the person's sex and any female indicators are present, then the record for this person should be assigned

female sex. For example, if the person being edited includes sufficient fertility items, then sex can be assigned as female. The fertility items include children ever born, children living in this household, children living elsewhere, children dead and children born alive in the last 12 months. Another possibility is that this person could be the mother of someone else in the household, so that this person's line number equals the line number of the mother of another person in the household. With electronic data collection method, this problem can be corrected automatically.

(f) When the sex code is invalid, and this person is spouse's husband

651. If the person is the husband of someone else in the household, based on an item showing the husband's line number, the entry for sex should be assigned male.

(g) When the sex code is invalid and there is insufficient information to determine sex

652. If the editing team does not use dynamic imputation at all, a value for unknown sex must be assigned. Unfortunately, this means that all tabulations would have to carry an extra column or an extra row or sets of columns or rows for persons of unknown sex. Sex values could be assigned stochastically with a certain probability.

(h) Note on imputed sex ratios

653. It is important to note that sex is more likely to be reported than fertility information. Nonetheless, female sex is likely to be assigned more often than males. Adult females are the only ones with fertility entries and their selection is skewed somewhat from random. For this reason, if insufficient information is available, a person with no information is more likely to be male than female. Consequently, it is important to consider developing imputation matrices that account for the overall proportions between the sexes.

2. Birth date and age (core topic)

654. Age is one of the most difficult characteristics to collect and to edit. However, it is probably the most important variable since virtually all population characteristics are analyzed based on age. Editing of age requires extensive comparison with other variables and other people in the house. In most cases, the imputed age should be based on stored differences between the ages of related persons. If age cannot be imputed on this basis, then other characteristics within the person's record should be used. The edit should probably require a series of imputation matrices, including age by sex, marital status, relationship and school attendance; age difference between mother and child; age difference between husband and wife; and age difference between head of household and spouse.

655. Because age is so important in the edit and later in tabulations, extreme care is needed to obtain the best entry. Edits on entry with electronic data collection technologies should be used to assure the relationship between age and date of birth. The edit on entry could also look at difference in ages of parents and children for checking possible entry errors. Other items which have a relationship with age, like school attendance, economic activity, and fertility should also be checked to ensure that these questions are asked to eligible persons. However, for some other variables that have a relationship with age, such as educational attainment by level of education and retirement, it would be best to do the full edit for age during office edit.

(a) Age and date of birth

656. The structure edit calculates age from date of birth. First, however, it is useful to review the difference between age and birth date. As stated in Principles and Recommendations (United Nations, 2017, para. 4.151 to 4.162), information on age may be secured either by obtaining the date (year, month and day) of birth or by asking directly for age at the person's last birthday.

657. In recent years, most countries collect information on date of birth in addition to age. The date of birth yields more precise information and should be used whenever circumstances permit. If neither the exact day nor even the month of birth is known, an indication of the season of the year might be substituted in appropriate for the country. However, clearly month and day of birth provide more accuracy than season. The question on date of birth is appropriate when people know their birth date, which may be established in accordance with the solar calendar or a lunar calendar or expressed in years numbered or identified in traditional folk culture by names within a regular cycle.

658. It is extremely important, however, that a clear understanding should exist between the enumerator and the respondent about which calendar system the date of birth is based on. If there is a possibility that some respondents will reply with reference to a calendar system different than that of other respondents, provision must be made in the questionnaire for noting the calendar system that was used. It is not advisable for the enumerator to attempt to convert the date from one system to another. The needed conversion can be best carried out as part of the data editing work (United Nations, 2017, para. 4.153).

659. The direct question on age is likely to yield less accurate responses for a number of reasons. Even if all responses are based on the same method of reckoning age, the respondent may not understand whether the age wanted is that at the last birthday, the next birthday or the nearest birthday. In addition, other problems can occur: age may be rounded to the nearest number ending in zero or five; estimates may not be identified as such, and deliberate misstatements can be made with comparative ease (United Nations, 2017, para. 4.155).

660. Many national census/statistical offices collect either date of birth or age, but not both. As noted in the Principles and Recommendations (United Nations, 2017, para. 4.151), age in completed years is very important: it is used for many of the edits and as a dimension for many of the imputation matrices. More importantly, many country policies are based on age, so every effort must be made to obtain the best quality age reporting. However, even in ideal situations, some ages will not be reported. Hence, efforts must be made to ensure that age is computed properly and is consistent with other responses for individual members of the household.

661. For some types of analysis exact date of birth is helpful. For example, if a country wants to compare births in the year before the census with births reported in vital status, the exact age of the mother at the birth of the last child can be determined through subtraction – the reference data of the census and the date of mother's birth. Those provide the denominators. The numerators would come from the exact age of the last birth at the time of the census if the full date of last birth is collected in enumeration.

(b) Relationship between date of birth and age

662. During the structure edit, age should be calculated if it was not collected separately from date of birth. The age edit during the individual edits will be a thorough test of consistency within and between records, but a first step is calculating the age from the date of birth and the census date. It is important to test the age as calculated based on date of birth to make certain it falls within the bounds of the census date.

663. This correspondence will be most important when electronic questionnaires are used for data collection. A message should be displayed if the keyed in age and birthday don't agree so the enumerator or the respondent can fix it on the spot. Hence, this relationship should be determined, if possible, at entry, and not in a later structure edit. The relationship between date of birth and age can be part of the entry program, so that the computed age is done automatically and checked against the reference date.

664. The age of children born during the census year but after the census date will be calculated as –1 and must be rectified. Babies enumerated after the census date must be dropped from the census. However, if after examination the date of birth is found to be erroneous because of enumeration or processing, other variables should be used to obtain a better age estimate.

665. Unless it is within a month of the reference date, it is likely that either the report is a year off or something similar. In some censuses, based on analysis of births and last births, it is clear that more often the problem is scanning or keying or misreporting than that the child was born after the census date.

(c)When calculated age falls above the upper limit

666. Censuses now record all four digits for year of birth. For those around 2020, the acceptable range will be in the 1900s or 2000s up to the census year. While three digits are enough for the computer to do its work, the use of three-digit years might confuse both enumerators and office workers. Sometimes the calculated age will fall above the upper bound of the census-defined ages and will need to be adjusted. If the census is in 2020, and a person reports being born in 1890, the computed age of 130 years is likely to be outside the acceptable range and will need to be changed. Some birthdates will be way out of range, so depending on the age upper limit, those too old would need to be changed.

(d) Age edit

667. The editing program should check the consistency of the reported age of the person with the reported age of the person's mother, father or child. The edit should provide for a minimum difference in years between the age of the mother or father and the age of the child. When the age is imputed, consistency checks should be made with entries such as years lived in the district (duration of residence) and highest grade of school completed (level of educational attainment). All such checks should be made before the age is changed or before an imputed age is assigned.

668. With electronic data collection, the check in difference of age between the head and spouse with the children and the parents with the head, can be done during enumeration. As this type of consistency control is performed based on multiple cases, it is better a approach to do edits after entering all data in the questionnaire. After completing the interview, the data collection application may produce a list of inconsistencies in age difference between parents and children. When the difference is too small, or potentially too great, like a female entered as 70 years older than her child, a message can appear. The editing team may decide to resolve this kind of issue in the field or wait until the office computer edit. In most cases, the ages of the individuals are very important, but the difference in ages is relatively less important. It should be noted that this kind of inconsistencies might be real cases for those who are adopted or are foster children.

669. The edit should begin with a check for validity. If the age is valid, specialists might want to check to see whether this person's age is consistent with his/her mother's age (if the person's mother is found in the household) and with the age of this person's children (if this person is a woman and has children in the household). If the ages are inconsistent, this person's age should be noted, and the age should be changed later.

(e) Age edit when the head of household and spouse are present

670. The next step in the edit is to determine whether a spouse is present. If so, the spouse's age should be checked for validity (at least X years old, depending on the country's defined minimum age at marriage). If age is inconsistent, and if dynamic imputation is used, the program will now use a special imputation value derived from the difference between the age of the husband and the age of the wife. Age differences vary less than the ages themselves, so an imputation matrix in the program will store the difference in age (from previous records) of a husband and wife. This value is added to or subtracted from the age of the spouse of this person to form a computed age.

671. To ensure that this computed age is consistent with other characteristics, the imputation matrix should also include marital status, duration of residence and highest grade of school completed. Exclusion of those variables can

result in a computed age that is less than the number of years the person has lived in the place, or less than the level of schooling implies. For example, the imputation matrix may give an age of 8, but the person may have recorded that they lived in the place for 10 years. Without the other variables, when the editing program carries out the years-in-place edit, another imputation matrix will change the years in residence from a correct value to an incorrect value.

(f) Age edit for head or reference person when his/her spouse is absent, but child is present

672. When comparison with the age of the spouse is not possible in determining the age of the head or the reference person of household, the program can then check relationship. If the relationship is "head or reference person of household", the editing program can check the other records of the household (if any) for a son or daughter having an age that is known to be correct. The program checks the son's or daughter's age and computes an age for this person using an "age difference" dynamic imputation similar to the technique described above for husband and wife. As before, the computed age takes duration of residence and highest level of educational attainment into account. The completed age will then be consistent with these variables and will avoid obvious errors by including the years lived in the district and the highest grade of school completed as part of the imputation matrix.

(g) Age edit for head when head's parent is present

673. When a person does not fall into one of the categories described above, the program can search for the person's parent in the household. If the person's parent is found, an age can be computed with an imputation matrix using the difference in age. The difference in age between child and parent generally varies much more than that between husband and wife. For this reason, the program applies this edit only after the husband/wife age difference technique fails. The computed age should take into account the educational characteristics, the highest grade of school completed, and the years lived in the district, marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or reports economic activity of any kind.

(h) Age edit for head when head's grandchild is present

674. When a person does not fall into one of the categories described above, the program can search for the person's grandchild in the household. If the person's grandchild is found, an age can be computed with an imputation matrix using the difference in age. The difference in age between the head of household and the grandchild varies much more than that between husband and wife, or between head and child.

675. For this reason, the program applies this edit only after the edit for husband/wife and head/child age difference fails. The computed age should take into account the educational characteristics, including the highest grade of school completed, including the years lived in the district, marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or participates in economic activity of any kind.

(i) Age edit for head when no other ages are available

676. When a person does not fall into one of the categories described above, the program can search for another relative or a nonrelative of the head. If such a person is found, and that person has a reported age, the editing team must decide whether to use whatever information is available with an imputation matrix using the difference in age. However, these differences in age between the head and other relatives or nonrelatives vary so much that the editing team may decide to abandon the effort altogether and simply to use other variables for the dynamic imputation of the head of household's age. In any case, the program applies this edit only after the husband/wife, head/child, head/parent and head/grandchild age difference techniques fail. However, the computed age is determined, it should take into account the educational characteristics, including the highest grade of school completed, as well as the years

lived in the district, marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or participation in economic activity of any kind.

(j) Age edit for spouse when head's age already determined

677. The age edit for spouse is usually performed at the same time as the age edit for the head of household, since information from both persons is needed for the joint edit. If, however, the edit is separate, when the spouse's age is invalid or inconsistent with other variables, a dynamic imputation matrix using the age difference with the head and other variables should be used to determine the best estimate for the spouse's age. As before, the computed age should take into account the educational characteristics, including the highest grade of school completed, and the years lived in the district, marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or participation in economic activity of any kind.

(k) Age edit for other married couples in the household when the age of one of the persons is known

678. The edit should first determine whether this record is that of a married person. If so, the program can search among the other records of the household for the person's spouse. If no spouse is found, the program goes to the next part of the edit. If a spouse is found, the spouse's age should be checked for validity (at least X years old, depending on the country's defined minimum age at marriage). If age is inconsistent and if dynamic imputation is used, the program will now use a special imputation value derived from the difference between the age of the husband and the age of the wife. Age differences vary less than the ages themselves, so an imputation matrix in the program would store the difference in ages (from previous records) of a husband and wife. This value is added to or subtracted from the age of the spouse of this person to form a computed age.

679. To ensure that this computed age is consistent with other characteristics, the imputation matrix should also include marital status, duration of residence and highest level of educational attainment. Exclusion of those variables could result in a computed age that is less than the number of years the person has lived in the place, or less than the level of schooling implies.

(l) Age edit for child when head/reference person's age already determined

680. If this is a son or daughter of the head or reference person of household, a computed age can be derived using the head of household's age, the age difference, the duration of residence, and the level of educational attainment. Again, the computed age should take into account the educational characteristics, including the highest grade of school completed, years lived in the district, and the marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or participates in economic activity of any kind.

(m) Age edit for parent when head/reference person's age already determined

681. If this is a parent of the head or reference person of household, a computed age can be derived using the head of household's age, the age difference, duration of residence and level of educational attainment. The computed age should take into account the educational characteristics, including the highest grade of school completed, and the years lived in the district, marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or participates in economic activity of any kind.

(n) Age edit for grandchild when head/reference person's age already determined

682. If this is a grandchild of the head or reference person of household, a computer age can be derived using the head of household's age, the age difference, duration of residence and educational attainment. Again, the computed age should take into account the educational characteristics, including highest grade of school completed, and years lived in the district, marital status, fertility and economic activity. The program should presume that a person is at least 12 years old if he/she has ever married, has children, or participates in economic activity of any kind.

(o) Age edits for all other persons

683. The editing team should determine appropriate imputation matrices for other related and nonrelated persons in the household. Guidelines will depend on the particular census or survey and the country's social and economic characteristics. For example, a person who has ever been married, has ever had children or participated in economic activity is likely to be at least as old as some country-defined minimum age. Based on that information, if dynamic imputation is used, the value received from the imputation matrix should not be below the minimum age. Similarly, if a person attends school, has any schooling or can read and write, but is not head of household, has never been married and has no economic activity, then this person should be placed in a group whose age is less than the minimum age for adults but greater than or equal to the minimum age to attend school. The imputation matrix value can then be found for those with less than the minimum age for school. Although not perfect, this technique limits the range of values that the imputation matrix can take.

3. Marital status (core topic)

684. In Principles and Recommendations (United Nations, 2017, paras. 4.164), marital status is defined as the personal status of each individual in relation to the marriage laws or customs of the country. The categories of marital status to be identified include, but are not limited to, the following: (a) single, (never married); (b) married; (c) married, but separated; (d) widowed and not remarried; and (e) divorced and not remarried. In some countries, it will be necessary to take into account customary unions, such as registered partnership and consensual unions, which are legal and binding under customary law. In countries with legal provision for registered or legal partnership (for opposite-sex couples or same-sex couples), or where same-sex couples can legally marry, subcategories should be included in the category of (b) married or in a legally registered partnership, namely (b)(i) opposite-sex marriage/partnership, (b)(ii) same-sex marriage/partnership. These sub-categories are crucial for developing an edit separately for same-sex and opposite-sex marriage/partnership.

(a) Marital status edits

685. The editing team must decide on the appropriate minimum age at first marriage for the census or survey. Minimum age at first marriage (some age X) may differ for different areas of a country or different ethnic group within the country for each sex. If, for example, the rural population marries earlier than the urban population, the editing rules should include this fact. Normally the national census/statistical office determines the age at earliest marriage before enumeration, so that only persons above the determined age are asked the question. Younger persons fall into the "never married" category automatically.

686. If everyone is asked the marital status item, however, the editing team must develop an edit for the whole population. This item refers to the current and absolute marital status, so the minimum age is used to distinguish those who must be "never married" from those who are "ever married". The edit on entry can test for the relationship between age and marital status if age has already been edited. Also, the edit on entry can either assign "unknown" when dynamic imputation is not used, or a valid value when dynamic entry is used, if the hot deck is developed and in place, embedded in the program.

(b) Marital status assignment when dynamic imputation is not used

687. Although marital status should be tabulated only for persons aged X years and older, where X is the earliest age at first marriage, editing teams must determine whether and how much to edit. If the country uses only "not stated" or "unknown" for invalid or inconsistent responses, then when invalid or inconsistent entries are found, the code for "not stated" should replace the inappropriate response. If, for persons under age X, the response "never married" is missing, it should be imputed; since statistical offices release samples of data to the public, it is important that items like marital status always have entries.

(c) Marital status assignment when dynamic imputation is used

688. If dynamic imputation is used, the edit for marital status should (a) impute a value when an entry is out of range and (b) check for consistency between reported marital status and relationship and age.

(d) Spouse should be married

689. All persons coded "spouse" in the relationship category should be coded as married.

(e) Spouse of a married couple pair

690. If the line number of person A's spouse (person B) is a variable, then person B should have person A given as the spouse; in addition, A and B should both be married and of the opposite sex. The edit will differ, of course, in countries that allow same-sex marriage.

(o) If spouse, head/reference person should be married

691. If no entry appears for marital status, but the entry for relationship to head/reference person of household is "head/reference person", the program should check to see whether the spouse is present (by checking relationship for other members of the household). If the spouse is present, the program assigns the marital status for the head/reference person of household as "married".

(p) Head/reference person, no spouse, without children

692. If the spouse is not present, and this person is male with children present, the program imputes marital status by age with children present. If no children are present, the program might impute marital status by age with no children present. A male who is head/reference person of household, but whose wife is not in the household, is most likely to be divorced, separated or widowed.

(q) If all else fails, impute

693. For persons with out-of-range codes who cannot be assigned a code based on the above tests, age should be checked next. If age has a valid entry of less than age X, "never married" should be assigned. In all other cases, an entry should be assigned using an imputation matrix. The imputation matrix should be set up by sex and age (two-dimensional); by sex, age and relationship (three-dimensional); or by sex, age, relationship and number of children ever born (four-dimensional). Again, the editing teams should have determined the order of the edit, so in developing the imputation matrices, it is important to remember which items have been edited and which have not been edited. If only sex and relationship have been edited before marital status, the imputation matrix must allow for "not reported" in the other items.

(r) Relationship of age to marital status

694. For all persons reporting a valid marital status other than "never married", a consistency check with age should be made. All ever-married persons must be X years of age or older, where X is the country-specific minimum age allowed for a person to be ever married. If age is less than X or blank, further consistency checks should be made based on other relevant variables (such as number of children ever born or economic activity). If the entries for these items are not valid, "never married" should be assigned to marital status; in all other cases marital status should not be changed.

4. Ethnocultural characteristics

695. Besides the variable on ethnic origin, some countries also collect information on ethnocultural characteristics (2017, para. 4.172). These characteristics will vary from country to country, and even within a country.

696. The edit, however, is basically the same edit as for ethnicity or religion. First, if the householder does not have a value for the ethnocultural characteristic studied, the computer will look for someone in the housing unit with a value for that characteristic. If no one in the unit has a value for the item, the item should probably be imputed from another householder with the same ethnicity and age and sex. All others in the unit will then receive the householder's ethnocultural characteristic.

5. Religion

697. For census purposes, religion may be defined as either (a) religious or spiritual belief of preference, regardless of whether or not this belief is represented by an organized group, or (b) affiliation with an organized group having specific religious or spiritual tenets. Each country that investigates religion in its census should use the definition most appropriate to its needs and should set forth, in the census publication, the definition that has been used (United Nations, 2017, para. 4.175).

(a) Religion edits

698. Religion is one of the variables fitting the standard edit examples introduced in chapter II. For the religion item, unlike others, a "nonresponse" is significant and needs accounting; some people may be reluctant to declare their religion. A valid value (including "no response) is obtained for an individual, either directly from another household member, if a value is available, or from another head of household with similar characteristics. Editing team should determine the logical editing scheme used for the other social variables. The head/reference person of household should be designated and edited first, whether or not he or she is the first person in the unit. If a person with an invalid or unknown religion is the head of household, the following steps should be taken:

i. No religion for head of household, but religion present for someone else in the unit

699. The first step is to determine if anyone else in the housing unit has a valid religion to assign the first valid religion.

ii. No religion for head, or for anyone else in unit

700. If religion is not reported for anyone in the household, either assign "unknown" (if this country does not use dynamic imputation) or impute a religion from the most recent head/reference person of household with similar characteristics including age and sex as well as language, birthplace and other variables as appropriate, considering the circumstances.

(b) For person other than head, without religion

701. If this person is not the head/reference person of household and reports no religion, assign the head/reference person's religion.

6. Language

702. Four types of language data can be collected in censuses (United Nations, 2017, para. 4.179), namely:

- Mother tongue, defined as the language usually spoken in the individual's home in his or her early childhood;
- Main language defines as the language that the person commands best.
- Usual language, defined as the language currently spoken, or most often spoken, by the individual in his or her present home;
- Ability to speak one or more designated languages.

(a) Language edits

703. Of the four different measures of language that may appear on the questionnaire the first two, mother tongue and usual language, are related. When both are present on a questionnaire, editing teams should consider editing them together. If either is invalid, the other can be used to supply an entry.

(b) Language edits: head/reference person of household

704. Language is another variable fitting the examples presented in chapter II. Editing teams should establish the logical editing scheme used for the other social variables, editing the head of household first. If the person with an invalid or unknown language (mother tongue or usual language) is the head/reference person of household, first determine whether anyone else in the housing unit has a valid language and assign the first valid language. When there is none, either assign "unknown" if dynamic imputation is not used or impute a language from the most recent head/reference of household with similar characteristics, including age and sex as well as other language variables, birthplace and other variables as appropriate under these circumstances.

(c) Language edits: persons other than head/reference of household

705. If the person is not the head/reference of household and the language is invalid, then assign the head/reference of household's language.

(d) Language edits: use of ethnic origin or birthplace

706. Language and ethnic origin, and sometimes birthplace, are closely related, and for some countries can be edited together. Also, editing teams should consider organizing codes to reflect the relationships among these variables. Depending on the number of digits in the code and the distribution of the country's languages and ethnic groups, correspondences can be developed to help in assigning unknown or inconsistent responses. For electronic devise edit on entry, both the ethnicity and language items would need to be entered before they can be used to assist each other; look up lists can be embedded in the electronic devise to assist in obtaining the appropriate response.

(e) Language edit: Mother tongue

707. If the mother tongue is unknown, for example, the person is Filipino and was born in the Philippines, an appropriate equivalent language – Tagalog, Ilokano or another language of the Philippines – can be assigned. Usually, only the head/reference person of household is assigned a language in this way, and the code for that language is assigned to the other members of the household, but each country's editing team needs to consider the particular circumstances, including geography (such as urban or rural residence), age or other items. For electronic devise entry,

when the head's value is ascertained, the other household members how do not have a valid entry can automatically receive the head's value.

(f) Language edits: Ability to speak a designated language

708. The ability to speak a designated language is a third variable fitting the examples presented in chapter II. Again, the head/reference person of household should be edited first. If the value for language for the head of household is invalid or unknown, the first step is to see whether anyone else in the housing unit has a valid ability to speak the language and assign the first valid one. Then, if no such person exists, either assign "unknown", if this country does not use dynamic imputation, or impute language ability from the most recent head of household with similar characteristics (e.g., age and sex, but also birthplace and other variables as appropriate, considering the circumstances). If the person is not the head of household, and the ability to speak a designated language is invalid, then assign the head of household's ability.

(g) Multiple languages

709. Many countries now allow for two languages (or nationalities). In this case, the edit must check for several possible errors. The same code may appear twice, the code may only appear in the second position with nothing or an illegal code in the first position, and whether similar languages should be considered the same language.

710. If the same code appears in both the first and second positions, the code in the first position should be kept, and the code in the second position should be blanked. Some countries will prefer to use a specific code in the second position (for example, 999) to designate that this is a single language. When that designation is used, that fact must be taken into account when making tables. Leaving the second position blank will make tabulation easier.

711. When a language appears only in the second place, or the first language is invalid, then the edit should move the language code in the second position and the second position should be made blank. This will create a single language entry for the person. The editing team may prefer to create a "look up" file of frequent invalid entries to substitute valid entries and so keep the multiple language response. Or, the language of someone else in the house could be used to substitute the invalid entry.

712. When similar entries appear in the two positions, that is, the two entries basically are the same language, say two very similar languages from the same area of a country, then the editing team may choose to drop the second language. It is important to establish rules of precedence in this case.

713. Very few countries permit more than two language responses. When they do, the above edits would need to be modified to account for these multiple entries.

714. For electronic data collection method, the data collection application should not allow an entry in only the second position. An embedded look up list should assist in telling the enumerator/respondent if the languages are too similar and should accept only the single language determined by the subject matter specialists. Determining the one or two languages while still the interview is going on. The enumerator and respondent can work together to arrive at the appropriate response or responses.

7. Ethnicity

715. The need for information about the national and/or ethnic groups within a population is dependent upon national circumstances. Some of the bases upon which ethnic groups are identified include ethnic nationality (country or area of origin as distinct from citizenship or country of legal nationality), race, color, language, religion, customs of dress or eating, tribe or various combinations of these characteristics. In addition, some of the terms used, such as

"race", "origin" and “tribe”, have a number of different connotations. The definitions and criteria applied by each country investigating the ethnic characteristics of its population must therefore be determined by the groups that it desires to identify. By the very nature of the subject, these groups will vary widely from country to country; thus, no internationally relevant criteria can be recommended (United Nations, 2017, para. 4.187).

716. The Principles and Recommendations (United Nations, 2017, para. 4.185) suggests taking particular care in identifying indigenous peoples, usually as a subset of for the ethnicity item. Care must be taken in developing the code lists that indigeneity is identified uniquely so that appropriate edits and tabulations can be developed to assist in planning and policy formation for indigenous peoples. For example, separate codes may be needed for the same group when they are nomadic compared to when they have settled into residential areas. Special edits can be developed, partially through “look up” files for particular groups of indigenous peoples to make certain that they are fully and properly identified for subsequent tables. Special imputation procedures can be developed for these groups, or additional categories within in existing hot decks can be used.

(a) Ethnicity edits

717. Several other variables, if collected, can assist in “determining” ethnicity when it is invalid or unknown. In many countries, a relationship exists between birthplace, both within the country and in foreign countries, and ethnicity. Similarly, “mother tongue” is often a good indicator of ethnicity for many countries since the categories, and therefore the codes, will be similar, if not the same. In CAPI or CASI entry, the values for the other variables can be available to assist in selecting the most appropriate category for ethnicity.

(b) Ethnicity edit: for head of household or reference person

718. Ethnic origin also fits the example introduced in chapter II. Editing teams should follow consider the scheme already described for the other social variables. The head of household/reference person should be edited first. If the person with an invalid or unknown ethnic origin is the head of household/reference person, look first for a valid ethnicity for anyone else in the housing unit, and assign the first valid ethnicity. If no such person exists, the next step is either to assign “unknown” or, if this country does not use dynamic imputation, to impute an ethnicity from the most recent head of household or reference person with similar characteristics (age and sex as well as language, birthplace and other variables that may be appropriate, considering the circumstances).

(c) Ethnicity edit: persons other than head of household or reference person

719. If the person is not the head of household or reference person and ethnic origin is invalid, then assign the head of household/reference person’s ethnic origin.

(d) Ethnicity edit: use of language and birthplace

720. Ethnic origin and language, and sometimes birthplace, are closely related, and for some countries can be edited together. Also, the editing teams should consider organizing their codes to reflect the relationships among these variables. Depending on the number of digits in the code and the distribution of the country’s ethnic groups and languages, correspondences can be developed that will help in assigning unknown or inconsistent responses.

721. For example, if ethnic origin is unknown, but, for example, the person speaks one of the languages of the Philippines and was born in the Philippines, an appropriate equivalent ethnic origin, Filipino, might be assigned. Usually only the head of household would be assigned ethnicity in this way (and the other members would be assigned that code), but each country’s editing team needs to consider their particular circumstances, including geography (such as urban or rural residence), age or other items.

(e) Multiple ethnicities

722. Many countries now allow for two ethnicities. In this case, the edit must check for several possible errors. The same code may appear twice, the code may only appear in the second position with nothing or an illegal code in the first position, and whether similar ethnicities should be considered the same ethnicity. For electronic data collection methods, the procedures for more than one language would also be applied.

723. If the same code appears in both the first and second positions, the code in the first position should be kept, and the code in the second position should be blanked. Some countries will prefer to use a specific code in the second position (for example, 999) to designate that this is a single ethnicity. When that designation is used, that fact must be taken into account when making tables. Leaving the second position blank will make tabulation easier.

724. When an ethnicity appears only in the second place, or the first ethnicity is invalid, then the edit should move the ethnicity code in the second position and the second position should be made blank. This will create a single ethnicity entry for the person. The editing team may prefer to create a "look up" file of frequent invalid entries to substitute valid entries (there should not be invalid entry for CAPI and CASI) and so keep the multiple ethnicity response. Or, the ethnicity of someone else in the house could be used to substitute the invalid entry.

725. When similar entries appear in the two positions, that is, the two entries basically are the same ethnicity, say two very similar ethnicities from the same area of a country, then the editing team may choose to drop the second ethnicity. It is important to establish rules of precedence in this case.

726. Very few countries permit more than two ethnicity responses. When they do, the above edits would need to be modified to account for these multiple entries.

8. Indigenous peoples

727. The Principles and Recommendations for Population and Housing Censuses, Revision 3 (2017, para. 4.188-4.191 discuss how indigenous peoples can be defined based on ethnicity origin or on indigenous identity and other characteristics. An assumption is made that indigenous households live in adjacent or nearby groups.

728. The edit for indigenous peoples should probably be the standard edit. If the householder does not have a value for this item, the missing value can be imputed using the value of the head of household or reference person or someone in the household with a value and assign that value to the householder. If no one in the unit has a value for indigenous people, then a response from a previous householder with similar characteristics and a value for this item would be taken. All other household members would then receive the same value for indigenous people.

9. Disability (core topic)

729. Disability status separates the population into those with and without a disability. A person with a disability is as a person who is at greater risk than the general population in experiencing restrictions in performing specific tasks or participating in activities. The UN recommends inclusion of six domains in assessing disability status: (1) walking, (2) seeing, (3) hearing, (4) cognition, (5) self-care and (6) communication. (United Nations, 2017, para. 4.194- 4.196).

730. The question identifies persons with disabilities with a list of broad disability categories so that each person checks the presence or absence of each type of disability. Many countries use the following list of disabilities based on the International Classification of Functioning, Disability and Health (ICF) monitors disability: (1) functioning and disability, including body functions and body structures (impairments) and activities (limitations) and

participation (restrictions), and (2) contextual factors, including environmental factors and personal factors. (United Nations, 2017, para. 4.193)

731. A census format offers only limited space and time for questions on any topic such as disability. Because of census requirements, often a large follow-on survey or even an independent survey should be used to provide information on disability. Formatting is particularly important to consider, given the several recommended disability categories.

(a) Disability census questions

732. The UN recommends asking questions about each domain of disability separately. The language should be clear, unambiguous and simple, without negative terms, and should be addressed of each household member separately.

(b) Disability edits

733. When someone does not respond to the disability questions asked, it is difficult to determine whether the item is left blank because of no disability or because of an unwillingness on the part of the respondent to answer, for whatever reason. A country's editing team must decide whether they want to edit the item in the usual way, by assigning unknowns when dynamic imputation is not used, or by using the responses of other individuals when dynamic imputation is used. It should be noted that if only those responses specifying that a disability is present is accepted, and invalid response is considered as "no disability", this decision will lead to introduce an error with an increase in people without disability.

(c) Multiple disabilities

734. Countries collecting information on multiple disabilities will need to modify the edit. The editing program will need to keep track of how many total disabilities are possible and of the duplication and distribution of those disabilities. As before, most countries will find it inappropriate to use data from other persons to assign disabilities, so "unknown" and even "unknown whether disability is present" may be needed in invalid cases. Edit on entry with electronic data collection applications can provide assistance when multiple disabilities are present through a set of predetermined rules.

735. Countries should consider developing derived variables for combinations of multiple disabilities and cases as well as people. For example, variables would be created so total persons with seeing disabilities as well as total cases of persons having seeing disabilities whether alone or in combination with others would be created. See the section on derived variables for more detail. Derived variables should be obtained in the office edit and not in the field.

(d) Cause of disability

736. A country's editing team must decide whether to edit the item in the usual way by assigning unknowns, when dynamic imputation is not used, or by using the responses of other individuals when dynamic imputation is used. Alternatively, the specialists may decide that only those responses specifying that a cause of disability is present will be accepted, and an imputation matrix will not be used.

E. FERTILITY AND MORTALITY

1. Children ever born and children surviving (core topic)

737. "Children ever born" is the total number of children ever born alive, thus excluding stillbirths, miscarriages and abortions. Sometimes, demographers use the expression "children ever born alive," but here the terms "children ever born" or "children born" will be used.

738. The universe for which data should be collected for each of the topics included in this section consists of women 15 (or some other minimum acceptable age) years of age and over, regardless of marital status or of particular subcategories such as ever married women. In countries that do not collect or tabulate data for women 50 years of age and over, efforts should be concentrated on collecting data from women between 15 and 50 years of age only; in the investigation of recent fertility it may be appropriate in some countries to reduce the lower age-limit by several years (United Nations, 2017 para. 4.216).

(a) Fertility items collected

739. The Principles and Recommendations for Population and Housing Censuses Revision 3 recommend obtaining information on three fertility items: children ever born, date of last child born alive and age of mother at birth of first child born alive. Responses to items on age, date or duration of marriage may improve fertility estimates based on children ever born. Also, many countries continue to collect information on children living, which helps, particularly in retrospective fertility analysis and in indirect measures of fertility and mortality.

740. Censuses and surveys collect information on fertility from all females, using a country-defined minimum age and sometimes a maximum age as well.

741. Because of the complexity of the fertility edit, there is a need to use a fairly sophisticated hot deck when using dynamic imputation. Often, models are needed to obtain the most likely fertility distributions in many countries. The information available at the time of the office edit will provide a better forum for the full fertility edit.

(b) General rules for the fertility edit

742. Females younger than the designated earliest age for fertility and all males should be checked, and any present fertility information should be blanked out. The purpose of the fertility edit is to make the entries consistent with each other and with age:

i. The total number of children ever born alive cannot be greater than the person's age plus some country-defined minimum age multiplied by a factor. That factor will be 1 when females are allowed one birth per year; the factor will be 1.5 for one and half years between adjacent children, and so forth. See the section below on "age at first birth" for the edit to determine the minimum difference in age between the mother and the eldest child born alive;
ii. The total number of children ever born cannot be greater than the sum of the number of children living in the housing unit, living elsewhere and dead. When the total number is greater than the sum of the parts, the editing teams must decide which takes precedence, so adjustments can be made;
iii. If data are collected for both children still alive and children deceased, the total number of these children cannot be greater than the number of children ever born;
iv. The number of children ever born cannot be smaller than the entry in "children born in last 12 months";
v. Depending on the country, and the actual number of children ever born and children still alive, an imputation matrix might be used for the item on children born in the last 12 months to allocate a response by age and children ever born. However, great care must be taken in assigning a value to children born in

the last 12 months when a blank appears. For most countries, a blank for this item means that no child was born. Allocated values might skew the data;

vi. Sometimes countries collect children ever born, children surviving and other fertility items by sex. In these cases, the edits presented here work in the aggregate, but the countries may want to add additional checks to account for the additional information available. These additional checks include making certain that the number of male children ever born is the sum of male children surviving and deceased male children, and the number of female children ever born is the sum of female children surviving and deceased. As for the edits for children not differentiated by sex, appropriate action needs to be taken when the sums are not equal to the parts.

(c) Relationship between children born and children surviving

743. The data on children ever born and children surviving are used for indirect estimates of both fertility and mortality. Results of the census or survey are organized by single year or five-year age groups of females. Various algorithms obtain constant or changing mortality estimates. However, in order to get the best results, editing teams must be careful in determining the appropriate edit for the available data.

744. Part of the problem with developing a general edit is that different countries request different types of information. For example, the following sets of information are collected in different countries:

i. Children ever born only
ii. Children ever born and children surviving (both sexes combined or separate sexes)
iii. Children ever born, children surviving and children who died (both sexes combined or separate sexes)
iv. Children ever born, children living at home, children living away and children who died (both sexes combined or separate sexes)

(d) Edit when only children ever born is reported

745. If the country does not use dynamic imputation, an invalid or missing value for "children ever born" should be assigned as "unknown". In countries using dynamic imputation, the specialists must decide whether they want to use dynamic imputation for all items. If the specialists use this method, children ever born can be obtained based on single year of age of the female and at least one other characteristic. It is also possible to use a single dimensional array for single year of age of mother only. The other characteristics might be items such as educational attainment or religion, since it is known that in many countries differential fertility exists for various levels of educational attainment or different religious affiliations.

(e) Edit when children ever born and children surviving are reported

746. If responses are present for both "children ever born" and "children surviving", the program needs to determine the following:

i. Whether the items are internally consistent (is the number of children ever born equal to or greater than the number of children surviving);
ii. Whether at each item agrees with the age of the female;
iii. Whether "children ever born" agrees with "children born in the last year" (or last birth), if collected.

747. Demographers use the items on children ever born and children surviving to obtain indirect mortality estimates. Because of this, the edit must maintain the relationship between the two items. Sometimes only one of the two items are reported, and the other is unknown. An easy edit would be to assume no deaths to children ever born

and make both items the same. However, in making the two items the same, the indirect mortality estimation would not take into account babies who might have died after birth, thus underestimating the mortality and overestimating the life expectancy. If few of these cases appear in the census or survey, little damage is done. However, if this occurs with some frequency, as would be expected in those countries using the indirect method, the effects could be substantial. An example is given in figure 27.

Figure 28. Illustration of household with fertility information

Person	Relation	Sex	Age	Children ever born	Children surviving
1	Head of household/ Reference person	1	60		
2	Spouse	2	60	5	99
3	Daughter	2	40	3	3
4	Granddaughter	2	20	1	1
5	Granddaughter	2	18	0	0
6	Granddaughter	2	1		

NOTE: 99 = Data missing or invalid

748. Here the spouse reports 5 children ever born, but for whatever reason, the number of children surviving did not get recorded. The respondent or the enumerator did not report the value, or the data entry operator mis keyed the information. Many countries develop an edit that would assign the value "5" to the children surviving based on the number of children ever born. However, in doing this, the data become skewed.

749. In fact, the value does not have to be changed at all. Those countries not using dynamic imputation may choose to leave the "unknown" value in place. Of course, this decision also creates a skewing, since that edit decides that the "unknown" and "known" responses have the same distribution for tabulations. If a country requires data on children ever born and children surviving to determine indirect estimates for mortality, it is also probably a country with reporting problems in the data. In this case, keeping unknowns in the data is likely to skew the final analysis. Females with an unknown for either children ever born or children surviving cannot be used in the determination of the mortality estimation since the difference between the children ever born and children surviving cannot be determined.

750. Those countries using dynamic imputation should consider determining the missing piece of information based on the other fertility item and the age of the female, at a minimum. The imputation matrices can be updated when valid information for age of female, children ever born, and children surviving is present and can be used when the item is missing. When children ever born is missing, the imputation matrix will have age of female and number of children surviving. When children surviving is missing, the imputation matrix will have age of female and number of children ever born.

751. Further, in developing the imputation matrices it is important to remember that the number of children ever born, and number surviving must conform to the age difference between mother and eldest child (if this information is present) and the total number of children ever born for a particular age of mother.

752. For example, the difference between the imputed number of children ever born and the mother's age might be at least 12. Then, an imputation matrix using 5-year age groups of females would almost certainly impute incompatible information in some cases.

753. The accompanying imputation matrix in figure 28 shows female ages across the top and the number of children ever born down the side. The entries are the imputed values for children surviving. Sometimes the responses will be appropriate, but sometimes they will not. If the program encounters a 19-year-old female with 5 children ever born, the value of 5 children surviving should probably pass the age difference criteria (an age difference of 15, based on children surviving and reported age.) However, for a 15-year-old, neither the 5 children ever born (age difference of 10) or 4 children surviving (age difference of 11) would be acceptable.

Figure 29. Initial values for determining children surviving when age and children ever born are valid

Children ever born	Age												
	15	16	17	18	19	20	21	22	23	24	25-29	30-34	35+
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0
2		2	2	2	2	2	2	2	2	2	1	1	1
3			3	3	3	3	3	3	3	3	2	2	2
4					4	4	4	4	4	4	3	3	3
5					5	5	5	5	5	5	4	4	4

754. The imputation matrix is better when single years of age apply for young females. Then, only valid age difference responses for that particular age would be entered in the imputation matrix, and only valid responses could be pulled from the imputation matrix.

755. The fertility hot deck should contain all fertility variables for a single year of age. All fertility variables must agree (Children ever born – CEB – must be greater or equal to children surviving – CS, so CEB => CS, TCEB (total children ever born) = MCEB (male children ever born) + FCEB (female children ever born), TCS=MCS+FCS, etc.) as well as information for last birth. Even then, when last birth is considered after editing the CEB and CS, problems can occur when the female considered has a different CEB and CS then the female in the hot deck.

(f) Edit when children ever born, children surviving, and children who died are reported

756. "Children ever born" is the sum of "children surviving" and "children who died". Any inconsistency may be resolved as explained below.

(i) When all three items are reported

757. If all three pieces of information are present, the program needs to determine:

- Whether the three items are internally consistent is that the number of children ever born the sum of the children surviving and children who died;
- Whether each of the three items is consistent with the age of the female;
- Whether the number of children ever born is consistent with number born in the last year (or the last birth), if collected.

758. If all of these are consistent, the edit is finished. However, if any are inconsistent, the edit must resolve them. The three items may not be internally consistent: for example, a female may have 5 children ever born, but only two children surviving and two deceased. The editing team should decide which variable takes precedence over the others. In many cases, the female is likely to remember all of the children she has ever borne, although she may forget the

exact number who died. Then, the editing team may choose to accept the number of children ever born and those surviving, and subtract to obtain a new, consistent value for deceased children.

(ii) When two items are reported

759. Since the category children ever born (CEB) is the sum of the children surviving (CS) and the children who died (CD), if any two of the three pieces of information are available, the computer program can determine the third variable:

If CEB and CS are known, CD = CEB – CS.
If CS and CD are known, CEB = CS + CD.
If CEB and CD are known, CS = CEB – CD.

760. These tests would normally be run first. Once the program determines that all three pieces of information are valid and consistent, the edit is finished.

(iii) When one item is reported

761. When only one of the three items is known, if the country does not use dynamic imputation, the other two items should be made "unknown". If the country uses dynamic imputation, editing teams need to determine a method of getting at least one more item and the third item should then be obtained through subtraction or addition. A two-dimensional matrix can be used to get the second fertility value, based on the first item and single year of age for the females. If children ever born is known, for example, children surviving can be obtained from the imputation matrix, as described above, and then dead children should be obtained by subtraction. Similarly, if children surviving is known, children ever born is obtained from the imputation matrix of single year of age of female and children surviving, and number of dead children is obtained by subtraction.

(iv) When none of the items is reported

762. When none of the three items is available, the editing team must make decisions about how to proceed. If the country does not use dynamic imputation, all items should become "unknown", and should not be used in the mortality or fertility indirect methods. In countries using dynamic imputation, the specialists must decide whether they want to use dynamic imputation for all items.

763. If the specialists decide to use dynamic imputation, children ever born can be obtained based on single year of age of the female and at least one other characteristic. It is also possible to use a single dimensional array for single year of age of mother only. The other characteristics might be items such as educational attainment or religion.

764. Once the first item is determined, to obtain the second fertility item it is possible to follow the steps outlined above for editing when only one item is reported. Then, the third item can be obtained from the first two items. The three items should be compatible because the imputation matrices should be updated only when all items are compatible. The fertility obtained should also be compatible with other females in the geographical area since information from those females is used to update the imputation matrix.

(g) Edit when both children ever born, children living at home, children living away and children who died are reported

(i) When all four items are reported

765. If all four pieces of information are present, the program needs to determine:

- whether the four items are internally consistent, so that the number of children ever born is the sum of the children living at home, children living away, and children who died;
- whether each of the four items is consistent with the age of the female;
- whether children ever born "is consistent with" children born in the last year (or last birth), if collected.

766. If all of these are consistent, the edit is finished. However, if any are inconsistent, the edit needs to resolve the inconsistencies. As in the case of the three items case described above, all four items may not be internally consistent. Again, the editing team should decide which variable takes precedence over the others. In many cases, the female respondent is likely to remember all of the children she has ever borne, although she may forget some of those who moved away or the exact number who died. Then, the editing team may choose to accept the number of children ever born and those surviving (the sum of the children living away and the children living at home), and subtract to obtain new, consistent values for other variables. The editing team may need to develop algorithms for various combinations of events.

(ii) When three of the four items are reported

767. The children ever born (CEB) is the sum of the children living at home (CLH), the children living away (CLA) and the children who died (CD). If any three of the four pieces of information are available, the computer program can determine the fourth variable:

If CEB, CLH and CLA are known, CD = CEB – CLH – CLA.
If CLH, CLA and CD are known, CEB = CLH + CLA + CD.
If CEB, CLH and CD are known, CLA = CEB – CLH – CD.
If CEB, CLA and CD are known, CLH = CEB – CLA – CD.

(iii) When two of the four items are reported

768. If only two of the items are known, then the editing team must decide what to do next. For example, in many countries, women do not report the number of children who died. The other item most likely to be omitted is information on children residing outside the housing unit, which also cannot be obtained directly. Hence, care must be taken in developing the questionnaire, in implementing the enumeration and in processing in order to obtain the best quality data for all of the fertility items.

769. The data for children residing in the unit (CLH) can be obtained by summing the children in the housing unit. As long as only one female in the unit has the appropriate relationship, a simple tally should give the number of children living in the unit. If more than one female has this relationship, the editing program might still be used, on the assumption that the children will immediately follow the mother during data collection. When all else fails, those countries using dynamic imputation could impute the number of children living in the unit from the age of the mother and one of the other known variables. (See the general rules below for imputing individual fertility items from other items and mother's age.) It is important to use single year of age of female whenever possible, as well as single number of children ever born, living in the unit, living away, or dead.

770. As an example, children ever born and dead children may be valid entries, but children living in the household and children living away may be invalid. In this case, the number of children living at home can be determined by summing the children with the appropriate relationship to the mother (assuming the mother is the head of the household). Then three out of the four items will be available, and the fourth, children living away, can be determined by subtraction: CLA = CEB – CLH – CD.

771. However, when only two items are known, it is more likely that children ever born and children living at home will need to be recoded. Females usually readily report children ever born, and information on children living

at home can usually be obtained by observation or by working with respondents while enumerating, but these solutions are not available for children living away or dead children. Then, the edit can use an imputation matrix with age of female and children ever born (CEB) or, even better, age of female, children ever born (CEB), and children living at home (CLH). The variables will obtain information from a similar female with the same characteristics for children living away (CLA).

772. Countries using only the two-dimensional matrix for age of female and children ever born (CEB) without also including the third dimension, children living at home (CLH), risk obtaining a value for children living away (CLA) that is not compatible with the other two. For example, if the female's age is 25 and CEB is 5, a value of 3 might be obtained from the imputation matrix for children living away. If the value for children living at home is 2, then the edit has no problem. The value for dead children should be 0, and the fertility items should be: CEB = 5, CLH = 2, CLA = 3, CD = 0.

773. However, the value of children living at home might actually be 4, with only the female's age and children ever born are used to determine the value for children living away. The value of 3 for children living away would then produce an incompatibility among the items. The value for children ever born (5) would be less than the sum of the living children (4 at home and 3 away, or a total of 7). Hence, a three-dimensional matrix should be used: for 5 CEB and 4 CLH, the value in the imputation matrix might be 1 for children living away (and the value of 0 should be determined by subtraction for dead children). Or, the value in the imputation matrix should be 0 for children living away (and the value of 1 should be determined by subtraction for dead children). Similar imputation matrices need to be developed for the other pairs of known information as in figure 30.

Figure 30. Sample imputation matrices to be developed for pairs of known information

If these are known..		Use dynamic imputation for one of these (and then subtract or add)	
Children ever born	Children living at home	Children living away	Dead children
Children ever born	Children living away	Children living at home	Dead children
Children ever born	Dead children	Children living at home	Children living away
Children living at home	Children living away	Children ever born	Dead children
Children living at home	Dead children	Children ever born	Children living away
Children living away	Dead children	Children ever born	Children living at home

774. In each case, two of the four items are available. The third item is obtained by dynamic imputation, and the fourth item by subtraction or addition. Editing teams must decide which is the best path to follow based on cultural circumstances.

(iv) When only one item is reported

775. When only one of the four items are known, the situation is even more problematic. Countries must decide how they want to proceed when this much information is missing. If dynamic imputation is used, the first imputation matrix would, as noted above, use an item such as single year of age of female and the one known item to create a two-dimensional matrix for imputation of any one of the other items. Once two items are determined, the other two remain unknown, by definition. Hence, continuing to use dynamic imputation for the third item should not create an incompatibility with the other items since they are unknown. The scheme discussed above for two known items and

two unknown items, is used to obtain a third item. Then, the fourth item is obtained by subtraction. All four items should be compatible.

(v) When none of the items is reported

776. When none of these four items is available, the editing team must decide how to proceed without any known items. If the country does not use dynamic imputation, all items should become "unknown", and should not be used in indirect methods for estimating mortality or fertility. In countries that do use dynamic imputation, the specialists must decide whether they want to use imputation for all items.

777. If the specialists decide to use dynamic imputation, values for children ever born can be obtained based on single year of age of the female and at least one other characteristic. It is also possible to use a single dimensional array for single year of age of mother only. The other characteristics might be items such as level of educational attainment or religion, since it is known that in many countries differential fertility exists for various levels of educational attainment or religious affiliation.

778. Once the first item is determined, the approach used above when only one item is known can be used to obtain the second fertility item. Then, the third item can be obtained from the first two items, and the fourth item can be obtained by subtraction. The four items should be compatible because the imputation matrices should only be updated when all items are compatible. The fertility obtained should also be compatible with other females in the geographical area since information from these females is used to update the imputation matrix.

(h) Special case of 5 or more items

779. As international migration has become more important in some smaller countries, additional information on children away is being collected. When the variable for "children away" is divided into "children away but in the country" and "children away internationally", the procedures for four variables – at home, away, dead, and total – must be expanded to take this additional information into account. And, it is a good idea, as noted, to have a single array line for each age of woman, with complete fertility information going in when all items are valid and internally consistent and consistent with her age. And, then when cases appear with fertility information being inconsistent (including with age), the whole, appropriate line is taken.

(i) Importance of a single donor source for all fertility items

780. So, if at all possible, it is very important to impute all items from one woman when nothing is known. In order to make certain that all of the information comes from the same female source, it may be necessary to develop imputation matrices that use all of the fertility information. In this case, the imputation matrices could be updated only when the editing program determined that all fertility items agreed. As the previous paragraph describes, it is better not to impute item by item, but when several items are amiss, to use another woman's total information.

(j) Relationship of own children to children in the house and children surviving

781. When countries use the own children method to assist in checking the fertility edit as it is developed and implemented, information from the children in the house and in the mother child matrix can assist in assessing the reliability of the results of the edit. Very few countries use this method to assist in the edit. So, it remains experimental, but results look promising.

2. Children living (core topic)

782. The edit for the number of children still alive is incorporated into the general fertility edit, and so is covered there.

3. Date of birth of last child born alive and Births in the 12 months before the Census (core topic)

783. Information on last births assists in providing estimates of current fertility just prior to a census or survey. One approach is to collect the date of birth (day, month and year) of the last child born alive and on the child's sex (and sometimes vital status). A second approach is to collect births in the 12 months before the census; this second approach is easier for enumerators and respondents because only a single "yes" or "no" is needed rather than an exact date. However, this approach by itself cannot generally be relied on to generate accurate estimates of current fertility because of errors and omissions commonly encountered in the reporting of live births within a 12-month retrospective period (United Nations, 2017, para. 4.242).

784. For the first item, during processing, the number of children born alive in the 12 months immediately preceding the census date can be derived (and then kept as a recode) as an estimate of live births in the last 12 months. For estimating current age-specific fertility rates and other fertility measures, the data provided by this approach are more accurate than information on the number of births to a woman during the 12 months immediately preceding the census (United Nations, 2017, para.4.238).

785. It should be noted that information on the date of birth of the last child born alive does not produce data on the total number of children born alive during the 12-month period. Even if there are no errors in reporting the data on the last live-born child, this item only ascertains the number of women who had at least one live-born child during the 12-month period, not the number of births, since a small proportion of women will have had more than one child in a year (United Nations, 2017, para.4.238).

786. The information needs to be collected only for women between 15 and 50 years of age who have reported having at least one live birth during their lifetime. In addition, the information should be collected for all the marital-status categories of women for whom data on children ever born by sex are collected. If the data on children ever born are collected for a sample of women, information on current fertility should be collected for the same sample (United Nations, 2017, 4.239).

787. The following edits should be included in the editing program: The date of birth of last child should be entered for all females between a country-defined minimum age and a country-defined maximum age. The program should check for a correspondence. For example, no information should appear for males and females not in the selected age group. Also, females in the selected age group with parity greater than zero should have a valid day, month and year of last birth (or an indication of whether a birth occurred in the last 12 months if that question is used). As with the general fertility items, edit on entry using electronic data collection technologies will improve the valid value due to checking automatically unknowns and consistency with the selected age group and sex. However, there will be edit during the office edit when all variables are available as well as other information, like vital records, to offer assistance in determining the most likely entry.

788. The editing team needs to decide whether the day and month must be valid: editing teams using dynamic imputation can impute day and month when they are missing; those not using dynamic imputation would assign "unknown" for day and month. If the subject matter specialists, usually in the form of demographers, want actual age of mothers at birth of their children as a recode for fertility analysis, then at least the month of last birth should probably be imputed if it is not present. The recode can then be obtained.

789. Similarly, some demographers want to analyze months since the last birth. Editing year and month of last birth provides the information needed to obtain completed months since the last birth. When day of last birth is also collected, it can also be used in the determination of the recode for months since last birth. (See Annex I for a method of obtaining a recode for months since last birth.)

790. If the information is missing or invalid, for the year of birth of the last child, countries not using dynamic imputation can assign "not stated" or "unknown". Countries using dynamic imputation can use other variables such as age and number of children ever born to obtain the date of birth of the last child.

791. Because of the importance of the use of date of last birth in providing a measure of the recent national, regional, and local fertility experience, additional checks should be considered. A useful edit is to check within the household for children zero years old and use the relationships of the mother and that child (or mother's person number for the child, in collected) to determine that the child is reported as a last birth for the mother. The checks should go both ways: the zero-year-old should be checked from mothers and the last births should be checked against the household listing.

792. Those countries also collecting deaths in the year before the census or survey may decide to also include a check of deaths to zero-year-old in the year before the census against last births when the last birth is reported as "deceased" or no longer alive. While this check will not work if the mother is not in the house because of death or movement, or the child may not be reported for whatever reason, some percentage of infant deaths could be checked in this manner.

4. Deaths among children born in the past 12 months (core topic)

793. Deaths among children born in the past 12 months should have only been asked if a birth had occurred in those 12 months. If no birth occurred this item should be left or made blank.

794. If a birth did occur, but the item has been left blank, the subject matter specialists will need to decide where to assume that the child is still alive (for example, a 0 year old child appears in the housing unit and therefore there was no death) or there was a child who might have died.

795. If vital statistics are available, an algorithm for determining the percentage of deaths might be used to decide on what value to assign. Otherwise, females of the same age with the same or similar number of children and last births might be used to assign a value.

796. If the value is blank or invalid but a date of death for the last child is present, than the child should be assigned the value for having died. If a date of death is present that date should be after the date of birth of the last child but before the date of the census.

5. Age at first marriage

797. According to Principles and Recommendations (United Nations, 2017, para. 4.247), "date of first marriage" comprises the day, month and year when the first marriage took place. In countries where the date of first marriage is difficult to obtain, it is advisable to collect information on age at marriage or on how many years ago the marriage took place (duration of marriage). Include not only contractual first marriages and de facto unions but also customary marriages and religious marriages. For women who are widowed, separated or divorced at the time of the census, "date of/age at/number of years since dissolution of first marriage" should be secured.

798. Information on dissolution of first marriage (if pertinent) provides data necessary to calculate "duration of first marriage" as a derived topic at the processing stage. In countries where duration of marriage is reported more reliably than age, tabulations of children ever born by duration of marriage yield better fertility estimates than those based on data on children born alive classified by age of the woman. Data on duration of marriage can be obtained by subtracting the age at marriage from the current age, or directly from the number of years elapsed since the marriage took place.

799. The date of first marriage should be entered for all ever-married persons (or, females only, following the Principles and Recommendations). The program should check for a correspondence: never-married persons should have no information, but ever-married persons should have a valid day, month and year. Editing teams need to decide whether day and month must be valid: countries not using dynamic imputation can assign "unknown" for day and month; countries using dynamic imputation can impute day and month when they are missing. The edit on entry for CAPI and CASI can be easily implemented because age will already have been edited. Length of time since the first marriage can be determined when the date of the marriage is provided.

(a) Age at marriage for never married persons should be blank

800. Persons who have never been married should not report age at first marriage. If a valid entry appears for a never-married person, the editing team must decide whether to change the marital status or blank the age at marriage for the person. If the marital status is to change, countries using only "not stated" will apply that code. Countries using dynamic imputation should probably use age and sex to obtain an appropriate marital status response.

(b) Ever married persons should have an entry

801. For the year of first marriage, countries not using dynamic imputation can assign "not stated" or "unknown". Countries using dynamic imputation can use other variables, such as age of spouse or age differences between spouses, number of children and children born in the last year, to determine an appropriate year of first marriage.

6. Fertility: age at first birth

802. The age of the mother at the time of the birth of her first live-born child is used for the indirect estimation of fertility based on first births and to provide information on the onset of childbearing. If the topic is included in the census, information should be obtained for each woman who has had at least one child born alive (United Nations, 2017, para. 4.249).

803. Age at first birth is determined either directly by an explicit item, "age at first birth," or by the age difference between the mother's current age and the age of the eldest child, if the eldest child's age is known. The earliest country-defined age for children is not the biological earliest age. If, for example, a country's earliest acceptable age at first birth is 13 years, respondents may report, or enumerators may record an age at birth of 11 or 12 for a person. Then, editing teams must decide whether to change the earliest acceptable age, delete the birth, or change either the mother's age or her age at first birth (using either a child's age or her age, depending on the variables used to determine the age difference). Similarly, editing teams must decide what "oldest age" is a maximum for age at first birth. While females are capable of having children into their 50s, this event does not happen very often, and, in order to correct mistakes, the edit must determine whether the outliers are real.

804. It is important to remember that the earliest or latest age at first birth (and the age difference between the mother and her eldest child resident in the household) must conform to country customs and traditions. The subject specialists must decide when a value is noise rather than a legitimate age at first birth. When the rules are established, then the specialists must decide how to correct the problem. If dynamic imputation is not used, the program should assign "unknown". When dynamic imputation is used, it can determine the age at first birth based on other females of similar age and similar number of children ever born. Specialists determining the imputation matrix may want to take into account such factors as urban/rural residence (if fertility differs between the two areas), the presence of the female in the work force (although current labor force status is not necessarily the same as status at her first birth) and level of educational attainment. As with the other fertility items, this item is better editing in the office than in the field, so edit on entry is not encouraged; however, like age at first marriage, the edit is straight-forward, so could be included.

7. Household deaths in the past 12 months (core topic)

805. Information on deaths in the past 12 months is used to estimate the level and pattern of mortality by sex and age in countries that lack satisfactory continuous death statistics from civil registration. In order for estimates derived from this item to be reliable, it is important that deaths in the past 12 months by sex and age be reported as completely and as accurately as possible. The fact that mortality questions have been included extensively in the census questionnaire in the past decades has resulted in an improvement in the use of indirect estimation procedures for estimates of adult mortality (United Nations, 2017, para. 4.250).

806. Ideally, mortality should be sought for each household in terms of the total number of deaths in the 12-month period prior to the census reference date. In cases where it is not possible to obtain information on deaths during the past 12 months, it is advisable at least to collect data on the deaths of children under one year of age. For each deceased person reported, name, age, sex and date (day, month, year) of death should also be collected. For respondents, care should be taken to specify the reference period clearly so as to avoid errors due to its misinterpretation. For example, a precise reference period could be defined in terms of a festive or historic date for each country (United Nations, 2017, para. 4.251).

807. When any person of household members is dead in the last 12 months, it is suggested that for all persons information on name, age and sex, and day, month and year of death for persons should be collected. Countries not using dynamic imputation can assign "unknown" for each of these variables when invalid. Countries using dynamic imputation might use age (in age groups), sex and year of death as the dimensions of the imputation matrices for the other variables. The actual imputation matrices probably are country-specific, and the editing team including subject specialist such as demographer will have to work together to develop the appropriate imputation matrices. The population structure of the country, or sub-national geographic levels can aid in developing the most appropriate edit.

808. Entering each deceased person by his/her basic characteristics with CAPI or CASI may significantly improve the data quality, especially for the date of death. Also, assisting the enumerator and respondent for verifying information for all items during the interview will decrease unknowns. However, it can be still difficult to obtain the data at all in some countries and it may be necessary to use other information, like vital records, for developing a model for the edit program to obtain the most likely entries.

8. Cause of death

809. Some countries are now collecting information on cause of death for deaths in the 12 months before the census. Because of the sensitivity of the question, and sometimes because of the difficulty in obtaining the information in the field, countries may ask a question "Was the death due to accident or violence?" to obtain indirect information on HIV/AIDS or other epidemics for selected age groups. The edit for this item will usually be to assume that if the information is not collected, or is invalid, the value will normally become "unknown". If a country chooses to use imputation, a hot deck using sex and 0, 1-4 and then 5-year age groups, would be appropriate.

9. Maternal mortality

810. In the current census round, more and more countries are also asking if the person deceased person was female, whether she was pregnant at the time of her death. This item assists in determining maternal mortality at the national and regional levels. The edit for this item could require "unknown" status for invalid or blank entries. However, if a country choose imputation, the hot deck would be for females only, obviously, and only for ages of likely pregnancy – probably 12 to 54 – and probably by single year, rather than 5-year age groups.

811. Many countries now collect all three conditions separately – during pregnancy, at childbirth, or within 42 days of birth. Frequently, the enumerator checks two of them or all three, which is a biological impossibility. An edit needs to select one and only one. Usually, the subject matter expert in fertility will select an order. That order is usually the same order as the conditions above, so that "during pregnancy" because the default option. If other vital statistical data are available, they can be used to assist in the order.

812. The bigger problem is when the item is completely blank. In this case, the usual assumption is that she did not die a maternal death when it is blank. Also, some countries collect "maternal death" as a cause of death which can aid in the determination. An edit would need to be written to cover this.

10. Infant mortality (core topic)

813. Finally, the Principles and Recommendations (United Nations, 2017) suggests collecting information on deaths among children born "in the past 12 months". Normally, this question would only be asked in conjunction with the item on births in the 12 months before the census. If the other current fertility item – date of last birth is not used, then this item probably should not be used either.

814. Some countries are now asking for deaths in the 12 months before the census. When the household reports a death to a person less than one year old, this information can assist in ascertaining whether a death to a baby born in the 12 months before the census occurred.

815. Data on deaths to births in the year before the census assists those countries with good vital registration to check their infant mortality rates, at both the national and regional levels. Those countries without good vital registration can use the information to obtain estimates of infant mortality. Once again, a check between last births who have died in the year before the census and death to zero-year-olds in that year is an appropriate edit check and could provide useful information for checking infant mortality.

816. A traditional infant mortality rate cannot be obtained directly in a census. Births in the last year are over the period of the 12 months before the census. The deaths are to those children. But if a child is born the day before the census, it has only one day of exposure, not the 12 months needed to obtain the infant mortality rate. However, indirect measures can be used to obtain an estimate of the rate.

817. Edits for this item require some thought and will tend to be country-specific. Ideally, information on children ever born and surviving can be used to check the reported information; with a single adult female in the household, the check is relatively easy. With several females in the unit, care must be taken to make sure the right children are connected to the appropriate women.

11. Maternal or paternal orphan-hood and mother's line number

818. For the collection of information on orphanhood, two direct questions should be asked: (a) if the natural mother of the person enumerated in the household is still alive at the time of the census and (b) if the natural father of the person enumerated in the household is still alive at the time of the census. The investigation should secure information on biological parents. Thus, care should be taken to exclude adopting and fostering parents (United Nations, 2017, para. 4.256).

819. It is preferable for these questions to be collected from every person in the household regardless of age (not just children under 18, which would otherwise make the information useless for estimating adult mortality). Not only is this important for estimating mortality at older ages, but also for estimating the extent of age exaggeration at the older ages. Whenever the context allows, the date of death should be collected to help to improve knowledge of the

timing of death, and in other contexts a simple follow-up question about whether the parent was still alive five years ago can help to narrow down the timing of death and to improve adult mortality measurement for recent years by analysing these data as successive cross-sectional enquiries (United Nations, 2017, para. 4.257).

820. Very few countries collect information directly on the eldest child. For those who do, a program would go through the house looking for the eldest child and noting it on his or her record. This will not solve the problem of children in different houses.

821. The edits for "mother living" and "mother's line number" items are connected and should be carried out together. For persons who report other than "yes" for mother living, the mother's line number should be checked for a valid entry; if a valid entry appears, the code for "yes" should be assigned for mother living. For persons who report other than "yes" for mother living, mother's line number should be checked to see whether it is 00 or whether it equals the line number of a female with age greater than or equal to 12 years. If either of these cases is true, the program assumes the person has a mother and assigns yes to mother's vital status. If the entry in line number of mother is not valid and mother living is coded "no" or "does not know", the entry in mother's line number should be eliminated. In all other cases, the code for "does not know" should be assigned to mother living, and any entry in line number should be eliminated.

822. Persons reporting "Don't know" or "unknown" whether alive or dead, can be imputed on the basis of the previous person of the same age and sex with complete information for this item. Particularly in the time of HIV/AIDS, real differences in mortality will pertain in different parts of the country. The original data will be kept on the ends of the records for demographers wanting to do further analysis. The original data on orphanhood should be kept at the ends of the records for future demographic research.

823. The country might choose not to edit the mother's line number for persons who reported "no" or "does not know" if mother is living. In all other cases, the line number should be checked for consistency or should be assigned using relationship of person and line number, sex, relationship and age of person who was reported as mother. Where inconsistencies exist, or mother cannot be determined, the code for "living elsewhere" might be assigned.

824. Edit on entry with electronic data collection methods is particularly useful for mother's vital status and mother's line number because, for CAPI the enumerator, while in the housing unit, and for CASI the respondent, can clarify cases where conflicting information is entered. For example, if the enumerator keys in a mother's information that is for a male or for a female who is too young to have had that child, a message could appear for the enumerator or the respondent to elucidate the situation. In some cases, the child could be an adopted child reported as a biological child or the mother could be an adopting mother – in this case the enumerator may need to work with the respondents to correct the relationship reporting. The structure edits may need to be revisited if the head or the reference person is not the first person and is then moved to the first person. In that case, mother's line number may need to be adjusted for one or more people.

F. EDUCATIONAL CHARACTERISTICS

825. Most countries ask questions on current schooling and educational attainment to assist in determining the current education of the population, and to predict needs for additional schools and school rooms. Many countries also collect information on literacy – whether the individuals can read and write.

1. Ability to read and write (literacy) (core topic)

826. Data on literacy should be collected for all persons 10 years of age and over. In a number of countries, however, certain persons between 10 and 14 years of age may be about to become literate through schooling and the literacy rate for this age group may be misleading. Therefore, in an international comparison of literacy, data on literacy should be tabulated for all persons 15 years of age and over. Where countries collect data on younger persons, tabulations for literacy should at least distinguish between persons under 15 years of age and those 15 years of age and over (United Nations, 2017, para. 4.260).

827. Each country must establish the minimum age for literacy tabulations; similarly, editing teams must decide on the minimum age for literacy edits, since additional tabulations for internal use may be needed. As the questionnaire is being developed, the editing teams should decide the minimum age for collection and at what educational level the question no longer needs to be asked. Therefore, if the respondent has already reached a certain level of schooling, the enumerator may not need to ask the question about literacy. But, the item should be filled during edit to assist researchers and others with the public use data.

828. Persons at a defined level of schooling should be considered literate. In cases where an invalid code for literacy is found, a value should be assigned. The entry should be either "not stated" or determined using an imputation matrix based on specified variables, such as highest grade and sex. The "highest level" will depend on the particular country's definitions of what is "literate."

829. When electronic data collection methods are used, the edit on entry can do the edit described above. However, when the question on literacy comes before the question on educational attainment, the latter item could not be used to obtain a value, so the edit would have to occur after that item is entered. Therefore, it is better to ask the question on literacy after educational attainment. Also, these items can slow down the enumeration if there are many variables in addition to educational attainment such as age are considered.

2. School attendance (core topic)

830. In principle, information on school attendance should be collected for persons of all ages. School attendance relates in particular to the population of official school age, which usually ranges in general from 5 to 29 years of age but can vary from country to country depending on the national education structure. When data collection is extended to cover attendance for pre-primary education and/or other systematic educational and training programmed organized for adults in productive and service enterprises, community-based organizations and other non-educational institutions, the age range may be adjusted as appropriate (United Nations, 2017, para. 4.266).

(a) School attendance edit

831. Each country's editing team must decide which ages are appropriate for the collection of data on school attendance. Since most countries also divide schooling into several levels (according to ISCED classification), if these levels are going to be compiled by age, the specialists must also decide which age groups are appropriate for various levels of schooling. If the editing program produces inconsistent responses for the category, either the age or school attendance must be changed. Usually age is set by the time this edit is performed, so it is the school attendance that is changed. Enumerators should be instructed to omit school attendance for persons above a predetermined age, if appropriate for that particular country. In cases where persons continue in tertiary education into middle age, it may not be appropriate to set upper limits for school attendance. Presumably, responses and combinations of responses are tested prior to the census through pretests, so these decisions may be made before the actual census.

832. See the section below on age and educational attainment. For those in school, each grade or level will have a minimum and maximum age set by the subject matter specialists in the country. When electronic data collection methods are used, the minimum and maximum age for each grade can be checked on entry, and unusual cases can be presented in a message for further analysis. During the office computer edit, additional variables can be used to help

assure accuracy. The result in compilations will be empty triangles in two corners, and a column of official ages and grades going diagonally across the middle of the table.

(b) Full-time or part-time enrolment

833. Some countries may want to obtain information on part-time or full-time attendance in school. In this item is included, it may need to be part of the school attendance edit, or it may be a separate edit.

(c) Consistency between school attendance and economic activity

834. Consistency edits with other major items, such as major economic activity and reason for being outside the labor force, should be performed first. If attending school is one of the entries for being outside the labor force, and a person reported his or her major activity as going to school, the code for "yes" should be assigned to school attendance and the reason should be "student". That is, the responses should be consistent. In all other cases, any valid response should be accepted.

835. With electronic data collection, this edit can only be performed after the data on economic characteristics are entered. Hence, the enumerator or the respondent will key in the education information and then the economics data, and then the edit on entry will check for consistency between the items. Again, because enumerators have already gone on from education, they may have to back track to change the education items, if necessary, otherwise, this type of edit should be done during the office computer edit.

(d) Assignment for invalid or inconsistent entries for school attendance

836. If the entry is out of range and the entry in highest grade completed is valid, an entry should be assigned using an imputation matrix based on age, sex and highest grade. If highest grade does not have a valid code, then the entry in literacy should be used to assign school attendance. If literacy does not have a valid code, then an entry for school attendance should be assigned based on age and sex alone.

837. Imputation matrices may need to reflect the different patterns of school attendance by sex and age (sometimes by single year of age or small age groups). The edit would be the same for electronic data collection entry, but the edit probably should be performed after the whole record is keyed.

3. Educational attainment (highest grade or level completed) (core topic)

(a) Edit for educational attainment

838. The edit for educational attainment (highest grade or level) should consist of (a) a consistency check between a valid entry and age, and (b) imputation of an entry when the original entry is out of range. As mentioned above, in countries that do not use dynamic imputation, the value should be "not stated". In countries that use dynamic imputation, sex and single year of age will be needed for young persons, and sex and small age groups will be needed for slightly older children. In countries whose data include both highest grade and highest level, multiple imputation matrices may be necessary (for definition of grade, see United Nations, 2017, para. 4.273).

839. Some countries collect both current grade (or level) and grade or level attained. Sometimes this redundancy is helpful, particularly in countries in the process of changing their education systems. However, sometimes the redundancy presents a situation of incompatibility between the two entries with little information to determine the actual situation. Most countries develop a derived variable from current school attendance and educational attainment as a derived variable. See Annex I for suggestions for a recode for "current grade" based on school attendance and highest grade attended.

(b) Minimum age for educational attainment

840. Each country's editing teams must decide the minimum age for entering school. When the minimum age is set, the highest level completed ordinarily should not exceed a person's age plus some constant (which represents that minimum of age for entering school). Again, it is important to use single year of age for children since updating the imputation matrices may introduce errors if the age groups are very broad. Electronic data collection entry should notify enumerators/respondents if they enter an age that is too young.

(c) Relationship of age to educational attainment

841. The editing team must also decide how much noise will be allowed in the dataset. Usually it is better to change a few exceptional cases where age and educational attainment conflict, rather than accept many responses that are truly inconsistent. Therefore, for cases where the original entry is out of range or inconsistent with age, an entry can be assigned. For countries not using dynamic imputation, "not stated" can be entered. For those using dynamic imputation, an entry can be obtained based on age (including single year of age for persons of school age), sex and school attendance.

842. UNESCO recognizes literacy as separate from educational attainment, so "ability to read and write" should probably not be used as a value in the imputation matrix.

843. The relationship between age and education – both for level of current schooling and level of educational attainment – are very important. Usually at least two imputation matrices are needed, one for current schooling and one for educational attainment. The current schooling edit would need to look at valid ages for each level (usually grade) of schooling, so a table crossing education by age should see blank triangles on either side of an angled column in between. The compiled table for education completed would have a single blank triangle for those too young for the educational level or grade.

4. Field of education and educational qualifications

844. Information on persons by level of education and field of education is important for examining the match between the supply and demand for qualified manpower with specific specializations within the labor market. It is equally important for planning and regulating the production capacities of different levels, types and branches of educational institutions and training programmes (United Nations, 2017, para. 4.281).

845. Persons who are younger than 15 (or other predetermined age) should not have information about field of education and/or educational qualifications. For persons 15 years and over, a relationship should exist between the level of educational attainment and the field of education and/or educational qualifications. In each case, when invalid entries occur, countries not using dynamic imputation can make the entry "unknown". Countries using dynamic imputation might want to consider using age, sex, educational attainment and, possibly, occupation to assign field of education and/or educational qualifications.

846. For electronic data collection, a look up list of likely qualifications could be embedded in the application and used as needed to obtain the appropriate code. Since all of the economic variables, including occupation, are collected after education, it is possible to check the consistency between economic variables, such as occupation and education characteristics. However, it is important to ensure these procedures do not slow the data collection and not subject to errors as the enumerator or the respondent moves back and forth across the record to change the entered data.

847. Most countries will use a code list derived from the occupation code list for simplicity in coding. Sometimes only broad categories of study are obtained, like "science", so general as well as specific codes should be considered.

G. ECONOMIC CHARACTERISTICS

1. Introduction

848. The United Nations has made changes in how work is defined, and with employment and unemployment, paid and own production work. Hence, analysts and programmers should work together to learn the new concepts, to implement them, and to make sure overlap between censuses and surveys occurs so that some of the trends can be determined (United Nations, 2017, para. 4.289-4.333).

849. Information on economic activity status should in principle cover the entire population, but in practice it is collected for each person at or above a minimum age, set in accordance with the conditions in each country. The minimum school-leaving age should not automatically be taken as the lower age-limit for the collection of information on activity status. Countries in which, normally, many children participate in agriculture or other types of economic activity (for example, mining, weaving and petty trade) will need to select a lower minimum age than that in countries where the employment of young children is uncommon.

850. Tabulation of economic characteristics should at least distinguish persons under 15 years of age and those 15 years of age and over; countries where the minimum school-leaving age is higher than 15 years of age and where there are economically active children below this age should endeavor to secure data on the economic characteristics of these children with a view to achieving international comparability at least for persons 15 years of age and over. The participation in economic activities of elderly men and women after the normal age of retirement is also frequently overlooked (United Nations, 2017, para. 4.306).

851. Each country must determine a minimum age for participation in economic activity. Countries interested in collecting data on child labor may need to choose a low minimum age but must remember that some noise will occur when children who are not in the labor force are erroneously enumerated as being in the labor force. After the minimum age is established, the items of economic activity, are edited to be tabulated for persons X years or older; therefore, editing for children under X years old will be necessary only to make certain that all entries are blank. In order to facilitate tabulations, any responses that may have been entered for children under age X should be eliminated.

2. Conceptual framework for work statistics

852. The United Nations Principles and Recommendations, Revision 3, para. 4.294 defines work:
Measurement of the economic characteristics of the population is based on the conceptual framework for work statistics (see the, box 4.1 in the Principles and Recommendations). In this framework, work is defined for reference purposes as “any activity performed by persons of any sex and age to produce goods or to provide services for use by others or for own use”.

853. The conceptual framework for work statistics identifies five mutually exclusive forms of work for separate measurement (United Nations, 2017, para. 4.297):

(a) Own-use production work, comprising production of goods and services for own final use;
(b) Employment work, comprising work performed in exchange for pay or profit;

(c) Unpaid trainee work, comprising work performed for others without pay to acquire workplace experience or skills;
(d) Volunteer work, comprising non-compulsory work performed for others without pay;

(e) Other work activities, including unpaid compulsory work performed for others, such as community service and work by prisoners, when ordered by a court or similar authority, and unpaid military or alternative civilian service.

854. To meet different objectives, countries may measure the economic characteristics of the population with respect to their participation in one or in several forms of work. In particular, in the population census, this may include measurement of the following (United Nations, 2017, para. 4.299):

(a) *Persons in employment* is essential as part of the preparation of labour force statistics that include unemployment and other measures of labour underutilization. It is needed to assess the labour market participation of the population and to classify the population according to their labour force status in a short reference period.

(b) *Persons in own-use production of goods* is especially important in countries where particular groups of the population engage in agriculture, fishing or hunting and gathering for own final consumption, including for subsistence, and to enable integration of the population census with the agricultural census.

(c) *Persons in unpaid trainee work* may be advisable where unpaid apprenticeships, internships and traineeships may be a main mechanism of labour market entry for particular groups such as youths or for specific occupations such as mechanics or tailors, given their likely overall small size in the country and limited availability of alternative statistical sources.

855. This definition of work differs from previous definitions, and of economic activities, so analysts must take care to obtain statistics under the current definition as well as making sure that at least some trends can be obtained.

856. Some people only do "own production" which the UN considers work. In the past, many of these people were placed in the category "in employed persons". But, the new suggested variables now include work in "own use production of goods (United Nations, 2017, paras. 4.376-4.381)."

857. The United Nations Principles and Recommendations, Version 3 suggest data collection for current activity. As a result, the labour force status of persons is established with regard to a short reference period of seven days or one week prior to the census reference date (United Nations, 2017, paras 4.307-4.311).

3. Labour force status (core topic)

858. Labor force status is determined by several economic variables, and comprises three categories: employment, unemployment and outside the labor force. These categories of labour force status are mutually exclusive and exhaustive. For establishing labour force status, priority is given to employment over other forms of work, and over employment; and to unemployment over outside the labor force (United Nations, 2017, para. 4.308).

859. As discussed, "current activity status" is the relationship of a person to economic activity, based on a brief reference period such as one week or seven days. The use of current activity is considered most appropriate for countries where the economic activity of people is not greatly influenced by seasonal or other factors causing variations over the year. This one-week or one-day reference period may be either a specified recent fixed week, the last complete calendar week or the last seven days prior to enumeration (United Nations, 2017, para.309).

(a) Categories related to labour force

860. For identifying the labor force status, usually a number of questions is asked for collecting enough information for each category of labor force status:

(1) Employed persons

861. According to the United Nations (2017 para. 4.313) the employed comprise all persons above a specified age who, during a short reference period of either one week or seven days, are in one of the following categories:

(a) Paid employment:

(i) At work: persons who during the reference period performed some work for wage or salary, in cash or in kind;

(ii) With a job but not at work: persons who, having already worked in their present job, were temporarily not at work during the reference period and had a formal attachment to their job as evidenced by, for example, continued receipt of wage/salary, an assurance of return to work following the end of the contingency or an agreement on the date of return following the short duration of absence from the job.

(b) Self-employment:

(i) At work: persons who during the reference period performed some work for profit or family gain, in cash or in kind;

(ii) With an enterprise but not at work: persons with an enterprise, which may be a business enterprise, a farm or a service undertaking, who were temporarily not at work during the reference period for some specific reason.

(2) Unemployed population

862. The unemployed population comprises, according to the United Nations (2017, para. 4.321), all persons above a specified age who, during the reference period, met the following conditions:

(a) Not in employment: they were not in paid employment or self-employment;

(b) Currently available for work: they were available for paid employment or self-employment during the reference period;

(c) Seeking work: they took specific steps in a specified recent period to seek paid employment or self-employment. The specific steps may have included registration at a public or private employment exchange; application to employers; checking at work¬ sites, farms, factory gates, markets or other places of assembly; placing or answering newspaper advertisements; seeking the assistance of friends and relatives; looking for land, building, machinery or equipment to establish one's own enterprise; arranging for financial resources; and applying for permits and licenses. It is useful to distinguish first-time job-seekers from other job-seekers in the classification of the unemployed.

863. The *edits for unemployment* — "on layoff", "looking for work", whether the person could take a job, and "year last worked" (if present) — should be done jointly. Also, they need to be compatible with the response for economic activity and, in most cases, should not be filled if the items for time worked, industry, occupation, class of worker, and place of work are filled. If the subject-matter specialists determine that an entry is needed for "on layoff" when the response is either blank or invalid, then an imputation matrix using age and sex, and perhaps educational attainment of the person, could be implemented.

864. The *edit for "looking for work"* should be done jointly with the edit for "on layoff" and "why not looking for work". Subject-matter personnel should develop edits using entries for these items to impute the other items. The edit should consider local and regional conditions as well as census or survey variables.

(3) Persons outside the labor force

865. The population that is "not currently active" comprises all persons not classified either as employed or as unemployed. It is recommended that persons outside the labor force may be classified by their main activity or reason for not entering the labour market into the following groups (United Nations, 2017, para.4.332):

(a) *Attending an educational institution* refers to persons outside the labour force, who attended any regular educational institution, public or private, for systematic instruction at any level of education, or were on temporary absence from the institution for relevant reasons corresponding to those specified for employed persons "not at work".

(b) *Performing unpaid household services* refers to persons outside the labour force engaged in the unpaid provision of services for their own household, such as spouses and other relatives responsible for the care and management of the home, children and elderly people. (Domestic and personal services provided by domestic employees working for pay in somebody else's home are considered as employed in line with paragraph 4.312 above).

(c) *Retiring on pension or capital income* refers to persons outside the labour force who receive income from property or investments, interests, rents, royalties or pensions from former employment.

(d) *Other reasons* refers to all persons outside the labour force who do not fall into any of the above categories (for example, children not attending school, those receiving public aid or private support and persons with disabilities).

866. The edits for "not currently active" have been incorporated into the above edits for economic activity.

867. This item should be edited only for persons who were recorded as "not looking for work"; all others should have a blank entry. Alternatively, if a valid entry appears in occupation, industry and status in employment, the code for "with a job but not at work" should be entered. This code designates economically active persons who were employed but were not at work during the reference period. In all other cases, if dynamic imputation is not used, "unknown" can be assigned. For countries using dynamic imputation, an entry can be allocated using age, sex and major activity.

(b) Editing for labour force status

868. As discussed, labour force status generally has the following categories:
(1) Employed, at work;
(2) Employed, not at work;
(3) Self-employed, at work;
(4) Self-employed, not at work;
(5) Looking for work;
(6) Student;
(7) Homemaker;
(8) Pension or capital income recipient;
(9) Other not in the labor force.

869. For this variable, the first five possibilities are for persons who are in labour force, and the second four categories are for persons who are outside the labour force. Persons who are "at work" (categories 1 and 3) and those who are "not at work" (categories 2 and 4) are employed.

(i) Employed persons

870. If one of the categories for active persons is selected (categories 1 to 4), the variables for time worked, occupation, industry, status in employment, and work place should be filled. If they are not filled, they should be edited and filled, either as unknowns, or with cold deck values or hot deck values. If a category from 1 to 4 is selected, the variables for on layoff, looking for work and year last worked should be blank. If they are filled, they should be changed to BLANK. For electronic data collection methods, the edit on entry can be applied according to the sequence of the questions. Common approach is first to identify labour force status of a person. If a person is employed, then automatically the related questions on occupation, industry, work place and so on will be asked. If a person is not employed but looking for a job, other following questions will be asked to clarify whether a person is unemployed or outside the labour force. And finally, if a person is outside the labour force the question on reason for not working will be collected. The use of CAPI and CASI can improve the data quality if a series of questions related to these three major groups of labor force status are properly designed without considering the space limitations as it exists for paper questionnaire. There should be an effort by subject specialists for asking additional questions for making clear the status of every individual.

(ii) Unemployed persons

871. If category of "looking for work" (5) is selected, the variables for "seeking work", "availability for work" and, if included, "year last worked" should be filled. If they are not filled with valid entries, they should be edited and filled, either as "unknowns", or with cold deck or hot deck values. If categories 5 through 9 is selected, the variables for time worked, occupation, industry, economic activity status, and work place should be blank. If they are filled, they should be made BLANK.

(iii) Students and retired persons

872. If category 6, student, is selected, the subject matter personnel need to decide whether the entry for the variable for school attendance must be "yes, in school". If category 8, pensioner, is selected, the subject-matter personnel need to decide whether persons must be of a certain age to be retired.

(iv) When labour force status is not valid and employed variables are reported

873. If the entry for labour force status is not valid, for example the person is retired or student, and if some of the variables for time worked, occupation, industry and workplace are reported, the respondent's status should be coded with a value from 1 to 4. An imputation matrix will probably be needed to select the appropriate response.

(v) When labour force status is not valid, and the unemployed variables are reported

874. If any of the variables for "looking for work", "available for work", "seeking work" and, if asked, "year last worked" are reported, the entry for economic activity should be coded with a value from 5 to 9. If the person is attending school, that value should probably be 6. If the person is elderly, the value should probably be 8. Otherwise, the subject-matter specialists may decide to use an imputation matrix to allocate an appropriate response.

(vi) When labour force status is not valid and none of the economic variables are reported

875. If no response appears for any of the labour force status items, the subject-matter specialists should consider using imputation matrices to determine the most appropriate response and then impute the other economic items.

(vii) Derived variables needed for development of "Economic Status Recodes"

876. As noted in the Principles and Recommendations and elsewhere in this volume, labour force status can be identified based on a number of questions asked to obtain information on whether a person is employed, unemployed or outside the labour force. Derived variables for labor force status can be generated in the data records to perform

the editing process systematically. All variables need to be edited before generating the derived variables. (See Annex III on Derived Variables).

4. Status in employment (core topic)

877. Status in employment refers to the status of a person with respect to his or her employment, that is to say, the type of explicit or implicit contract of employment with other persons or organizations that the person has in his/her job. The basic criteria used to define the groups of the classification are the type of economic risk, an element of which is the strength of the attachment between the person and the job, and the type of authority over establishments and other workers that the person has or will have in the job. Care should be taken to ensure that an employed person is classified by status in employment based on the same job(s) as used for classifying the person by occupation, industry and sector (United Nations, 2017, para.4.339).

878. The economically active population should be classified by status in employment (United Nations, 2107, para. 4.340), as follows:

(a) Employees, among whom it may be possible to distinguish between employees with stable contracts (including regular employees) and other employees;
(b) Employers;
(c) Own-account workers;
(d) Contributing family workers;
(e) Members of producers' co-operatives;
(f) Persons not classifiable by status.

879. Owner-managers of incorporated enterprises, who would normally be classified among employees, but whom one may prefer to group together with employers for certain descriptive and analytical purposes should be identified separately.

880. This item should be edited only for persons who are “employed” or “self-employed”. If dynamic imputation is not used, blank, zero or invalid responses can be changed to “not reported”. If dynamic imputation is used, minimal variables for the imputation matrix include age groups and sex, but other variables such as educational attainment or industry major categories can also be used.

5. Occupation (core topic)

881. Occupation refers to the type of work done during the time-reference period by the person employed (or the type of work done previously, if the person is unemployed), irrespective of the industry or the status in employment in which the person should be classified (United Nations, 2017, para. 4.352).

882. This item should be edited only for persons who are “employed” or “self-employed” and if a question on previous work experience is asked, this question can be edited for persons who have worked before but currently are unemployed. If dynamic imputation is not used, blank, zero or invalid responses should be changed to “not reported”.

883. Codes for occupation tend to be developed so that different digits represent major and minor occupation codes. Open-ended question (write-ins), which is almost unavoidable for occupation when the paper questionnaire is used, will add to the coding burden.

884. If dynamic imputation is used, minimal variables for the imputation matrix include age groups and sex, but other variables such as educational attainment or industry major categories can also be used. For electronic data collection, look up lists will assist in obtained the most specific occupation based on information about the job, so is an advantage over an office edit which would have much less information. However, some countries may prefer to

design this question in open-ended format to type in a series of characteristics of an occupation and then perform coding after data collection.

6. Industry (core topic)

885. According to the United Nations (2017, para. 4.356) "industry refers to the activity of the establishment in which an employed person worked during the time-reference period established for data on economic characteristics (or last worked, if unemployed). For guidance on the selection of the job/activity to be classified, see paragraph 4.357 in Principles and Recommendations.

886. This item should be edited only for persons whose economic activity was "employed" or "self-employed". If dynamic imputation is not used, blank, zero or invalid responses should be changed to "not reported".

887. Codes for industry tend to be developed so that different digits represent major and minor industry codes. Open-ended format of this question, which are almost unavoidable for this item, will add to the coding burden. As with occupation, embedded lookup codes for industry will assist in obtaining the best selected industry. If dynamic imputation is used, minimal variables for the imputation matrix include age groups and sex, but other variables such as educational attainment or occupation major categories can also be used.

888. Some jobs and establishments are primarily public sector and others private sector. This information could also be incorporated into the dynamic imputation.

889. These larger imputation matrices are sometimes difficult to manage and must be thoroughly checked out before the census edit. If mistakes are made in large – 4 or more variables – matrices, the results could be skewed and harder to interpret. As noted earlier, large matrices should be avoided.

7. Place of work

890. "Place of work" of persons in employment includes two main topics, the "*type of workplace*" and its "*geographic location*". While the information on place of work can be used to develop area profiles in terms of the employed labor force (as opposed to demographic profiles by place of residence), the primary objective is to link place-of-work information to place of residence (United Nations, 2017, para.4.361).

891. Since "place of work" is used for statistics on commuting, it is important for any changes to the reported information to reflect the specific geographical areas considered. Hence, country editing teams may want to consider assigning "unknown" for invalid cases, and analyses only the "known" cases.

892. Coding operations for this item will increase in time and complexity if write-ins are accepted and must be coded. If a hierarchy is determined for the digits, for example, the first digit representing the province, the second the district and so on, the coding operation will probably be more efficient and more accurate.

893. For imputation matrices, the data processors need to make certain that only likely geographic places are assigned to the matrices. It may be wise to start a new cold deck for each civil division or other geographical area to make certain that previous values cannot be selected. For the imputation matrices themselves, age and sex, and perhaps modified major occupation or industry major categories, can be included. Also, different imputation matrices may be needed for work inside and outside the country.

8. Institutional sector

894. The Institutional sector of employment relates to the legal organization and principal functions, behavior and objectives of the enterprise with which a job is associated (United Nations, 2017, para. 4.366).

895. A relationship exists between some of the possible industries and occupations and the institutional sector of employment (corporation, Government, nonprofit, household or other). Some countries may choose to check for these relationships among the variables to make certain that tabulations do not show inconsistencies when these variables are cross-tabulated.

896. For the edit, countries not using dynamic imputation will have to assign "unknown" for the institutional sector when it is not known. Countries using dynamic imputation should consider using age and sex, and perhaps major industry or occupation of similar persons in the geographical area.

9. Working time

897. Information on two distinct concepts of working time can be collected in a population census: *hours actually worked* and *hours usually worked*. Measurement of hours actually worked in employment, in the context of the population census, is usually collected using one direct question; it is optimally measured using a set of questions, requesting hours separately for each day of the week. For employed persons not at work in the short reference period, it is possible to have a value for hours actually worked of zero (for persons away on leave) or reduced (if a part of the reference period was taken off for sickness, holiday, or other purpose) (United Nations, 2017, paras 4.371-4.375).

898. Time worked should also include time spent in activities that, while not leading directly to produced goods or services, are still defined as part of the tasks and duties of the job, such as preparing, repairing or maintaining the workplace or work instruments. In practice, it will also include inactive time spent in the course of performing these activities, such as time spent waiting or standing by, and in other short breaks. Longer meal breaks and time spent not working because of vacation, holidays, sickness or industrial disputes should be excluded. (United Nations, 2017, para. 4.372).

899. This item should be edited only for persons whose response for economic activity was "employed, or "self-employed". For some countries, time worked should also be included for homemakers. Categories that are predetermined by the editing team should be accepted. If dynamic imputation is not used, blank, zero or non-numeric codes should be changed to "not reported", and the subject-matter specialists might want to change the economic activity variable to "not working", if reported hours equal zero.

900. If dynamic imputation is used, the minimal variables for the imputation matrix includes age groups and sex, but other variables such as educational attainment, occupation or industry major categories can also be used.

10. Participation in own use production of goods (core topic)

901. The Principles and Recommendations for Population and Housing Censuses Revision 3 has added a new suggested core item for the 2020 Round censuses, taking into consideration the new definition of employment. Countries where production of goods for own final use (such as foodstuffs from agriculture, fishing, hunting and gathering, water, firewood and other household goods) represents an important component of the livelihood of a part of the population, whether as a main or secondary activity, will need to consider collecting information in the population census on the number of persons engaged in this form of work (previously included within the concept of employment). Such information is essential for benchmarking purposes, especially where household surveys are not frequent, for comprehensive sectoral analysis, particularly of work in agriculture, forestry and fishing, and to enable integration of the population census with the agricultural census (United Nations, 2017, para. 4.376).

902. Persons in own-use production of goods are all those above a specified age who, during a specified reference period, performed "any activity" to produce goods for own final use. The notion of "for own use" is interpreted as production where the intended destination of the output is mainly for final use by the producer in the form of capital formation, or final consumption by household members, or by family members living in other households (United Nations, 2017, para. 4.377).

903. This item is collected independently of paid or unpaid labor for others or voluntary work. So, individuals could be doing work other than producing and/or processing for own use, as well as for own use. Hence, in editing the data, the independence of the two items should be maintained.

904. However, the edit should consider age, schooling, and other economic activities. Unless the relationship between these variables and own production activities are inconsistent – for example a person living in a high rise in the middle of a large city – the response, if valid, should be accepted. But even in the case where potential inconsistencies might seem present, a person growing herbs in a window box might be producing for own use, and so the response might be accepted.

905. In cases of clear inconsistencies, or when the item is blank, countries not using imputation, could assign a value for "unknown" or "not stated".

906. When imputation is used, several possibilities should be considered. If the housing unit is clearly in a rural area, persons are much more likely to be doing own production than in highly urban areas. So, if someone in the unit is recorded as doing own production, others in the unit could also be assumed to be doing own production as well, depending on age, sex, current schooling, and other economic activities.

907. But, if no one in the housing unit is doing own production – that is, the item is blank for everyone – then the edit should probably consider imputing data from the head of household or reference person on the basis of age and sex, and possibly other economic activities. Once the information is obtained for the first person in the unit, invalid responses for other people in the unit could be imputed with the same value for own production.

908. The danger in this method is that every unit should have some source of livelihood; producing for own use is one course of livelihood. If the head of household or reference person is very aged, he or she might be imputed as not doing own production activities, but others in the housing unit must be producing in some way for the unit to survive as a unit. Hence, the edit must take demographic variables into account and a hot deck using relationship to head, age and sex might be used to obtain an appropriate response.

909. So, the general edit would be to accept a response when one is provided, but to impute the information for others in the unit when the head of household or reference person has "own production" filled. When the head of household must be imputed, others would be provided the same response as the head of household or be imputed themselves based on age and sex.

11. Income

10. The census topics relating to economic characteristics of the population presented in Principles and Recommendations for Population and Housing Censuses, Revision 3 focus on the economically active population as defined in the recommendations of the International Labour Organization (ILO), where the concept of economic production is established with respect to the System of National Accounts 2008 (SNA) (United Nations, 2017, para. 4.296).

11. Within this framework, income may be defined in terms of (a) monthly income in cash and/or in kind from the work performed by each active person or (b) the total annual income in cash and/or in kind of households regardless of source. Collection of reliable data on income, especially income from self-employment and property income, is extremely difficult in general field inquiries, and particularly for population censuses. The inclusion of non-cash income further compounds the difficulties. Collection of income data in a population census, even when confined to cash income, presents special problems in terms of burden of work and response errors, among other concerns. Therefore, this topic, including the broader definition of income, is generally considered more suitable for use in a sample survey. Depending on the national requirements, countries may nonetheless wish to obtain limited information on cash income. As thus defined, the information collected can provide some input into statistics on the distribution of income, consumption and accumulation of households, in addition to serving the immediate purposes of the census.

12. Principles and Recommendations identifies two types of income: individual income and household income. Both items require similar edits. For individual income, if dynamic imputation is not used, invalid income responses should be assigned "not stated" or "unknown". If dynamic imputation is used, age, sex, educational attainment, industry, occupation and other qualifiers might be used to form the imputation matrix for income.

13. Household income is the sum of all income earned by the household and is entered on the housing record. The edit with dynamic imputation is about the same, but, using age, sex, and level of educational attainment of the head of household, rather than that of each individual. See further discussion of household and family income recodes in Annex III.

H. AGRICULTURE

14. Some countries may want to use the population census to identify households engaged in own-account agricultural production. This information is useful for agriculture-related analysis of the population census and for use as a frame for a subsequent agricultural census or survey. In this case, information should be collected for all households on whether any member of the household is engaged in any form of own-account agricultural production activities.

1. Introduction

15. Where possible, information should be collected to identify whether the household is engaged in any form of own-account agricultural production, covering the main agricultural activities important to the country (which can include crops, livestock and related activities). Information may also be collected on forestry, fishery and aquaculture activities in case they are important for a country.

16. Where aquaculturally production is important at the household level, information can also be collected on whether or not any member of the household is engaged in any form of own-account aquaculturally production activities.

2. Own-account agriculture production

17. To overcome the problem of someone having both commercial and own-account agriculture, information should be collected on all persons who carried out agricultural activities during the year preceding the population census day. The information to be collected should include the occupation and status of employment of all agricultural jobs and could be expanded to cover working time and whether the job was performed as a main or secondary activity. Given the newly adopted conceptual framework for work statistics, information should also be collected on participation in own-use production of agricultural goods, particularly in countries where subsistence agriculture is practiced by part of the population (see United Nations, 2017, paragraphs 4.387-4.396).

18. Information on occupation and status in employment of all agricultural jobs (main and secondary), and on participation in own-use production of agricultural goods, can be used as an alternative way to facilitate identification of households engaged in own-account agricultural production activities (see United Nations, 2017, paragraphs 4.389–4.392).

3. Characteristics of all agricultural activities during the last year

19. An agricultural job or work activity is defined as a job or work activity in the agricultural industry as defined (ISIC Rev. 4.0):

Group 011: Growing of non-perennial crops
Group 012: Growing of perennial crops
Group 013: Plant propagation
Group 014: Animal production
Group 015: Mixed farming.

20. Since agriculture is collected for individuals in the occupation and industry questions, as noted in the Principles and Recommendations, an edit must control the consistency between the two topics.

21. When dynamic imputation is not used, then unknowns can be put in the own-account agriculture items. But, when dynamic imputation is used, age, sex, and regular economic information (e.g., occupation, economic status, employment status) could be used to obtain a best guess for the information. When certain areas of the country are more likely to have own-account agriculture than others, this factor (e.g., urban vs rural areas) might be considered as well.

VI.HOUSING EDITS

A. INTRODUCTION

22. The specifications for housing edits check the validity of individual items as well as consistency between items. Knowledge of specific relationships among items for a given country makes it possible to plan consistency edits to assure higher quality data for the tabulation. For example, a housing unit should not have a cement roof when the walls are constructed of bamboo. Similarly, units should have piped water inside the house in order to have a flush toilet or a bathtub or shower inside the structure.

23. The edits described below are for occupied housing units. However, vacant housing units and occupied housing units sometimes have different characteristics and will not use the same edits. The national census/statistical office editing team will need to develop different edits for each type of unit when, as is usually the case, not all housing items are collected for vacant housing units. The editing team will need to pay attention to the imputation matrix variables since these are most likely to differ.

24. As with population items, for missing or invalid items the editing team must decide whether to assign "not stated," a static imputation (cold deck) value for "unknown" or other value, or a dynamic imputation (hot deck) value based on the characteristics of other housing units. As before, in many cases, dynamic imputation is preferred since it eliminates the kind of imputation required at the tabulation stage, when only the information in the tabulations themselves is available to make decisions about the unknowns. The imputation matrices thus established supply entries for blanks, invalid entries, or resolved inconsistencies when no other related items with valid responses exist.

25. Some countries may have some variation in housing characteristics across the nation, but very little within most localities. Other countries may have considerable variation for particular items between localities, particularly urban and rural areas. This variation must be considered when developing imputation matrices, and particularly for the initial cold deck values.

26. Except when a country lacks housing information for collective (group) quarters, one (and only one) housing record should be assigned to each serial number (see "Structure edits" chapter IV). The chapter on structure edits outlines a series of quality assurance procedures. Depending on the decisions of the editing team, the editing program can create a housing record if it is missing. Similarly, the program can remove one or more records when duplicate or multiple records occur.

27. Ideally, each housing record should be edited selectively for applicable items only. The edited items may differ depending on urban/rural, climatic, and other conditions. However, in practice few countries have the time or expertise to develop and implement multiple arrays to change missing or inconsistent data. Even fewer countries implement selective editing.

28. Nonetheless, for aesthetic, more than for technical reasons, and particularly for housing items, as editing has become more sophisticated and detailed, more emphasis is now put on making sure selected geographical areas have only "appropriate" responses. For example, if certain geographical areas of a country do not have electricity, they also should not have air conditioners, electric refrigerators or electric stoves. An edit can be written to address issues like these in certain geographic areas to make sure that no anomalies slip through into the final data set. The best approach is probably to remove cases that may actually be extraneous.

29. The information collected on the questionnaire will also depend on the type of living quarters (housing unit or group quarters) and whether the housing unit was vacant or occupied. For collectives or group quarters, the edit

can be limited to only those items collected at group quarters or those collected at both group quarters and other housing units.

30. By definition, housing records do not exist for homeless persons. If these records do exist because the country chooses to have identifiers for them, the country may treat such records in the same manner as those for collective quarters, or it may require a completely different edit, or none at all.

31. Sometimes a "not reported" entry should be allowed for a particular item. This may occur when the country's editing team lacks a good basis for imputing responses for a given characteristic. The decision to leave "not reported" responses must be balanced against the requirement to produce appropriate, tabular characteristics for planning and policy use. When planners need selected information, as long as the "not reported" cases have the same distribution as the reported cases, allocating the "not reported" cases should pose no problem.

32. If the "not reported" cases are somehow skewed, however, the post-compilation imputation could be problematic, particularly for small areas or particular types of conditions. For example, respondents living in country-defined "substandard" housing may refuse to reveal some of their housing characteristics. If the enumerator does not report them, planners may not be able to introduce remedial programs to alleviate the substandard conditions.

33. Housing edits tend to be simpler than population edits because cross-tabulations are generally much less complicated. Most countries compile individual housing characteristics only by various levels of geography. As indicated above, countries choosing not to use dynamic imputation should determine an identifier for "unknown" to use when invalid or inconsistent responses occur.

34. For countries that use dynamic imputation, the editing team should develop simple imputation matrices with dimensions that differentiate housing characteristics. For most countries a variable on "type of living quarters", whether housing unit or collective living quarters, including type of unit within these categories, is the best primary variable for dynamic imputation.

35. For some countries, geographical areas can be used as one dimension of these imputation matrices. Tenure can also be used. For example, if the country has about half its units rented and half owned, tenure is suitable for inclusion as one of the dimensions of the imputation matrix. However, if only 5 per cent of the units are rentals, some other characteristic would be more appropriate. Tenure is often a useful variable to use in imputation matrices, particularly in countries having large percentages of the major types of tenure. Other characteristics to consider include the type of walls and the presence of electricity.

36. For each country, the particular variables included as dimensions of the imputation matrices must correspond to the variables in the dataset, so for the housing items, care must be taken that the individual items as well as the combinations of items distinguish among the characteristics.

37. This chapter looks at the housing variables recommended in the *Principles and Recommendations for Population and Housing Census, Revision 3*. No country should be using all of these variables. The selected variables and their relationships with the other variables should be thoroughly tested in pre-filed and field tests for reliable and complete responses. Housing variables are important for their own use as parts of a wealth index to assess well-being in all or parts of a country.

B. CORE AND ADDITIONAL TOPICS

38. The units of enumeration in housing censuses are (a) buildings; (b) living quarters; and (c) occupants of living quarters. The United Nations has developed a list of basic editing topics of general interest and value that are also of importance in enabling comprehensive statistical comparisons at the international level. For the convenience of the

users, suggested codes for these and a number of additional topics are given below. The topics are shown by type of units of enumeration.

39. Most of the housing topics are not cross-tabulated with other housing characteristics, except that they are compiled by geography. Hence, edit-on-entry with the use of electronic data collection methods will go smoothly. But, a few inconsistencies should probably cause a message, with the use of electronic data collection an application. For example, housing units that do not have electricity should not have air conditioning. Similarly, units with thatch walls probably should not have concrete roofs. Only those items below that could require a different approach with electronic data collection will be noted as such.

1. Living quarters: type of living quarters (Core topic)

40. The classification outlined below describes a system of three-digit codes designed by the United Nations (2017, paras. 4.421-4.462) to group in broad classes housing units and collective living quarters with similar structural characteristics. The distribution of occupants (population) among the various groups supplies valuable information about the housing accommodations available at the time of the census. The classification also affords a useful basis of stratification for sample surveys. The living quarters may be divided into the following categories:

1 Housing units
- 1.1 Conventional dwellings
  - 1.1.1 Has all basic facilities
  - 1.1.2 Does not have all basic facilities
- 1.2 Other housing units
  - 1.2.1 Semi-permanent housing units
  - 1.2.2 Mobile housing units
  - 1.2.3 Informal housing units
  - 1.2.4 Housing units in permanent buildings not intended for human habitation
  - 1.2.5 Other premises not intended for human habitation

2 Collective living quarters
- 2.1 Hotels, rooming houses and other lodging houses
- 2.2 Institutions
  - 2.2.1 Hospitals
  - 2.2.2 Correctional institutions (prisons, penitentiaries)
  - 2.2.3 Military institutions
  - 2.2.4 Religious institutions (monasteries, convents, etc.)
  - 2.2.5 Retirement homes, homes for elderly
  - 2.2.6 Student dormitories and similar
  - 2.2.7 Staff quarters (e.g., hostels and nurses' homes)
  - 2.2.8 Orphanages
  - 2.2.9 Other
- 2.3 Camps and workers' quarters
  - 2.3.1 Military camps
  - 2.3.2 Worker camps
  - 2.3.3 Refugee camps
  - 2.3.4 Camps for internally displaced people
  - 2.3.5 Other
- 2.4 Other

41. Editing teams should develop edits that make certain that all collective living quarters and housing units have internally consistent information. If the value for type of living quarters is unknown or invalid, editing teams might want to develop an edit that looks at other variables to assign type of living quarters. Otherwise, if the value is invalid,

"unknown" should be assigned when dynamic imputation is not used. National statistical/census offices choosing dynamic imputation for invalid values should use at least two characteristics, such as type of building, tenure, number of rooms, floor space or vacancy status, to obtain "known" information from similar housing units in the geographical area.

2. Living quarters: Location of living quarters (Core topic)

42. Location of living quarters is a geographical variable and is presented with the structure edits in Chapter IV.

3. Occupancy status (Core topic)

43. Occupancy status refers to whether or not a conventional dwelling is occupied at the time of the census. For those dwellings not occupied, countries should collect data for the reason for not being occupied (such as vacant or in secondary use) (United Nations, 2017. Para. 4.471).

44. The decision to record conventional dwellings whose occupants are temporarily absent or temporarily present as "occupied" or "vacant" will need to be considered in relation to whether the census is conducted based on usual resident population or present population . In either case, it would seem useful to distinguish as far as possible dwellings used as a primary residence from those that are used as a second residence. This is particularly important if the second residence has markedly different characteristics from the primary residence, as is the case, for example, when persons in agricultural households move during certain seasons of the year from their permanent living quarters in a village to rudimentary structures located on agricultural holdings (United Nations, 2017, para. 4.471-4.475). The recommended classification for conventional and basic dwellings is as follows:

1. Occupied
2. Vacant
   - 2.1 Seasonally vacant
     - 2.1.1 Holiday homes
     - 2.1.2 Seasonal workers' quarters
     - 2.1.3 Other
   - 2.2 Non-seasonally vacant
     - 2.2.1 Secondary residences
     - 2.2.2 For rent
     - 2.2.3 For sale
     - 2.2.4 For demolition
     - 2.2.5 Other

45. If the housing unit is occupied, the number of occupants and the count of population records must not be zero. When using electronic data collection methods, the relationship between occupancy and presence of population should be determined, and housing units with persons should usually be made "occupied" automatically.

46. If no persons are recorded, either the unit is vacant or the persons are missing. As noted earlier in the structural edits, specialists must create procedures for determining whether the unit is vacant. If it is listed as occupied, but is actually vacant, then a method must be developed to determine the type of vacancy, either by listing it as "unknown" or by using dynamic imputation. If the unit is listed as vacant, but it can be determined that it is actually occupied because of information available in number of occupants or the count of population records, then the occupancy status must be changed to "occupied".

47. If the value is invalid, the value for number of occupants is zero and no population records are present, "unknown vacant" should be assigned when dynamic imputation is not used. If the value is invalid, but the number of occupants is not zero or population records are present, "occupied" should be assigned. Countries choosing

dynamic imputation for invalid values (to impute type of vacancy) should use at least two characteristics to obtain "known" information from similar housing units in the geographical area, or, alternatively, "unknown vacant" can be assigned.

4. Type of ownership (Core topic)

48. The type of ownership refers to the type of ownership of the housing unit itself and not of that of the land on which the units stand (United Nations, 2017, paras 4.476-4.481). Type of ownership should not be confused with tenure. Information should be obtained to show whether the housing units are owned by the public sector (central Government, local Government, public corporations) or whether the units are privately owned (by households, private corporations, cooperatives, housing associations or other). The question is sometimes expanded to show whether the living quarters are fully paid for, being purchased in installments or mortgaged. The classification of housing units by type of ownership is as follows:

1. Owner-occupied
2. Non owner-occupied
   2.1 Publicly owned
   2.2 Privately owned
   2.3 Communally owned
   2.4 Cooperatively owned
   2.5 Other

49. If ownership is related to tenure, this should be taken into account in developing the edit; if it is not related, then the type of ownership is probably independent of other housing variables. If the value for "type of ownership" is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics which might include construction material of walls, tenure, type of housing unit and number of rooms, in order to obtain "unknown" information from similar housing units in the geographical area.

5. Number of rooms (Core topic)

50. A room is defined as a space in a housing unit or other living quarters enclosed by walls reaching from the floor to the ceiling or roof covering, or to a height of at least two metres, of an area large enough to hold a bed for an adult, that is, at least four square metres. The total number of types of rooms therefore includes bedrooms, dining rooms, living rooms, studies, habitable attics, servants' rooms, kitchens, rooms used for professional or business purposes and other separate spaces used or intended for dwelling purposes, so long as they meet the criteria concerning walls and floor space. Passageways, verandas, lobbies, bathrooms and toilet rooms should not be counted as rooms, even if they meet the criteria. Separate information may be collected for national purposes on spaces of less than four square metres that conform in other respects to the definition of 'room' if it is considered that their number warrants such a procedure (United Nations, 2017, para. 4.482-4.484).

51. Since the number of rooms may be independent of the other housing variables, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, construction material of walls, tenure and vacancy status, to obtain "known" information from similar housing units in the geographical area.

6. Number of bedrooms

52. In addition to enumerating the number of rooms, a number of national censuses collect information on the number of bedrooms in a housing unit, which is the unit of enumeration for this topic. A bedroom is defined as a room equipped with a bed and used for night rest (United Nations, 2017, paras. 4.485-4.486).

53. Sometimes enumerators report a value for the number of bedrooms that is greater than the value for the number of rooms. If both rooms and bedrooms are present, they should be edited together, and the number of bedrooms should not exceed the number of rooms. Since the number of bedrooms is an "additional" topic, the edit is implemented only when both are present. With electronic data collection, a message should appear to let the enumerator/respondent that an illegal combination of rooms and bedrooms has been entered and that enumerators and respondents should probably do a literal count of rooms and bedrooms to resolve the issue.

54. When the number of bedrooms entered is greater than the number of rooms, if the country uses "not stated" only for invalid or inconsistent responses, "not stated" could appear for number of bedrooms. If dynamic imputation is used, bedrooms should be "estimated" from an imputation matrix with number of rooms as one of the elements. In this way, the number of bedrooms will not be greater than the number of rooms, because the value for bedrooms will be updated only when the values for rooms and bedrooms agree. The simplest case would be a linear array with the number of rooms as the cells and the value for bedrooms in the cells. A more complex imputation matrix might include the number of persons in the housing unit and the type of structure.

55. Otherwise, if the value for bedrooms is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics (with one of them being number of rooms) to obtain "known" information from similar housing units in the geographical area.

7. Useful floor space

56. Floor space refers to the useful floor space in housing units: that is, the floor space measured inside the outer walls of housing units, excluding non-habitable cellars and attics. In multiple-dwelling buildings, all common spaces should be excluded. The approaches for housing units and collective living quarters should differ, considering that information on the useful floor space per occupant of the set of collective living quarters is more meaningful (United Nations, 2017, paras. 4.487-4.489).

57. Floor space may relate to number of rooms and/or number of bedrooms, so country editing teams may want to take these into account when developing the edits. Other useful items for dynamic imputation include number of occupants and occupants per room. For the most part, floor space is independent of other housing edits. A unit of measurement, such as square metres, may need to be specified. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including type of housing unit, construction material of walls, tenure and vacancy, to obtain "known" information from similar housing units in the geographical area.

8. Water supply system (Core topic)

58. According to the United Nations (2017, paras. 4.490-4.493), the basic information to be obtained in the census regarding a water supply system is whether housing units have or do not have a piped water installation. This information will show whether piped water is inside the unit provided by pipes from a community-wide system or by an individual installation, such as a pressure tank or pump. The unit of enumeration for this topic is a housing unit. It is also necessary to indicate whether the unit has a tap inside or, if not, whether it is within a certain distance from the door. The recommended distance is 200 metres (not more than 200 metres), assuming that access to piped water within that distance allows the occupants of the housing unit to provide water for household needs without being subjected to extreme efforts. Besides the location of the tap, the source of available water is also of special interest. Therefore, the recommended classification of housing unit by water supply system is as follows:

1. Piped water inside the unit;
    1.1. From the community scheme;
    1.2. From a private source;

2. Piped water outside the unit but within 200 metres;
    2.1. From the community scheme;
        2.1.1. For exclusive use;
        2.1.2. Shared;
    2.2 From a private source;
        2.2.1. For exclusive use;
        2.2.2. Shared;
3. Other (see category of the topic on drinking water)

59. A community scheme is one that is subject to inspection and control by public authorities. Such schemes are generally operated by a public body, but in some cases they are generated by a cooperative or private enterprise.

60. The items on water facilities—water supply system, drinking water, toilet and sewerage facilities, bathing facilities and availability of hot water—should probably be edited together. Since these are closely related, when one is missing or invalid, the others can be used to generate a value. In geographical areas without running water, specialists may need to use specialized edits for the units. Otherwise, other units in the area will probably have similar characteristics, and these items are recommended for dynamic imputation when the latter is used.

61. Countries using electronic data collection methods could develop edits on entry that will assess the relationships between the variables, both as they are entered and when all of the relevant items have been entered, and appropriate messages can appear in case of inconsistencies between the above-mentioned topics. This type of edits will also help to find out data entry errors.

62. If the value for water system is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics. These might include as a rule, type of housing unit, and then toilet and sewerage facilities, and bathing facilities, to obtain "known" information from similar housing units in the geographical area.

9. Drinking water – main source of (Core topic)

63. Drinking water should be edited with water system. Many of the criteria described above also apply here. Bottled and other non-traditional sources of drinking water will normally be included on the questionnaire, so must also be included in the edit (United Nations, 2017, para. 2.494-4.495).

64. If the value for drinking water is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics. These might include as a rule, type of housing unit, and then water system, toilet and sewerage facilities, and bathing facilities, to obtain "known" information from similar housing units in the geographical area.

10. Toilet facilities (Core topics) and
11. Sewage disposal (Core topic)

65. Toilet facilities and sewerage should be edited together with the other plumbing variables to obtain the most consistent results. The 2000 Census Principles and Recommendations combined the two variables, but they have been separated for 2010 and 2020. Nonetheless, these items should be edited together, and use the same dynamic imputation matrices, if possible. As noted above, electronic data collection methods will allow checking for inconsistencies in water use during entry, so the relationship between toilets and sewage should be checked then and corrections made with the assistance of the respondents, if necessary.

66. Some countries have found it useful to expand the classification for non-flush toilets so as to distinguish certain types that are widely used and indicate a certain level of sanitation. The United Nations (2017, para. 4.498) recommendations for classification of housing unit by toilet facilities include the following:

1. With toilet within housing unit
   - 1.1 Flush/pour flush toilet
   - 1.2 Other
2. With toilet outside housing unit
   - 2.1 For exclusive use
     - 2.1.1 Flush/pour flush toilet
     - 2.1.2 Ventilated improved pit latrine
     - 2.1.3 Pit latrine without ventilation with covering
     - 2.1.4 Holes or dug pits with temporary coverings or without shelter
     - 2.1.5 Other
   - 2.2 Shared
     - 2.2.1 Flush/pour flush toilet
     - 2.2.2 Ventilated improved pit latrine
     - 2.2.3 Pit latrine without ventilation with covering
     - 2.2.4 Holes or dug pits with temporary coverings or without shelter
     - 2.2.5 Other
3. No toilet available
   - 3.1 Service or bucket facility (excreta manually removed)
   - 3.2 Use of natural environment, for example, bush, river, stream

67. The type of toilet facilities and sewerage (2017, para. 4.500) are other housing items having to do with water, and should be part of a joint edit with other water-related items. Values such as "private," "shared," "exclusive use" and so forth, could be used in determining whether values are consistent, and, if they are not, what edit paths to follow to fix the problem. When one or more other water-related variables is present, an estimate for unknown or inconsistent information may be developed without resorting to use of "unknown" or dynamic imputation. However, if this does not supply a valid value, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including type of housing unit, as a rule, as well as water supply, construction material of walls, tenure and vacancy status, to obtain 'known" information from similar to housing units in the geographical area.

12. Solid waste disposal – main type of (Core topic)

68. According to *Principles and Recommendations* (United Nations, 2017, para.4.501-4.502), this topic refers to the collection and disposal of solid waste generated by occupants of the housing unit. The unit of enumeration is a housing unit. The guidelines for classifying housing units by type of solid waste disposal are given below:

1. Solid waste collected on a regular basis by authorized collectors
2. Solid waste collected on an irregular basis by authorized collectors
3. Solid waste collected by self-appointed collectors
4. Occupants dispose of solid waste in a local dump supervised by authorities
5. Occupants dispose of solid waste in a local dump not supervised by authorities
6. Occupants burn solid waste
7. Occupants bury solid waste
8. Occupants dispose solid waste into river/sea/creek/pond
9. Occupants compost solid waste

10. Other arrangement

69. Solid waste is independent of the other housing variables. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics. These might include, as a rule, type of housing unit, and then construction material of walls, tenure, vacancy status or kitchen facilities, to obtain "known" information from similar housing units in the geographical area.

13. Bathing facilities (Core topic)

70. According to the United Nations (2017, para. 4.503-4.505), information should be obtained on whether or not a fixed bath or shower is installed within the premises of each set of housing units. The unit of enumeration for this topic is also a housing unit. Additional information may be collected to show if the facilities are for the exclusive use of the occupants of the living quarters and if there is a supply of hot water for bathing purposes or cold water only. However, in some areas of the world the distinction proposed above may not be the most appropriate for national needs. Instead, it may be important, for example, to distinguish in terms of availability among a separate room for bathing in the living quarters, a separate room for bathing in the building, an open cubicle for bathing in the building and a public bathhouse. The recommended classification of housing units by availability and type of bathing facilities is as follows:

1. With fixed bath or shower within housing unit;
2. Without fixed bath or shower within housing unit;
   2.1. Fixed bath or shower available outside housing unit;
      2.1.1. For exclusive use;
      2.1.2. Shared;
   2.2. No fixed bath or shower available.

71. Type of bathing facilities should be part of a joint edit with other water-related items. Values such as "private," "shared," or "exclusive use" can be used to determine whether values are consistent, and, if they are not, to establish the edit paths to follow to fix the problem. When one or more other water-related variables is present, an estimate for unknown or inconsistent information may be developed without resorting to use of "unknown" or dynamic imputation. However, when all else fails, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics. These include, as a rule, type of housing unit and then water supply, construction material of walls, tenure or vacancy status, to obtain "known" information from similar housing units in the geographical area.

14. Kitchen – availability of (Core topic)

72. According to *Principles and Recommendations* (United Nations, 2017, para. 4.506-4.509) the collection of data on the availability of a kitchen may provide a convenient opportunity to gather information on the kind of equipment that is used for cooking, such as a stove, hotplate or open fire, and on the availability of a kitchen sink and a space for food storage so as to prevent spoilage. The recommended classification of housing units by availability of a kitchen or other space reserved for cooking is as follows:

1. With kitchen within housing unit
   1.1 For exclusive use
   1.2 Shared
2. With other space for cooking within housing unit, such as kitchenette
   2.1 For exclusive use
   2.2 Shared

3. Without kitchen or other space for cooking within housing unit
    3.1 Kitchen or other space for cooking available outside housing unit
        3.1.1 For exclusive use
        3.1.2 Shared
    3.2 No kitchen or other space for cooking available

73. The edit for cooking facilities uses values such as "private," "shared," "exclusive use" and so forth, to determine whether values are consistent, and, if they are not, which edit paths to follow to fix the problem. When one or both cooking variables are present, an estimate for unknown or inconsistent information may be developed without resorting to use of "unknown" or dynamic imputation. However, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including, as a rule, type of housing unit, and then water supply, construction material of walls, tenure and vacancy status, in order to obtain "known" information from similar housing units in the geographical area.

15. Fuel used for cooking (Core topic)

74. In the context of the need to monitor closely the use of natural resources, national housing censuses include the topic of cooking fuel. The unit of enumeration is a housing unit; "fuel used for cooking" refers to the fuel used predominantly for preparation of principal meals. If two fuels (for example, electricity and gas) are used, the one used most often should be enumerated. The classification of fuels used for cooking depends on national circumstances and may include electricity, gas, oil, coal, wood and animal waste. It is also useful to collect this information for collective living quarters, especially if the number of sets of collective living quarters in the country is significant (United Nations, 2017, para. 4.510).

75. Response for type of cooking fuel should be edited with those for cooking facilities. When electronic devises are used, the edit can be done on entry – housing units that do not have cooking facilities should not have cooking fuel. The editing team determines the relationship between the two variables and develops an edit to check for consistency between them. Values having to do with "private," "shared," "exclusive use" and so forth, will probably be used in determining whether values are consistent, and, if they are not, which edit paths to follow to fix the problem. When one or both cooking variables are present, an estimate for unknown or inconsistent information may be developed without resorting to use of "unknown" or dynamic imputation. However, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including cooking facilities, type of building, construction material of walls, tenure and vacancy status, to obtain information similar to housing units in the geographical area.

16. Lighting and/or electricity – type of (Core topic)

76. Information should be collected on the type of lighting in the housing units, such as that provided by electricity, gas or oil lamp or by some other source. If the lighting is by electricity, some countries may wish to collect information showing whether the electricity comes from a community supply, generating plant or some other source, such as an industrial plant. In addition to the type of lighting, countries should assess the information on the availability of electricity for purposes other than lighting (such as cooking, heating water and heating the premises). If housing conditions in the country allow this information to be derived from the type of lighting, there is no need for additional inquiry (United Nations, 2017, para.4.511-4.512).

77. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics including, as a rule, type of housing unit, construction material of walls, tenure and vacancy status, to obtain "known" information from similar housing units in the geographic area.

17. Type of heating and energy used for heating

78. This topic refers to the type of heating of housing units and the energy used for that purpose. The units of enumeration are all housing units. This topic is irrelevant for countries where, owing to their geographical position and climate, heating is not provided in housing units. Type of heating refers to the kind of system used to provide heating for most of the space. It may be central heating serving all the sets of housing units or serving a set of housing units, or it may not be central, with the heating provided separately within the housing units by a stove, fireplace or other heating body. "Energy used for heating", is closely related to the type of heating and refers to the predominant source of energy, such as solid fuels (coal, lignite and products of coal and lignite, wood), oils, gaseous fuels (natural or liquefied gas) and electricity (United Nations, 2017, para. 4.513-4.514).

79. The type of heating and the energy used for heating are related to each other, as well as to the availability of hot water and to other utilities used in the housing unit, such as electricity and piped gas. Editing teams should consider the availability of these items in developing the editing specifications for heating type and energy for heating. Heating type may be independent of other housing items so may have to be edited separately. However, when "energy used for heating" is unknown or inconsistent, the program can check the type of energy used for lighting. Finally, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, construction material of walls, tenure and vacancy status, to obtain "known" information from similar housing units in the geographical area.

18. Availability of hot water

80. This topic concerns the availability of hot water in living quarters. Hot water denotes water heated to a certain temperature and conducted through pipes and tap to occupants. The information collected may indicate whether hot water is available within the living quarters or outside the living quarters, for exclusive or shared use, or not at all (United Nations, 2017, para. 4.515).

81. The availability of hot water may be related to the means for heating the water, although the use of solar energy for heating water may not be related to other housing items. The editing teams must decide on the appropriate edits, depending on other housing items and geographical location. In the end, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as those for piped water, to obtain "known" information from similar housing units in the geographical area.

19. Piped gas-availability of

82. This topic refers to the availability of piped gas in the housing units. Piped gas is usually defined as natural or manufactured gas that is distributed by pipeline and whose consumption is recorded. This topic may be irrelevant for a number of countries where a developed pipeline system or sources of natural gas are lacking (United Nations, 2017, para. 4.516).

83. Piped gas is not related to other housing items except for type of lighting and cooking fuel. Editing teams must determine the appropriate editing path as well as how to check for consistency. If the value remains invalid or inconsistent, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as energy used for heating, type of building, type of housing unit, construction material of walls, tenure and vacancy status, to obtain "known" information from similar housing units in the geographical area.

20. Use of housing unit

84. "Use of a housing unit" indicates whether a housing unit is being used wholly for habitational (residential) purposes or not. The housing unit can be used for habitational as well as for commercial, manufacturing or other purposes (United Nations, 2017, para. 4.517-4.518).

85. "Use of housing unit" is independent of the other housing items. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, construction material of walls, tenure and ownership, to obtain "known" information from similar housing units in the geographical area.

21. Occupancy by one or more households (Core topic)

86. Occupancy by more than one household is independent of other housing items. If the value is invalid, a country should count the number of heads of household and use that number (United Nations, 2017, para. 4.519-4.523). It is important to note that this edit must come after the structure edit determining the household head or reference person.

22. Number of occupants (Core topic)

87. Each person usually resident in a housing unit or set of collective living quarters should be counted as an occupant. Therefore, the units of enumeration for this topic are living quarters. However, since housing censuses are usually carried out simultaneously with population censuses, the applicability of this definition depends upon whether the information collected and recorded for each person in the population census indicates where he or she was on the day of the census or whether it refers to the usual residence. For persons occupying mobile units, such as boats, caravans and trailers, care should be exercised to distinguish those who use them as living quarters from persons who use these units as a means of transportation (United Nations, 2017, para.4.524-4.525).

88. "Number of occupants" is related to the number of population records and the two should be identical. Data entry with electronic data collection technologies should check for compatibility in these two items. If not, measures must be taken to correct the number of occupants item or the number of population records. Normally, the number of occupants will be adjusted to equal the number of persons in the unit. This item should not be "unknown" nor should it be imputed.

23. Building type (Core topic)

89. The following classification by type is recommended by the United Nations (2017, para. 4.526-4.534) for buildings in which some space is used for residential purposes.

1. Residential buildings
    - 1.1. Buildings containing a single housing unit
        - 1.1.1. Detached
        - 1.1.2. Attached
    - 1.2. Buildings containing more than one housing unit
        - 1.2.1. Up to 2 floors
        - 1.2.2. From 3 to 4 floors
        - 1. 2.3. From 5 to 10 floors
        - 1.2.4. Eleven floors or more
    - 1.3. Buildings for persons living in institutions
    - 1.4. Other residential buildings
2. Non-residential buildings

90. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, which might include construction material of outer walls, period of construction, and/or type of housing units in the building, to obtain "known" information from similar housing units in the geographical area.

24. Year or period of construction

91. The year or period of construction refers to the age of the building in which the sets of living quarters are located. It is recommended that the exact year of construction be sought for buildings constructed during the intercensal period immediately preceding if it does not exceed 10 years. Where the intercensal period exceeds 10 years or where no previous census has been carried out, the exact year of construction should be sought for buildings constructed during the preceding 10 years. For buildings constructed before that time, the information should be collected in terms of periods that will provide a useful means of assessing the age of the housing stock. Difficulty may be experienced in collecting data on this topic because in some cases the occupants may not know the date of construction (United Nations, 2017, 4.535).

92. Some countries, even those using dynamic imputation, accept an "unknown" response for the item on year or period of construction. When this occurs, the country may choose not to use dynamic imputation for this item, even if it uses imputation matrices for other variables. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including type of building, construction material of outer walls and/or type of housing units in the building, to obtain "known" information from similar housing units in the geographical area.

25. Number of dwellings in the building

93. Editing for the number of conventional dwellings in a building (United Nations, 2017, para 4.540) is explained in Chapter IV as part of the structure edits.

26. Position of dwelling in the building

94. Some countries may want to collect information on the position of the dwelling or housing unit in the building (United Nations, 2017, para 4.541-4.543). This information can be used as an indicator of accessibility to dwellings, possibly in conjunction with information on the accessibility to the dwellings.

95. The following classification of dwellings by position in the building is recommended:

1.0 Dwelling on one floor only
- 1.1 Dwelling below the ground floor
- 1.2 Dwelling on the ground floor of the building
- 1.3 Dwelling on the 1st or 2nd floor of the building
- 1.4 Dwelling on the 3rd or 4th floor of the building
- 1.5 Dwelling on the 5th floor of the building or higher

2.0 Dwellings on two or more floors
- 2.1 Dwelling on the ground floor of the building or below ground level
- 2.2 Dwelling on the 1st or 2nd floor of the building
- 2.3 Dwelling on the 3rd or 4th floor of the building
- 2.4 Dwelling on the 5th floor of the building or higher

96. For dwellings on two or more floors, information should be provided with reference to the lowest floor level of the dwelling. When face-to-face interview method is used, the enumerators would be able to determine the appropriate response by looking at the structure; however, the information should be confirmed by the respondent.

97. Some countries, even those using dynamic imputation, accept an "unknown" response for the item on position of dwelling in the building. When this occurs, the country may choose not to use dynamic imputation for this item, even if it uses imputation matrices for other variables. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including type of building, number of dwellings in the building, and/or type of housing units in the building, to obtain "known" information from similar housing units in the geographical area.

27. Accessibility to dwelling

98. The following classification of accessibility to the front door of the dwelling or housing unit is recommended, based on the presence of ramps, steps and lifts:

1. Access with no steps or ramp
2. Access by ramp
3. Access by disabled stair lift
4. Access using lift only (though the building may have staircases as well)
5. Access by using only steps
6. Access only by using both lift and steps

99. Note that these categories are not necessarily mutually exclusive (United Nations, 2017, para 4.544).

100. Some countries, even those using dynamic imputation, accept an "unknown" response for the item on accessibility to the dwelling. When this occurs, the country may choose not to use dynamic imputation for this item, even if it uses imputation matrices for other variables. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including type of building, number of dwellings in the building, and/or position of the dwelling in the building, to obtain "known" information from similar housing units in the geographical area.

28. Construction material of outer walls (Core topic)

101. Construction material of the external (outer) walls of the building refers to the walls in which the sets of living quarters are located. If the walls are constructed of more than one type of material, the predominant type of material should be reported. The types distinguished (e.g., brick, concrete, wood, adobe) will depend upon the materials most frequently used in the country concerned and on their significance from the point of view of permanency of construction or assessment of durability (United Nations, 2017, para 4.545-4.547).

102. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as period of construction and/or type of housing units in the building, to obtain "known" information from similar housing units in the geographical area.

29. Construction material of floor and roof

103. In some cases, the materials used for the construction of roofs and floors may be of special interest and can be used to assess further the quality of dwellings in the building. This topic refers to the material used for roof and/or floor (although, depending on the specific needs of a country, it may refer to other parts of the building as well, such as the frame or the foundation). The unit of enumeration is a building. Only the predominant material is enumerated and, in the case of a roof, it may be tile, concrete or metal sheeting, palm, straw, bamboo or similar plant material; or mud, plastic sheeting or some other material (United Nations, 2017, para. 4.548).

104. Sometimes the response on construction material for outside walls does not agree with the response on construction material of the roof; this might occur, for example, if the construction material identified for the walls is not strong enough to support the roof. As noted above, when this occurs, the specialists must decide whether to change one of the two variables, or use "unknown". If CAPI is used for entry, a routine should check to make sure that the walls and roof are compatible; if not, a message should appear, and the enumerator can rectify the relationship while till in the housing unit.

105. If a value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of building, construction material of outer walls, type of housing unit, construction material of walls, tenure and vacancy status, to obtain "known" information from similar housing units in the geographical area.

106. The reported construction material of the floor may or may not be consistent with the construction of the roof and walls. If the country editing team finds inconsistent or invalid combinations, it must decide whether to assign "unknown" or to use imputation matrices to change one or more responses. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of building, construction material of outer walls, type of housing unit, tenure and vacancy status, to obtain "known" information from similar housing units in the geographical area.

30. Availability of elevator

107. This topic refers to the availability of an elevator (an enclosed platform raised and lowered to transport people and freight) in a multi-storey building. The information is collected on the availability of an elevator for most of the time: in other words, one that is operational for most of the time, subject to regular maintenance (United Nations, 2017, para. 4.549-4.550).

108. Many countries will not be collecting the number of floors in their housing units. When they do, a relationship should be checked between the number of floors and presence of an elevator. That is, if the housing unit is in a building of only one floor, it should not have an elevator. When number of floors is collected as well as presence of elevator, when electronic data collection methods are used for entry, a check can be built into the edit on entry to compare the two entries and issue a message, if needed, so that the enumerator or the respondent can correct the problem.

109. If the building has only one floor or is a single, detached unit, an elevator should not be present. If an elevator is present, the editing team must decide which takes precedence, the number of floors or the fact that an elevator is present. If the elevator takes precedence, the number of floors must be changed, either by making the value "unknown" or by using dynamic imputation to obtain another value. If the number of stores takes precedence, and the building has only one floor, the response on "presence of an elevator" must be changed to "no".

110. When an elevator is present, if it requires electricity, a check should be made to be certain that electricity exists in the building. With electronic data collection, housing units without electricity should not have elevators, so a check should be made.

111. Finally, if the value for elevator is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as the type of building, availability of electricity and construction material of outer walls, to obtain "known" information from similar housing units in the geographical area

31. Farm building

112. Some countries have found it necessary for their national censuses to specify if an enumerated building is a farm building or not. A farm building is one that is part of an agricultural holding and is used for agricultural and/or housing purposes (United Nations, 2017, para. 4.551).

113. Farm buildings are independent of the other housing items. Countries may choose to check for correspondences with the population items for occupation and industry. Otherwise, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, to obtain "known" information from similar housing units in the geographical area.

32. State of repair

114. State of repair refers to whether the housing unit or the building in which the housing unit is located in need of repair and to the kind of repair needed. The unit of enumeration is a housing unit. The classification of housing units according to the state of repair may include:

1. Repair not needed,
2. In need of repair,
    2.1. Minor repair,
    2.2. Moderate repair",
    2.3. Serious repair and
3. Irreparable.

115. Minor repairs refer mostly to the regular maintenance of the building and its components, such as repair of a cracked window. Moderate repairs refer to the correction of moderate defects such as missing gutters on the roof, large areas of broken plaster or stairways with no secure handrails. Serious repairs are needed in the case of serious structural defects of the building, such as shingles or tiles missing from the roof, cracks and holes in the exterior walls or missing stairways.

116. The term "irreparable" refers to buildings that are beyond repairs, they have so many serious structural defects that it is deemed more appropriate to tear the buildings down than to undertake repairs. This term is most often used for buildings with only the frame left standing, without complete external walls and/or a roof (United Nations, 2017, para. 4.552-4.553).

117. The state of repair of the building is independent of the other housing variables. Hence, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of building, year or duration of construction, construction of outer walls and type of housing unit, to obtain "known" information from similar housing units in the geographical area.

33. Age and sex of the reference person of household (core topic)

118. The characteristics of the head of household or reference person are obtained from the population records to assist in developing cross-tabular information for planning and analysis (United Nations, 2017, para. 4.554-4.555)

119. These items – age and sex – should be required, but other items might include head's ethnic origin, religion or income. These items provide demographic information, but also assist in determining differential social status or need. If electronic data collection methods are used for capture, the keyed information about the head can be moved automatically, and no further edit would be needed. However, if the head's or reference person`s information is not complete, the population edit for the appropriate variable might be applied and the transfer then made. Whether

housing is edited first or population is edited first, the move of the head's information will come afterwards as derived variables. No further editing should be needed.

34. Tenure (Core topic)

120. According to the United Nations (2017, para. 4.556-4.559), tenure refers to the arrangements under which the household occupies all or part of a housing unit. The unit of enumeration is a household occupying a housing unit. The classification of households by tenure is as follows:

1. Household owns housing unit
2. Household rents all or a part of housing unit
    2.1 Household rents all or a part of housing unit as a main tenant
    2.2 Household rents a part of housing unit as a subtenant
3. Household occupies housing unit partly free of rent
4. Household occupies housing unit wholly free of rent
5. Household occupies housing unit under some other arrangement

121. Units occupied free of cash rent, with or without the permission of the owner, especially where this practice is prevalent, should be considered separately (United Nations, 2017, para. 4.557).

122. Tenure may relate to type of ownership, so the editing team may need to consider the relationship between the two items when developing the edits. Otherwise, if the value for tenure is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, rent and vacancy status, to obtain "known" information from similar housing units in the geographical area.

35. Rental and housing costs

123. The item for rental and owner-occupied housing costs is independent of the other housing variables except that, obviously, rental costs should occur only for rental units and owner costs should occur only for owner-occupied units. Since a relationship exists between tenure and housing costs, for electronic data collection, the appropriate items should be checked, and a message appearing if there are any problems. In the office during computer edit, the editing team must look at each case and determine the most appropriate relationships between these variables. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including ownership, as a rule, type of housing unit to obtain "known" information from similar housing units in the geographical area (United Nations, 2017, para. 4.560-4.562).

36. Furnished or unfurnished

124. The item on whether the unit is furnished or unfurnished is new. Editing teams should consider testing the item, if it is included, to determine the best items to use in dynamic imputation, if that method is used to resolve invalids or inconsistencies (United Nations, 2017, para. 4.563).

37. Information and communication technology devices – availability of (core topic)

125. The importance of availability of information communication technology (ICT) devices is increasing significantly in contemporary society. These devices provide a set of services that are changing the structure and pattern of major social and economic phenomena. The housing census provides an outstanding opportunity to assess the availability of these devices to the household. The choice of topics should be sufficient for understanding the place

of ICTs in the household, as well as for use for planning purposes by government and private sector to enable wider and improved delivery of services, and to assess their impact on the society (United Nations, 2017, para. 4.564-4.571).

126. The recommended classification is:

1. Household having radio
2. Household having television set
3. Household having fixed-line telephone
4. Household having one or more mobile cellular telephones
5. Household having a personal computer
6. Household accessing the Internet from home
    6.1. Landline connection
    6.2. Mobile connection
7. Household accessing the Internet from elsewhere other than home
8. Household without access to the internet

127. Information and Communication technology (ICT) devices are new items. Items requiring electricity should only occur where electricity is available. As solar power, wind power and other "renewables" become more frequently used, however, that factor must be considered in developing edits for this item. Country edit teams should thoroughly test the item and its imputation matrices before the census or survey. Useful items for the hot decks include social level of the household (as determined by a wealth index, for example), and age of household head or age of reference person.

128. These topics refer to the availability of the item within the housing unit. For example, a telephone denotes a telephone line rather than a physical telephone, since more than one telephone can be connected to a single telephone line (United Nations, 2017, para. 4.568). Telephones are not related to other housing items during the edit. However, if certain geographical areas do not have telephones, the editing team should take this into account in developing the edits. If the value for "telephone" is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, construction material of walls and tenure, to obtain "known" information from similar housing units in the geographical area.

38. Number of available cars

129. "Number of cars" refers to the number of cars and vans normally available for use by the occupants of a housing unit. The term "normally available" refers to cars and vans that are either owned by the occupants or used under a more or less permanent agreement, such as a lease (United Nations, 2017, para. 4.572).

130. The number of vehicles is independent of the other housing variables. If the country has areas without any vehicles, specialists might want to consider special edits for these selected geographic areas. Otherwise, if the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, construction material of walls, tenure, and ownership, or, in this particular case, number of adult occupants, to obtain "known" information from similar housing units in the geographical area.

39. Availability of durable household appliances

131. Information is collected on the availability of durable appliances such as washing machines, dishwashing machines, refrigerators, deep freezes and microwave cookers depending on national circumstances (United Nations, 2017, para. 4.573).

132. For most appliances, electricity must be available in the unit for the appliance to function. When these items appear, the editing team should consider an edit that checks for electricity (with the possible exceptions of a refrigerator that might be gas-powered or an "ice box"). Further, if running water is required in the specific country to run a washing machine or a dishwasher, the edit needs to account for this as well. Edits can be used to determine whether a particular item should be present, depending on the availability of electricity and water, and appropriate actions should be taken when inconsistencies appear.

133. Also, particular parts of a country may not have electricity or running water, and specialists may need to acknowledge this as they develop their edits. If the value is invalid or inconsistent, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as, type of housing unit, electricity, construction material of walls and tenure, to "obtain" "known" information from similar housing units in the geographical area (because the social levels of the households should be similar).

40. Access to outdoor space

134. This topic refers to the availability of outdoor space intended for recreational activities of the members of a household occupying a housing unit. The classification can refer to the outdoor space available (United Nations, 2017, para. 4.574):

1. As part of a housing unit (for example, the backyard in the case of a detached house),
2. Adjacent to a building (for example, backyards and playgrounds placed next to an apartment building),
3. As part of common recreational areas within a walkable distance (less than 10 minutes) from the housing unit (for example, parks, lakes, sport centres and similar sites)
4. Beyond a 10-minute walk from the housing unit, .

135. The amount of outdoor space available for household use is independent of other housing items. However, in certain geographical areas or certain types of buildings, no outdoor space may be available. Editing teams may need to consider the specific circumstances as they develop their edits. If the value is invalid, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, for example, type of building and type of housing unit, to obtain "known" information from similar housing units in the geographical area.

ANNEX I - OVERVIEW OF REAL-TIME EDITING APPLICATIONS FOR ELECTRONIC DATA COLLECTION

1. The editing process for electronic data collection with CAPI and CASI in censuses and editing applications on entry have been discussed throughout the previous chapters. This Annex will provide general considerations in designing real time editing rules, in other words, checking for invalid entries and inconsistencies during data collection.

2. While electronic questionnaires collect information in the same way as paper questionnaire, they also can detect and reconcile inconsistencies during the interview. To do the latter objective, electronic questionnaires must be designed to resolve the greatest number of inconsistencies while paying attention to the fluency of the interview and not frustrating the respondent. The error management should make the interviewer/respondent understand sooner what kind of error happened and which questions were involved in it. The customization of texts, in this regard, helps the enumerator/respondent to be reminded of the information previously collected, thus facilitating the resolution of inconsistencies and errors. The discrepancies should be resolved by either correcting a data entry error or by clarifying a response directly with the respondent.

3. A number of considerations have to be made when building edits-on-entry for CAPI and CASI used in censuses. Countries using CAPI or CASI must decide how much editing to do in the field and how much should be reserved for the office computer edit. As with most of the recommendations in this handbook, no perfect editing procedure can be developed. As with all other procedures, the amount of edit on entry should be tested before implementation.

4. Excessive edits on entry may slow up the enumeration. When few edits are implemented with CAPI or CASI, the enumerator or the respondent will not see much effect, but when many edits are implemented, the process will slow down. As was noted in earlier chapters, some variables are more important than others and enormous care should be given to these variables as data are edited on entry.

5. Electronic data collection applications used with CAPI and CASI have numerous built-in checks for identifying obvious discrepancies so that they can be resolved during data collection. There are several types of consistency checks. These include:

(i) Range checks – where the entered answer falls outside a valid range of responses-e.g. age is 150
(ii) Logic checks – where a combination of responses to certain questions are not logically consistent (e.g. to check that the sum of male and female children ever born equals to the total number of children ever born)
(iii) Consistency checks – to determine whether linked responses in different parts of the form are consistent with each other (e.g. female aged 60 give a birth in one year prior to the enumeration)
(iv) Unknown checks – where no information is entered.

6. These checks are implemented with two approaches:

(i) Hard checks – where the enumerator or the respondent cannot continue with the interview until they have changed the data entered in some way to remove the inconsistency. Hard checks are used when the inconsistency is impossible- e.g. male with children ever born.
(ii) Soft checks (signals) – where the enumerator or respondent is told about the error but they can ignore it and move on to the next question. Soft checks are used when an answer is unlikely but possible, e.g. if a respondent says they have more than 5 bathrooms. These checks are used to get the e respondent to confirm that the answer is correct and is not a data entry error.

7. In theory, all types of consistency checks can be done with hard checks, but this is not easy and will not necessarily improve data quality. It is usually very difficult to define a hard check for every individual variable and also for every type of consistency relationships between census variables during the field operation. Hard checks would create a burden on the user and slow down the interview. Therefore, careful testing is needed before making a decision on what types of consistency checks will be done with hard checks.

8. Soft checks are much easier to deal with in the field by the user—the enumerator in the case of CAPI or the respondent in the case of CASI. Soft checks are more along the lines of "Are you sure of that answer?". They are called 'soft' because they can be overridden, and the apparently inconsistent data left to stand if the user is convinced with the response. Presumably these soft checks often identify simple mis-keys by the interviewer or by the respondent.

9. Range checks are probably the simplest to implement, being effectively hard checks on numerical type data. Where the data-entry routine has been programmed to take an answer within certain limits the check merely makes those limits visible to all. However, this procedure can provide a false sense of security since recording an answer within the acceptable range does not ensure that it is correct, only that it is not grossly incorrect.

10. The decision to use hard-checks may be more problematic, and countries may have different opinions on what variables require hard checks. The decision to apply hard-checks should be based on a clear assessment of its impacts on data quality. However, it is important to apply hard checks for key census variables. This Handbook suggests that hard checks be applied for the following variables:

a. Variables that will be used for controlling census coverage. These variables may differ depending on the method of enumeration, however, at least the following variables should be checked with hard checks;
    i. Type of living quarters;
    ii. The result of enumeration for each unit, such as completed, not completed;
    iii. Occupancy status;
    iv. Listing household members.

b. Key variables which affect significantly the quality of other variables. The following variables could be controlled with hard checks:
    i. Age;
    ii. Sex;
    iii. Relationship to the reference person/head of household;

11. Age and sex are the most important variables in any population census since so much cross-tabulations leading to planning and policy formation depends on these two variables. So, the entry should not allow a sex out of range, and extreme ages should be checked against other variables and cultural norms. Application of hard checks will allow for better skipping for the questions on fertility, education and economic characteristics.

12. Skip patterns are another important consideration when designing electronic data collection applications. The choice of a 'system-controlled skipping' or a 'user-controlled skipping' should be carefully examined for each skip pattern in the questionnaire. Some census questions may be programmed by a user-controlled skipping. Especially for key census questions such as sex, user-controlled skipping may be preferred to ensure fertility questions will not be skipped automatically due to choosing wrong category of sex. A system-controlled skipping can speed up the enumeration, but they can also cause a bogging down if the enumerator/respondent must go backwards to recover missing data.

13. The census questionnaire includes different types of questions that can be responded using drop-box menu or entering the response in empty field as a numerical or textual information. As a result, census database will have numerical and textual information and some empty fields due to skipping some questions or missing or unknowns.

The numerical fields such as number of children born and surviving, will sometimes have "0" value. In order not have any problem to distinguish different types of information, it is very important to ensure that the database system allows to distinguish a deliberate entry of "0" value from empty fields.

14. Basic applications of editing on entry are summarized below.

SEX

15. Normally sex will be obvious. In case of face-to-face interview, most countries allow the enumerator to record the sex through asking a confirmation if a person is responding to the question. Traditionally, the only problem in obtaining the sex of an individual occurs either for small babies or the person is not present, and they have an ambiguous name, such as Francis or Kelly. Then, the enumerator should ask for the person's sex.

16. When gender is controlled with hard checks, the enumerator or the respondent should not be allowed to go to the next item until sex is filled in.

17. More and more countries have transgender individuals or people who are determining their sex or are transitioning. But not all countries recognize these types of sex for the census. When self-identification is used, the enumerator can accept what the respondent says – in this case the enumerator will have asked rather than simply filled the item by sight. If the person is transgender, rules will determine what sex is accepted. Since sex is one of the first items in the population census, it is important not to rile the respondent, so sometimes the enumerator must accept a response even when they personally feel uncomfortable. As before, the enumerator should not be allowed to go to the next item until sex is determined.

RELATIONSHIP BETWEEN HOUSEHOLDS MEMBERS AND MARITAL STATUS

18. Relationship and marital status are obtained in almost every census. Data are not usually compiled and presented for the relationship codes, but they are used to make sure everyone in the housing unit is included, and can be used to construct types of families – married couple families, female-headed households, etc.

19. For most countries these items can be used in an edit on entry to improve the accuracy of the sex reporting. For example, in case of opposite-sex marriage/partnership, when the head of household is one sex, the spouse should be the opposite sex. Again, usually the data will be inputted properly, but sometimes an error occurs. When a mistake happens on entry. The edit should only accept male and female, but an enumerator may accidentally key in one when they should have keyed in the other. In this case, the edit on entry will catch the issue, and a message "You just keyed in a husband and wife as the same sex" would appear and remedial action taken.

20. This type of real time editing will also check compatible relationships. Married children with the spouse in the housing unit should have the opposite sexes. Siblings and siblings-in-law should be opposite sex. Parents should have opposite sex, as should parents-in-law when different codes are used for these relationships. The edit on entry can provide a message when something is amiss, and if the sex is wrong, it can be corrected on site In countries with same sex marriages or codable relationships this edit may not be possible.

AGE

21. Because age is so important in subsequent planning and policy formation, every effort should be made to record the actual age. That is, like gender, the age needs to be hard checked, and not left to later office editing. And, although most countries do not make tabulations for relationship to head, the variable is very important in assisting in determining age when it might be ambiguous or missing, and so hard coding relationship should be considered.

22. Collection and edit for sex is straight forward. The edit on entry for age is much more complicated. For example, while the sex of the head and spouse must be opposite each other, their ages can vary considerably and still be valid. In first marriages, the ages are usually close, but not always. So, consideration of other items, like educational attainment, fertility, and economic variables might be needed to get a good fit.

23. Obvious checks appear when age at first marriage or age at first birth is included in the census. In this case, as the data are keyed for an individual, the electronic data collection application can check to make sure that the age at first marriage is above some minimum age and that it is also the same or less than the age recorded. Similarly, age at first birth can be used for females who have had at least one birth. However, it is important to remember that most people are more likely to know their current age than their age at first marriage or age at first birth, so other items might be needed in the checking.

24. These days most people can provide either their exact age, or, in the case of very old people, an estimate of their age. Historical events or age of relations in the housing unit can assist older people. If an older person does not know their age, someone else in the housing unit may be able to provide an estimate. The UN Principles and Recommendations expect the enumerator to provide an age before continuing with the enumeration.

25. Some edits are easy. Most countries put lower limits on the ages of the head and spouse, so if the enumerator enters age lower than the minimum, a message something like "The age is too low for head or spouse" and then remedial action can be taken. And, as noted, school attendance, language use, military status, educational attainment and economics (and fertility for females) can be used to make sure the recorded age is within appropriate range.

26. As with the sex edit, the relationships are used – both to make sure the age range with respect to the head is ok but also when the age of one or the other is not available, although at least an estimate of the head's age would be keyed at the beginning of the population enumeration.

27. Some countries may decide to estimate a spouse's age based on the head's age. This method may lead to errors with respect to other items, although if the estimate is like the head's age, it is safer.

28. For other ages, relationship to head can be used to assist in determining an appropriate age when the respondent cannot provide one. One problem is that the electronic devices are usually programmed to do a top down approach. So, the head's age is established first, followed by the spouse if there is one, followed by children, and then other relatives.

29. As noted earlier in the text, the head or reference person should be placed in the first position. With CAPI and CASI methods, the age difference between the head/reference person and their parents and children can be checked to identify simple mis-keys. Applying such edits on data entry would be possible after entering the data on age for all persons. Therefore, some of the edits on entry can be left after completing the interview by providing a list of errors to the enumerator or the respondent at the end of the interview.

30. For example, for a son or daughter who is 5 years younger than father or mother, an error list can provide this inconsistency to the attention of the respondent or the enumerators to be solved before the interview is completed.

31. Because age is so important for planning purposes, the edit on entry needs to do a fair amount of checking for consistency. For example, usually a minimum age of 3 is needed for to start school. The electronic data collection application may have an automatic skip pattern that jumps out if the age of a child has been keyed as 0, 1, or 2. If the application is not programmed to skip automatically but programmed to skip with the control of the enumerators or the respondents, then, a message should appear saying that the age and schooling are incompatible.

32. At the other end of the spectrum, people over a certain age should probably not be in school – or at least the enumerator should question the respondents if an incompatibility occurs. If a respondent says a person 50-years old

is currently attending school, that person might be taking college courses. The electronic data collection application should be programmed to display an error message for some maximum age for schooling but allow for an override if one is needed.

Use of relationship coding in determining age

33. If respondents report their ages, usually the enumerator should accept what they say unless there is a clear problem with the relationship variables. In most cases a "child" of head – unless it is actually a stepchild, should be not only younger than the head, but also a generation younger. Similarly, the parent of a head should be one generation older than the head. A grandchild should be two generations younger. Different countries will have various upper and lower limits which need to be respected.

34. One of the advantages of edit on entry is that the application can tell the enumerator when one of the "rules" is violated. If the household is unusually large, for example, the grandchild might appear some distance down the list and the enumerator (or respondent) may have forgotten the age of the head/reference person. The application can keep tract of the relationships and tell the enumerator when action must be taken.

35. The clearly generational relationships should not present problems for the enumerator – parents, children and grandchildren. Other relatives could be more difficult and might need a different level of triggering. A sibling could be as much as 30 years older or younger than the head, so looking at the same generation might or might not be helpful. And the method cannot be used for "other" relatives.

36. The CAPI application will insist that the enumerator enter an age before going on to the next item. Most respondents will know their age, and the enumerator can then record it directly. But some older people might not know their ages and so require assistance in determining one. Also, some people may misunderstand or misrepresent their age.

37. When an age is recorded, it does not have to remain permanently as inputted when CAPI or CASI is used. If, for example, a female – a head or spouse – is reported as 25 years old. Then, if the first child is reported as 15 years old, the edit on entry should report "You keyed an age difference of 10 between parent and child."

DATE OF BIRTH

38. The date of birth gives a more accurate measure for detailed age than age itself, since offices can use the information to obtain exact ages. In most cases in current census taking, exact day, month, and year of birth can be obtained for each person. The date of birth should be checked against the age and an error message presented when they are inconsistent.

39. When the day or month or both are not present, the office edit might impute values so that exact age can be determined when needed. The edit on entry should present a message when the birth year is clearly out of range, so it can be determined with the respondents.

MARRIAGE AND FAMILY

40. Although sex and age should be hard checked when possible, most of the other items on the census can usually be checked with soft-edits or left blank when they are not known.

41. Edit on entry can be used to verify that the data on marital status are consistent with the relationship reporting directly and the use of it in determining comparable marital statuses of couples. That is, the head and the spouse should both either be married or consensually married. A message can be delivered when an inconsistency appears.

Many countries have age limits for people to have any marital status than "never married", so when this is violated, a message should appear, and the problem fixed immediately.

FERTILITY AND MORTALITY

Children ever born and surviving

42. The number of children ever born should never be more than the children surviving, so the data collection application should check these on entry and report any problems. Also, the number of male children ever born should be greater or the same as the male children surviving, and the number of female children ever born should be greater or the same as the female children surviving.

43. The edit on entry should also check the sums of the individual items when more than just the summary measures are presented. If children in the household, children away, and children who have died are collected, then the children surviving should be the sum of those in the house and those away. The number of children ever born should be the sum of the children surviving and the sum of those who have died. These checks should also be made for each sex if the fertility is collected by sex.

44. The mother should be the right age for the number of children she has ever borne. This information is obtained by an edit that compares the mother's age with an upper limit of the number of births at that age. Countries usually expect births to be between age 15 and 49 or 54, so if the number of children ever born is greater than the upper limit for that age, the enumerator/respondent should be informed so the items can be corrected.

45. The upper limit for number of births will differ from country to country and in urban vs rural residences, or by religion or ethnicity. Any of these may need to be accounted for when the edit on entry is developed and run.

Births in the year before the census or last births

46. Children who are born after the census reference date should not be included in the census. The edit on entry should immediately inform enumerators if they key in a date of birth of the last child that is after the census date. Because most censuses begin after the reference date, babies may be born between the reference date and the date the enumerator arrives at the unit to do the enumeration. The decision should be to delete this information and collect information for the previous birth when date of last birth is requested.

47. Although children born in the last year may not be in the unit, in most cases they are unless adopted out. So, the edit on entry should check to make sure that the date of birth of the last child is the same as the date of birth for the child when the child appears in the roster or as enumerated.

48. Usually the mother of the child appears in the listing before the child, so her information about the date of birth of the last child can be saved and checked against subsequent family members as they are entered. If, , the date of birth of a child born in the year before the census is not matched with the household members are entered, a message could state that fact. The sex of the child and of the last birth should be compared.

49. Vital status of the child should be considered when doing these comparisons. Many countries now collect information about date of death of the last child. In this case, the edit must check to make sure that the date of death of the last child is the same or later than the date of birth.

50. The mother should be the right age for the date of birth of the last child. This information is obtained by an edit that subtracts the mother's birth date from date of last birth entered to obtain her age at that birth. Usually countries expect births to be between age 15 and 49 or 54, so if the age is out of range, the enumerator/respondent should be informed so it can be corrected.

51. If a female has had no children ever born (or, in some cases, children surviving), then she should not have had a last birth. If a last birth occurs for females with no children ever born, a message should appear, and a decision made. In most cases last children must be biological children for the demographers to do their analysis, so mothers reporting adopted, or foster children should not be reporting these as last births.

Mother's Person Number

52. Some censuses ask mother's person number from everyone or from a sample based on age. This item is used in fertility estimation, particularly the own child method, which gives retrospective fertility. The enumerator asks respondents for their mother's name and line number from the roster of members of the household and records it. Sometimes, if the mother is dead, 99 is entered.

53. This item is often important in obtaining fertility trends at the national and sub-national levels. So, as the value is entered, the edit on entry should check to make sure that the number represents a female and that the relationships are appropriate (head/spouse and child, child and grandchild, head and parent). The mother should have been appropriate age to have the child, and the age difference should be between some lower limit (usually 15) and some upper limit (usually 50).

54. When the checks fail, the edit on entry might try to find an appropriate mother if there is one in the housing unit. If they can be associated, then that mother's line number should be entered.

Parental orphanhood

55. The items on mother's and father's vital status are used (1) to obtain mortality estimates based on the percentages alive at ascending ages and (2) orphans.

56. The items ask whether the parent is alive or dead. Most countries also accept unknown as well. Demographers often analyze the known data and ignore the unknowns. The data can be edited with a dynamic imputation in the office but the original responses should be kept in the application so that they can be analyzed.

Deaths in the year before the census

57. Deaths in the year before the census is a fairly new item on censuses. The item is used to obtain age specific mortality estimates. The items collected include age and sex of the deceased. Completeness is a problem since many people are unable to provide the age at death. It is recommended to allow unknown and keep all available information for imputation and further analysis in the office, even if some data are not consistent with other variables, such as an infant died within the year before the census but not declared as a death for the question of vital status of last child born in the year before the census.

MIGRATION

Place of birth

58. An item on place of birth is asked in most censuses. But the level of detail varies from country with some only reporting down to the major civil division and others reporting to the minor civil division or lower.

59. As the data are entered, the machine can assist the enumerator in obtaining the codes. A code list could be embedded or displayed on the screen for the enumerator/respondent to record the appropriate response. Or the enumerator or the respondent could key in the first few letters of the place, and the machine could obtain the code, display, and ask for confirmation by the user.

60. Alpha characters inevitably slow down the enumeration process, so they should be avoided when possible. However, if a country decides to have several levels of geography or has many major or minor civil division, reference files could be very bulky, so entering letters might be appropriate.

61. The data collection application can assist when the respondent has never moved since birth. When a code is used for "never moved" the application can automatically enter the appropriate geography. If the census questionnaire asks if the person was in the same exact house as well as the same geographic area, an additional coding would be needed.

Residence at a specific point in the past

62. An item on residence at a specific point in the past – usually 5 or 10 years – is asked in most censuses. Sometimes the item refers to a specific event like a new constitution or a natural disaster. While most countries ask for this short-term migration, the level of detail varies from country to country with some only reporting down to the major civil division and others reporting to the minor civil division or lower.

63. As with place of birth, as the data are entered, the application can assist the enumerator/respondent in obtaining the codes. A code list could be embedded or displayed on the screen for selecting the appropriate response. Or the enumerator/respondent could key in the first few letters of the place, and the application could obtain the code, display, and wait for the enumerator/respondent to accept.

64. For this item also, as explained for "place of birth" it is suggested not to use alpha characters. Also, if a code is used for "never moved", it is suggested to enter automatically the appropriate geography. If the country wants to know if the person was in the same exact house as well as the same geographic area, an additional coding would be needed.

Previous residence and length of current residence

65. Items on previous residence and length of current residence are asked in some censuses. This is a different way of measuring migration. As before, the level of detail varies from country to country with some only reporting previous residence down to the major civil division and others reporting to the minor civil division or lower.

66. As with place of birth, as the data on previous residence are entered, the machine can assist the enumerator in obtaining the codes. A code list could be embedded or displayed on the screen for the enumerator to record the appropriate response. Or the enumerator could key in the first few letters of the place, and the machine could obtain the code, display, and wait for the enumerator to accept.

67. The data collection application can assist when the respondent has not moved and so the residence at the point in the past is the same as the place of birth and the current residence. When the person has never moved, some censuses require the length of stay to be skipped. In this case, part of the edit on entry might be to insert the age into the length of stay.

68. The edit on entry needs to check to make sure that persons do not migrate before they are born. This problem frequently occurs in censuses when the age variable is distant from the length of residence item. A message should be shown when an illegal entry is provided.

Parental place of birth

69. Many countries ask where the mother and father were born. These variables show generational migration. The edit on entry cannot use place of birth or residence in the past to assist since the parents' place of birth could be very different from the individual. The most likely edit on entry is to make illegal or unknown entries as "unknown".

ETHNOCULTURAL CHARACTERISTICS

Ethnicity

70. When paper is used, the enumerator either can write out each ethnicity or an abbreviation for the ethnicity or use a code list to enter the ethnicity. The use of CAPI or CASI has the advantage of providing pull down menus or lists of possible ethnicities on the screen. The enumerator can then enter the code.

71. The drawback of this procedure, of course, is that once the code is keyed in, it is there forever, whether it was coded correctly and entered correctly. An edit might analyze the household and report when an ethnicity looks strange or out of place. For example, if everyone in the unit except one child or grandchild, for example, has the same ethnicity, the electronic data collection application might be programmed to report the possible error.

72. As is noted in several places in this handbook, each check takes time, and while computers are extremely fast and can provide possible errors like single people without ethnicity or with a different ethnicity than the others, it should be kept in mind that this kind of error messages aim to find out data entry error not checking inconsistencies. Most of the time, all members of a household would belong to the same ethnicity or religion, but there will be some households where only one person or few persons would belong to different ethnicity or religion.

73. Often with paper enumeration, once the head and spouse report ethnicity, the enumerator leaves the children blank because they have the same ethnicity as the parents. With the use of CAPI and CASI, it is possible to alleviate some of this problem.

Religion

74. The edit on entry for religion will be very similar to that of ethnicity, and for the same reason. Most of the people in a housing unit will have the same religion just as most of the people in the house have the same ethnicity.

75. Again, a code list can be embedded into the data collection application as a dropdown list or on the screen for coding rather than having to write out the religion or an abbreviation for the religion.

Literacy and language

76. Literacy is often asked in countries where many of the people cannot read and write in any language. If literacy is language-specific (in multilingual countries, the census questionnaire may inquire into the languages in which a person can read and write) then the data collection application can make sure the relationship is correct and trigger a note when it is not.

77. Most countries have a minimum age for asking the literacy item and a minimum age for children to be literate. A skip pattern is usually used according to the minimum age defined for this question.

78. A relationship also exists between educational attainment and literacy. Students who have reached a certain grade level can be assumed to be literate. If the enumerator/respondent keys that the student is illiterate, an error message should appear.

79. The relationship works the other way as well. Students or those who left school at certain levels should be assumed to be literate, so if a student is in a high grade but listed as illiterate, a question should appear.

EDUCATION

School attendance

80. School attendance usually has a minimum age which the data collection application should respect. After that, when countries ask both "never attended" and "no longer attending", it is important that the distinction be made, usually at least the first of the two based on age.

81. Different grades will have lower and upper age limits, which can be programmed into the edit on entry, and then checked as the grade for those attending school is entered. Sometimes outliers occur but the range should help keep some of the extraneous information out.

Educational attainment

82. Educational attainment is asked in almost every census. The educational attainment for those in school will differ for those out of school so parallel edits are needed. The edit can check, for those who are "in school", if educational attainment is the lower level of education compared to the level which a person is currently attending For those "not in school", this information can be checked with the literacy for people who were not completed any level of education.. People out of school could be of any age, but they have a lower limit for each grade.

Field of study

83. The field of study is normally asked of only a segment of the population with sufficient education to be either in technical school or in an academic institution. Sometimes technical education occurs in high school, so a field of study could be asked for those individuals. The edit on entry should skip those persons not getting the item but expect a response from those who should have a field of study.

84. A code list of fields of study could be embedded in the edit and referred to either by several letters keyed or a code. The more detailed the code list, the longer it will take to make certain the correct entry is made. The list of fields of study could also be on the screen if there are not too many of them.

85. The edit can check the field of study against the occupation, but obviously only after the occupation has been entered. When a mismatch occurs, a message can be generated. However, this edit can slow down the interview, therefore it might be better to do after data collection in the office.

DISABILITY

Type of disability

86. Disability is frequently a problematic item from the point of view of the edit. It is internationally recommended that a separate question should be asked for each domain and the edit on entry should check if there is a response for each domain of disability.

87. In many cases, edit on entry will force the enumerator or the respondent to enter a response, so in those cases the ambiguity will not occur (unlike on paper where the enumerator could just skip those items and leave blank if the person does not have any difficulty). If there is a possibility of unwillingness in responding to some or all parts of questions on disability, an additional category "do not know" can be added to deal with these cases after data collection.

ECONOMIC CHARACTERISTICS

Work last week

88. As noted earlier in the text, the United Nations looks at both paid and unpaid economic activities as work. Countries define the earliest age to be considered in the labor force in each case – some consider children as young as 5 as possibly being in the labor force. In any case, the data collection application should be programmed to collect information only from those old enough to be considered.

89. Countries may have different interests for measuring economic characteristics of the population with respect to their participation in one or in several forms of work. In particular, in the population census, the Principles and Recommendation for Population and Housing Census, Revision 3 suggests measuring the following groups:

a. *Employed persons* are all those above the specified age who during a short reference period of seven days or one week were engaged in any activity to produce goods or provide services for pay or profit;
b. *Unemployed persons* are all those above the specified age who (i) were not in employment, (ii) carried our activities to seek employment during a specified recent period and (c) were currently available to take up employment given a job opportunity;
c. *Persons outside the labor force* comprise all those who in the short reference period were neither employed nor unemployed as defined above, including persons below the minimum age specified for the collection of economic characteristics;
d. *Persons in own-use production of goods* are all those above the specified age who, during a specified reference period, performed "any activity" to produce goods for own final use.

90. Countries ask different types of questions measuring people in the labor force (employed and unemployed people) or outside the labor force or involved in any activity for producing goods for their own final use. Some countries ask one question consisting of many categories for measuring groups of (a), (b) and (c), while others ask several questions for collecting the data for the same topics. The types of questions may have an impact on the editing procedures and the steps. In general, it is important to design the edits on entry for ensuring the consistencies between the questions which collect data on the above-mentioned groups.

Employed population

91. A correspondence exists between paid work (or subsistence work) and hours worked, occupation, industry, and status in employment. If these are present the person should be listed as working. If they are not present that items like whether on layoff, looking for work, and availability to work should occur so that the person would be considered in the labor force but not working – so unemployed. People who are neither employed nor unemployed should be listed as not in the labor force.

92. People who were engaged in any activity during the seven days prior to the census day to produce goods or services for pay or profit or people who were not engaged but had paid job or business from which they were temporarily absent should answer the questions on status in employment, occupation and industry. The edit on entry should check the consistency between these variables. When a proxy respondent provides the information, the respondent may not give correct information for some of the questions asked to the employed people, especially the questions on place of work and industry. In this case, these field should be left empty to deal with these cases during data processing in the office. It should be noted that persons who are not engaged in any economic activity should not answer these questions. Therefore, if there is any error in identifying the person as being outside the labor market, the following questions with regard to unemployment and reason for not entering the labor market would not be meaningful questions, so the enumerator or respondent will be able to go back to the question for correcting employment status of the person. Sometimes the edit on entry can change an errant response, but other times the enumerator may have to work with the respondent to determine the best response.

Hours worked

93. Hours worked is a non-core topic and may not be included in the census questionnaire. If it is included, it should only be asked of those who are employed. Hours worked is almost always asked for paid work. Sometimes hours of non-paid work are also collected.

94. The number of hours worked should not be 0 because the person should have been working in the week before the census to get to this point. If enumerators key 0, then the data collection application should prompt the enumerator/respondent to check again the person's status.

95. At the other end of the spectrum, each country will need to determine a maximum number of hours that could be worked in a week and put that in the edit on entry. Some countries pay biweekly and when 80 hours is entered, a notice should ask if this was for a two-week period.

96. When the hours worked cannot be determined because the respondent is not there or is reluctant or doesn't know, the item could be left blank for later imputation in the office.

Unemployment variables

97. Most countries collect data on unemployment which is asked to all those above the specified age who are not in employment and looking for work. Unemployed people must satisfy three criteria: (1) not in employment, (2) seeking a job, and (3) currently available to work.

98. A person who works even one hour for pay or for profit will be considered as employed and will not be asked the unemployment questions. That is, the data collection application will skip these items.

99. When the employment items are keyed – economic activity being some form of employment and hours worked entered, then these items should be skipped. When the economic activity is not working, then the hours worked should be skipped, and these items entered (assuming the follow hours worked).

100. Some countries will expect "availability" to be keyed whether or not the person is looking for work. If the country wants availability for all people who are not employed, then this item will not be skipped; if wants availability only from those looking for work, then the item should be expected to be keyed.

Occupation and industry

101. In most cases, occupation and industry need to be treated similarly when CAPI or CASI is used. The code for the occupation can be entered directly by the enumerator or responded from the list of the classification of occupation and industry, and no further coding is needed. However, countries may prefer to do coding in the office as this operation requires a skill staff especially for 3 or 4 digits level of coding.

102. Most data collection applications will allow for external files for reference during enumeration. Like for geographic places, these are embedded in the application and can be called up as needed. Hence, in the case of the use of CAPI, the enumerator could key in the first few letters of the occupation or industry and that occupation or industry and its code will pop up, and the code can be entered. In this way, the detail is kept.

103. Some countries may choose to have the enumerator or respondent (it is preferable for CASI) key in the alphabetic responses for occupation for later coding. This method will require a later procedure for obtaining a code for the occupation or the industry from the alphabetic response. Part of the office edit may be reserved for this or

keyers in the office could determine the appropriate code and enter it. This latter method will greatly slow up the total process and may hold up the results of the census.

104. Some countries may decide to use this procedure – later coding of occupation or industry – and so may choose to release the demographic and social data at one time and the economic activity data later.

Status in employment

105. In many cases, a relationship exists between status in employment and occupation or industry or both and also with the completed level of education. For example, it is unlikely that teachers would be family workers or graduated from primary school. It would be difficult to use edits on entry for these kinds of consistency checks requiring more in-depth investigation across several variables. During data collection, it might be a better approach to focus on valid data entry and leave detailed analysis for after data collection.

Work in the year before the census

106. Many countries ask about work in the year before the census to obtain information about long-term employment compared to the employment collected in the week before census.

107. Usually three items are collected – (1) whether the person did any work in the year before the census (sometimes the 12 months before, sometimes the calendar year), (2) how many weeks worked, and (3) the usual number of hours worked per week. The weeks can be multiplied by the hours and, with income, can get hourly rates by occupation, industry, age, etc.

108. The edit on entry should either be programmed to skip or check to be sure that if the respondent says he/she worked in the year before the census, that the weeks and hours are filled. And, if "no" is selected then no weeks and hours occur.

109. The number of weeks should be between 1 and 52 – 0 should not be accepted – and the number of usual hours should be between 1 and some upper limit defined by the country's statisticians. If the weeks or hours cannot be determined, a dynamic imputation might be used to supply them based on a similar worker.

Income

110. Most countries do not ask income at all – using income and expenditures surveys to obtain the data on income – or use categories to collect it. Income is the trickiest item because it is the most intrusive, so care must be taken in asking about it.

111. To assist in obtaining the income item, staff might have provided upper- and lower-income levels for various occupations, which could guide the enumerator in questioning the respondent. Of course, some respondents will be doing work which is out of range.

HOUSING

112. Most countries now ask population and housing items in the same census. Unlike the population items which have many skips, and which require inter-record checking as well as intra-record checking, most censuses have a single housing record for each housing unit, and few skip patterns. Most of the standard skips are discussed in Chapter VI on Housing Edits in this handbook.

113. Many countries include both vacant and occupied units in the housing census. In this case, skip patterns will be needed when information concerning the interior of the housing unit is required. For example, the type of toilet or

presence of a refrigerator may not be known because it cannot be seen. So, in most censuses, only some items will be obtained at vacant units and the other items will need to be programmed to be skipped.

114. In some cases, the enumerator will visit the unit several times and still not be able to obtain the housing information (or, for that matter, the population) information. In fact, in some cases, even after conferring with neighbors, the enumerator may not be able to tell whether the unit is vacant or not. When this happens, the enumerator may be required to key in a few variables (such as type of living quarters and constructions materials) , sometimes called "last resort" information to be used for completing the enumeration of the unit, and sometimes to be used to impute the other variables.

115. Most structure items should be entered from the order on the questionnaire. Some of the skips are obvious. If the housing unit has no cooking facilities, then, as noted in the Housing Chapter, the type of cooking fuel should be skipped. Similarly, units without electricity should not be asked about air conditioners or other electrical devices.

116. When both rooms and bedrooms are on the questionnaire, the edit on entry should check to make sure that the number of bedrooms is not greater than the number of rooms and issue a message if this is keyed in.

117. The number of occupants should be the same as the sum of the keyed population in the housing unit. When it is not, a message should appear. Complete coverage is most important in any census. When the number of males and females is entered these figures should correspond to the numbers of males and females in the unit from individual population records.

118. Tenure – whether owned or rented – is usually collected in housing censuses. When owned, if value of unit is collected, then the edit on entry should skip the amount of rent paid. Similarly, if the unit is rented, rent should be obtained, but if value of unit is on the questionnaire, it should be skipped.

OTHER CONSIDERATIONS

119. The role of the supervisor. The role of the field supervisor changes when CAPI is used, as noted earlier. It is not always possible for the supervisor to go over completed questionnaires side-by-side with the enumerators with both looking at the completed questionnaire. Depending on the type of data transfer protocols adopted, the supervisor should be able to review the status and quality of completed electronic questionnaires on a daily basis.

120. In addition to checking the questionnaire itself, electronic data collection can provide useful para-data information such as that on the amount of time spent by the enumerator in the housing unit, the per capita time, number of entry key strokes, number of missing items, error rates, etc. Such information will assist the supervisor and operations control at headquarters to monitor progress. When too many questionnaires are transferred it could mean that not enough care is being done on each one; when too few come in, the enumerator might need to be checked to make sure they are completing their assignments.

121. Inconsistent relationships during the interview. As with the example of the 5-year-old in the second grade, the editing team needs to decide which edits on entry are appropriate for the census. Any decision about rectifying inconsistencies between two items starts with a 50 percent chance of being correct. A proper edit increases the chances of obtaining the correct relationship. And, sometimes, no edit can provide certainty.

122. Inconsistencies between variables can be a major problem. Because edit on entry is so straight-forward in most cases, some programmers tend to over-edit on entry. Some countries now seem to feel that they don't need an office edit at all, that everything can be done in the field, resulting in an extremely unwieldy edit being installed in data collection applications. This getting carried away with the edit can be unwieldy and hard to check for accuracy before implementation. Sometimes, the over-editing will decrease the overall quality of the enumeration.

123. Simple relationships that can be corrected easily probably should be included in the edit on entry, particularly for crucial items like age and sex, and semi-crucial items, like fertility and marital status. But others should wait for the office edit.

ANNEX II : EDITED VERSUS UNEDITED DATA

1. Countries perform census edits to improve the quality of data and its presentation. In this section, the *Handbook* highlights a problem facing national census/statistical offices when unedited census data are released. The issues are illustrated using a hypothetical set of data.

2. The national census/statistical office of a fictional country faces the dilemma of trying to serve multiple users. Some users may want unknown entries included for analysis or research and some others (usually government workers responding to requests from users) may want data with minimum noise (possible error) for their planning or policy purposes. If the national census/statistical office disseminates an unedited table, such as that on the left side of table 1, both the analysts and the policy makers will have to make assumptions when using the data. Table 1 illustrates this point with only a small number of persons. It shows that for 23 persons in this country sex or gender was not reported and for 15, age was not reported. These omissions may have resulted from non-responses or from keying errors. Of these, two cases reported neither sex nor age.

FIGURE A.II.1. SAMPLE POPULATION BY 15-YEAR AGE GROUP AND SEX, USING UNEDITED AND EDITED DATA

	Unedited data				*Edited data*		
Age group	*Total*	*Male*	*Female*	*Not reported*	*Total*	*Male*	*Female*
Total	4,147	2,033	2,091	23	4,147	2,045	2,102
Less than 15 years	1,639	799	825	15	1,743	855	888
15 to 29 years	1,256	612	643	1	1,217	603	614
30 to 44 years	727	356	369	2	695	338	357
45 to 59 years	360	194	166	0	341	182	159
60 to 74 years	116	54	59	3	114	53	61
75 years and over	34	12	22	0	37	14	23
Not reported	15	6	7	2			

3. Most users would make their own decisions about what to do with the unknowns. A logical, possibly naïve, approach would be to distribute the unknowns in the same proportion as the known values. If the national census/statistical office chooses to impute for the unknowns, the editing team may decide to have 12 males and 11 females, a figure that is about half-and-half, but skewed because the census enumerated more females. The results will then be consistent with the edited data shown on the right side of table 1.

4. Other options are available for handling the unknowns. For example, the editing team may decide to impute based on the sex distribution alone, ignoring other available information, such as the relationship between spouses, whether a person of unknown sex is reported as a mother of another person or whether a person of unknown sex has a positive entry for number of children ever born. An alternative imputation strategy would be to take one or more of these other variables into account.

5. Another alternative the national census/statistical office could choose would be to base the imputation on the age distribution. For sample population illustrated in table 1, a total of 15 cases occurred with unreported age. These data could also be distributed in the same proportions as the known values, again, a logical strategy for imputation. Still, the editing team could probably obtain better results by considering other variables and combinations, such as the relative age of husband and wife, of parent and child or grandparent and grandchild, or the presence of school age children, retirees and persons in the labour force.

6. In table 1, the edited data on the right are "cleaner" because the unknown cases have been imputed (see columns under "edited data"). This side of the table has no unknowns, since the program allocates them to other responses. Nevertheless, many demographers and other subject-matter specialists have traditionally wanted to have the unknowns shown in the tables, as in the unedited data of table 1. They believe that this procedure allows them to perform various kinds of evaluations on the figures to measure the effectiveness of census procedures or to assist in planning for future censuses and surveys. Both objectives can be accomplished—an edited table for substantive users and an unedited one for evaluation—by making tabulations both with and without unknowns.

7. Statistical offices should make every effort to maintain the original, collected data, whether collected using electronic data collection technologies or on paper than keyed or scanned. A complete set of the original, keyed data should be archived, both as part of the historical record, but also for reference if staff make decisions about re-editing any part of the data set from the beginning. However, original values of crucial items, like age, sex and fertility, should be kept somewhere on each record to allow demographers and others to analyze the results of the edits.

8. Another problem with the use of unknowns in the published tables is that the unknowns may affect the analysis of trends. The new technology makes this analysis much easier than it used to be. For example, table 2 shows an age distribution from two consecutive censuses. The number of unknowns decreased for this small country, from 217 or about 6.5 per cent of the reported responses in 2010, to only 15, or less than one per cent of the responses in 2020.

9. Here the national census/statistical office must deal with how inconsistent numbers of unknowns affect the individual census and the change between censuses. For example, the 6.5 per cent unknown for the 2010 census makes it difficult to compare the change in percentage distributions for the 15-year age groups in the two censuses. The percentage of persons 15 to 29 years seems to increase from only 27 per cent to 30 per cent during the decade, but the distributed unknowns could change the analysis.

FIGURE A.II.2. POPULATION AND POPULATION CHANGE BY 15-YEAR AGE GROUP WITH UNKNOWNS: 2010 AND 2020

Age group	*Numbers*		*Number Change*	*Percent Change*	*Per cent*	
	2020	2010			2020	2010
Total	4,147	3,319	828	24.9	100.0	100.0
Less than 15 years	1,639	1,348	291	21.6	39.5	40.6
15 to 29 years	1,256	902	354	39.2	30.3	27.2
30 to 44 years	727	538	189	35.1	17.5	16.2
45 to 59 years	360	200	160	80.0	8.7	6.0
60 to 74 years	116	89	27	30.3	2.8	2.7
75 years and over	34	25	9	36.0	0.8	0.8
Not reported	15	217	-202	-93.1	0.4	6.5

10. The revised table, table 3, shows the unknowns distributed, either proportionally or through some method of imputation. Here it is much easier to see both the numeric and percentage changes as well as the distribution of the age groups in the two censuses. Of course, in order to obtain accurate, reliable results, the editing teams must make sure the edits are consistent between the two censuses and/or surveys, as well as internally consistent. The row for "not reported" is dropped.

FIGURE A.II.3. POPULATION AND POPULATION CHANGE BY 15-YEAR AGE GROUP WITHOUT UNKNOWN DATA: 2010 AND 2020

Age group	*Numbers*		*Number change*	*Per cent change*	*Per cent*	
	2020	2010			2020	2010
Total	4,147	3,319	828	24.9	100.0	100.0
Less than 15	1,743	1,408	335	23.8	42.0	42.4
15 to 29 years	1,217	952	265	27.8	29.3	28.7
30 to 44 years	695	578	117	20.2	16.8	17.4
45 to 59 years	341	230	111	48.3	8.2	6.9
60 to 74 years	114	109	5	4.6	2.7	3.3
75 years and over	37	42	-5	-11.9	0.9	1.3

ANNEX III - DERIVED VARIABLES

1. In order to get the best use out of their census or survey data, countries often need variables that are combinations and variations of other variables. Rather than having to develop a program to derive the information each time the national census/statistical office wants a special tabulation, data processing specialists can write a program to make the recode variables once, store the derived information on the person's record, and then use it for further tabulations.

2. Derived variables should be developed during data processing. The derived variables often need information from several persons or the housing information as well as the person information.

3. Many variables can be created in this way. For example, if date of birth is reported, but not age, then age can be determined one time by subtracting the date of birth from the census reference date, and this information will be stored on the record. Similarly, household income can be obtained by summing the types of income for each person, and then all the individuals, and then placing the sum on the housing record for later use.

4. Sometimes derived variables come from a combination of one or several entries in a single record, or sometimes from several records. For example, the classification "Not in labour force–going to school" may require looking at the responses for as many as four items. When developing table formats or planning supplementary tables, the use of derived variables will make programming easier and more efficient, as well as help to make data comparable over time. Some examples of possible derived records are given below.

A. DERIVED VARIABLES FOR POPULATION RECORDS

1. *Geographic variable: Urban/Rural and Region*

6. Almost all variables in censuses are disseminated by geographical areas, especially by urban and rural and by a set of regions. In order to easily disaggregate data by population and housing topics, some geographic variables can be created based on geographic codes of individual records which usually contain codes for state, province, district, village and enumeration area.

7. There is no standard definition for urban and rural areas, but usually it is defined according to population size. For example, if urban areas are defined as places with a population of at least 2000, rural areas will be defined as places with less than 2000 population. Once population of all administrative units are finalized, derived variables can be generated for all unit of enumerations, especially for personal records and records of building, housing unit and households. Similarly, derived variables can be added to census database for different definition of regions (such as geographic regions of a country, NUTS regions of Eurostat, etc.).

2. Labour force status

8. A derived variable for economic status can be very useful for the tabulations, but it requires information from several variables. In following the categories of *Principles and Recommendations for Population and Housing Censuses*, reconfiguration of several variables is necessary. The derived variable might consist of nine categories:

Persons in labour force
Employed
1. At work
2. With job, not at work
3. In Armed Forces

Unemployed

4. Looking for work
5. Seeking work
6.Currently available for work

Persons outside the labour force

7. Unavailable job seeker (seeking employment but not currently available)
8. Available potential jobseekers (not seeking employment but currently available)
9. Willing non-job seekers (neither seeking employment nor currently available but want employment)

Reason for not entering the labour force market:

10. Homemaker
11. Student
12. Unable to work
13. Retired
14. Other

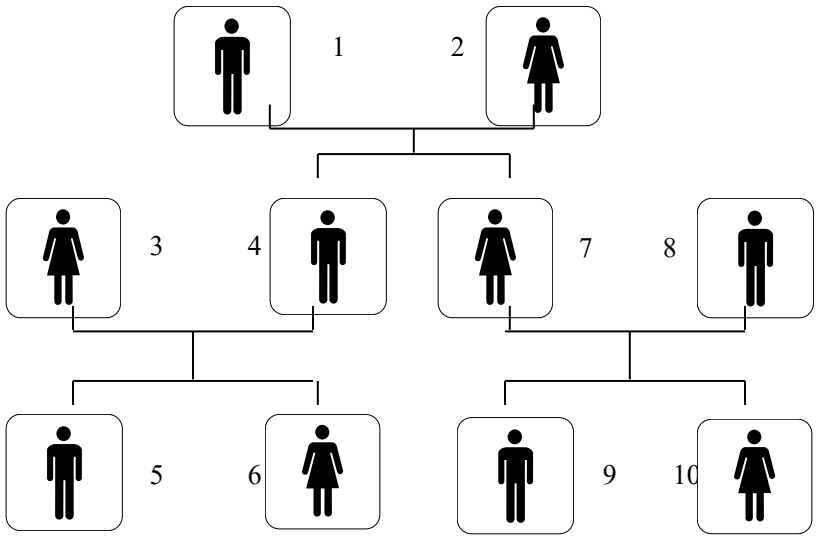
9. Since the various classifications of economic activity are used in many of the related tables, the editing team should consider adding a derived variable on to the data records rather than having the data processors reclassify labour force status during tabulation. Reclassification during tabulation may introduce errors since different data processors might develop the reclassification in slightly different ways; even a single program might reclassify differently depending on the requirements of the edit or tabulation. Specialists in economic characteristics should prepare the specifications for the derived variable.

10. When countries do have large parts of the adult population doing only subsistence activities, or subsistence activities in addition to paid work, then a different derived variable would be needed to reflect the subsistence activities.

3. Family nucleus

11. All countries have extended as well as nucleus families. Consider the following extended family, the household has a reference person and spouse (numbers 1 and 2), with two children (numbers 4 and 7). Their son (person 4) is married to their daughter-in-law (person 3), and they have two grandchildren through their son (persons 5 and 6). Their daughter (person 7) is married to their son-in-law (person 8) and they have two grandchildren through their daughter, persons 9 and

Figure A.III.1. Illustration of an extended family



12. In most censuses, if the editing team wants to study the structure of a family such as the one illustrated in figure A.III.1, it may be difficult to distinguish between the grandchildren, since persons 5, 6, 9, and 10 will all have "grandchild" recorded as their relationship to the reference persons. Recodes for family nucleus and their members will permit a more detailed analysis of the family structure.

13. Definition of a family nuclei is "a married couple (husband and wife enumerated as members of the same household) with or without never-married children." The editing team might want to add to this "one parent with one or more never-married children, living in a household. The number of family nucleus is not included in the count of households, since family nucleus members are counted as part of the reference person`s family.

14. The derived variables for family nucleus, including both the number and the type of relatives, can be defined during the processing of the data. A special edit can be developed to assist in determining family nucleus based on relationships within the household. As each family nucleus is determined numbers are assigned – in order – to each family nucleus. Code numbers then can be assigned to the various relationships:

Family "Husband/wife" is code 1,
"Spouse" of family is code 2, and
"Child" of family is code 3.

15. A family nucleus will exist when at least one pair of family relatives exists: either a "husband/wife" and "spouse" (codes 1 and 2), or a "one of parent" and "child" (codes 1 and 3).

16. In developing the derived variables, the relationship to the reference person is used to determine the relationships within the family; therefore, the more detailed the relationship coding in the census, the better the match for the family nucleus. For example, if the relationship "child-in-law" has its own code, the program will be able to match a "son/daughter" with a "son/daughter-in-law" of the opposite sex to create a family. Without this additional information, the match might still be made, but "other relative" may be ambiguous when matched, or may be matched erroneously. Similarly, the program will match codes for "sibling of reference person of household" with "spouse-in-law" and with "niece/nephew".

17. The example given in figure A.III.2. shows a household with three family nucleous: family 1 consists of household members 1 and 2, family 2 consists of household members 3,4.5 and 6 and family 3 consists of household members 7,8,9 and 10.

Figure A.III.2. Sample household with two subfamilies

Person number	*Relationship*	*Sex*	*Family nuclei*	
			Number	*Relation*
1	Head of household	M	1	1
2	Spouse	F	1	2
3	Son	F	2	1
4	Daughter-in-law	M	2	2
5	Grandchild	M	2	3
6	Grandchild	F	2	3
7	Daughter	F	3	1
8	Son-in-law	M	3	2
9	Grandchild	M	3	3
10	Grandchild	F	3	3

4. Own children

18. Sometimes countries want to produce information about "own children", who are the biological children of the reference person and/or spouse. An "own child" might be, for example, a never-married child under 18 years who is a son or daughter by birth, a stepchild (of one of the parents), or an adopted child of the reference person. Tables could show "own children" further classified as living with two parents or with one parent only. In this scheme, "own children" of the reference person living with two parents are, by definition, found only in married-couple families.

19. In a nucleus family, "own child" might refer to a never-married child younger than 18 years old who is a son, daughter or stepchild, or an adopted child of (1) a mother in a mother-child nucleus family, (2) a father in a father-child subfamily, or (3) either spouse in a married-couple nucleus family.

20. The derived variable for "own children" might be the sum of the number of own children of a particular person, usually a female, following the definitions selected by the editing teams. Sometimes users need more detailed information on own children by age. For the United States, for example, derived variables are developed for number of own children less than 6 years old and for those 6 to 17 years old. These values are placed on the records of all females. The information is used to determine the characteristics of females in the labour force with own children.

5. Parents in the house

21. The variable for parents in the house allows researchers to look at the characteristics of children in single-parent families (or no-parent families) compared to housing units in which both parents reside. The edit obtains this derived variable by determining how many parents of a person are in the house, using the relationship codes. The program looks at the relationship code for each child and uses that information in combination with the information on family nucleus to determine how many parents are living in the housing unit.

6. Current year in school

22. Many countries ask two basic questions about education:
    (a) if the person currently attends school;
    (b) the highest level of educational attainment.

23. In these countries, editing teams often find a mismatch between the two items when a person is attending school at the time of enumeration. Sometimes this circumstance may cause the person's highest level of attainment to be one year less than the current year in school. If the person is in the middle of a series of grades or levels, the statistics will be unaffected. However, if the person is attending the first grade in a series for a level, a match with data from other sources might not be possible. For example, a person attending the first grade will be recorded as being in school but having no educational attainment. Similarly, a person entering secondary school will be recorded as being in school, but the level of attainment will be the highest grade (or level) of primary school.

24. A derived variable called "current year in school" can be developed for this combination of items. If the person is not currently attending school, the code will be the same as the highest level of educational attainment. If the person is currently attending school, the edit will add one to the grade (or level) for educational attainment, and assign that to "current year in school."

25. Some countries ask three questions for education, the two items above, and a third item on whether the highest grade was completed. If this information is also obtained, it should be used as well in determining "current year in school."

7. Months since last birth

26. If the census asks adult females about the date of last birth – day, month and year or month and year of last birth – a recode can be created to get indirect estimates of year by year age specific and total fertility. The recode takes the date of enumeration, usually the month and year, and then converts this to all months, and the date of last birth to all months and then subtracts to obtain the number of months since the last birth. This figure is saved on the woman's record to assist in determining year by year fertility estimates. The raw figures can be grouped into years before the census to assist in tabulations. Note, though, that the data will be skewed when a female has more than one birth in a year or births in adjacent years before the census.

8. Disability status

27. Many countries ask questions for identifying people with disabilities in censuses. The United Nations recommends six domains (seeing, hearing, walking, cognitions, self-care and communications) for measuring disability. Of the six domains, the first four domains are considered essential. In addition to domains, scaled response categories (no difficulty; yes-some difficulty; yes-a lot of difficulty; and cannot do at all) are suggested for improving the quality of data. It is recommended to ask separate question for each domain, therefore there is no screening question for identifying whether person has any difficulty or not.

28. For this type of information, a derived variable for disability status can be generated based on any response that is "yes-a lot of difficulty" or "cannot do at all" for any type of domains (UN Principles and Recommendations for Population and Housing Censuses Revision 3, paragraphs 4.193-4.213). For example, disability status variable can be generated as following:

a. It can be coded as 1, if the person with no disability meaning that the person has either "no difficulty" or "some difficulty" for all domains;
b. It can be coded as 2, if the person with disability meaning that the person has either "a lot of difficulty" or "cannot do at all" for at least one domain
c. It can be coded as 9, if there is no enough information for identifying disability status, for example if some fields are missing. The editing team may decide to impute this kind of cases.

29. Derived variable on "disability status" can be used for calculation of some indicators related to disability, such as the prevalence of disability (people with disability status "2" divided by total population) , distribution of population by disability status, the age-sex pyramid of people with disability, sex ratio of persons with disabilities.

B. DERIVED VARIABLES FOR HOUSING RECORDS

1. Household income

30. The derived variable for household income is the sum of the income obtained in all categories of income for all persons in a household. Categories of income information might include wages, own business income, interest and dividends, social security and retirement income, remittances, royalties and rentals. If total income is also collected, during the edit each person's total income should be checked by summing the individual categories. This total is then checked against the recorded total income. If the summed income does not equal the reported total income, editing teams must develop a plan for correction. Either the total must be changed to reflect the sum of the parts or one or more of the individual income categories must be changed. When the total incomes are set for all individuals in a household, the variable for household income is obtained by summing the individual incomes.

31. The editing team must consider the situation in which one or more persons in the household has negative income because of a business failure or other reasons. In such a case, the total household income will be decreased, rather than increased, by this person's income.

2. Family income

32. The derived variable for family income is the sum of income obtained in all categories of income for all persons in a family. Families, unlike households, usually include only related individuals, but this variable will depend on the country's definition of a family. For some countries, households and families will be the same, so a derived variable for family income will be unnecessary. Categories of family income information might include wages, own business income, interest and dividends, social security and retirement income, remittances, royalties or rentals. If total income is also collected, during the edit each person's total income should be checked by summing the individual categories. This total is then checked against the recorded total income. If the summed income does not equal the reported total income, the editing team must develop a plan for correction. Either the total must be changed to reflect the sum of the parts or one or more of the individual income categories must be changed. When the total income is established for all individuals, the family income is obtained by summing the individual incomes within the family.

33. As for household income, the editing team should consider the situation where any member of the family has negative income because of a business failure or other reasons. In this case, the total family income will be decreased due to negative income.

3. Family type

34. Sometimes it is useful to identify "family type" for certain tabulations. For example, a derived variable for family type might range from 1 to 8, representing the type and composition of a family. The derived variable for family type could be used to look at the impact of various characteristics on family structure.

35. As stated in *Principles and Recommendations for Population and Housing Censuses, Revision 3* (United Nations, 2017 paras. 4.140-4.145), the family within the household is defined as those members of the household who are related, to a specific degree, through blood, adoption or marriage. According to this definition, not all households contain families since a household might comprise a group of unrelated persons or one person living alone.

36. Subject-matter specialists might classify families and households from different points of view but for census purposes, it is recommended that the primary aspects considered should be that of the family nucleus (United Nations, 2017, para. 4.140).

37. A simpler method of identifying family categories is to obtain a derived variable called "reference person (or head of household), married with spouse present". The marital status of the reference person of household can be recoded according to whether his/her spouse is enumerated in the household. In each housing unit, the population records are scanned for a person with spouse as relationship. A single code for "yes" or "no" is placed in the housing record in the appropriate field. For collective quarters, this variable can be left blank, or another code can be assigned. Then, for population tables, married persons with spouse listed will be identified during tabulation.

4. Family nucleus

38. For household composition, the *Principles and Recommendations Revision 3* developed a code for family nucleus, defined as one of the following, with recode suggestions in parentheses:

1. Married couple without children (householder and spouse or a couple living in consensual union)
2. Married couple with one or more unmarried children (as above, but, through a search the household, or a recode for number of unmarried children in the housing unit, at least one unmarried child)
3. A father with one or more unmarried children (male householder, no wife in the household, with at least one unmarried child determined as above)

4. A mother with one or more unmarried children (female householder, no husband in the household, with at least one unmarried child determined as above)

39. Note that other relatives, such as grandparents in skip-generation households may also be included as part of family nuclei, depending on the country's situation. The family nucleus excludes other relatives, like siblings, and nonrelatives.

5. Type of household

40. The *Principles and Recommendations Revision 3* includes general conditions for various types of households to assist in developing a recode for household composition. Countries may choose to make a single recode, or a series of recodes, depending on potential use of the data.

41. A first recode could identify type of household, as represented by the following items, including definitions. The suggested recodes follow in the next section:

1. One-person household
2. Nuclear household – a single family nucleus, so married couple family or partner in consensual union with or without child(ren) or lone parent with child(ren)
3. Extended household – a single family nucleus *and* other people related to the householder, two or more family nuclei, or two or more persons related to each other but not part of a family nucleus
4. Composite household
5. Other types of households

6. Household composition

42. *Single person households* are households, but not families, so should be included as a separate category in the household composition recode.

43. *Nuclear family households*. Nuclear family households can be divided into (and received individual codes for):
(21) married-couple family with children,
(22) married-couple family without children,
(23) partners in consensual union with children,
(24) partners in consensual union without children,
(25) fathers with children, and
(26) mothers with children.

44. To determine the appropriate code, the sex of the householder is used, then searches of the household for a spouse and children will provide the appropriate code. If the value 2 is used for the first of two digits (the code 1 reserved for single person households), the type of nuclear household could be a two-digit code; so, code 21 would represent a married-couple family with children.

45. *Extended households.* Extended households can also be divided into categories which would include (based on the above designations):
(31) a single-family nucleus and other persons related to the nucleus,
(32) two or more family nuclei related to each other without any persons,
(33) two or more family nuclei related to each other plus other persons related to the nuclei, and
(34) two or more persons related to each other, none of whom constitute a family nucleus.

46. The actual codes would be determined by searching the household for numbers of nuclei and relationships among the persons in the household. If a household is already coded as nuclear, the procedure would not be done.

47. *Composite households.* All other households would be composite households. Using the same scheme as before, we would have the following:

(41) a single-family nucleus with other persons, some of whom are related to the nucleus and some of whom are not,

(42) a single-family nucleus with other persons, none of whom is related to the nucleus,

(43) two or more family nuclei related to each other plus other persons, some of whom are related to at least one of the nuclei and some of whom are not related to any of the nuclei,

(44) two or more family nuclei related to each other plus other persons, none of whom is related to any of the nuclei,

(45) two or more family nuclei not related to each other, with or without any other persons,

(46) two or more persons related to each other but none of whom constitute a family nucleus, plus other unrelated persons; and

(47) non-related persons.

48. Once again, a series of searches and summaries will permit the appropriate designation for each type of household.

7. Family composition

49. Families are a subset of households, so the recode for family composition will include those categories appropriate for families above. A one-person household does not constitute a family so will not be included in the recode for family composition. Similarly, composite households are households but not families, so also will not be included. Individual countries will then decide whether they want to include a single recode for all families (nuclear and extended families together) or separate recodes for nuclear and extended families, with the understanding that these recodes will not overlap (although the case could be made to include nuclear family households with extended families for all families).

8. Household and family status

50. Household and family status represents how a person relates to other household or family members. The person's relationship to other family and household members benefits from the creation of the subfamily number and relationship described in the next section below. The approach for household and family status differs from the traditional approach of classifying household members solely according to their relationship to the head or reference person.

51. The *Principles and Recommendations for Population and Housing Censuses Revision 3* (see para. 4.148) includes the following suggested coding scheme for household status. The first set of codes is for persons in households with at least one family nucleus (that is, the household is also a family). Suggested determination of the recode is included:

1.1 Husband (male head/reference person or male spouse)

1.2 Spouse (female head/reference person or female spouse)

1.3 Partner in consensual union or cohabiting partner (from relationship codes, if present, or from combination of relationship codes and marital status)

1.4 Lone mother (determined on the basis of husband not being present for a female, but with children present)

1.5 Lone father (determined on the basis of wife not being present for a male, but with children present)

1.6 Child living with both parents (child of householder, with both parents in the house)

1.7 Child living with lone mother (child of householder, but father of child is not present)

1.8 Child living with lone father (child of householder, but mother of child is not present)

1.9 Not a member of a family nucleus (any other relative). This category can be divided into two more groups – (1) living with relatives and (2) living with non-relatives

52. The second set for the recode is for persons in households without any family nucleus, persons living alone, or with other relatives and/or non-relatives not including spouse or child of householder. These include:

2.1 Living alone (single person household)
2.2 Living with others (person living in a housing unit without a spouse or child of householder). This category is further divided into the person living (1) with siblings, (2) with other non-sibling relatives, and (3) living with non-relatives.

53. A single variable should be developed from these categories since they are mutually exclusive. The variable would be two digits. Some statistical agencies may want the first digit to be independent of the second digit – that is, the first digit will indicate whether the household status is for a family nucleus or not, and the second will identify which kind of household status the person has.

54. The *Principles and Recommendations for Population and Housing Censuses Revision 3* also include criteria for family status (see para. 4.148). These include:

(1) whether the person is male or female in a householder-spouse pair and whether they have children,
(2) whether a lone parent, by sex,
(3) whether a child of householder, and, if so, whether of a married couple or a single parent (by sex of parent), and whether not a member of the family nucleus (unrelated or related, and if related, how related).

55. The determinations shown above for household status can be used for family status.

9. Household structure under stress situations

56. The impact of the HIV/AIDS epidemic affected many countries, so a recode assists in describing different kinds of housing units. For example, if the recode describes missing generation households (grandparents and grandchildren only), households with heads under 18, widow-headed households, and so forth, this information could be used to assess the social and economic impact of the epidemic, albeit very indirectly. Children in and out of the work force, structure of the work force within these households, etc., assists government planners in fully describing the impact of the HIV/AIDS situation.

57. The HIV/AIDS epidemic is currently under control so countries may not want this derived variable. But, as noted, it shows missing generation households, households headed by persons under 18 and so forth.

10. Related persons

58. Related persons are those persons who are related to the reference person/head of household in some way. The derived variable for related persons is the sum of all persons related to the reference person/head of household. This value is particularly important in situations where large numbers of persons who are not related are living together in housing units. When many unrelated persons live together in this manner, they are often classified as living in "collective quarters" or "group quarters."

59. When developing datasets, national statistical offices often develop derived variables for different sets of related persons by age. For example, derived variables might be developed for specific groups of related persons. For example, a country may want recodes for related children 0 to 5 years old, related children 5 to 17 years old, related children 6 to 17 years old, related children 0 to 17 years old, related persons 65 years of age and over, and related persons 75 years of age and over.

60. "Related children" in a family might include, for example, the head of household's own children and other persons under 18 years of age in the household, regardless of marital status, who are related to the reference person/head of

household, except the spouse of the reference person/head of household. Related children may or may not include foster children since they are not related to the reference person/head of household, but this decision would depend on the country's situation.

11. Workers in family

61. Sometimes countries want to compare household variables by number of workers in the housing unit, such as income distributions by household size and workers per dependent. The country would obtain the derived variable for the number of workers in the family by summing the number of persons who worked at least one hour in a reference period, such as a week or a year (either a calendar year or the last 12 months). The country could use the number of persons performing work "last week", if data are collected only for that period.

12. Complete plumbing

62. Several items on the census questionnaire are used to obtain data on plumbing facilities. These items are usually related to the presence of piped water, a flush toilet, and a bathtub or a shower and are usually obtained at both occupied and vacant housing units. A derived variable for complete plumbing can assist in comparing socio-economic conditions between areas or groups at one point in time, or over time. The derived variable for complete plumbing might be obtained, for example, when three facilities—piped water (either hot or cold), flush toilet, and bathtub or shower—are present (either inside the unit or outside the building in which the unit was located). The editing team will need to determine the most appropriate set of variables for complete plumbing.

63. In this example, the derived variable can be obtained when the three items are asked separately, and during the editing operation, the sum of the presence of all three items will be determined. If the housing unit has piped water, a flush toilet and a bathtub or shower, then it "has complete plumbing". Without all three items, it "lacks complete plumbing." Codes can also be developed for one or two of these items.

13. Complete kitchen

64. Censuses are used to obtain data on kitchen facilities from questionnaire items concerned with cooking equipment, refrigerator and sink; these items are gathered for both occupied and vacant housing units. A unit might be considered to have "complete kitchen facilities" when cooking facilities (electric, kerosene or gas stove, microwave oven and non-portable burners, or cook stove), a refrigerator, and a sink with piped water are in the same building as the living quarters being enumerated. They need not be in the same room.

65. The derived variable is obtained when the above three items are asked separately and, during the editing operation, the sum of the presence of all three items is determined. "Lacking complete kitchen facilities" includes those conditions when all three specified kitchen facilities are present, but the equipment is in a different building; some, but not all facilities are present; or none of the three specified kitchen facilities is present in the same building as the living quarters being enumerated. Codes could also be developed for when one or two of the items is present.

14. Gross rent

66. Countries may collect data on cash or contract rent. Cash rent usually excludes the cost of utilities. Sometimes countries also need information about gross rent. Gross rent is the cash or contract rent plus the estimated average monthly cost of utilities (electricity, gas and water) and fuels (including oil, coal, kerosene and wood) if payment of these is the responsibility of the renter. Gross rent is intended to eliminate differentials resulting from varying practices with respect to the inclusion of utilities and fuels as part of the rental payment. Renter units occupied without payment of cash rent may be shown separately as "no cash rent" in the tabulations.

67. The derived variable for gross rent is obtained by summing the amount of rent paid and the amount paid for utilities and fuels, if these are collected separately. The result can be a specific amount or could appear as categories. Specific amounts are general better because they can be aggregated into categories, but the reverse is not true – categories cannot be disaggregated after the fact. But sometimes, only category data are available.

15. Wealth index

68. The 'wealth index' is a measure of well-being in a country, or parts of a country. The index is built from the household assets, in most cases. Often, factor analysis is used to obtain the best set of items and the variants within those items. Usually, the items are assigned binary values – 1 for present and 0 for absent – and then the values are summed. The higher the value, the more 'wealth'. So, for example, having a TV would be coded 1 for presence, 0 for absence. But toilet might be coded 1 for outhouse, 2 for gravity flush, 3 for flush (so three sets of binary variables). The various items might be weighted in the sums.

69. Quintiles can then be created by taking each fifth part of the distribution of the values of the wealth index. The lowest $1/5^{th}$ would be the poorest households, the highest $1/5^{th}$ would be the wealthiest households. Deciles or quartiles or some other division could also be created.

ANNEX IV - RELATIONSHIP OF QUESTIONNAIRE FORMAT TO KEYING IN PAPER QUESTIONNAIRE

1. The two most common questionnaire formats for population items in a census or survey are person pages and household pages.

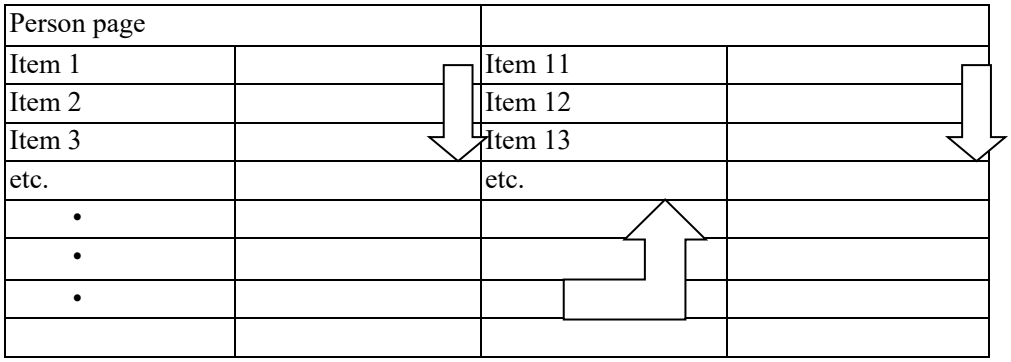
2. Person pages contain one page or two facing pages of population information, with separate pages for each person. This method is useful because all of the information for one person appears on one page, making it easy to collect. Also, this format makes it easy to check for internal consistency during enumeration. Person pages may be combined in a bound booklet for ease of handling in the field as in figure A.IV.1.

Figure A.IV.1. Sample questionnaire form with person pages

Person page for person X		*Person page for person X+1*	
Item 1	Item 10	Item 1	Item 10
Item 2	Item 11	Item 2	Item 11
.	.	.	.
.	.	.	.
.	.	.	.

3. Coding and keying for items on person pages is basically a mechanical operation, in which the coder/data entry operator is not expected to evaluate the validity of the information supplied but rather assign its appropriate code or keystroke. Figure A.IV.2. illustrates the flow of information for a given person recorded on a single page. It is easier to enter data on a single page for that person than to key by turning pages. Validity checks are implemented later during the computer edits.

Figure A.IV.2. Example of flow within a questionnaire with person pages



4. Household pages have all information for a household on one page, if possible, or on a series of pages with all household members listed on each page. Listing the household members in this way is useful because the questionnaire items do not have to be printed for each individual, thus saving space. In addition, the enumerator can compare entries between household members as the data are collected.

Figure A.IV.3. Sample questionnaire, household page with all persons on same page

Household page					
Person	*Item 1*	*Item 2*	*Item 3*	*Item 4*	*Etc.*
1					
2					
3					
4					
5					
.					
.					
.					

5. A third method is to have separate forms for each person, with the enumerator then assembling a loose booklet during or after enumeration. This method is efficient since the enumerator collects only the exact number of forms (pages) necessary for the household. The disadvantage is that the forms may separate during transfer or other handling, creating many potential editing and coverage problems if the census office is unable to reassemble them for the correct household.

6. The physical size of questionnaire pages is also a consideration, not only for enumeration, but also for keying. During coding and keying, the document must lie flat on the surface of the worktable, and coders or data entry operators must be able to locate and handle items on the form easily.

7. When all information is on a single page, staff can easily key the household pages as well, and it will obviously be faster since the data entry operator does not have to turn pages. Figure A.IV.4. illustrates the flow of information on a household page.

Figure A.IV.4. Example of flow for a questionnaire with household pages, with multiple persons per page

Household page					
Persons	*Item 1*	*Item 2*	*Item 3*	*etc.*	
1	⇨				↩
2	⇨				↩
3	⇨				↩
4					
5					
•					
•					
•					

8. Problems can occur in keying population or housing information that extends over more than one page. To solve the problems, the national statistical office is likely to take either of the two approaches outlined below:

9. Data may be entered one person at a time. The data entry operator may key the line of information for a person on the first page of the series of pages, and then turn to the second and subsequent pages. At the end of the first person's pages, the data entry operator then turns back to the first of the household pages for that household, and keys the second person, third person and so forth. This type of keying works as long as the data entry operator can

remain on the proper line throughout the keying. Although computer editing programs can be created to disentangle information when person items are erroneously keyed on another person's line, the program itself is very difficult to prepare.

10. Data may be entered one page at a time. The data entry operator may key a whole page of information before moving on to the next page. Here, the data entry operator keys all information on the first page regardless of the number of persons. Then, the data entry operator turns the page and keys the next part of the information for all persons. Skip patterns may or may not be included here, depending on the type of keying (with or without computer editing). In any case, during the computer edit, the records from the various sets of keyed data will have to be assembled, and any miskeys of person numbers will have to be dealt with then.

11. In the following example (figure A.IV.5), the household's demographic information poses no unusual keying problems since the census obtained a response for all items for all persons.

Figure A.IV.5. Example of a household page with multiple persons, without keying problems

Household page					
Persons	*Relation*	*Sex*	*Age*	*Etc.*	
1	Head of household	M	40		
2	Spouse	F	35		
3	Child	F	18		
4	Child	M	12		
5	Sibling	M	35		
6	Sibling of spouse	F	30		
7	Sibling child	M	5		
8	Sibling child	F	3		
etc.					

12. Many times a country must use the household form because of cost or space constraints. However, when the population is small, or the country can afford the additional expense, the form with person pages is likely to contain fewer matching errors through miskeying than occur with the household forms.

ANNEX V – SCANNING VERSUS KEYING

1. Several simultaneous and different methods for data capture are being used in a census when paper questionnaire is used. They include basically keyboard data entry and scanning technology. Many countries are using scanning equipment, either optical mark reading (OMR) or optical character recognition (OCR)/intelligent character recognition (ICR). Each of these has advantages over keying when the operation is smooth and efficient and when the costs are not great. On the positive side, many countries use the scanners obtained for the census for a number of purposes, including other surveys and such administrative records as entry and exit forms. However, unlike keying where the skills transfer easily to other applications, the basic skills involved in feeding documents into a scanner transfer only if the same or similar machines are used.

2. When a country is deciding whether to use scanning equipment for its census, it should also decide whether the country will continue to use the machines. Multi-purpose machines continue to be useful long after the census. However, national census/statistical offices that key their data will find that the skills learned transfer and the machines themselves continue to be useful either in the national statistical office or elsewhere in the Government. The continued use of the equipment should be factored in when making a decision about keying or scanning. Also, countries should consider out-sourcing the scanning which can be cheaper and more efficient than buying scanners and trying to do all for the work internally.

3. The quantity and type of data entry equipment required will depend on the method of data capture selected, the time available for this phase of the census, the size of the country, the degree of decentralization of the data capture operations and other factors. For keyboard data entry, the average input rates usually vary between 5,000 and 10,000 keystrokes per hour. Some operators stay well below that range, while others surpass it significantly. Among the factors that affect operator speed are (a) the supporting software and program; (b) the complexity of the operators' tasks; (c) the ergonomic characteristics, reliability and speed of the equipment; (d) the question whether work is always available; (e) the training and aptitude of the recruited staff; and (f) the motivation of the workers (United Nations, 2017, para. 3.180).

A. ENTERING THE DATA

1. Scanning

4. Imaging techniques or scanner devices, together with optical character reading and intelligent character recognition software, have been used by several countries for data capture. Experience shows that significantly low error rates are achieved at an optimum cost using these techniques. The efficiency is greater in the case of numerical and alphanumerical characters written by trained enumerators. However, alphanumerical characters are prone to higher error rates. The use of scanning method is also dependent on the availability of local maintenance and support capabilities. In addition to the benefits of the scanning technology for capturing the information, an important by-product of scanning census questionnaires is that this allows for the possibility of digitally filing and naming the scanned questionnaires. This increases the efficiency of storage and retrival of the questionnaires for future use, particularly during subsequent data-editing operation (See UN Principles and Recommendations for Population and Housing Censuses Revision 3, paragraphs 3.174-3.181).

5. Countries who choose to key their data, however, have several choices, depending on how quickly they need the data keyed and how much manual checking is needed. Each of the options depends on the requirements of the editing teams, the skills of the data entry operators and the sophistication of the editing program. In large census operations, the biggest problem is getting the data keyed at all. The method producing the fastest results should be decided based on extensive testing of alternative approaches. .

2. Keyboard data entry

6. Keyboard data entry takes two forms. The first is keying all data items as they are encountered with no skip patterns. In this case, keying proceeds more quickly since data entry operators do not have to stop when invalid or inconsistent information is encountered. It may also be more accurate since keyers perform the task more mechanically. The second form of keyboard data entry entails stopping to check the questionnaires for invalid or inconsistent results, so the process will go more slowly and will require much more expertise on the part of the keying staff. The high price in terms of speed must be seriously considered. Paradoxically, accuracy may also be improved with this method if data entry operators find that the data were recorded correctly but were miscoded. Miskeying itself may sometimes be immediately challenged because the editing package provides for automatic checking.

7. After developing keying instructions, national census/statistical offices must have actual data entry operators test the keying instructions before deciding on the actual operation, whether or not heads-down keying is used. As the keying instructions are tested, bugs can be worked out of the system, and optimum keying can be obtained.

(a) Keyboard data entry without skip patterns

8. When all entries are keyed, or skipped manually, a particular rhythm can be maintained, and certain skip patterns will not obviate valid but temporarily inconsistent information. For example, if a person is recorded as male, most editing teams will require that the whole section on fertility be skipped. In this case, a data entry operator will key through, (use the space bar or arrow to move through a male's or young female's record) because all fields will be blank. However, this takes time, and the spacing may not be completely accurate. For example, the data entry operator might go too far or not far enough, and other items might be miskeyed because they are improperly aligned. If all fields are keyed in this way, then this information can be keyed when no skip patterns are present.

9. For example, when the data entry operator encounters an adult female with fertility (a female for whom such items as children ever born, children surviving or children born in the last year have been collected and coded), all items are keyed. If the fertility information is keyed, the computer editing program can determine which item or set of items is valid and which must be changed. When the edit determines that the person is an adult female, but the fertility information is blank, then dynamic imputation or other appropriate means has to be used to obtain fertility information for the tabulations. If the actual information has been lost because of the skip patterns, the editing team must decide whether the loss is worth the increased efficiency and speed. If skip patterns are present, the data entry operators can still move backwards through the screens to the appropriate position for corrections. Although the data entry operators will waste some time spacing through items they do not key, with this form of data entry, inconsistencies between sex, age and fertility can be attacked during the edit rather than at the time of keying.

(b) Keyboard data entry with skip patterns

10. A second method of keyboard data entry involves keying with skip patterns in place. Again, if the editing team requires skip patterns, usually to represent the way the enumerators collected the data, keying is easier and faster if the skip patterns are easy to follow and if the data entry operators learn the keying patterns quickly. If the skip patterns are very complicated, data entry operators may become confused and persistently key in the wrong places. The most efficient keying with skip patterns occurs when limited patterns that cover large parts of the record being keyed are used.

11. The editing team will need to determine the appropriate skip patterns for their country's census or survey. For example, it makes sense, to skip all of the employment items for children, that is, persons below the country's defined age for potential employment. Often, these are half of the population items, so it is efficient to skip them for children, except for special situations such as for children whose age is borderline, or when the country may be interested in child labour.

12. The editing team decides on an item-by-item basis which items will be included for which age groups. Staff can group the items to manage the skip patterns easily.

13. It is not always easy to have clear-cut decisions about skip patterns. For example, consider the following sequence:

---

1. What is this person's place of birth?
   - Born in this country (skip to item 3)
   - Another country, please specify ….

2. What is this person's year of entry?
3. NEXT ITEM

---

14. A skip pattern could be created to skip from 1 to 3, that is, skip the item on year of entry, for persons born in the country. However, sometimes data entry operators violate the skip pattern, either because the enumerator or coder makes a mistake, or because of miskeying. The many factors involved include the skill level of the data entry operators, the cultural circumstances, the layout of the questionnaire and the layout of the screens. The editing team often works together to determine whether a skip pattern in a case such is this case is reasonable.

3. Interactive keying

15. Interactive keying may be used during census input but is more appropriate for surveys, particularly for small surveys where allocated items could affect the results of the survey. Interactive keying may involve manual or automatic corrections, depending upon the information available to make changes or corrections.

16. Consider the case of a small survey. For small surveys, every response is important. If a country takes a 1 per cent sample survey, for example, each response represents 100 persons, housing units, or agricultural holdings. A few invalid or inconsistent cases could have a considerable impact on the results of the survey. In these cases, the demographers and other social scientists usually want to have considerable control over the processing.

17. Control may be established in several ways. The demographers and other specialists may key the data themselves, checking for extraneous, invalid or inconsistent responses as they go along, using the information as recorded on the data collection forms. They can often resolve conflicts, miscodes, or other inconsistencies immediately, while looking directly at the collected information. Sometimes they may opt to send incomplete or invalid questionnaires back to the field. This type of interactive keying gives the best results since the demographer also provides guidance to data entry operator, but it is by far the most expensive, and not many countries can afford this method.

18. The editing teams may develop very detailed edit rules to determine what data entry operators must do when particular cases occur during keying. For each unresolved invalid code, they can decide what the data entry operator will key. The editing team may resolve cases not covered by the detailed rules and may modify the rules (although at the risk of having inconsistencies between the first part and later parts of the keying).

19. Skip patterns which play an important role in keyboard data capture, are important here, too. As with keying, data entry operators must be aware of and learn any skip patterns in use. As mentioned above, skip patterns can increase the speed of keying, but usually with some loss of quality. For interactive keying, a common rule of thumb is that the fewer the skips the better the quality.

B. VERIFICATION

20. The national census/statistical office must also decide what level of verification is appropriate. For keyed data, several countries use 100 percent verification, while many countries rely on a sample of data for verification. In the first case, all items are rekeyed (or keyed over the existing information) to make certain that the data collected are the data that go into the machine for computer processing. Often, however, total verification is not practical, because the country does not have the time, financial or human resources to rekey all of the data. The percentage sample verified should be larger for beginning keyers, but less for more experienced keyers. Also, if the tested error rate for the keying is very low, with the data entry operators making very few errors, then complete verification is probably not necessary.

21. In any verification operation, it is first necessary to determine what information is needed. Does the country want to track individual operators? A team of keyers? Determination of whether skills are being acquired, or maintained? The units of control could also be important, including daily, weekly, monthly reporting, and so forth, to determine the flow of work and the skills gained.

22. Finally, it is very important that verification be independent, that a different set of keyers do the verification from the data entry, or, at least different parts of the same team. The different sets of keyers allows for independence in the operations, and, therefore, better results.

23. For scanned data, some countries also perform verification to make sure that the scanning was comprehensive and complete. It should be kept in mind that, even when the systems are thoroughly tested with pilot or pretest data, changes in paper quality, actual printing of forms in various places, storage, etc, can cause problems that need to be addressed through verification.

24. If errors are systematic and can be removed through the edit program, keyers and verifiers should not be making judgments about correction. However, the keyers and verifiers are responsible for finding the errors. These errors could include inadequate testing of the scanning equipment causing systematic errors for certain items or combination items, confusing in reading certain digits (for example, interchanging 2s and 3s, or 8s and 9s), continuation check-off boxes, and so forth.

25. Misreading of continuation checkoffs has been a continuing problem in recent years, and can only sometimes be addressed in the edit. If the forms are not contiguous, other procedures will be needed, most likely during the structure edits, to resolve issues. As noted earlier, a completely sound, structured file is needed before content editing begins.

26. There are two approaches for verification:

a. Dependent verification :

In dependent verification, data entry operators key over data previously keyed by other staff. When the key strokes differ, the software package informs the data entry operator, and, depending on the program, the data entry operator either overrides the previous data, or a note is made of the discrepancy. Since the data are keyed from the original questionnaires, usually the data entry operator himself or herself can make an informed decision about whether the original keying was in error.

b. Independent verification

In independent verification, data entry operators rekey the data from scratch; they create a completely independent file of the keyed data, using the original questionnaires. The two resulting files, the original keyed data set and the verification data set, are then compared, using a computer program, to test for discrepancies. Presumably, some manual operation is used to rectify invalid and inconsistent key strokes.

C. EDITING CONSIDERATIONS WITH SCANNED DATA

27. More and more countries now scan their data. In the early 2000s, many of these countries were surprised to find that scanning introduces different types of errors than keying. Part of the problem with editing scanned data involves lack of quality control during the scanning process. Because the technology was so new in the early part of the 2000s, many statistical offices did not have the background or facility to develop appropriate quality control for all items. Many of the countries that did develop appropriate quality control procedures, did not end up developing them for all items, and so some of the items at the end of the question – particularly the fertility items – ended up being invalid or inconsistent.

28. Of course, many of the inconsistencies found in keyed data also occur in scanned data, and this handbook primarily addresses problems in keyed data since as of this writing most surveys and many censuses continue to be keyed. However, it is useful to take some space to discuss the special problems evolving from the use of scanned data.

29. Because scannable questionnaires require markers to assist the machine in reading them, sometimes items can be displayed in ways that cause problems for enumerators and respondents during data collection. These items must be addressed systematically. When the items are closely related to other items, like sex and fertility, the regular edits described in the text can be used.

30. However, care must be taken when items needed for planning and policy have problems. The item for sex usually does not have problems because of it has only the two possibilities. But, as noted above, while keying usually restricts the keyer to only be able to key a 1 or a 2 (or a code for 'unknown'), any value can appear in the columns for sex – other digits, or alpha-characters or other characters. So, some edit needs to be added to what was once done for keyed data to account for these miscellaneous values.

31. **Relationship codes** are a good illustration of the problem. If you have a single digit for relationship codes, as shown in the text, you normally won't have any problems. But if two digits are used, then sometimes a problem will occur in scanning when the first digit is either coded incorrectly or picked up incorrectly by the scanner. Normally, if codes 1 to 12 are used, the keyer will be restricted to keying these codes only, and the entry package will "complain" when an illegal code is entered. With scanning, almost anything will be accepted (although newer scanning packages can be programmed to "complain" as well.) Then, erroneous codes must be changed during edit, or they will cause all sorts of problems during tabulation stage.

32. **Age** sometimes is an issue, particularly when 3 columns are used (to allow for people to be more than 100), so a digit by digit analysis may be needed – that is, looking separately at the ones, the tens, and the hundreds digits – to do a proper edit. Once it is established that the age has been properly captured, the regular edit can be used.

33. However, when both **age and date of birth** are present, misleading information can cause problems when one item takes precedence over the others. Usually, subject matter specialists prefer use of date of birth with the census or survey reference date to produce an exact age (by subtraction) to compare with the reported age. When one or several digits are missing, care must be taken to be certain that all remaining digits are used properly to obtain the best estimate of the computed age for comparison. When the scan does not pick up a single digit, for example, the edit can take this into account to provide a best guess of what would have been there. This type of problem does not usually occur with keying.

34. The items with the largest problems in the early 2000s resulting from scanning involve **fertility** – both the numbers of children born and surviving as well as children born in the last year or over time. For most countries, the main problem has been lack of quality control during scanning, resulting in strange items in the data capture. When a country has 17, 18, or 19 dead female children, for example, without editing the data are useless for planning.

35. **Mortality** information also can present problems in the scanned data. For keyed data, if you have a series of items for deaths in the year before the census (sex and age of the deceased, and whether the person died a natural death, and whether it was a maternal death), the keying proceeds even over erasures and strike-outs. However, with scanned data, erasures would normally not be read, and the scanner will leave blank information before continuing with the capture. The edit program must move the information into appropriate spaces for tabulation and subsequent analysis. It should be noted that newer scanning operations can do these moves during and just after capture.

36. Unfortunately, each country's problems depend on the particular programming and functioning of the individual scanners, so general guidelines are difficult. However, in all cases so far, the scanning problems have been systematic. That is, when staffs determine the algorithm to alleviate the problems, fully edited data sets could be produced.

ANNEX VI - SAMPLE FLOW CHARTS

1. One of the tasks of the editing team is to develop a relational structure for the variables used in the editing process. Flow charts facilitate the identification of various linkages among variables, and help in the development of clear and concise editing specifications. These specifications for relational linkages help both subject-matter and data processing specialists to visualize the editing process from beginning to end and facilitate communication between the two groups.

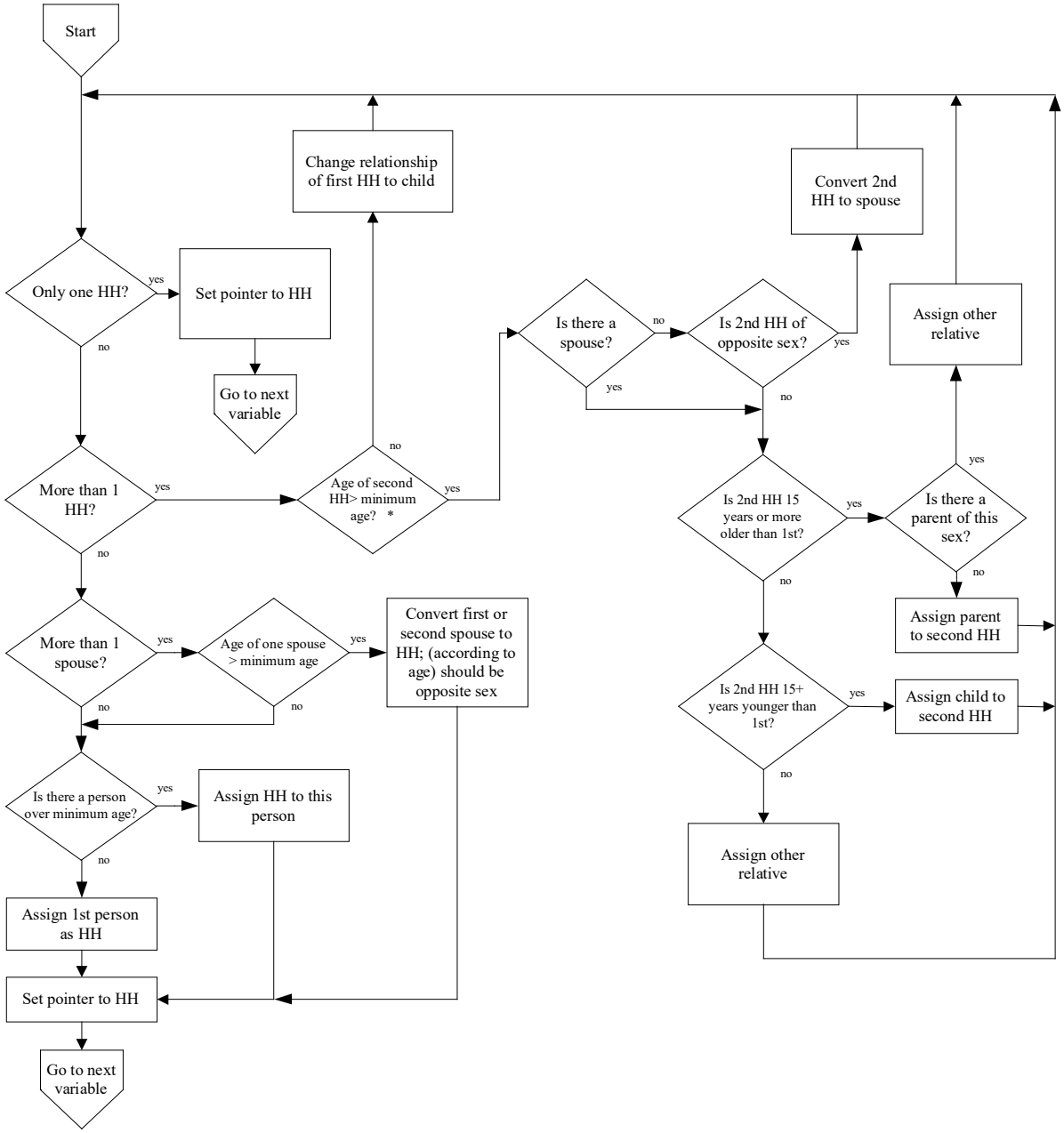
2. Three sample flow charts are presented on the following pages:

(a) Flow chart to determine the head of household or reference person;
(b) Flow chart to determine a spouse in the households;
(c) Flow chart to edit sex variable for head of household/reference person and spouse.

3. These sample flow charts are provided for illustrative purposes only and should be treated accordingly. The editing team may modify further the sample flow charts as necessary based on the situation in the country.

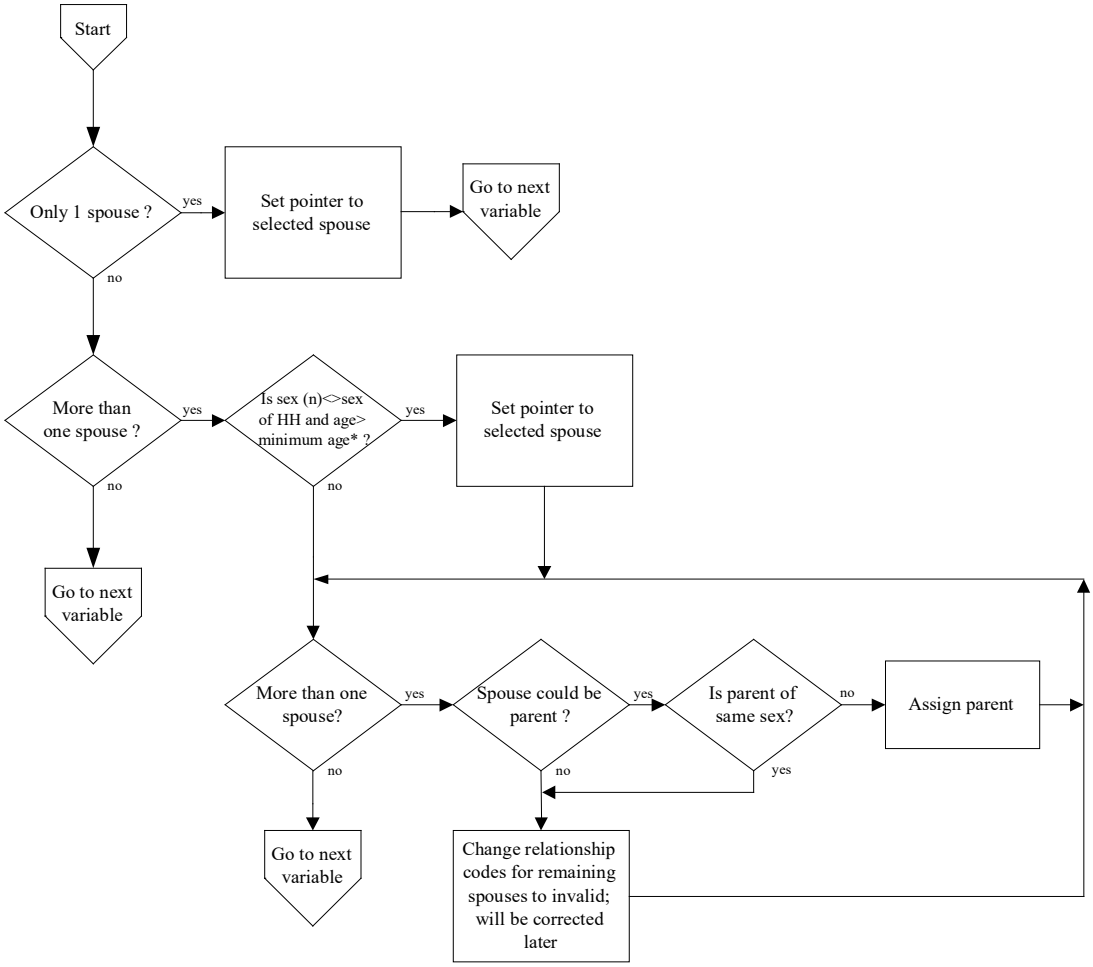
4. Editing flow charts should be developed for each variable in a census. The editing team should work together on the development of the flow charts, and the data processing specialists should use them with the editing specifications to develop computer programs to edit the census data. The flow charts and editing specifications should be properly documented for use in future census and survey data processing.

Figure A.VI.1 Sample flow chart to determine reference person or head of household (HH)



* - minimum age to be specified by the editing team

Figure A.VI.2. Sample flowchart to determine presence of spouse in household



Note: HH = Head of household * - minimum age to be specified by the editing team

Figure A.VI.3. Sample flowchart to edit sex variable for reference person or head of household and spouse for opposite sex marriage/partnership

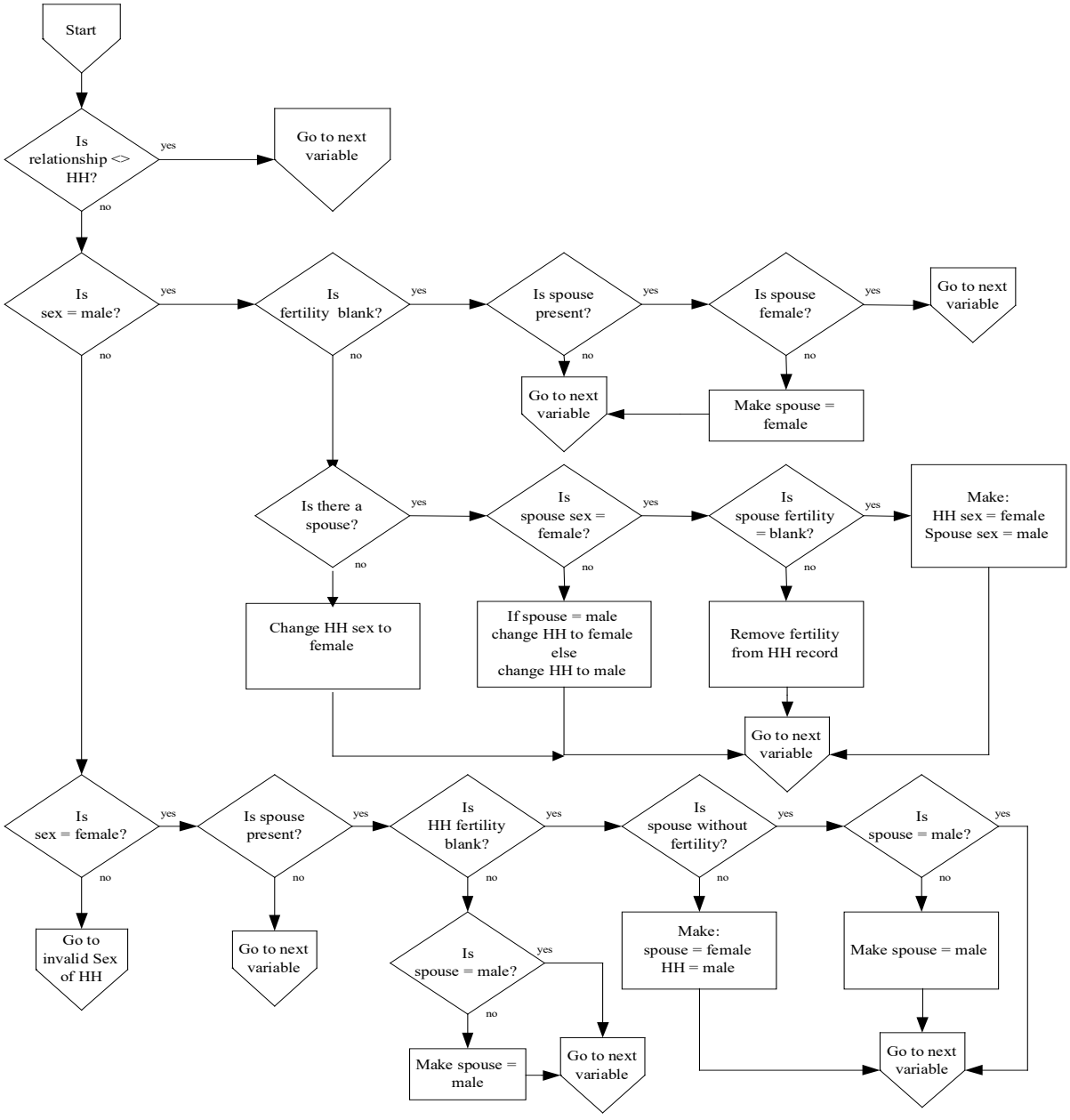
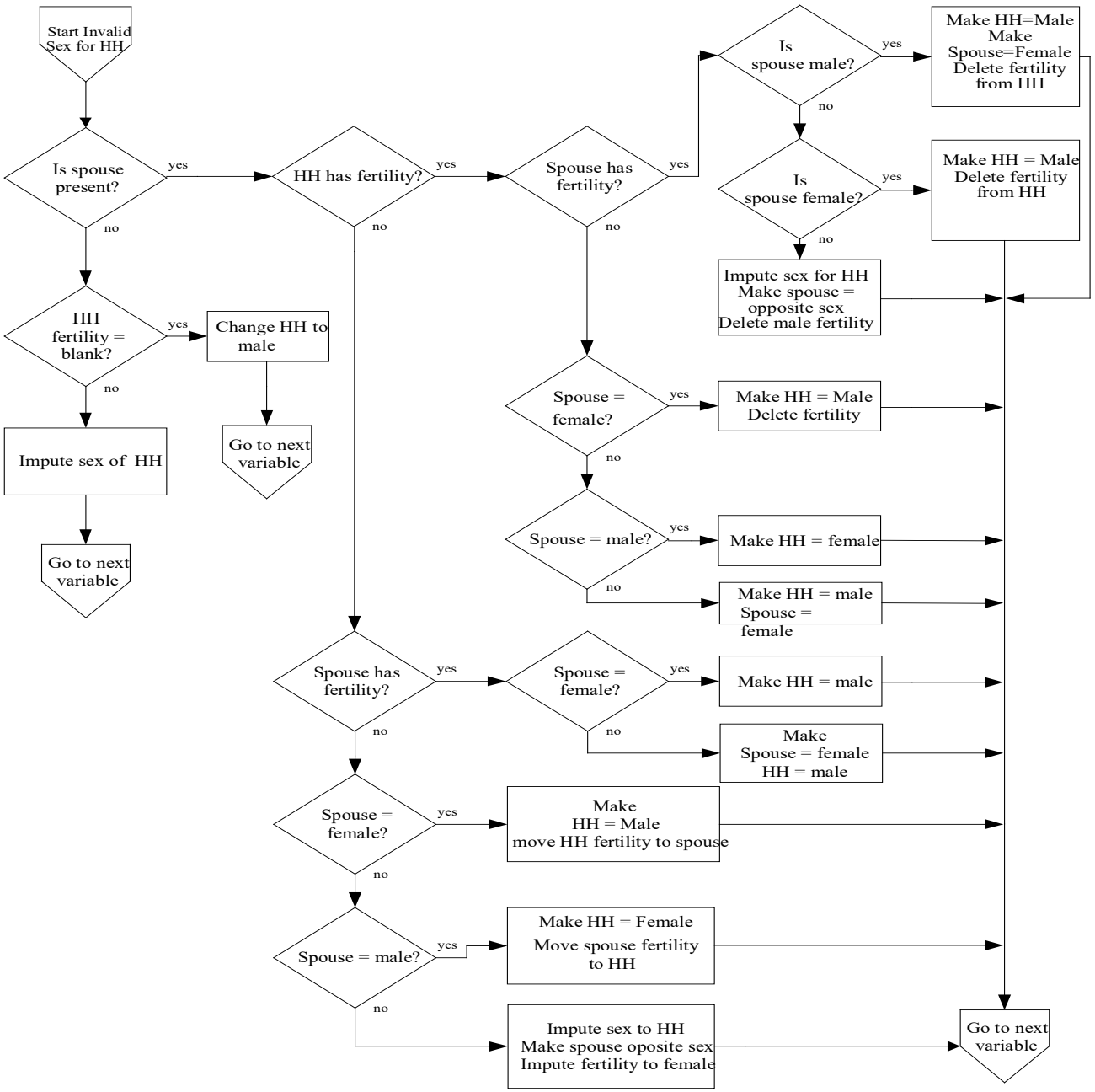


Figure A.VI.3. (continued)



ANNEX VII - IMPUTATION METHODS

1. A number of imputation methods have been developed. Most methods described below are reviewed in papers by Kalton and Kasprzyk (1982, 1986); Sande (1982); and Särndal, Swensson and Wretman (1992).

2. Imputation methods can be classified as either stochastic or deterministic, depending upon the degree of randomness in the imputed data.

3. Deterministic **imputation methods** include deductive imputation; model-based imputation methods such as mean imputation and regression imputation; and (if appropriate) nearest-neighbour imputation.

4. Deductive **imputation** is a method that allows a missing or inconsistent value to be deduced with certainty. Often this will be based upon the pattern of responses given to other items on the questionnaire.

5. More commonly, the imputation technique must substitute a value that is not certain to be the true value. Some common imputation procedures are outlined in the following paragraphs.

6. With the exception of single donor dynamic imputation algorithms, the methods described below involve the imputation of one item at a time. So, within each imputation class, the items on the record are considered one after the other in a sequential fashion. Commonly, this is done by considering only those edits pertaining explicitly to the item in question or to a small set of closely related variables. Because there may be explicit or implicit edits that link the item(s) in question to other items to be considered later in the process, this procedure may cause an imputed value, while passing the edits currently being considered, to bring about failures on other edits to be considered later in the process. Only when a complete set of edits, including all implied edits, is considered can it be assured that imputed values will pass all edits. An implied edit is one that can be derived by logically combining two or more of the explicit edits.

7. In the following descriptions, “passed edit records” refers to those which have passed all edits pertaining to the item(s) in question. “Failed edit records” refers to those that have failed at least one edit pertaining to the item(s) in question.

8. **Overall mean** imputation assigns the item mean for passed edit records to the missing or inconsistent item for all failed edit records. This method may produce reasonable point estimates but is less appealing if variance estimates are to be computed using a standard variance estimator. Variance estimates can be severely underestimated unless the imputation rate is very low or a variance estimator modified to account for imputation is used.

9. Class **mean imputation** uses imputation classes defined to create groups of records having a degree of similarity. Within each class the item mean for passed edit records is imputed for the missing or inconsistent item for all failed edit records. This is much like overall mean imputation, but the impact upon the distribution and problems with variance estimation are likely to be less severe.

10. Regression **imputation** or, more generally, **model-based imputation** uses data from passed edit records to regress the variable for which imputation is required on a set of predictor variables. The predictors in the regression can be items from the questionnaire or auxiliary variables. The regression equation is then used to impute the values for the missing or inconsistent item values. This is a special case of model-based imputation. This method is frequently used for continuous variables in business survey applications where previous occasion data can often predict well current occasion values satisfactorily.

11. Nearest-**neighbor imputation** or **distance function matching** assigns an item value for a failed edit record from a "nearest" passed edit record where "nearest" is defined using a distance function in terms of other known variables. This method can be applied within imputation classes. It is usually considered appropriate for continuous variables but can also be applied with non-numeric variables.

12. **Stochastic imputation methods** include regression, or any other deterministic method, with random residuals added and hot deck or cold deck methods.

13. For each deterministic method there is a stochastic counterpart. This can be achieved by adding a random residual from an appropriate distribution to the imputed value from the deterministic imputation. This procedure will help to better preserve the frequency structure of the data file. Kalton and Kasprzyk (1986) review some approaches to this technique.

14. **Dynamic** and **static imputation** attempt to create a more realistic variability in the imputed values than deterministic methods can. Hot deck imputation procedures replace missing or inconsistent values with values selected (at random) from passed edit records in the current survey or census. Cold deck imputation procedures impute based on other sources, often historical data such as earlier occasions of the same survey or census. There are a number of different forms of hot deck and cold deck imputation.

15. Random **overall imputation** is the simplest form of hot deck imputation. For each failed edit record, one passed edit record is selected at random from the set of all passed edit records and its reported value for the item in question is imputed for the failed edit record.

16. Random **imputation within classes** again uses imputation classes to constrain the random selection of the donor record to a set considered to have some similarity to the record requiring imputation.

17. Sequential **dynamic imputation** also uses imputation classes and has the advantage that a single pass through the data file is sufficient to complete the imputation process. The procedure starts with a cold deck value for each imputation class and the records in the data file are considered in turn. When a passed edit record is detected, its value for the item in question replaces the stored value for the imputation class. When a failed edit record is detected its missing or inconsistent value is replaced by the stored value. The number of imputation classes cannot be excessively large as it must be ensured that donors are available in every imputation class. If the order of records in the data file is random, this method will be nearly equivalent to random imputation within classes. A disadvantage of this procedure is that it often leads to multiple uses of donors and so can adversely affect the item's distribution and variance estimates.

18. Hierarchical **dynamic imputation** is an enhancement of sequential hot deck imputation in which a large number of imputation classes are used. When a donor cannot be found in the initial imputation class, classes are collapsed in a hierarchical fashion until a donor is found.

19. The objective of **single donor dynamic imputation** algorithms is to impute data for a failed edit record from a single donor. Hence they allow for the joint imputation of all item values on a record identified as problematic by the edits. Often in practice, the objective is to use a single donor for each section of closely related variables in the record. This approach provides the significant advantage of better maintaining not only marginal distributions, as the above hot deck imputation methods, but also joint frequency distributions. Another advantage of single donor dynamic imputation methods is that they reduce the problem of imputing values that will fail edits considered in subsequent sections of variables. In the context of single donor hot deck imputation methods, a passed edit record is one that has passed all edits applying to the section. A failed edit record is one that has failed at least one of those edits.

20. The Fellegi-**Holt edit and imputation method** (Fellegi and Holt, 1976) considers all edits concurrently. A key feature of the Fellegi-Holt edit and imputation method is that the imputation rules are derived from the corresponding edits without explicit specification. For each failed edit record, it first proceeds through a step of error localization in which it determines the minimal set of variables to impute as well as the acceptable range(s) of values to impute and then performs the imputation. In most implementations, a single donor is selected from among passed edit records by matching on the basis of other variables involved in the edits but not requiring imputation. The method searches for a single exact match and can be extended to take account of other variables not explicitly involved in the edits. Occasionally no suitable donor can be found and a default imputation method must be employed.

21. **The** New **Imputation Methodology (NIM)** (Bankier, Luc, Nadeau and Newcombe 1996; Bankier, Lachance and Poirier, 1999) is similar to the Fellegi-Holt method in that it considers all edits concurrently, does not explicitly specify imputation actions and imputes from a single donor. For each failed edit record it identifies minimum-change imputation actions conditional on the potential donors available. This guarantees that a donor will be available. Unlike Fellegi-Holt, NIM first searches for donors and then determines minimum-change imputation actions. NIM searches for donors by matching, using all variables (including those potentially to be imputed) involved in the edits, and can be satisfied by near matches for numeric variables plus matches for most, but not necessarily all, other variables. Imputation actions based on each potential donor are determined and those that are minimum-change imputation actions are identified. The method also considers near minimum-change imputation actions; these can sometimes yield more plausible imputed records. Finally, one of the minimum-change and near minimum-change imputation actions is selected at random and the imputation is performed.

22. Although both Fellegi-Holt and NIM are computationally demanding, efficient algorithms are available so that their implementation and application are feasible with modern computers. This is particularly true for NIM, which can readily handle somewhat larger editing and imputation problems than can the Fellegi-Holt method.

23. All of the above imputation methods produce a single imputed value for each missing or inconsistent value. All will distort to some extent the usual distribution of values for the item in question and can lead to inappropriate variance estimates when standard variance estimators are used. The extent of distortion varies considerably depending on the amount of imputation and the method used.

24. **Multiple** imputation is a method, proposed by Rubin (1987), that addresses this problem by imputing several times (*m)* for each value requiring imputation. Then, from the completed data set *m* estimates can be produced for the item. From these, a single combined estimate is produced along with a pooled variance estimate that will express the uncertainty about which value to impute. A disadvantage of the multiple imputation method is that it requires more work for data processing and computation of estimates.

25. In most imputation systems a mix of imputation methods is used; typically, deductive imputation is used where possible and is followed by one or more other procedures. Most national statistical offices use some form of dynamic imputation method for census editing and imputation. Sequential hot deck imputation and the Fellegi-Holt method are currently the most commonly used. Of the national statistical offices presently using the Fellegi-Holt method, one is changing to NIM and a number of others are considering it. However, given the expected primary readership, this *Handbook* focuses on a form of sequential hot deck imputation.

ANNEX VIII - COMPUTER SOTWARE AND APPLICATIONS FOR DATA EDITING

1. Microcomputers and software have made it possible for countries to edit census and survey data thoroughly as well as in a timely manner. This handbook discusses background related especially to the use of software to edit census data and factors to consider in making decisions about when deciding on how editing will be implemented within a census program. Most of the software packages or applications discussed here were designed for larger scale editing operations but the information provided in this annex may be applicable to any census or survey. This annex is meant to provide basic information for those countries who may be interested in both automated approaches to data editing and interactive editing in an application.

2. When census or survey files get larger, the use of some statistical methods, such as regression and multiple variate analyses, may make more demands on the computer hardware and software in use, and may result in operational difficulties such as longer times to process or even difficulties in completing intended processing. The complexity of edit rules to be used, and factors such as the expected impact of item non-response, may vary according to a countries' census or survey context. Hence, each country office must consider carefully its overall computing environment in determining the best packages or programs for its needs. This must include testing various packages within the computing environment to be used.

3. In the past, each country had to write its own editing program in a custom fashion, requiring expensive debugging and processing time. As standard editing software have been developed and become available, it may be that one of these may be suitable to a country's editing needs, and greatly reduce the development time to create a custom editing program. The need to develop expertise in the use of standard software editing packages or programs may still be a factor to be considered during decision making concerning the best editing software to use. It is expected, for instance, that edit rules will still need to be input and tested within any software used.

4. One advantage of using editing software is that when properly used, data will be output in a standard, reliable fashion for all inputs. This may aid subsequent steps such as tabulations. Generic software packages such as SAS and SPSS, or other high-level languages, can be used to write editing programs in the customized fashion noted above. Or, a country can choose to use software applications written specifically for some stage of editing of census and survey data. For most countries, using an already developed and generalized edit software may be more efficient than developing a custom application, and allow expertise to be more fully focused on editing decisions as opposed to software development.

5. Communication between subject matter specialists and those responsible for implementing the edit rules they describe is an important aspect of a successful editing operation. A key consideration in selecting the right editing software is ensuring that effective communication can be supported. For instance, a good editing package should allow the placement of narrative or pseudocode that is transparent, and aids in the development of test strategies for edits, and later in resolving issues if they are discovered. Specialists need to be able to understand how their specifications have been implemented within the software being used.

6. Any editing software that a country might consider for use will need to meet a variety of requirements across entire edit process. For instance, it will need to produce reports for the various checks, tests and imputations required for editing census data. It should be noted that not all requirements need be satisfied by a single edit software package. In Canada, for example, edits are applied during the data capture phase by one software application, but imputation is carried out by another (CANCEIS, discussed below).

7. Regardless of whether a custom editing program is developed, a single editing package is used, or a number of different applications make up the overall editing process, all editing requirements for the country must be met. Understanding how each requirement will be met will help clarify the software selection process. The following list of requirements illustrates the range of consideration needed in making decisions about software use (note that these requirements are discussed at some length throughout the handbook and annexes):

(a) The processing operation may to key and/or verify entry data. Depending on how this is done, skip patterns may need to be considered. For example, the editing team may decide that fertility information must be skipped for males;
(b) Structural edits may be implemented at various points in an editing process, which make it possible to determine whether the types of records that should be present are in fact present, including, for example, a housing record for each serial number;
(c) The processing operation may, respecting subject matter rules, generate records if they are missing and/or add weights to existing records;
(d) The processing operation may require that all or part of previously edited records be retained;
(e) The processing operation may require that consistency between two or more characteristics in the same record and between records be tested. A subset of this is to test the consistency within households, checking responses with those of previous household members.
(f) The processing operation may impute values by the hot deck technique, if the country chooses to use dynamic imputation;
(g) The processing operation may use several values within a record or from multiple records to construct a derived variable and insert the derived variable in the appropriate record;
(h) Duplicate records will need to be eliminated;
(i) The processing operation may require reports documenting errors found and changes to parameters by small geographical area.

8. Consideration should be given to whether editing software will need to edit one record at a time, or require inter-record checking, particularly within housing units. It may not be the case that all software would support both requirements.

9. As noted in the text, until the Fellegi/Holt (1976) method and its follow-ups, almost all editing used a top-down approach. That is, items were edited in order, usually – but not always – in the same order as they were collected. The first population item, for example, is usually relationship, so it would edited, then sex would be edited on the basis of that item, then age could use both sex and relationship, and so forth.

10. In the last few decades, several minimum change imputation systems based on Fellegi/Holt have been developed at Statistics Canada, each incorporating incremental improvements to either function or performance, and there have been several "precursors" to the current system used at Statistics Canada. These (and other) tools were developed based on another approach to minimum change (the NIM – new imputation methodology), by Mike Bankier at Statistics Canada. NIM was first implemented in CANEDIT then replaced by the CANadian Census Edit and Imputation System (CANCEIS) (Bankier 2005, Chen 2007). CANCEIS has been in use at Statistics Canada since 2001, and is used to process data after certain edits during collection and data capture have been applied such as validity edits in electronic questionnaires and checks for record duplication.

11. In a review of the CANCEIS, the Canadian authors summarize: "CANCEIS, with its highly efficient editing and imputation algorithms, shows great promise for solving very general imputation problems involving a large number of edit rules and a large number of qualitative and quantitative variables when minimum change donor imputation is appropriate. The Fellegi/Holt minimum change edit and imputation algorithm, however, should still be the method of choice for smaller imputation problems if there may not be sufficient donors available or if it is

more appropriate to use another method to perform imputation". (Bankier, Lachance and Poirier 2000. p.10) Since then, CANCEIS has proven itself within the Canadian Census context and is now used by other agencies as well.

12. The 1996 Canadian Census (and others) used a different approach, called Nearest Neighbor (NIM). The 1996 version imputed responses for age, sex, marital status and relationship for all persons in a house simultaneously (Bankier 1999). The method was improved and expanded for the 2001 and subsequent Canadian statistical activities (Bankier et al, 2000, 2001) and CANCEIS now processes all Census variables for the Canadian Census of population (from both short and long form populations).

13. The NIM approach searches for nearest-neighbor donors first and then determines the minimum change imputation actions based on these donors. While the Fellegi/Holt method involves imputing the fewest variables and preserving the integrity of the subpopulations, the NIM, which reverses the order of the operation – starting with donors and then moving to minimum variables to change – provides computational advantage and is data driven. But the NIM can only carry out donor nearest neighbor imputation while Fellegi/Holt can be used with other methodologies (like top-down). Statistics Canada incorporated the NIM into its Canadian Census Edit and Imputation System (CANCEIS) for the 2001 and it has been part of its overall editing strategy since then.

14. National Statistical Institute of Italy (INSTAT) was developed Data Imputation and Editing System (DIESIS), in collaboration with academic researchers (Department of Computer and Systems Science of the University of Roma "La Sapienza") (Bruni *et al*., 2001). The DIESIS system allows to deal with qualitative and quantitative variables simultaneously, at household and individual level. After a rigorous statistical evaluation of its, the DIESIS system was successfully used for imputing nonresponse and resolve inconsistent responses for the 2001 and 2011 Population Census (Bianchi *et al*., 2005; Bianchi *et al*., 2008).

15. Two editing approaches are implemented in the DIESIS system, the *data driven* and the (theorical) *minimum change*, through the *first donors then fields* and the *first fields then donors* algorithms.

16. The *first donors then fields* algorithm first identifies a subset of potential donors and then determines the minimum number of variables to impute on the basis of these donors. The potential donors are the passed edit households as similar as possible to the failed edit household. The similarity between each failed edit household and each passed edit household is calculated by a function defined as the weighted sum of the distances (for quantitative variables) or similarities (for qualitative variables) for each household variable over all the persons. The algorithm selects, from the potential donors, the minimum (weighted) set of values to impute so that the new adjusted household will pass all the edits (minimum change given the potential donors). By using this algorithm, the imputed values for a household come from a single donor household.

17. The *first fields then donors* algorithm first determines the minimum (weighted) number of variables to impute and identifies the potential donors (as previously described). Then, for each recipient person, the algorithm takes the values to impute from the donor person as similar as possible to the recipient one. This algorithm imputes the variables of one person in turn. If possible, the variables inside the person are imputed simultaneously. Note that the imputed values for a household may come from two or more donor households.

18. The two algorithms were jointly used for the treatment of the demographic variables, in order to balance the plausibility of the imputation actions with the preservation of the collected information. The *first donors then fields* algorithm was selected as default one, with the option to turn to the *first fields then donors* algorithm when, for a given failed edit household, the number of changes proposed by the first algorithm was exceedingly high in comparison with the number of changes proposed by the second algorithm (the extent was set on the basis of the household size).

19. The *first donors then fields* algorithm was mainly used to process the households having common structure, that are usually those having smaller household size. For these households it was generally possible to find enough potential donors. Otherwise, in the treatment of households having uncommon structure, usually those with largest size, few donors were generally available, and often they were not very similar to the failed edit household. In these cases the data driven imputation action would have required many changes to obtain an adjusted household passing the edits, therefore the minimum change approach was preferred.

20. Other methods for imputing unknowns besides using actual cases are exist. Average measures are sometimes used. Also, some countries use other approaches, such as regression models (Russia, 2000). Regression was also used for imputing age on the 2000 U.S. short form (Williams, 1998).

21. Some systems integrate coding, keying, and editing into a single system, particularly for surveys. Among these has been Brazil's CRIPTAX system (Hanono and Barbosa, n.d.) which describes a method of editing on entry. Others, including the Census and Survey Processing System (CSPro), have aspects of interactive editing. As noted elsewhere, country statistical offices must consider the overall editing strategy in making the best choice of editing software. Factors to consider may include the expected volume of edits to be processed, the type of data being treated (numerical, character, categorical), the need for customized software development, and the computing environment in which the editing package or program may be used.

22. As for the top-down approach, the United States Census Bureau developed the Integrated Microcomputer Processing System (IMPS), a DOS based data processing package that was widely used for the 1980 through 2000 Censuses. During the late 1990s and the 2000s, the Census Bureau developed the Census and Survey Processing System (CSPro) for the Windows platform as a replacement for IMPS. CSPro is a software package for entry, editing, tabulation, and dissemination of census and survey data. CSPro combines the features of IMPS and the Integrated System for Survey Analysis (ISSA) in a Windows environment. Batch edit applications in CSPro can be developed for the Nearest Neighbor method, as well (for advanced users) as the Fellegi/Holt method.

23. CSPro can be used to process censuses and surveys of any size and has been used to process the census data for both Djibouti (less than one million people) and Bangladesh (more than 163 million). Examples of uses of CSPro include: censuses (population and housing; agriculture; and economic); demographic and labor force surveys; household income and expenditure surveys; major international projects such as the Demographic and Health Surveys (DHS), Living Standards Measurement Study (LSMS), and Multiple Indicator Cluster Survey (MICS).

24. CSPro allows the user to create, modify, and run data entry, batch editing, and tabulation applications from a single, integrated development environment. It processes data on a case basis (one or more questionnaire), where a case can consist of one or many data records. CSPro contains a powerful procedural language to implement data entry control and edit rules. Specifically, the batch editing function of CSPro identifies and reports structure, value, and consistency errors in questionnaire data. The package can change (impute) data values based on simple or complex methods. CSPro batch editing features make it easy to implement hot deck or cold deck imputations; generate imputation statistics; and produce summary or detailed reports of errors and corrections.

25. CSPro is developed and supported by the U.S. Census Bureau and ICF Macro, the organization that implements the Demographic and Health Surveys (DHS). Funding for the development and maintenance of CSPro is primarily provided by the United States Agency for International Development (USAID). CSPro is available at no cost and can be downloaded from https://www.census.gov/data/software/cspro.Download.html.