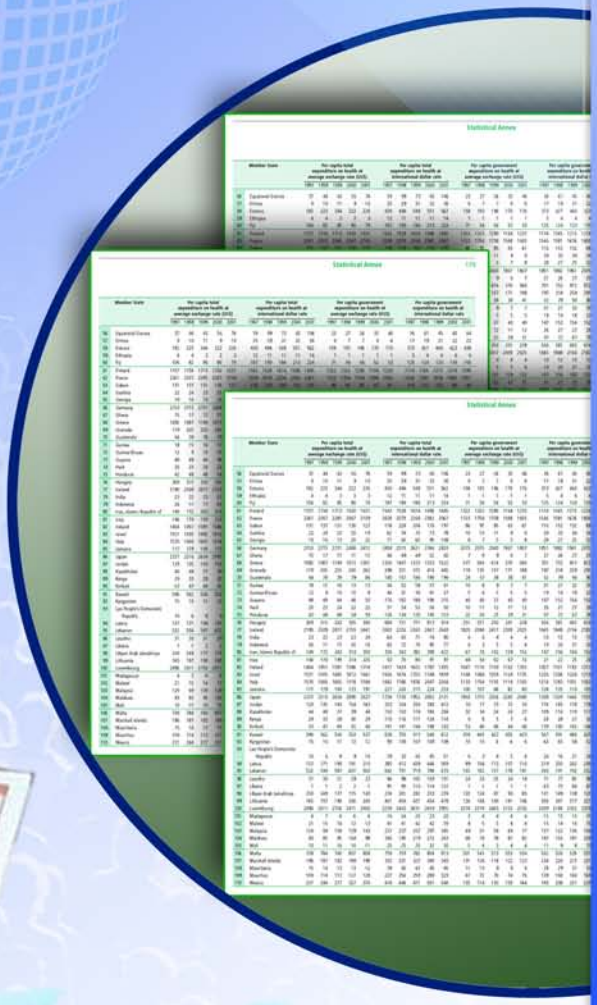
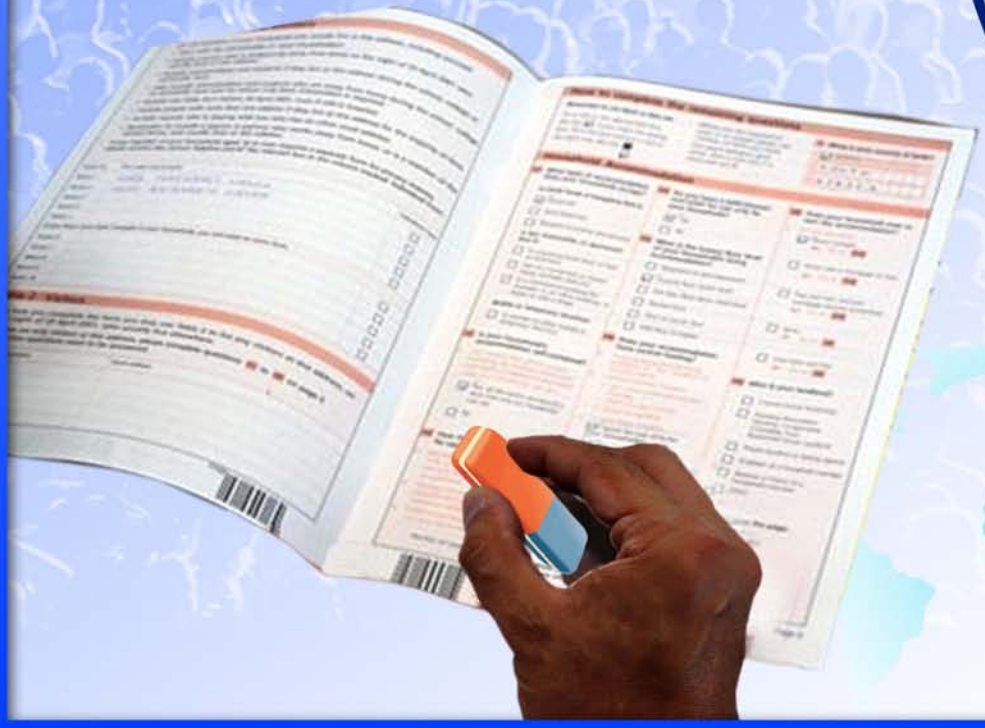
Economic Commission for Africa
African Centre for Statistics

# Africa Census Editing Handbook





Economic Commission for Africa
African Centre for Statistics

# Africa Census Editing Handbook

# II. Africa Census Editing Handbook

TABLE OF CONTENTS

## FOREWORD

The programme of work of the United Nations Economic Commission (UNECA), through its African Centre for Statistics (ACS), include a significant component aimed at improving the statistical capacity of its member states to effectively conduct the 2010 Round of Population and Housing Censuses in recognition of the importance of the latter in the developments process of African Countries. This programme includes the promotion of the adoption of agreed international methodological guidelines and standards. It also ensures the exchange of national experiences and know-how to contribute to the efficiency and effectiveness of census operations. UNECA has conducted workshops training, and advisory services on various census related topics, including cartography, Geographical Information Systems (GIS), census operations management, data processing, dissemination and archiving of census data.

The African Census Editing Handbook provides principles and techniques of editing of census data, which is one of the crucial elements of data processing. It presents in details the various quality control measures that need to be applied by the census offices during the process of enumeration, at the stage of coding and data capture and finally during editing. The rapid and continued technological improvement in various editing software has opened up possibilities of sophisticated and quicker computer based editing of vast volume of data such as in census. The Handbook has adequately dealt in computer based editing with concrete examples from Africa. The Handbook has also dealt with edit logic/rules that needed to be followed for each of the core variables in census and also to ensure inter-consistencies between them.

The Handbook will therefore enhance the capacity of countries to systematically produce quality census information through appropriate editing processes to be implemented during data collection, capture and processing. This will help countries in improving the census data quality, which is crucial for effective socio-economic planning, especially at the local level.

The Economic Commission for Africa's Statistics Division's is interested in providing guidelines for Africa's National Statistical Offices as they develop their Data Capture, Editing, and Tabulation and Dissemination procedures. This is in accordance with the various recommendations and resolutions proposed from various forums mandating the UNECA to produce technical documents in support of countries participating in the 2010 Round of Population and Housing Censuses.

The ACS has taken necessary steps to provide Africa's NSOs with a series of guidelines as they develop their Data Capture, Editing, and Tabulation and Dissemination procedures. This is in accordance with the various recommendations and resolutions proposed from various forums mandating the UNECA to produce technical documents in support of countries participating in the 2010 Round of Population and Housing Censuses.

We hope that his Handbook will contribute to the improved participation in the 2010 Round and beyond.

**Dimitri Sanga**
**Director**
**African Centre for Statistics (ACS)**
**United Nations Economic Commission for Africa (UNECA)**

**ACKNOWLEDGMENTS**

We would like to thank Mr. Michael J. Levin, Senior Census Trainer, Harvard Center for Population and Development drafting this handbook. We would also like to thank Mr. Anthony Matovu, Uganda Bureau of Statistics and Mr. Steven Lwendo, Harvard University, for their contributions.

Under the overall supervision of Mr. Dimitri, Sanga Director of ACS, Mr. Raj Gautam Mitra, (Chief), Ayenika Godheart Mbiydzenyuy (Statistician) and Mr. Oumar SARR (Statistician) of the Demographic and Social Statistics Section of ACS provided technical support by reviewing the entire handbook.

## II.1. **INTRODUCTION**

### **II.1.1. PURPOSE OF THIS PART OF THE HANDBOOK**

1. A well-designed census or survey[1], 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.

2. 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 have to reinvent the process used in earlier data collection activities for a new census or survey.

3. Every Census Editing Process should: (1) give users high quality census data; (2) identify the types and sources of error; and (3) provide adjusted census results. If the census editing process achieves these three goals – goals that we will stress throughout this handbook, the census editing will have been successful.

4. The *African Census Editing Handbook* is designed to bridge this gap in census and survey data editing methodology and to provide information for officials on the use of various approaches to census editing. It 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 the current census or survey in order to avoid duplication of effort in the future. This handbook is also expected to stimulate dialogue amongst specialist in developing a holistic strategy for census data capture, editing, tabulation and dissemination.

5. The *Handbook* is a reference for both subject-matter [2] and data processing specialists as they work as teams to develop editing specifications and programs for censuses and surveys. It 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.

### **II.1.2. THE EDITING TEAM**

6. 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 that includes census managers, subject-matter specialists and data processors. This team 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. 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.

7. 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 particular 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 at special tables after census processing. 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.

8. 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.

9. 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 metadatabases properly. The editing team must also use available administrative sources and survey registers efficiently in order to improve subsequent census or survey operations.

---

[1] A census is a full count. A survey usually enumerates a smaller proportion of the total population. The edits described here should work for either activity.

[2] As defined in this *Handbook*, subject-matter specialists include demographers, social scientists, economists and others who are working in population, housing and other related fields.

10. Communication between subject-matter and data processing specialists was limited when national statistical/census offices used mainframe computers. This division continued for some time after the advent of microcomputers, but computer program packages have become more user-friendly, and now many subject-matter personnel can actually develop and test their own tabulation plans and edits. While subject-matter specialists usually do not process the data, they often understand the steps the data processing specialists take to process the data.

### II.1.3. EDITING PRACTICES: EDITED VERSUS UNEDITED DATA

11. Countries perform census edits to improve the data and its presentation. In this section, the *Handbook* highlights a problem facing national census/statistical offices when unedited census data is released. The issues are illustrated using a hypothetical set of data.

12. 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 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[3] 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.

[3] 'Sex' and 'gender' are used interchangeably in this publication.

TABLE 1. SAMPLE POPULATION BY 15-YEAR AGE GROUP AND SEX, USING UNEDITED AND EDITED DATA

| *Age group* | *Unedited data* | | | | *Edited data* | | |
|---|---|---|---|---|---|---|---|
| | *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 | | | |

13. 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.

14. 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.

15. 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.

16. In table 1, the edited data on the right are "cleaner" because the unknowns have been suppressed (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. However, the latter may not be put in public domain and can be shared with demographers and subject matter specialist on request.

17. Statistical offices should make every effort to maintain the original, collected data. 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. But, 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.

18. 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 2000, to only 15, or less than one per cent of the responses in 2010.

19. 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 2000 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.

TABLE 2. POPULATION AND POPULATION CHANGE BY 15-YEAR AGE GROUP WITH UNKNOWNS: 2000 AND 2010

| *Age group* | *Numbers* | | *Number* | *Percent* | *Per cent* |
|---|---|---|---|---|---|

| | 2010 | 2000 | *Change* | *Change* | 2010 | 2000 |
|---|---|---|---|---|---|---|
| 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 |

20. 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 have to 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.

TABLE 3. POPULATION AND POPULATION CHANGE BY 15-YEAR AGE GROUP WITHOUT UNKNOWN DATA: 2000 AND 2010

| *Age group* | *Numbers* | | *Number change* | *Per cent change* | *Per cent* | |
|---|---|---|---|---|---|---|
| | 2010 | 2000 | | | 2010 | 2000 |
| 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 |

### II.1.4. THE BASICS OF EDITING

21. **Editing is the systematic inspection of invalid and inconsistent responses, and subsequent manual or automatic correction (using "unknowns" or dynamic imputation) according to predetermined rules.** Some editing operations involve manual corrections, which are corrections made manually in the office. Other editing operations involve electronic corrections, using computers. Census publications are likely to contain a certain amount of meaningless data if national census/statistical offices do not edit the census or survey results. Editing reduces distorted estimates, facilitates processing, and increases user confidence.

22. In determining the final edits for a census, subject-matter personnel should investigate the edits developed for pilot censuses and those developed during processing to make sure that individual edits have the expected cost of benefits. These investigations need to be part of the census evaluation.

23. Another set of techniques and terminology relates to micro-editing and macro-editing. As noted, census and survey editing detects 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. Macro-editing checks aggregated data to make sure that they are reasonable, but will not be covered here.

24. 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.

25. 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 has to be re-formed and re-

edited. Editing and imputation may or may not improve the quality of the data, but a clean dataset greatly facilitates analysis and use.

26. 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, 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.

27. *How over-editing is harmful.* 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.

***(1) Timeliness. 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 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.***

***(2) Finances. Similarly, the costs of the census process increase as the time increases. Each national census/statistical office has to 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.***

***(3) Distortion of true values. 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.***

***(4) A false sense of security. 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.***

28. *Treatment of unknowns.* The editing team must decide early in census planning how to handle "not stated" or unknown cases. Columns or rows of unknowns in tables are neither informative, nor useful, so planners in most countries prefer to have 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.

29. *Spurious changes.* National census/statistical offices do not usually work with models when they develop their editing rules. Editing teams should develop rules that fit the actual population or housing characteristics. All data should pass the edit rules. 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 actually 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.

30. *Determining tolerances.* 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, 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 should 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.

31. 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. 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.

32. 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 actually 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.

33. *Learning from the editing process.* As the data are edited, detailed analyses of positive and negative feedback need to be recorded to improve the quality of the both current census or survey and future censuses and surveys. The editing team has to 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.

34. *Quality assurance*. Quality assurance is important in all census operations. Consequently, formal quality assurance mechanisms should certainly be in place to monitor the progress of the computer editing and imputation phase. Audit trails, performance measures, and diagnostic statistics are crucial for analysis of the quality of the edits and the rapidity of processing.

35. *Costs of editing.* 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. For most countries, editing activities take a disproportionate amount of time and funding, so each country must determine the return on its investment. Excessive editing can delay census results. Rather than over-editing the data, National census/statistical offices might better spend their funding on obtaining a higher quality census or survey enumeration in the first place.

36. *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 micro-data and in possession of good auxiliary information do the best imputation.

(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.

37. *Archiving.* 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. The documentation should be complete enough for census or survey planners to be able to reconstruct the same processes at a later date 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.

38. 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.

39. The two sets of errors 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.

## II.2. EDITING APPLICATIONS

40. 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[4]. 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.

41. Whether a census data set is scanned or keyed, 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, which should be carefully chosen. Also, in some cases, however, the scanned data require a second automated coding operation to fill in items like birthplace, industry, and occupation.

42. In either case, the unedited should appear in a form allowing the computer programs to develop the **structure edits** (as described in Chapter 3 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 within 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.

43. 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.

44. After the structure is set, it is not really set. 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.

45. 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 4 and 5 cover the various population and housing items in the U.N. Principles and Recommendations, Revision 2.

46. 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.

[4] When this handbook was originally written for the 2000 Censuses, almost all countries keyed their data. Now, most countries scan, sometimes with keyed follow-up. The current handbook attempts to take scanning into account for the structure and content edits. Even as this handbook is being written, new technologies are emerging: the use of Personal Digital Assistants (PDAs) and use of Internet for data collection and interactive editing. Just as technologically developing countries had problems with scanning in the early 2000s, many countries are also finding that PDA usage still needs refinements. See Part I of this handbook, on data capture.

47. 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.

48. 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.

49. To implement the computer editing phase of the process the editing team prepares written editing instructions or specifications, decision tables, flow charts and pseudo-code. Pseudo-code is a set of written editing instructions or specifications as shown in figure 8.

50. Flow charts help the subject-matter specialists to understand the various linkages among the variables and make it easy to write editing instructions. 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.

51. 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.

### II.2.1. CODING CONSIDERATIONS

52. Countries keyed their data during much of the second half of the 20th century. Most countries now scan their censuses, but frequently continue to key surveys. Even when forms are scanned, certain variables still need to be translated from words to numbers. The process of making machine-readable numbers and alpha-numerics is called **coding.** Some editing packages can easily accept and work with alphanumeric data, but most packages have some problems categorizing, summing, and getting percentages, medians, etc., when non-numeric data are included. Codes that are completely alphabetic characters or alphabetic characters combined with numbers (called alphanumerics) should be avoided whenever possible. When forms are scanned, alphanumerics 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.

53. 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. But, 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.

54. When you have two columns, however, say codes 1 to 10, then you introduce a whole new realm of errors. Instead of legal values 1 to 9, you now have 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.

55. 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 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 boy 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. 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.

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

1. Head of household (or householder)
2. Spouse
3. Child
4. Adopted or step-child
5. Sibling
6. Parent
7. Grandchild
8. Other relative
9. Nonrelative

This set of standard codes covers the majority of relationships for most countries. Some countries add a "0" code for head of household and can then add a 10th category to the others.

57. Even these codes can be used to obtain household composition as shown in the section on derived variables. However, many countries, particularly those experiencing the HIV/AIDS epidemic need much more detailed information than can 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. 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:

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

58. 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.

59. 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.

60. 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, the U.S. Census Bureau, and others, already have the levels imbedded in the codes, so the statistical office does not have to do any additional work.

61. 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, work place), language, ethnicity/race, and citizenship are very similar. A common coding scheme for "place" might 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 also use country numerical codes developed by international organizations such as the United Nations Statistics Division (United Nations, 1999). 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.

62. 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.

63. Figure 1 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.

64. 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 1. 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 |

65. 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.

66. 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.

### II.2.2. MANUAL VERSUS AUTOMATIC CORRECTION

67. Manual editing of a census may 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.

68. 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 computer 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.

69. 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 not as time-consuming as manual correction.

70. 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.

71. 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.

72. This example 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 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.

73. 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 census or 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.

74. 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.

75. 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.

76. 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.

77. 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.

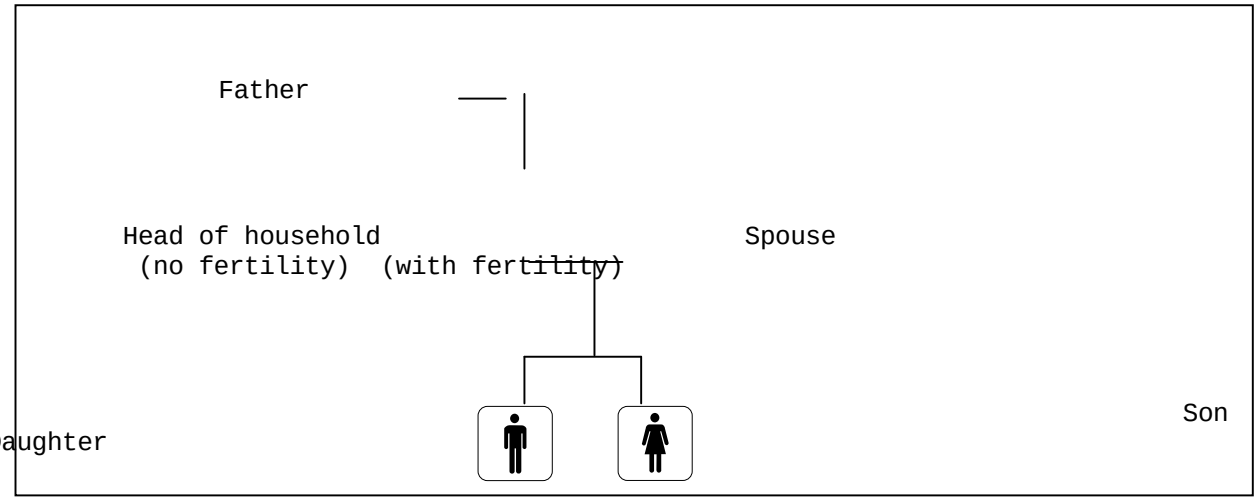
78. The edit, or audit, trail shows the changes made to each variable. The trail is used to trace the history of the responses from the receipt of the data through the editing and imputation process.

### II.2.3. GUIDELINES FOR CORRECTING DATA

79. 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. In fact, three major considerations pertain: (1) Make the fewest required changes possible to the originally recorded data; (2) Eliminate obvious inconsistencies among the entries; and, (3) Supply entries for erroneous or missing items by using other entries for the housing unit, person, or other persons in the household or comparable group as a guide, always in accordance with specified procedures. On some occasions, the category "not reported" is appropriate for certain items.

80. Consider the following diagram (figure 2) 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 2. A typical hypothetical household including relationships, sex and fertility of the members**



81. 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 3? 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.

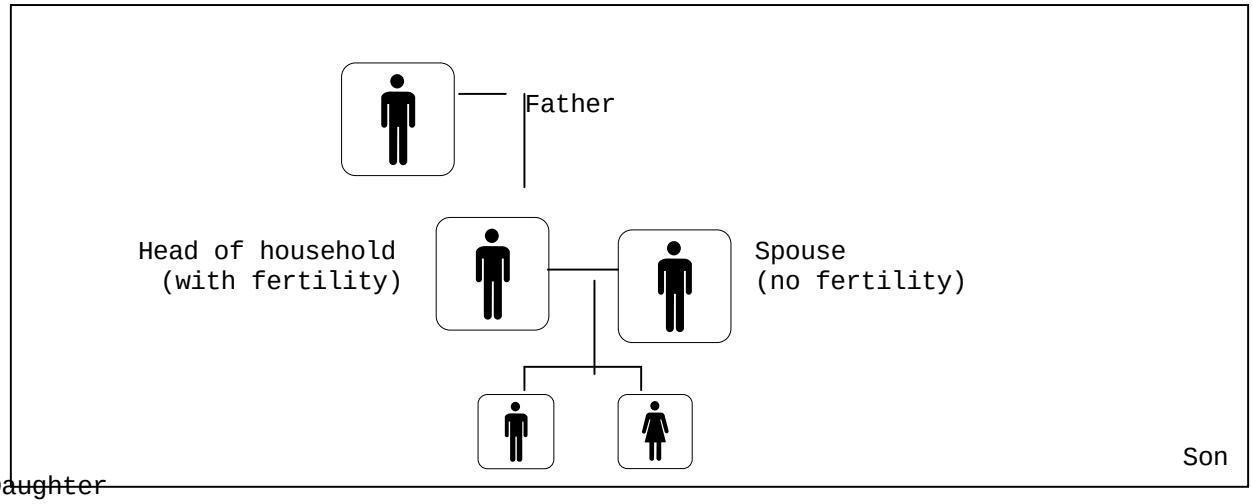
82. If the head of household in this case happens to be the wife rather than the husband, then the editing rule adopted would be wrong and the national census/statistical office would end up with four errors:

(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.

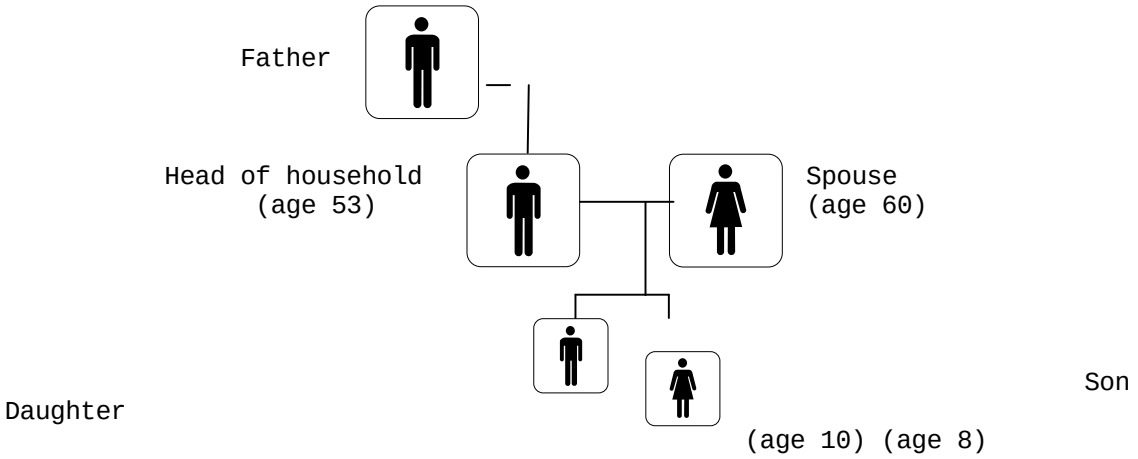
83. In contrast, when a good editing procedure finds that the head and spouse have the same sex, it then checks both persons for fertility. Since only the head has fertility, the head becomes the female. The editing rules for these items are then satisfied.

**Figure 3. Example of household with head and spouse of the same sex**




84. Another example, in figure 4, 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.

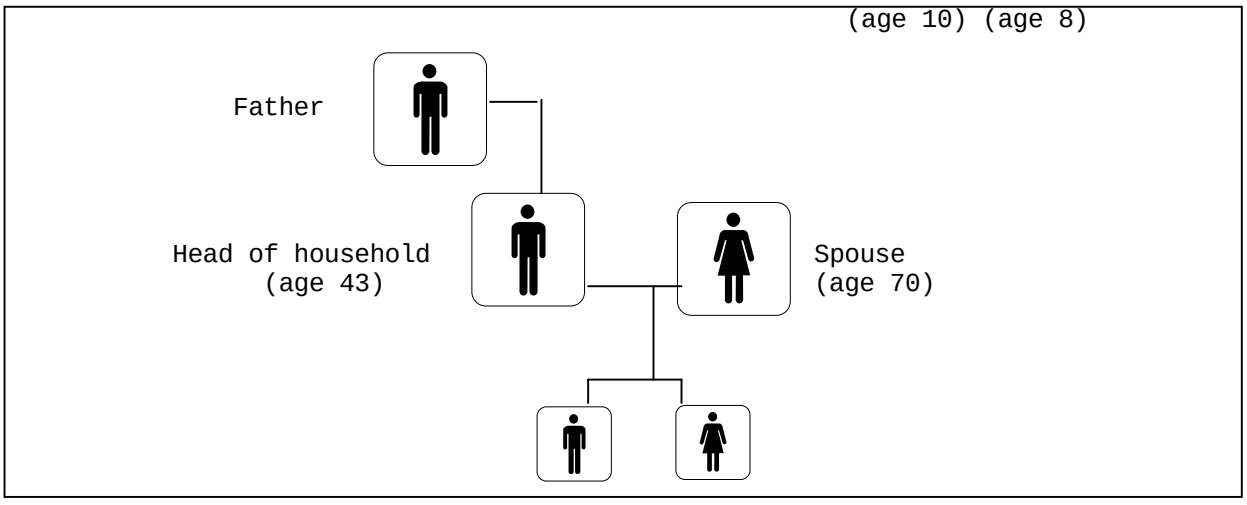
**Figure 4. Example of household with ages of some household members**




85. Figure 5 offers another possible scenario. Suppose the edit finds a woman 70 years old with children aged 10 and 8 as in figure 5. 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.

86. 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 5. Example of household with potential inconsistencies in age reporting**



### II.2.4. VALIDITY AND CONSISTENCY CHECKS

87. One of the major requirements in editing is that no item may contain invalid values and editing. 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. 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 look silly, even if the few cases would not affect actual planning for a country.

88. 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.

89. 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*

90. This procedure 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.

91. 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.

92. 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".

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

94. 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 staffs 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*

95. 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.

96. 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 the section on 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.

97. The 11 declarative statements in figure 6 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.

98. Figure 6 shows a simple case where, for some reason, both the head and spouse have the same sex – as it turns our, 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 6. 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. | | | | | |

99. In the example in figure 7, 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 information 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 7. 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 |

100. The editing specifications for this edit can be written as shown in figure 8. 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 8. 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
```

101. The figure below (figure 9) 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 9. 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 |

102. 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.

103. 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 rerun 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.

### II.2.5. METHODS OF CORRECTING AND IMPUTING DATA

104. 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.

105. 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 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.

*1. Static imputation or "cold deck" technique*

106. In static or cold deck imputation, the editing program assigns a particular response for a missing item (and not inconsistent/invalid values) 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.

107. Static imputation is a stochastic method, as is dynamic imputation, but the values do not change over time. 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*

108. Another method of ridding the data of unknowns is the dynamic imputation or hot deck technique, which is used to allocate values for unavailable, unknown, incorrect or inconsistent entries. 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.

109. 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 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 "hot deck".

110. The values stored in the hot 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 hot 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 or other unusual statistical activity. Similarly, housing characteristics are more likely to be similar within a village or set of villages than to other parts of the country.

111. 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.

112. One of the problems with the dynamic imputation (hot deck) 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.

113. The data below (figure 10) illustrate a household for a set of ten individuals. 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 10. 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

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

115. 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. 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

116. 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

117. 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 edit will almost certainly replace them before using them. Also, the edit calls for imputation of many values, so few initial values affect the final tabulations. These values might be as shown in figure 11.

**Figure 11. 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 |

118. Consider again the 10 individuals introduced in figure 10. 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 then contains the values displayed in figure 12.

**Figure 12. 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 |

119. 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.

120. 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).

121. 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.

122. 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.

123. After the tenth person, the imputation matrix values are given in figure 13. 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 13. 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*

124. *Geographical considerations*. 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.

125. *Use of related items*. 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.

126. *How the order of the variables affects the matrices*. 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 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 hot decks. Subsequent imputation matrices can use the data items after editing. However, whenever possible, statistical offices should consider excluding edited data from the imputation matrix.

127. 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 edited data is unavoidable because of other factors. It is important to note that most countries ignore this suggestion, and impute from previously imputed values.

128. *Complexity of the imputation matrices.* 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.

129. 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[5]. 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.

130. 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.

131. *Imputation matrix development*. 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.

132. 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.

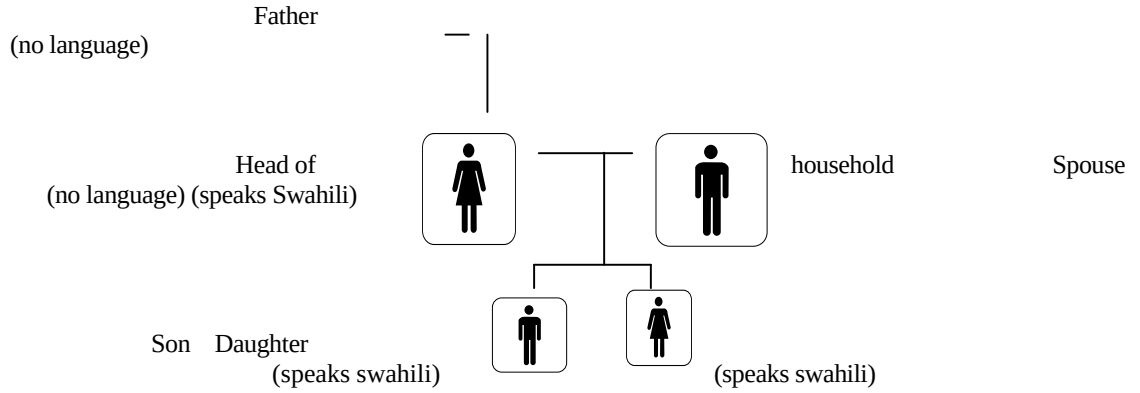
133. Although for the examples in this *Handbook*, each cell in the imputation matrices has one value, some editing teams keep more than one possibility for each cell. You can imagine this as a two dimensional matrix, with a third dimension, like going back into a blackboard. These cells provide an extra 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.

134. *Standardized imputation matrices*. 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. 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.

135. In figure 14 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 15 is female.

[5] 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.

**Figure 14. Example of head of household and head's father without assigned language**




136. 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 15.

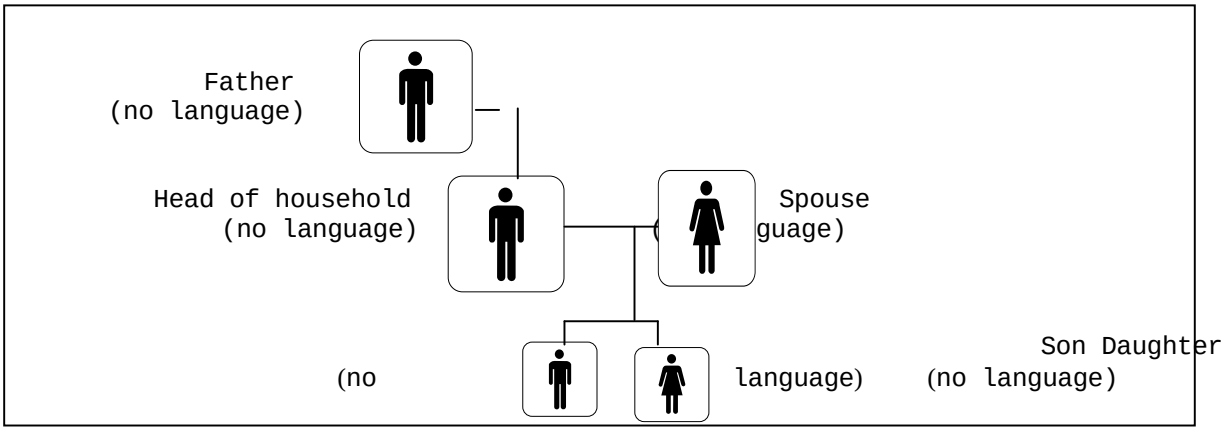
**Figure 15. Initial values for a dynamic imputation matrix for language**

| **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 | Language 1 | Language 1 | Language 1 | Language 1 | Language 2 |
| Female | Language 1 | Language 1 | Language 1 | Language 1 | Language 1 | Language 2 |

137. 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.

138. 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 language spoken in the general community.

**Figure 16. Example of members of a household without an assigned language**




139. 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.

140. 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.

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

142. 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.

143. 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.

144. 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.

145. 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.

146. 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.

147. 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.

148. *When dynamic imputation is not used*. 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 17:

**Figure 17. 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

149. 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).

150. 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.

151. 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 18. 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 |

152. In figure 18 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 set, then the grade can also be determined, and the latter should thereby be consistent with both age and working status.

153. 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 form 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.

154. 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.

155. *How big should the imputation matrices be?* Most computer packages can accept multidimensional imputation matrices. The following points should be taken into consideration before setting up the imputation matrices.

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

157. (ii) Understanding what the imputation matrix is doing. 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.

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

159. (iv) Items that are difficult for imputation matrices. 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*

160. The basic structure of the imputation matrix in an editing software package should look something like the display in Figure 19. 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***

161. 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 19. Sample set of values for a cold deck array and sample imputation code**

```
.22  A01-AGE-FM-SEXRL (2,6)
23.   Head of household   Spouse   Child   Other relative   Parent   Not reported   .Sex
24.          40              40       10          20           65          20        .Male
25.          40              40       10          20           65          20        .Female


.
.
.
40 if AGE = 0:98
41   let A01-AGE-FM-SEXRL (SEX,RELATIONSHIP) = AGE
42 else
43   message 'Age is unknown, so imputed' AGE
44   write ' Age is unknown, so imputed, Age = ' AGE
45   impute AGE = A01-AGE-FM-SEXRL (SEX,RELATIONSHIP)
46   message 'AGE is now known' AGE
47 end-if
.
.
.
```

***(b) Messages for errors***

162. 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. A third type of summary measures is frequency distributions. In good edits, using hotdecks, the distribution of attributes of the unedited item, the edited item, and the imputed values should be about the same. That is, if the distribution of males and females is about 51-49 in the unedited values, it should also be about 51-49 when edited, and the imputed values should also be approximately in this distribution as well.

163. One source of information is the display of a message, as seen above in figure 19. 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 of the questionnaires, a summary report might look something like figure 20:

**Figure 20. 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 | 2669xx |
| 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.

164. A report organized by questionnaire (figure 21) 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.

165. 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 meaningless 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 21. 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***

166. 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 22. 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 22. 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 |

167. This procedure assists the editing team in determining whether the edit is taking the proper paths. 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 over and over again in a single file. The procedure continues as outlined below:

1. 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;
2. The analyst then checks all of the paths early in the editing process;
3. 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;
4. Finally, the data processors run the whole dataset.

168. 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 20.

169. 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.

170. 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.

171. 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.

172. Let's look at a few cases to show how these household listings can aid in obtaining the highest quality edit. In the example below, we find a 6 person household, with the people identified separately in the column PN for person number. Next, the sex of each person is shown, and then, because we are concerned with an edit for fertility, all of the available fertility items. Here, these include total children ever born (CEB), male children born (MCB), female children born (FCB), children surviving (CS), male children surviving (MCS), female children surviving (FCS), month of last birth (MLB), year of last birth (YRLB), sex of last birth (SLB), and vital status of last birth (VLB). It is important to list all of the variables to make certain that the edit is going to do what you want it to do – that you can check that it is actually doing that.

```
PN SEX AGE CEB MCB FCB  CS MCS FCS MLB YRLB SLB VLB
01   1 041
02   2 038  04  02  02  04  02  02  08 1991   2   1
03   2 022  71  01  00  01  01  00  06 1999   1   1
04   1 012
05   2 009
06   1 001
V.27: problems detected in fertility info ... PN= 03
V.27b: imputing TCEB = MCEB+FCEB PN= 03 TCEB=71 MCEB=01 FCEB=00
PN SEX AGE CEB MCB FCB  CS MCS FCS MLB YRLB SLB VLB
01   1 041
02   2 038  04  02  02  04  02  02  08 1991   2   1
03   2 022  01  01  00  01  01  00  06 1999   1   1
04   1 012
05   2 009
06   1 001
```

173. So, while the spouse – person 2 – has consistent fertility information, the 3rd person, a 22 year old daughter does not. In fact, the scanner or keyer (or, of course the respondent or the enumerator) has attributed 71 children ever born to her. So, under the first listing for the household – the listing *before* the edit starts, we see the 71 CEB reported. In the middle of the total listing, we see the statement "V.27: problems detected in fertility info … PN= 3", indicating that the fertility for person 3 is inconsistent. But, in the next line we find we can resolve the discrepancy by adding the MCB and the FCB

together to get the new CEB – that is, 01 male children + 00 female children = 01 CEB. Then, the household listing is repeated, with the corrected information. In this case, that the female had only 1 CEB, and not 71.

174. The following example shows two other aspects of the use of the display. First, we find that for whatever reason, the month and year of last birth is blank. This could be a scanning or keying error. Or, it could be that the enumerator did not report the information. Or it could be that the respondent didn't remember the information. However, we find that female had 4 children ever born, and a check of the listing shows that all four of them are in the house – they are code 3 in the Relationship item, with ages, 20, 14, 12, and 5. We also see that the last child was a female (SLB = 2) and that the last child's vital status is 1, so alive. Given all this information, we are able to determine that the last child listed – person 6 in the listing – can provide the information about the month and year of last birth. And while we are not showing the month and year of birth of that child – only the age, in this listing – her information would go on to the end of the mother's record for month and year of last birth.

```
PN REL SEX AGE CEB MCB FCB  CS MCS FCS MLB YRLB SLB VLB
01   1   2 054
02   2   2 035  04  01  03  04  01  03            2   1
03   3   2 020  00
04   3   2 014  00
05   3   1 012
06   3   2 005
V.27: problems detected in fertility info ... PN=02
V.27POST: LAST info blank, imputing from youngest child PN= 02
(updates FCEB, TCEB, FCS, TCS)
V.27e: imputing fertility data from AFERTILITY PN=03
V.27e: imputing fertility data from AFERTILITY PN=04
PN REL SEX AGE CEB MCB FCB  CS MCS FCS MLB YRLB SLB VLB
01   1   2 054
02   2   2 035  04  01  03  04  01  03  11 1995   2   1
03   3   2 020  00  00  00  00  00  00
04   3   2 014  00  00  00  00  00  00
05   3   1 012
06   3   2 005
```

175. However, the subject matter specialists in this country have decided that although the enumerators were to report 00 for females with no children, and leave the remaining fertility information blank, for tabulation and other purposes, they wanted the zeros filled for those having no children. We see this happening for persons 3 and 4 in the listing – females who were 20 and 14, respectively, at the time of the census. This procedure makes a cleaner data set.

176. The following example shows several issues in African census work. First, we note that all of the ages and dates of birth are inconsistent. This can happen in several ways – the respondents could give wrong information and the enumerator could get it wrong. Or, the enumerator could collect only the dates of birth and then determine the ages later, but erroneously. Of the census could have been extended over a long period of time, so that the ages reported were at the time of enumeration rather than with respect to the census date. Sometimes, this is a real issue since the census is supposed to supply a "snapshot" of the population, and this is difficult to do if the enumeration goes over a long period.

```
PN SEX DOB MOB  YOB AGE REL MAR SPN CEB  CS MPN FPN
01   2  01  09 1986 015  01   5      00      90
02   2  09  06 1990 011  06   5
03   1  01  09 1991 010  06   5              99
04   2  01  09 1994 007  06   5              99
V.2b4b: age and DOB inconsistent, age <= DOB, Age = 015 Date = 01/09/1986
V.2b4b: age and DOB inconsistent, age <= DOB, Age = 011 Date = 09/06/1990
V.2b4b: age and DOB inconsistent, age <= DOB, Age = 010 Date = 01/09/1991
V.2b4b: age and DOB inconsistent, age <= DOB, Age = 007 Date = 01/09/1994
V.3a1: head is younger than 16, Age = 014
V.3a3: no older relatives found; keep young head
PN SEX DOB MOB  YOB AGE REL MAR SPN CEB  CS MPN FPN
01   2  01  09 1986 014  01   5      00      90
02   2  09  06 1990 010  06   5
03   1  01  09 1991 009  06   5              99
04   2  01  09 1994 006  06   5              99
```

177. A second issue in this household is that all of the people are 15 years old or younger, with the oldest becoming 14 because of the edit. Sometimes this happens when the ages are not properly recorded. Sometimes it happens when a continuation form gets disassociated from the first part of the questionnaire, and these children become "orphaned." But, because of the on-going HIV/AIDS epidemic, sometimes this type of household appears in transition, as children living alone, but who will be absorbed into some other type of housing. In the case here, the first person is made the head of household.

178. The following case shows a situation with multiple heads. Person 6, who is clearly not the head, still is listed as code 1 in the relationship entry. The edit notes that more than one head is appearing, and corrects the second one, making him an "other relative."

```
PN SEX DOB MOB  YOB AGE REL MAR SPN CEB  CS MPN FPN
01   1  11  01 1950 051  01   1  99          99  01
02   1  17  07 1977 023  03   5                  01
03   2  04  04 1985 005  03   5      00      09  01
04   1  24  10 1987 011  03   5              53  01
05   1  01  07 1990 010  03   5              49  01
06   1  20  02 1994 007  01   5              99  01
07   1  20  02 1994 007       5              99  01
V.2b4b: age and DOB inconsistent, age <= DOB, Age = 005 Date = 04/04/1985
V.2b4b: age and DOB inconsistent, age <= DOB, Age = 011 Date = 24/10/1987
V.3: either no heads or > 1= 0002
V.3h: more than 1 head =
V.3i: multiple heads, making oldest= 0051
V.3k: multiple heads, making excess other rel
V.9g: Relation invalid, has a dad, impute  Rela
PN SEX DOB MOB  YOB AGE REL MAR SPN CEB  CS MPN FPN
01   1  11  01 1950 051  01   1  99          99  01
02   1  17  07 1977 023  03   5                  01
03   2  04  04 1985 015  03   5      00      09  01
04   1  24  10 1987 013  03   5              53  01
05   1  01  07 1990 010  03   5              49  01
06   1  20  02 1994 007  11   5              99  01
07   1  20  02 1994 007  03   5              99  01
```

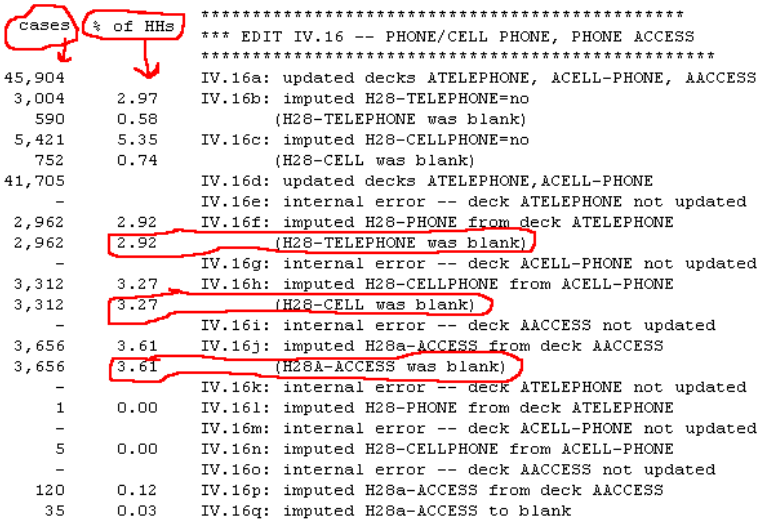
179. Later, in the discussion of the population items, we will discuss the "standard" edit. The standard edit accepts the head's information for an item, if it is available, and assigns it to others in the household if they don't have information for that item. In the example below, person 4 does not information about ethnic group, language, or religion. Often this happens when everyone in the household has the same information, but the enumerator neglects to record it. In the specific case below, we find a newborn, and this is another situation where this type of information is often neglected.

```
PN SEX AGE REL GRP LAN RGN RSA PRV CNT CTZ URS PERMPLAC SM
01   1 073  01   1  06  55   1  09       1       1
02   2 063  02   1  06  55   1  09       1       1
03   2 025  11   1  06  55   1  09       1       1
04   2 000  11              1  09       1       1
V.13e: Ethnic group invalid, impute from head PN=04 Group=Head Group= 1
PN SEX AGE REL GRP LAN RGN RSA PRV CNT CTZ URS PERMPLAC SM
01   1 073  01   1  06  55   1  09       1       1
02   2 063  02   1  06  55   1  09       1       1
03   2 025  11   1  06  55   1  09       1       1
04   2 000  11   1  06  55   1  09       1       1
```

180. The example below shows the case where the head's information is unknown (for ethnic group), and no one else in the housing unit has the information either. In this case, a hot deck, usually by age and sex, assigns a value for ethnic group to the head. Then, that information is used to obtain similar information for others in the housing unit. Note that because the hot deck is constantly being updated, it is most likely that the ethnic group assigned is one from an adjacent household, but almost certainly from someone in the village or surrounding area.

```
PN SEX AGE REL GRP LAN RGN RSA PRV CNT CTZ URS PERMPLAC SM
01   1 045  01      01  32   1  01              1
02   2 048  02      01  32   1  01              1
V.13b: Ethnic group invalid, impute from deck
V.13e: Ethnic group invalid, impute from head
PN SEX AGE REL GRP LAN RGN RSA PRV CNT CTZ URS PERMPLAC SM
01   1 045  01   4  01  32   1  01       1       1
02   2 048  02   4  01  32   1  01       1       1
```

181. Earlier we discussed the error listings. These error listings are generated by use of the command ERRMSG in CSPro, and other packages have similar commands. As will be seen in the sample edits below for population and housing, the ERRMSG command generates statistics for individuals and households, and also presents summary statistics, as seen in the figure below. If a denominator is implemented, percentages are also generated. If we look at the figures for presence of a cell phone, we find that 3,312 times the information about cell phones was not obtained in the sample, which was 3.27 percent of the cases.



```
 cases   % of HHs   ***********************************************
                   *** EDIT IV.16 -- PHONE/CELL PHONE, PHONE ACCESS
                   ***********************************************
45,904             IV.16a: updated decks ATELEPHONE, ACELL-PHONE, AACCESS
 3,004     2.97    IV.16b: imputed H28-TELEPHONE=no
   590     0.58          (H28-TELEPHONE was blank)
 5,421     5.35    IV.16c: imputed H28-CELLPHONE=no
   752     0.74          (H28-CELL was blank)
41,705             IV.16d: updated decks ATELEPHONE,ACELL-PHONE
     -             IV.16e: internal error -- deck ATELEPHONE not updated
 2,962     2.92    IV.16f: imputed H28-PHONE from deck ATELEPHONE
 2,962     2.92          (H28-TELEPHONE was blank)
     -             IV.16g: internal error -- deck ACELL-PHONE not updated
 3,312     3.27    IV.16h: imputed H28-CELLPHONE from ACELL-PHONE
 3,312     3.27          (H28-CELL was blank)
     -             IV.16i: internal error -- deck AACCESS not updated
 3,656     3.61    IV.16j: imputed H28a-ACCESS from deck AACCESS
 3,656     3.61          (H28A-ACCESS was blank)
     -             IV.16k: internal error -- deck ATELEPHONE not updated
     1     0.00    IV.16l: imputed H28-PHONE from deck ATELEPHONE
     -             IV.16m: internal error -- deck ACELL-PHONE not updated
     5     0.00    IV.16n: imputed H28-CELLPHONE from ACELL-PHONE
     -             IV.16o: internal error -- deck AACCESS not updated
   120     0.12    IV.16p: imputed H28a-ACCESS from deck AACCESS
    35     0.03    IV.16q: imputed H28a-ACCESS to blank
```

182. The following listing provides one case from the summary listing above, that is, for one of the 3,312 cases. Here, the abbreviation "CLL" stands for cell phone. We see that the information was either not collected or was not picked up in the scanning or keying. Hence, the information is "imputed". Either a hot deck is used, or, the subject specialists decide to use the "default" position of "no" when the information is left blank.

```
DWL MLT RMS SHR TEN WAT SRC TLT COK HET LIT RAD  TV CMP FRG TEL CLL ACC RFS
 01   2 006       1   4   7   4   1   5   1   1   1   2   1   2       2   4
IV.16c: impute cell phone = no  Phone2 Cell=   Access= 2
DWL MLT RMS SHR TEN WAT SRC TLT COK HET LIT RAD  TV CMP FRG TEL CLL ACC RFS
 01   2 006       1   4   7   4   1   5   1   1   1   2   1   2   2   2   4
```

***(d) How many times to run the edit?***

183. 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.

184. 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:

185. 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. Values from the actual data set, used as cold deck (or initial) values are more likely to be accurate and current. Edits use only about two percent of this initial static matrix; the rest are dynamic imputation values. That is, the initial values are used only about 2 percent of the time. In all other cases, actual values have been placed in the hot deck as they have been encountered in the top-down editing approach.

186. The **second edit run** performs the actual editing. The second edit run consists of several repeat runs in order to cover all situations. During the development of the final edits, 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.

187. 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. When this happens, the data

processor must go back to the "second run" above to correct remaining errors. In some packages, like CSPro, the hot decks may need to be seeded again.

*5. Imputation flags*

188. 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. Inconsistent information poses special problems: when a woman aged 70 has a 3 year old child, one or the other item will be changed. When 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.

189. 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. For example, sometimes data processing specialists accidentally use only cold deck values because the program neglects to update the imputation matrix. As in the example above for age of mother and age of child:

```
AGEDIF = AGE (SPOUSEPTR) – AGE;
If AGEDIF < 12 or AGEDIF> 55 then
   Impute (AGE,AAGEDIF (AGE (SPOUSEPTR));  {Imputing based on spouse's age and previous difference}
Else
  AAGEDIF (AGE (SPOUSEPTR)) = AGE;  {Updating the bot deck}
Endif;
```

If the hotdeck is not updated, then the same value would be assigned in every case, probably skewing the data.

190. If the country conducts a census pilot, the editing team may need to investigate the relationships between some of the variables after the pretest in order to finalize the questionnaire. Before microcomputers with large hard drives 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.

191. 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, a country might want to place imputation flags for each variable at the end of each record, for both housing and population records. For each 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. Hence, the program above would have an additional line:

```
FLAG_AGE = 0;
AGEDIF = AGE (SPOUSEPTR) – AGE;
If AGEDIF < 12 or AGEDIF> 55 then
   Impute (AGE,AAGEDIF (AGE (SPOUSEPTR));  {Imputing based on spouse's age and previous difference}
   FLAG_AGE = 1;
Else
  AAGEDIF (AGE (SPOUSEPTR)) = AGE;  {Updating the bot deck}
Endif;
```

192. If a country wants to save the actual, original value – the unedited value – it can do that as well. Then the program might look something like this:

```
ORIG_AGE = AGE;
FLAG_AGE = 0;
AGEDIF = AGE (SPOUSEPTR) – AGE;
If AGEDIF < 12 or AGEDIF> 55 then
   Impute (AGE,AAGEDIF (AGE (SPOUSEPTR));  {Imputing based on spouse's age and previous difference}
   FLAG_AGE = 1;
Else
  AAGEDIF (AGE (SPOUSEPTR)) = AGE;  {Updating the bot deck}
Endif;
```

The assumption here is that the child's age is going to be changed, since the way the edit is written, we are currently editing only the child's record. If both the mother's age and child's age are to be considered for changing (as in the example below), then we need to take that into account in the edit.

193. That is, other methods are available to save unedited responses. In the example in figure 23, 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. This eliminates the need fort the 0 or 1 in the Flags area of the record.

**Figure 23. 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 | |

194. 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.

195. Figure 24 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.

**Figure 24. 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 |

196. 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 record 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 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.

197. 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. When the edited data are followed by the flags for each item AND the unedited data for each item, the data set gets three times as large as the original. For many countries, this would be unacceptable for long-term storage. However, the original data and the edits could be stored for later reconstruction. The results will look something like this:

| Edited data | Flags | Unedited data |
|---|---|---|
| 1 XXXX | XXXX | XXXX |
| 2 XXXX | XXXX | XXXX |
| 3 XXXX | XXXX | XXXX |
| 4 XXXX | XXXX | XXXX |

198. Countries with very large populations might prefer to use imputation flags on a sample basis for research purposes. For example, a country might want to create a data set with every 100th 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.

199. Even for internal use, particularly for subject matter specialist who need to do quick runs for ministries or researchers inside or outside the statistical office, a data set without the flags and the unedited would both make the process of obtaining a table faster, but is also likely to decrease the number of errors when the analyst accidentally picks up the flag or unedited data when intending to use only the edited data in a table. Also, as noted elsewhere in this handbook, if a country collects both *de facto* and *de jure* data, two data sets may be needed to make sure that some people are not included twice.

### II.2.6. OTHER EDITING SYSTEMS

200. 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 C.2). 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.

201. 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.).

202. 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.

203. The New Imputation Methodology uses donors for items, with the hope that all missing or inconsistent information can come from a single donor or a few donors. In order 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.

204. The objectives of an automated hot deck imputation methodology should be as follows:

1. The imputed household should closely resemble the failed edit household;
2. 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;
3. 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).

205. Under the New Imputation Method 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.).

206. This chapter discussed general editing ad tabulation procedures. The next chapter will cover structure edits, the first, and most important of the computer editing tasks since it establishes that each housing unit appears, and appears only once, and appears in its proper place in the hierarchy of the geography of the country. After that, the chapters cover population items, and then housing items, and finally recodes.

## II.3. STRUCTURE EDITS

207. Structure edits check coverage and determine how the various records fit together. These structure edits must assure that (a) all households and collective quarters records within an enumeration area are present and are in the 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.

208. 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 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 geographic hirrarchy;
(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.

209. 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. Therefore, this chapter examines the more general issue of item validity and the relationship of items between and within records. Chapters IV and V deal with specific individual population and housing items.

### II.3.1. GEOGRAPHY EDITS

*1. Location of living quarters (locality)*

210. A locality, according to *Principles and Recommendations for Population and Housing Censuses*, is defined as "a distinct population cluster… in which the inhabitants live in neighbouring 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 the *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. When editing for location the geographical codes must be absolutely 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.

*2. Urban and rural residence*

211. The traditional distinction between urban and rural areas within a country was based on the assumption that urban areas, no matter how they were defined, provided a different way of life with pre-dominant non-agricultural activity 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.

212.. Most countries determine which geographical areas are "urban" and which are "rural" before the census, mostly during the cartographic phase, 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 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.

213.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.

### II.3.2. COVERAGE CHECKS

*1. De facto and de jure enumeration*

214. *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.

215. Given this definition of usual residence, National census/statistical offices tend to collect censuses *de facto* (where persons are found on census night) or *de jure* (where they are usually found). 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 *de facto* and *de jure* information. 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. *De facto* tabulations on a *de facto* basis use only (1) and (2) if all three types are present. *De jure* tabulations use only (1) and (3) if all three types are present.

216. National census/statistical organizations implementing these categories must be very careful in their use, not only during data collection and processing, and not make tables of the entire data set but also during later analysis. When these three categories are used, users must be aware of the selected population since analysing all of the dataset will result in including some persons twice. If a de facto population is required, the tabulation must exclude category (3), persons temporarily away; if a de jure population is required, the tabulation must exclude category (2). 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.

217. The editing program should be designed to ensure that, when all three record types are present, the correspondences are appropriate. If *de facto* records have few answers, it may indicate that they are in reality absent residents or that another enumeration problem requiring special treatment is present.

```
Malawi 2008
PROC RESSTAT

{. ***************************************************************************
 . *****************               RESIDENTAL STATUS             ****************
             What is the situation of residence of ...?
             [If visitor 93), go to the next person.
. ***************************************************************************}

  if RESSTAT in 1:3 then
    if RELATIONSHIP = 1 then
      ARESSTAT (AGE5A,SEX) = RESSTAT;
      HEADRESSTAT = RESSTAT;
    endif;
    if RESSTAT in 2:3 then
      N01 = 0;
      if RELIGION<> NOTAPPL then N01 = N01 + 1; endif;
      if TRIBE <> NOTAPPL then N01 = N01 + 1; endif;
      if PREVRES <> NOTAPPL then N01 = N01 + 1; endif;
      [Continue for all other population variables]
      if N01 >= 5 then
        errmsg("*P08-0a* RESSTAT [%2d] RESIDENT from other
          variables", RESSTAT) denom = POPCNT summary;
        impute (RESSTAT, 1); {Made resident by responses}
      endif;
    endif;
    {Checking for resident and absent people}
      if $ in 2 then
       if TYPEOFHH in 2:3 then
        if N01 > 5 then
          errmsg("*P08-0a1* RESSTAT [%2d] RESIDENT from
            household type [%1d]", RESSTAT,TYPEOFHH)
            denom = POPCNT summary;
          impute (RESSTAT, 1);
            {Made visitor because of responses}
        else
          if RESSTAT (HEADPTR) = 1 then
            errmsg("*P08-0a2* RESSTAT [%2d] RESIDENT from
              household type [%1d]", RESSTAT,TYPEOFHH)
              denom = POPCNT summary;
          impute (RESSTAT, 1);
            {Made regular because head is regular}
        else
          errmsg("*P08-0a3* RESSTAT [%2d] RESIDENT from
            household type [%1d]", RESSTAT,TYPEOFHH)
            denom = POPCNT summary;
          impute (RESSTAT, 3);
            {Made visitor because of responses}
        endif;
       endif;
      endif;
    endif;
  else   {Residential status unknown}
      N01 = 0;
      if RELIGION<> NOTAPPL then N01 = N01 + 1; endif;
      if TRIBE <> NOTAPPL then N01 = N01 + 1; endif;
      if PREVRES <> NOTAPPL then N01 = N01 + 1; endif;
      [Continue for all other population variables]
      if N01 >= 5 then
        errmsg("*P08-0a3* RESSTAT [%2d] RESIDENT from other
          variables", RESSTAT) denom = POPCNT summary;
        impute (RESSTAT, 1);
          {Made resident because of responses}
      else
        errmsg("*P08-0a4* RESSTAT [%2d] VISITOR from lack of
          other variables", RESSTAT) denom = POPCNT summary;
        impute (RESSTAT, 3);
          {Made visitor because of responses}
      endif;
  endif;
```

*2. Hierarchy of households and housing units*

218. Chapter II.5 examines the relationships between households, housing units and living quarters. 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.

*3. Fragments of questionnaires*

219. 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, in order to delete them. The most common case will be a record that contains geographical codes but no population or housing items.

```
// Delete blank person records

do varying i = totocc(PERSON_EDT) until i <= 0 by (-1)
    if RELATIONSHIP(i) in notappl,0,missing and  // blank relationship
        SEX(i) in notappl,0,missing and  // blank sex
        AGE(i) in notappl,missing then  // blank age
        delete (PERSON_EDT(i)); { remove "blank" person records }
        errmsg("Removed blank person records")denom = denomPop summary;
    endif;
enddo;
```

### II.3.3. STRUCTURE OF HOUSING RECORDS

220. 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. 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).

### II.3.4. CORRESPONDENCE BETWEEN HOUSING AND POPULATION RECORDS

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

*1. Vacant and occupied housing*

222. 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”. 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.

223. 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:

224. (a) Choosing to leave a housing unit vacant. In this case, the editing team decides that vacant housing units coming in from the field should be left as vacant, so no values are imputed. Housing edits for vacant units are described in section II.5.

225. (b) Revisiting the housing unit several times to complete questionnaires. The national census/statistical office may choose to implement procedures requiring enumerators to keep returning to the data on vacant units until they are certain that these units are either vacant or are occupied and until the enumerators have collected 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". Depending on what the editing team decides is "minimal" information, the regular edit described in section II.4 is applied, or data from donor records are supplied for "missing" persons, as described above.

226. (c) Substituting another housing unit for missing persons. Procedures for substituting whole households or individual missing persons are described elsewhere in this chapter. These procedures require assuming 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.

*2. Duplicate households and housing units*

227. Duplicate housing units occur for a variety of reasons. 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. 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.

*3. Missing households and housing units*

228. Similarly, 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 hot deck imputation to create information about household members since this method often produces variables inconsistent with each other.

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

229. 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. Chapter V elaborates on the housing edit for number of occupants.

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

230. 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.

231. 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. 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.

232. 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, when the program corrects all individual entries, 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.

233. (b) Checking numbers of persons by sex. Sometimes the number of occupants is reported by sex on the housing record. In this case, the 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.

```
MALES = count (POP where SEX = 1);
FEMALES = count (POP where SEX = 2);
If MALES <> MALES_IN_HOUSE then
  Errmsg ("*P00-1* Sum of males <> males in house") summary;
Endif;
If FEMALES <> FEMALES_IN_HOUSE then
  Errmsg ("*P00-1* Sum of males <> males in house") summary;
Endif;
```

234. (c) Sequence numbering. 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. 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.

```
if PERSON_NUMBER <> CUROCC (POPULATION_EDT) then
  errmsg ("*P01-01* Person number out of order, PN = [%2d], cur occ = [%2d]",PERSON_NUMBER,CUROCC (POPULATION_EDT))
      denom = denomPOP summary;
  FPOP();
  write ("*P01 -01* Person number out of order, PN = [%2d], cur occ = [%2d]",PERSON_NUMBER,CUROCC (POPULATION_EDT));
  impute (PERSON_NUMBER,CUROCC (POPULATION_EDT));
endif;
```

*5. Correspondence between occupants and type of building/household*

235. 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 take into account the size of the household and other variables.

```
INST_MEMBERS = count (POP where RELATIONSHIP = 99);
If INST_MEMBERS <> totocc (POP) then
   Errmsg ("*P00-99*  Supposed to be collective but others present") summary;
Endif;
```

### II.3.5. DUPLICATE RECORDS

236. 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 actually be twins when reference is made to their names.

237. 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.

238. 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:

1. 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.
2. If 3 or more characteristics are different, the line number will be changed.[6]

### II.3.6. SPECIAL POPULATIONS

*1. Persons in collectives*

239. 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.

240. *When collectives are a different record type.* When the national census/statistical office uses a separate record type, 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.

241. *When a variable distinguishes collectives from other records.* 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.

242. *When the "type of collective" code is missing.* 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.

243. *When the collective code is present, but all of the persons are related.* 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

[6] Traditionally, duplicate records were tracked down and corrected manually, but more and more, these are at least partially automated. A recent paper, "Data Quality: Automated Edit/Imputation and Record Linkage" (Winkler 2006) begins to look automating structure and content edits together.

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.

244. *Distinguishing various types of collectives*. 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*

245. *Seasonal migrants*. 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. 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.

246. *Homeless persons*. By definition, the record of a homeless person will not have housing information. However, 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. During the cartographic phase the geo-coding of the areas in which the homeless people stay is important as the EAs with such persons can easily be identified in the editing process

247. *Nomads and persons living in areas to which access is difficult.* 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.

248. *Civilian residents temporarily absent from the country*. 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* some indicator should show persons who are temporarily absent to allow for both *de jure* and *de facto* populations 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.

249. *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*. For a *de facto* census, everyone resident 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 their place of residence at the time of the census, and edited their, 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 for the hot decks.

250. Foreign persons only in the country temporarily, presumably are not included for *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 reside in the country usually. If a ship is usually harbored in the country, then presumably the persons on the ship would be included as usual residents and would be edited as such.

251. *Refugees*. Refugees 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.

252. *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.* 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. For a *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.

253. *Civilian foreigners who cross a frontier daily to work in the country.* Civilian foreigners who cross a frontier daily to work in the country would normally not be included in either the *de jure* or *de facto* censuses because they do not reside in the country on the reference date, 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.

254. *Civilian residents who cross a frontier daily to work in another country.* 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.

255. *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).* 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.

256. *Conflict areas*. In Africa there are conflict zones within many countries, and these areas should be identified during the census cartographic period and countries have to device special methodologies of enumeration in these special EAs.. The editing procedure should take into the methodology followed in each case.

### II.3.7. DETERMINING HEAD OF HOUSEHOLD AND SPOUSE

*1. Editing the head of household variable*

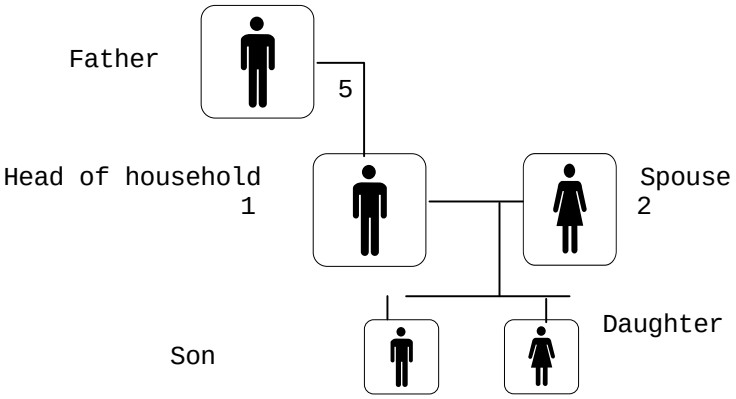
257. In identifying the members of a household, it is traditional to identify first the head of household or reference person and then the remaining members of the household according to their relationship to the head or reference person. The head of the household is defined as that person in the household 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, 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, the concepts and definitions that are used.

258. *The order of the relationships*. The order of the relationships in the unit has an effect on the edits since many of the edits assume that the head of household 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 has valid information for any of these variables, that information is imputed for any other person in the household where it is missing, miscoded or miskeyed (refer to Chapter II.4). 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.

259. *When the head is not the first person*. 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 25.

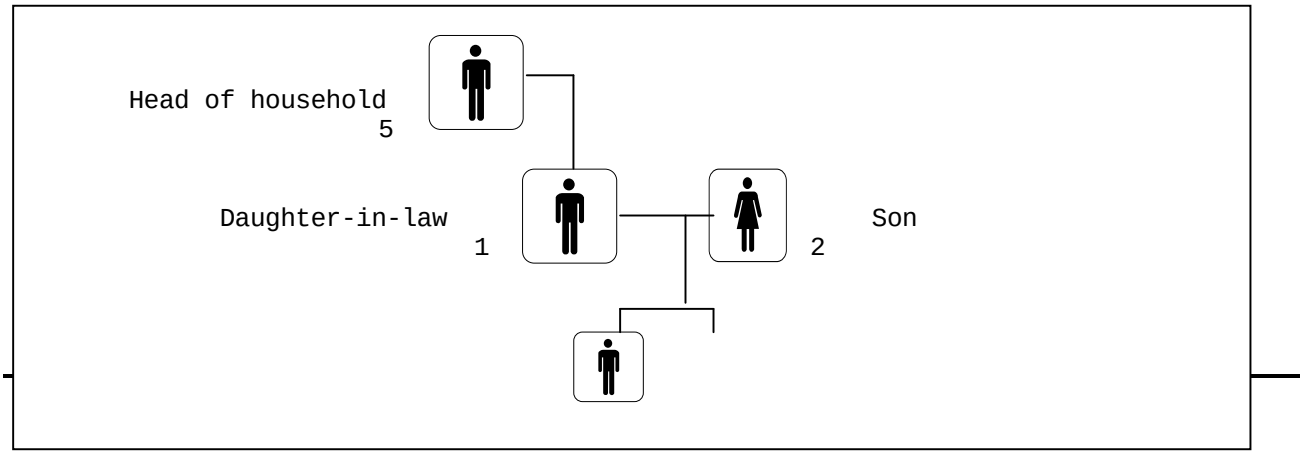
**Figure 25. Example of household with head of household listed as first person**




260. 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.

261. 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 26. 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 26. Example of household with head of household listed as fifth person**



Grandson 3  Granddaughter 4

262. Obviously as illustrated by theses 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: (a) a pointer can be used to note which person is head, and the pointer can be used throughout the edits and tabulations; (b) 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, (c) the relationship codes can be changed to have the first person as head, no matter what the other relationships. See "when the head has to be first" in the Relationship section below.

263. (i) Assigning a pointer for the head's record. In the editing procedures regarding the head of household, 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 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".

264. (ii) Making the first person the head. 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.

265. (iii) Reassigning relationship codes to make the first person the head. 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.

266. 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.

267. 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.

268. *More than one head*. When more than one head of household is found, the editing team must determine who is to be designated as the head of household. The edit 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.

269. 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. However, the current Principles and Recommendations (2.117)

include a provision for joint heads. If a county 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.

270. *No head*. Similarly, 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.

```
HEADS = count(POPULATION_EDT where RELATIONSHIP in 0:1 );    {.
Count the heads}
HEADPT = 0;

{ If one head, note head's line number with head pointer }

 if HEADS = 1 then
   for i in POPULATION_EDT do
     if RELATIONSHIP (i) in 0:1 then
       HEADPT = i;
     endif;
   enddo;
   {exit;}
 endif;

{ If no heads, make the first person the head of household }

 if HEADS = 0 then
   do varying N02 = 1 while N02 <= totocc (POPULATION_EDT);
     if AGE (N02) > 14 and HEADPT = 0 then
       errmsg ("*P00 -02* No head of household, first person
         14+ becomes head, pn [%01d]",PERSON_NUMBER(N02))
         denom = denomPOP summary;
       impute (RELATIONSHIP (N02),  1 )
         {. Make the first person the head};
       HEADPT = N02;
       break;
     endif;
   enddo;
   if HEADPT = 0 then
      errmsg  ("*P00 -03* No head 14+, first person becomes
        head - 1 pn [%01d]",PERSON_NUMBER(1))
         denom = denomPOP summary;
         impute (RELATIONSHIP (1),  1 )
          {. Make the first person the head};
         HEADPT = 1;
         break;
       endif;
endif;

{ If more than one head, find one and only one head }

if HEADS > 1 then
  errmsg("*P00 -04* Too many heads of household - 1 ")
    denom = denomPOP summary;
  do varying i = 1 while i <= totocc (POPULATION_EDT)
    if RELATIONSHIP (i) in 0:1 then
      HEADPT = PERSON_NUMBER (i);
      if totocc (POPULATION_EDT) > 1 then
        do varying j = i + 1 while j <=
          totocc (POPULATION_EDT)
          if RELATIONSHIP (j) in 0,1 then
            N02 = PERSON_NUMBER(j);
            errmsg   ("*P00 -05* Remaining heads made other
              RELATIONSHIP- 1 pn [%01d]",N02)
                denom = denomPOP summary;
            impute (RELATIONSHIP (j),7);
            HEADS = HEADS - 1;
          endif;
        enddo;
      endif;
    endif;
  enddo;
endif;
```

*2. Editing the spouse*

271. *When exactly one spouse is found in monogamous societies*. 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).

```
SPOUSES = 0;
Do varying J = 1 while J <= totocc (POP)
   If RELATIONSHIP (J) = 2 then
      SPOUSES = SPOUSES + 1;
      SPOUSE-PTR = J;
   Endif;
Enddo;
[if SPOUSES = 1, then the Spouse-ptr will be pointing to the line number of the spouse]
```

272. *When more than one spouse is found in monogamous societies*. In a monogamous society, if more than one spouse is found in the dataset, then 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.

```
SPCOUNT = 0;
If SPOUSES > 1 then
   Do varying J = 1 while J <= totocc (POP)
     If RELATIONSHIP (J) = 2 then
       If SPCOUNT = 0 then
         SPOUSE-PTR = J;
         SPCOUNT = 1;
         SPOUSES = 1;
       Else
         RELATIONSHIP (J) = 9: {Make other relative}
       Endif;
     Endif;
   Enddo;
Endif;
[SPOUSES should be 1, and the Spouse-ptr will be pointing to the line number of the spouse,
 and all other previous spouses will be other relative.]
```

273. *Spouses in polygamous societies*. 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.

274. *Other characteristics of Heads and Spouses*. 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. (Note that the edit below is a very simplified one, and more checking should be done, like which one is mother to the children. For religion, ethnicity, and language see the edits further on.)

```
If RELATIONSHIP = 2 then
  If SEX = SEX (1) then
     If CEB <> NOTAPPL then
        Impute (SEX,2);
        Impute (SEX(1),1);
     Else
        Impute (SEX,1);
        Impute (SEX(1),2);
     Endif;
  Endif;
Endif;
```

### II.3.8. AGE AND BIRTH DATE

275. *When date of birth is present, but age is not.* 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.

```
CENSUS_YEAR = 2010; CENSUS_MONTH = 7; CENSUS_DAY = 15;
AGE = CENSUS_YEAR – BIRTHYEAR;
If BIRTHMONTH >= 8 then
   AGE = AGE – 1;   {Have yet to get to birthday}
Endif;
If BIRTHMONTH = 7 and BIRTHDAY > 15 then
  AGE = AGE – 1;  {Have yet to get to birthday}
Endif;
```

276. *When the age and date of birth disagree*. When the census or survey obtains both 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 by more than one year from the reported age, 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.

Malawi 2008

```
{.************************************************************************
. **************                                          ****************
. **************                Age and Birthdate         ****************
. **************                                          ****************
. ************************************************************************}

if not YRBTH in 1908:2008 then
  if AGE in 0:99 then
    errmsg("*P05-1* Age good for unknown year of birth ")
      denom = popcnt summary;
    YRBTH = 2008 - AGE;
  else
    errmsg("*P05-2* Age no good for unknown year of birth ")
denom = popcnt summary;
    if AGE (HEADPT) in 15:99 then
      if RELATIONSHIP in 2 then   {Spouse,sibling}
        errmsg ("*P05-3* Age and year of birth of spouse
          from age of head") denom = popcnt summary;
        impute (AGE,AGE(HEADPT));
        YRBTH = 2008 - AGE;
      endif;
      if RELATIONSHIP in 3 then    {Child, child-in-law}
        errmsg ("*P05-4* Age and year of birth of child
          from age of head") denom = popcnt summary;
        impute (AGE,(AGE(HEADPT)-15));
        if AGE > 55 then impute (AGE,50); endif;
        YRBTH = 2008 - AGE;
      endif;
      if RELATIONSHIP > 3 then      {Grandchild}
        errmsg ("*P05-5* Age and year of birth of other
          relative from age of head") denom = popcnt summary;
        impute (AGE,30);
        YRBTH = 2008 - AGE;
      endif;
    endif;
  endif;
else
  if AGE in 0:99 then
  else
     if YRBTH in 1900:1908 and RELATIONSHIP in 3:5 then
      YRBTH = YRBTH + 100;
    endif;
    errmsg("*P05-6* Age from year of birth ")
      denom = popcnt summary;
    impute (AGE , 2008 - YRBTH);
  endif;
endif;
```

### II.3.9. COUNTING INVALID ENTRIES

277. 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 one or two 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. 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 the program uses values repeatedly, then the matrices must be larger and more varied.

## II.4. EDITS FOR POPULATION ITEMS

278. Chapter II.4 covers edits for population items, including those related to demographic, migration, social and economic characteristics. 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 tabulation. For example, population records should not have 15-year-old females with 10 children or 7-year-old children attending tertiary school.

279. 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.

280. In many 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.

281. All population records should have serial numbers to assist in data processing. The structural edits described in Chapter II.3 check for correspondence between the sequence number and the order of serial numbers.

282. 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 change missing or inconsistent data. Even fewer countries actually implement this added procedure.

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

284. 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.

285. Population edits tend to be more complicated than housing edits because cross-tabulations are generally much more complicated. 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.

286. 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, and 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 and employment status.

287. 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.

### II.4.1. DEMOGRAPHIC CHARACTERISTICS

289. Data on relationship, sex, age and marital status for each person are basic to any census and should probably be edited together. The age and sex structures of populations or subpopulations are fundamental for almost all planning based on population censuses. These items are also essential to the production of meaningful tabulations since virtually all other analyses are based upon cross-tabulations of other variables by age and sex.

290. The multiple-variable (Fellegi-Holt) approach to editing population and housing data was introduced in section II.2 of this *Handbook*. Since the demographic variables are integral for all planning based on a population census, 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. Relationship*

291. 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 and no more than one spouse (in non-polygamous societies) is covered in the structure edits.

292. *Relationship edits*. 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 endeavour. We discuss the recode variables "family type", and subfamily number and subfamily relationship recodes in the recode section.

293. 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.

294. *When the head must appear first.* If the head 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.

```
{Makes head first person}
if heads = 1 and isHH then
  if headPtr <> 1 then
    errmsg ("[PERSON-0.1] Head is not first person")denom = denomPop summary;
    ok = swap(PERSON_EDT,1,headPtr);
    if ok then
      errmsg ("[PERSON-0.2] Head moved to first person")denom = denomPop summary;
    endif;
  endif;
endif;
```

295. *When the relationships are coded upside down*. 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. The procedure for inverting the relationships usually involves running a "look up" file containing the original relationships and the inverted relationships, taking into account sex of the respondents.

```
count (PARENTS where RELATIONSHIP = 6); {Parent}
if PARENTS > 2 then
   [Invert the relationship is the household, so that parents become children and children parents,
     And grandparents become grandchildren, and so forth]
endif;
```

296. *When polygamous spouses are present.* The structure edits, if performed as indicated in chapter III, will have already checked for "one and only one" head and "no more than one spouse" for monogamous households. For polygamous households where more than one spouse is actually living in the housing unit, the editing team should decide when polygamous relationships are permitted and when they are not. (And whether only males can be polygamously married.) Sometimes households that seem to have polygamous relationships are actually mistakes. 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. See the relationship edit below for editing multiple spouses.

297. *When multiple parents appear*. 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.

```
PARENTS = count (POP when RELATIONSHIP = 6); {Relationship of parent is 6}
If PARENTS > 1 then
  If PARENTS = 2 then
      [use do loop to find line numbers of the two parents]
      If SEX (PARENT1) = SEX (PARENT2) then
         [use fertility or orphanhood or line numbers to determine which is female]
      Endif;
```

```
   Endif;
Endif;
```

298. *When censuses collect sex-specific relationships*. 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 is not a good idea to use sex-specific relationships since redundancy does not clarify the relationships, they only fog them, and require additional editing.

```
If RELATIONSHIP = 3 then  {This is a son}
   If SEX = 2 then
      Impute (SEX,1);  {Change female sons to male sons – sex takes precedence over relationship}
   Endif;
Endif;
If RELATIONSHIP = 4 then  {This is a daughter}
   If SEX = 1 then
      Impute (SEX,2);  {Change male daughters to female daughters – sex takes precedence over relationship}
   Endif;
Endif;
```

299. *When relationship and marital status do not match*. Relationship and marital status should agree when they overlap: 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).

```
If RELATIONSHIP = SPOUSE then
   If MARITAL_STATUS <> MARRIED then
       Impute (MARITAL_STATUS, MARRIED));
   Endif;
   If MARITAL_STATUS (HEADPTR) <> MARRIED then
       Impute (MARITAL_STATUS (HEADPTR),MARRIED))
    Endif;
Endif;
```

300. 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.

```
if RELATIONSHIP = UNMARRIED_PARTNER then
    if MARITAL_STATUS = MARRIED then
       impute (MARITAL_STATUS,NEVER_MARRIED);
    endif;
endif;
```

301. 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.

302. In the old days – like at the turn of this century – countries did not release microdata, and so users would not be crossing relationship with other variables, like age. However, now that microdata sets are released, it is important to decide how far to go in the edit. The following pseudo code provides methods of checking for children and grandchildren who are too old, and parents who are too young.

```
   if RELAT in 3:4 then           {. Child}
     if AGE >= 55 then
       errmsg("*P02 -3B* Child who is too old, pn= [%02d] RELAT= [%01d] age= [%01d]",PN,RELAT,AGE)  summary;
       impute (RELAT,8);
     endif;
     exit;
   endif;

   if RELAT = 7  then           { . Grandchild}
     if AGE >= 35 then
       errmsg("*P02 -3C* Grandchild too old, pn= [%02d] RELAT= [%01d] age= [%01d]",PN,RELAT,AGE)  summary;
        impute (RELAT,8);
```

```
    endif;
    exit;               {. See Age edit below}
  endif;

   if RELAT = 6  then           { . Parent}
    if AGE < 40 then
      errmsg("*P02 -3D* Parent too young, pn= [%02d] RELAT= [%01d] age= [%01d]",PN,RELAT,AGE)  summary;
      AGEDIF = AGE (HEADPT) - AGE;
      if AGEDIF in 0:15 then
        errmsg("*P02 -3D1* Parent too young, pn= [%02d] RELAT= [%01d] age= [%01d]",PN,RELAT,AGE)  summary;
        impute (RELAT,8);
      endif;
      if AGEDIF in 16:39 then
        errmsg("*P02 -3D2* Parent too young, RELAT becomes child,pn= [%02d] RELAT= [%01d] age= [%01d]",PN,RELAT,AGE)  summary;
        impute (RELAT,3);
      endif;
      if AGEDIF in 40:60 then
        errmsg("*P02 -3D3* Parent too young, RELAT becomes grchild,pn= [%02d] RELAT= [%01d] age= [%01d]",PN,RELAT,AGE)  summary;
        impute (RELAT,5);
      endif;
      if RELAT = 6 then
        errmsg("*P02 -3D4* Parent too young,RELAT becomes other relative,pn= [%02d] RELAT= [%01d] age= [%01d]",PN,RELAT,AGE)  summary;
        impute (RELAT,7);
      endif;
    endif;
    exit;               {. See Age edit below}
  endif;
```

*2. Sex*

303. 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 analysed 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. 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.

304. *When the sex code is valid but the head and spouse are the same sex*. 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, 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.

```
If SEX (HEADPT) = SEX (SPOUSEPT) then
 If CEB (SPOUSEPT) <> NOTAPPL then
    If SEX (HEADPT) = 1 then impute (SEX (SPOUSEPT),2);
    Else impute (SEX (HEADPT),1);
    Endif;
 Elseif CEB (HEADPT) <> NOTAPPL then
    If SEX (HEADPT) = 2 then impute (SEX (SPOUSEPT),1);
    Else impute (SEX (HEADPT),2);
    Endif;
 Else [Do something else – this didn't help!]
 Endif;
Endif;
```

305. *When a male has fertility information or an adult female does not*. 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.

```
if SEX = MALE or (SEX = FEMALE and AGE < 12) then
   if CEB <> NOTAPPL then
      [make decision about whether fertility takes precedence over sex, or vice versa, then act]
   endif;
else
   if CEB = NOTAPPL then
      impute [Fertility variables]
   else
       [Update array of fertility variables]
```

```
   endif;
endif;
```

306. *When the sex code is invalid and a spouse is present.* 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 is "head of household", 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.

```
if not SEX (HEADPTR) in 1:2 then
   if SEX (SPOUSEPTR) in 1:2 then
      impute (SEX (HEADPTR),3-SEX(SPOUSEPTR));
   endif;
endif;
```

307. *When the sex code for spouse is invalid*. If the relationship of the person to the head of household is "spouse", and the sex of the head of household is given, the program assigns to this person the sex opposite that of the head of household.

```
if not SEX (SPOUSEPTR) in 1:2 then
  impute (SEX (SPOUSEPTR),3-SEX(HEADPTR));
endif;
```

308. *When the sex code is invalid and female information is present.* 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 we are editing 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.

```
if not SEX in 1:2 then
    if FERTILITY <> NOTAPPL then   {The whole fertility block is checked}
       impute (SEX,2);   {Make sex "female" because fertility is present}
   endif;
endif;
```

309. *When the sex code is invalid and this person is spouse's husband.* 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. (Note that if the census collects "spouse's line number" rather than husband's line number, then both sexes must be checked – that is, the spouse of a female must be a male, and the spouse of a male must be a female.)

```
if HUSBANDS_LN <> NOTAPPL then
    if SEX (HUSBANDS_LN) <> 1 then
       impute (SEX(HUSBANDS_LN),1);   {Make husband's sex "male"}
   endif;
else
  impute (SEX(HUSBANDS_LN),1);   {Make husband's sex "male"}
endif;
```

310. *When the sex code is invalid and there is insufficient information to determine sex.* 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. Since sex is a binary variable, values can be assigned alternately, starting with either, using the opposite sex for the second invalid entry and continuing in this fashion.

{Sex of head and sex of spouse}

```
N01 = 0;
if HEADPT > 0 and SPOUSEPT > 0 then
 if SEX (HEADPT) in 1:2 then
   if not SEX (SPOUSEPT) = 3 - SEX(HEADPT) then
     impute (SEX (SPOUSEPT),3 - SEX(HEADPT));
   endif;
 else
   if SEX (SPOUSEPT) in 1:2 then
     impute (SEX(HEADPT),3 - SEX (SPOUSEPT));
   else
     impute (SEX(HEADPT),9);
     impute (SEX(SPOUSEPT),9);
   endif;
  endif;
```

```
if SEX (HEADPT) = SEX (SPOUSEPT) then
  K = 0;
  if MCEB (HEADPT) <> NOTAPPL or
     FCEB (HEADPT) <> NOTAPPL then
       K = 1;
  endif;
  if K = 1 then
      impute (SEX(HEADPT),2);
      impute (SEX(SPOUSEPT),1);
      N01 = 1;
  else
    K = 0;
    if MCEB (SPOUSEPT) <> NOTAPPL or
       FCEB (SPOUSEPT) <> NOTAPPL then
       K = 1;
```

```
      endif;
      if K = 1 then
        impute (SEX(SPOUSEPT),2);
        impute (SEX(HEADPT),1);
        N01 = 1;
      endif;
    endif;
```

```
      if N01 = 0 then
        impute (SEX(SPOUSEPT),2);
        impute (SEX(HEADPT),1);
      endif;
    endif;
endif;
```

311. *Note on imputed sex ratios*. Female sex is likely to be assigned more often when cold deck imputation is used. 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 take into account the overall proportions between the sexes.

**WHAT STAFF PROBABLY SHOULD NOT DO:**

```
if not SEX in 1:2 then
   if CEB in 0:20 then
       impute (SEX,2); {Sex imputed female because fertility is present}
   else
       impute (SEX,SEXASSIGN);   {Half of remaining unknowns go to each sex}
       SEXASSIGN = 3 – SEXASSIGN;
   endif;
endif;
```

*3. Birth date and age*

312. 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 analysed 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.

313. The edit and imputation for age should do the following:

(a) Assign age where age is blank;
(b) Check for minimum age of ever-married persons;
(c) Check for minimum age of head of household;
(d) Check for minimum age of parents; and
(e) Carry out any other country specific checks.

314. *Age and date of birth*. 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. Often, the structure edit calculates age from date of birth. First, however, it is useful to review the difference between age and birth date. 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. The question on date of birth is appropriate when people know their birth date, which may be estimated by using a solar calendar or a lunar calendar, or expressed in years numbered or identified in traditional folk culture by names within a regular cycle.

315. It is extremely important, however, that a clear understanding should exist between the enumerator and the respondent about which calendar system provides the date of birth. If some respondents might 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 used. It is not advisable for the enumerator to convert the date from one system to another. Programmers can best carry out the needed conversion as part of the computer editing work.

316. The direct question on age is likely to yield less accurate responses. Even if all responses are based on the same method of determining age, the respondent may not understand whether the age wanted is that at the last birthday, the next birthday or the nearest birthday. 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.

317. Many national census/statistical offices collect either date of birth or age, but not both. 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.

318. *Relationship between date of birth and age*. 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.

319. 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 should probably 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.

```
do varying i = 1 while i <= totocc (POPULATION_EDT)
  if not DOBY_YEAR (i) in 1908:2010 then    {Year of birth is not known}
    if AGE (i) in 00:98 then           {Age is known, so get from year of birth}
      errmsg ("*P00 -1x* Person [%01d] Year of birth unknown [%04d], so obtained from age [%2d]",
         PERSON_NUMBER (i),DOBY_YEAR (i),AGE(i))  summary;
      DOBY_YEAR (i) = CENSUS_YEAR - AGE (i);
    endif;
  else                                  {Year of birth is known}
    if not AGE (i) in 00:98 then       {So get age from year of birth}
      errmsg ("*P00 -1y* Person [%01d] Year of birth known [%04d], so obtained age obtained [%2d]",
         PERSON_NUMBER (i),DOBY_YEAR (i),AGE(i))  summary;
      AGE (i) = CENSUS_YEAR - DOBY_YEAR (i);
    endif;
  endif;
enddo;
```

320. *When calculated age falls above the upper limit*. For censuses in 2000 and beyond, most countries will choose to record all four digits for year of birth. For those around 2010, 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 2010, and a person reports being born in 1860, the computed age of 150 years is likely to be outside the acceptable range, and will need to be changed.

```
If YEAROFBIRTH in 1900:2010 then
  AGEX = CENSUSYEAR – YEAROFBIRTH;
  If  AGEX <> AGE then
     Impute (YEAROFBIRTH,(CENSUSYEAR – AGE));
     {Note that this is not a complete edit – depending on the month (and sometimes the day)
      The age could be one year off from this calculation.  This needs to be included}
  Endif;
Else
  If AGE in 0:110 then
    Impute (YEAROFBIRTH,(CENSUSYEAR – AGE));
  Endif;
Endif;
```

Also, it is important to note that with scanning, sometimes one or another of the digits is not picked up. Usually programmers will parse the 4 digits into their parts and work to with those in error to come up with a definite 4-digit response, sometimes using the reported age to help. Then, the edit above would run.

321. *Age edit*. 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. See paragraphs below for edits for age difference between head and child and head and parent.

322. 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.

323. *Age edit when the head of household and spouse are present.* 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.

```
if RELATIONSHIP = 1 then
   if not AGE in 15:99 then
      {   . ************************** Head's Age from Spouse ************************** .}
     if SPOUSEPT <> 0 then
       if AGE (SPOUSEPT) in 15:99 then
         errmsg("*P05 -13* Head's Age from Spouse's Age,pn= [%02d] spousepn= [%02d] agesp= [%02d], agehd= [%02d]",
             PN, PN (SPOUSEPT), AGE, AGE (SPOUSEPT)) denom = denomPOP summary;
         recode AGE (SPOUSEPT) => AGE5A;
                   0-4 =>  1;    5-9 =>  2;   10-14 => 3;    15-19 => 4;     20-24 => 5;
                  25-29 => 6;  30-34 =>  7;   35-39 => 8;    40-44 => 9;     45-49 => 10;
                  50-54 => 11; 55-59 => 12;   60-64 => 13;   65-69 => 14;    70-74 => 15;
                  75-79 => 16; 80-84 => 17;   85-89 => 18;   90-94 => 19;          => 20;
         endrecode;
         impute( AGE , ASPAGE (AGE5A,SEX));
         DOBY_YEAR = 2010 - AGE;
         exit;
       else
         if AGE (HEADPT) in 15:96 then
           errmsg("*P05 -14* Spouse's Age from Head's Age,pn= [%02d] headpn= [%02d] agehd= [%02d], agesp= [%02d]",
             PN, PN (HEADPT), AGE, AGE (HEADPT)) denom = denomPOP summary;
           recode AGE (HEADPT) => AGE5A;
                    0-4 =>  1;    5-9 =>  2;   10-14 => 3;    15-19 => 4;     20-24 => 5;
                   25-29 => 6;  30-34 =>  7;   35-39 => 8;    40-44 => 9;     45-49 => 10;
                   50-54 => 11; 55-59 => 12;   60-64 => 13;   65-69 => 14;    70-74 => 15;
                   75-79 => 16; 80-84 => 17;   85-89 => 18;   90-94 => 19;          => 20;
           endrecode;
           impute( AGE , ASPAGE (AGE5A,SEX));
           DOBY_YEAR = 2010 - AGE;
           exit;
         endif;
       endif;
     endif;


       if SPOUSEPT <> 0 then
         if AGE (SPOUSEPT) in 0:99 then
           recode AGE (SPOUSEPT) => AGE5A;
                     0-4 =>  1;    5-9 =>  2;   10-14 => 3;    15-19 => 4;     20-24 => 5;
                    25-29 => 6;  30-34 =>  7;   35-39 => 8;    40-44 => 9;     45-49 => 10;
                    50-54 => 11; 55-59 => 12;   60-64 => 13;   65-69 => 14;    70-74 => 15;
                    75-79 => 16; 80-84 => 17;   85-89 => 18;   90-94 => 19;          => 20;
           endrecode;
            ASPAGE (AGE5A,SEX) = AGE;
         endif;
       endif;
```

324. To ensure that this computed age is consistent with other characteristics, the imputation matrix might 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.

325. *Age edit for head when the head's spouse is absent, but child is present* When comparison with the age of the spouse is not possible in determining the age of the head of household, the program can then check relationship. If the relationship is "head 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.

```
if not AGE (HEADPTR) in 15:99 then   {Age of head is unknown}
   if RELATIONSHIP = CHILD then
      if AGE (CHILD_LINE_NUMBER) in 0:75 then   {Child is present with age}
         [recode child age to AGEX]
         impute (AGE,AAGEFROMCHILD(AGEX,SEX));
      endif;
   endif;
else          {Age of head is known}
  if RELATIONSHIP = CHILD then
    if AGE (CHILD_LINE_NUMBER) in 0:75 then    {Age of child is known}
      AGEDIF = AGE (HEADPTR) – AGE (CHILD_LINE_NUMBER);
      if AGEDIF > 13 then      {Age difference is acceptable, so put into hotdeck}
        [recode child age to AGEX]
        AAGEFROMCHILD(AGEX,SEX) = AGE;
      endif;
    endif;
  endif;
endif;
```

326. *Age edit for head when head's parent is present*. 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. The example for child, above, should be adapted for parent.

```
if not AGE (HEADPTR) in 15:99 then   {Age of head is unknown}
   if RELATIONSHIP = PARENT then
      if AGE (PARENT_LINE_NUMBER) in 30:98 then   {Parent is present with age}
         [recode parent age to AGEX]
         impute (AGE,AAGEFROMPARENT(AGEX,SEX));
      endif;
   endif;
else          {Age of head is known}
  if RELATIONSHIP = PARENT then
    if AGE (PARENT_LINE_NUMBER) in 30:98 then    {Age of parent is known}
      AGEDIF = AGE (HEADPTR) – AGE (PARENT_LINE_NUMBER);
      if AGEDIF > 13 then      {Age difference is acceptable, so put into hotdeck}
        [recode parent age to AGEX]
        AAGEFROMPARENT(AGEX,SEX) = AGE;
      endif;
    endif;
  endif;
endif;
```

327. *Age edit for head when head's grandchild is present.* 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. 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 in school, 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. The example for child above should be adapted for grandchild.

328. *Age edit for head when no other ages are available.* 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.

329. *Age edit for spouse when head's age already determined.* 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.

330. *Age edit for other married couples in the household when the age of one of the persons is known.* 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.

331. 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.

332. *Age edit for child when head's age already determined.* If this is a son or daughter of the head 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, 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.

```
if RELATIONSHIP = CHILD then
   if AGE in 0:75 then
      AGEDIF = AGE (HEADPTR) – AGE;
      if AGEDIF in 15:60 then
         AAGE_CHILD_FROM_HEAD (AGEX,SEX) = AGE;
      else
         impute (AGE,AAGE_CHILD_FROM_HEAD (AGEX,SEX));
      endif;
   else
      impute (AGE,AAGE_CHILD_FROM_HEAD (AGEX,SEX));
   endif;
endif;
```

333. *Age edit for parent when head's age already determined.* If this is a parent of the head 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.

```
if RELATIONSHIP = PARENT then
   if AGE in 30:99 then
      AGEDIF = AGE – AGE (HEADPTR);
      if AGEDIF in 15:60 then
         AAGE_PAR_FROM_HEAD (AGEX,SEX) = AGE;
      else
         impute (AGE,AAGE_PAR_FROM_HEAD (AGEX,SEX));
      endif;
   else
      impute (AGE,AAGE_PAR_FROM_HEAD (AGEX,SEX));
   endif;
endif;
```

334. *Age edit for grandchild when head's age already determined.* If this is a grandchild of the head of household, a computed 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.

```
if RELATIONSHIP = GRANDCHILD then
   if AGE in 0:50 then
      AGEDIF = AGE (HEADPTR) – AGE;
      if AGEDIF in 30:60 then
         AAGE_GRCH_FROM_HEAD (AGEX,SEX) = AGE;
      else
         impute (AGE,AAGE_GRCH_FROM_HEAD (AGEX,SEX));
      endif;
   else
      impute (AGE,AAGE_GRCH_FROM_HEAD (AGEX,SEX));
   endif;
endif;
```

335. *Age edit for all other persons.* 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. See the code above for suggestions for "other relatives."

*4. Marital status*

336. Marital status is 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) widowed and not remarried; (d) divorced and not remarried; and (e) married but separated. In some countries, category (b) may require a subcategory of persons who are contractually married but not yet living as man and wife. In all countries, category (e) should comprise both the legally and the de facto separated, who may be shown as separate subcategories if desired. Regardless of the fact that couples who are separated may be considered to be still married (because they are not free to remarry), neither of the subcategories of (e) should be included in category (b). In some countries, it will be necessary to take into account customary unions (which are legal and binding under customary law) and extralegal unions, the latter often known as de facto (consensual) unions.

337. *Marital status edit.* 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 parts of a country or different ethnic groups. 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 get the question. Younger persons fall into the "never married" category automatically. If everyone is asked the marital status item, however, the editing team must develop an edit for the whole population.

338. *Marital status assignment when dynamic imputation is not used.* 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.

```
if AGE < 15 then
   if MARITAL_STATUS <> NEVER_MARRIED then
       impute (MARITAL_STATUS,NEVER_MARRIED);
   endif;
else
   if MARITAL_STATUS in 1:5 then
   else
```

```
      impute (MARITAL_STATUS,UNKNOWN);
   endif;
endif;
```

339. *Marital status assignment when dynamic imputation is used.* 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.

```
  // Lets make sure that ppl < 15 are never married
  if AGE < 15 then
     if MARITAL_STATUS <> 5 then
        impute(MARITAL_STATUS,5);
        errmsg("[PERSON-0.9] imputed 'never married' for people < 15")denom = denomPop summary;
     endif;
     // What about those that are 15 and older
  else
    if MARITAL_STATUS in 1:5 then
      AMARITAL_STATUS (RELATIONSHIP,AGE10);
    else
      impute(MARITAL_STATUS, (AMARITAL_STATUS (RELATIONSHIP,AGE10);
      errmsg("[PERSON-0.10] impute marital of others based on sex and age")denom = denomPop summary;
    endif
  endif;
```

240. *Spouse should be married.* All persons coded “spouse” in the relationship category should be coded as married.

```
If RELATIONSHIP = SPOUSE then
   If MARITAL_STATUS <> MARRIED then
       Impute (MARITAL_STATUS, MARRIED));
   Endif;
   If MARITAL_STATUS (HEADPTR) <> MARRIED then
       Impute (MARITAL_STATUS (HEADPTR),MARRIED))
    Endif;
Endif;
```

341. *Spouse of a married couple pair.* 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. (Note that the edit below is simplified and should be more rigorous.)

```
If SEX (SPOUSELN) = SEX then    {the sex of the person designated as this person’s spouse is the same sex}
 If CEB <> NOTAPL then
   If SEX = 1 then
       Impute (SEX,2);
       Impute (SEX (SPOUSELN),1);
   Endif;
 Else
    If SEX = 2 then
       Impute (SEX,1);
       Impute (SEX (SPOUSELN),2);
    Endif;
 Endif;
Endif;
```

342. *If spouse, head should be married.* If no entry appears for marital status, but the entry for relationship to head of household is “head”, 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 of household as “married”.

```
// if this person is the spouse then both the spouse and the head has to be married
if spousePtr then
  if MARITAL_STATUS(spousePtr) <> 1 then
     impute(MARITAL_STATUS(spousePtr), 1);
     errmsg("[PERSON-0.11] imputed 'married' for spouse")denom = denomPop summary;
  endif;
  if MARITAL_STATUS(headPtr) <> 1 then
     impute(P7(headPtr), 1);
     errmsg("[PERSON-0.12] imputed 'married' for head")denom = denomPop summary;
  endif;
endif;
```

343. *Head, no spouse, without children.* 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 of household, but whose wife is not in the household, is most

likely to be divorced, separated or widowed.

344. *If all else fails, impute.* 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.

345. *Relationship of age to marital status for young people.* 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.

```
{.
. ****************************************************************************
. **************                                                ****************
. **************        Marital status                          ****************
. **************                                                ****************
. ****************************************************************************
.}
if FORMID (1) = 1 then
   {Children under 10 should have no applicable for marital status}
  IF S3Q6 < 10 then
    if $ <> NOTAPPL then
      IMPUTE (S3Q25,NOTAPPL);
    endif;
  else
    {Making sure heads and spouses are married when both are present}
    if RELATIONSHIP = 2 then  {this is a spouse}
      if $ <> 2 then  {this spouse is not married}
        errmsg ("P25-1 Spouse %d not married %d, PN = %2d",RELATIONSHIP,S3Q25,S3Q1) denom = denomPop summary;
        F1F3();
        write ("P25-1 Spouse %d not married %d, PN = %2d",RELATIONSHIP,S3Q25,S3Q1);
        impute ($,2);
      endif;
      if $ (1) <> 2 then {the head for this spouse is not married}
        errmsg ("P25-2 Head with Spouse %d not married %d",RELATIONSHIP(1),S3Q25(1)) denom = denomPop summary;
        F1F3();
        write ("P25-2 Head with Spouse %d not married %d, PN = %2d",RELATIONSHIP(1),S3Q25(1),S3Q1(1));
        impute ($(1),2);
      endif;
    endif;

    {All other marital statuses}
    If !($ in 1:6) THEN
       errmsg("P25-3 Marital status code Invalid,S3Q25 = %d",S3Q25) denom = DenomPop summary;
       F1F3();
       write ("P25-4 Marital status illegal %d, PN = %d",$,S3Q1);
       errmsg ("P25-4 Marital status illegal %d, PN = %d",$,S3Q1) denom = denomPop summary;
       impute ($,AMARITAL (AGEX,S3Q5));
    else
       AMARITAL (AGEX,S3Q5) = $;
    endif;
  endif;
endif; {FORMID}
```

*5. Age at first marriage*

346. The "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. 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.

347. 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.

348. *Age at marriage for never married persons should be blank.* 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 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.

349. *Ever married persons should have an entry.* 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.

[Southern Sudan]

```
if Q04_Age < 12 then {marital status must be blank if age < 12 }
else
  if Q24_Mar_Stat = 1 then
    {never married, so set age to not applicable }
    if Q25_Age_First_Mar <> notappl then
       errmsg("*P25-1* PN %d, age %d: Q25 (Age at first marriage) changed from %2d to not applicable",
         curocc(Person), Q04_Age, Q25_Age_First_Mar) denom = AgeGE12 summary;
       impute(Q25_Age_First_Mar, notappl);  {Must be blank for never married}
    endif;
  else
   {married or previously married }
    if Q24_Mar_Stat in 2:4  then
      {Age at first marriage is greater than 11 but less than or equal to current age}
      if Q25_Age_First_Mar > 11 and
         Q25_Age_First_Mar <= Q04_Age then
            AAGEFIRST (Q04_AGE) = Q25_AGE_FIRST_MAR;
      else
        {Age at first marriage is out of range}
        if not Q25_Age_First_Mar in 12:95 then
          errmsg("*P25-2* PN %d, age %d: Q25 (Age at first marriage) changed from %2d to 98 (Not reported)",
            curocc(Person), Q04_Age, Q25_Age_First_Mar) denom = AgeGE12 summary;
          impute(Q25_Age_First_Mar, AAGEFIRST (Q04_AGE));
         else
           {Age at first marriage is too low}
           if Q25_Age_First_Mar < 12 then
             errmsg("*P25-3* PN %d, age %d: Q25 (Age at first marriage) %2d imputed", curocc(Person),
               Q04_Age, Q25_Age_First_Mar) denom = AgeGE12 summary;
             impute(Q25_Age_First_Mar, AAGEFIRST (Q04_AGE));
           else
             {Age at first marriage is too high}
             if Q25_Age_First_Mar > Q04_Age then
               errmsg("*P25-4* PN %d, age %d: Q25 (Age at first marriage) %2d imputed)", curocc(Person),
                  Q04_Age, Q25_Age_First_Mar) denom = AgeGE12 summary;
               impute(Q25_Age_First_Mar, AAGEFIRST (Q04_AGE));
             endif;
           endif;
        endif;
      endif;
    endif;
  endif;
endif;


PROC YEARSSINCE1STMAR
{This variable holds the number of years since the first marriage}
if Q25_AGE_FIRST_MAR in 10:90 then
  $ = Q04_AGE - Q25_AGE_FIRST_MAR;
```

```
endif;
```

*6. Fertility: children ever born and children surviving*

350. "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.

351. 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.

352. *Fertility items collected.* In Principles and Recommendations for Population and Housing Censuses, recommends 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. Censuses and surveys collect information on fertility from all females, using a country-defined minimum age and sometimes a maximum age as well.

353. *General rules for the fertility edit.* 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:

a) 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;
b) 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;
c) 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;
d) The number of children ever born cannot be smaller than the entry in "children born in last 12 months";
e) 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;
f) 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.

354. *Relationship between children born and children surviving.* 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.

355. 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:

(a) Children ever born only
(b) Children ever born and children surviving (both sexes combined or separate sexes)

(c) Children ever born, children surviving and children who died (both sexes combined or separate sexes)

(d) Children ever born, children living at home, children living away and children who died (both sexes combined or separate sexes)

356. *Edit when only children ever born is reported.* 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.

```
array ACEB (99);

if SEX = 2 and AGE >= 12 then
   if CEB in 0:20 then
     if [age agrees with CEB]
      ACEB (AGE) = CEB;
     else
      Impute (CEB,ACEB (AGE));
     endif;
   else
      impute (CEB,ACEB (AGE));
   endif;
endif;
```

357. *Edit when children ever born and children surviving are reported.* If responses are present for both "children ever born" and "children surviving", the program needs to determine the following:

(a) Whether the items are internally consistent (is the number of children ever born equal to or greater than the number of children surviving);

(b) Whether at each item agrees with the age of the female;

(c) Whether "children ever born" agrees with "children born in the last year" (or last birth), if collected.

358. 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 is 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 27. Illustration of household with fertility information**

| *Person* | *Relation* | *Sex* | *Age* | *Children ever born* | *Children surviving* |
|---|---|---|---|---|---|
| 1 | Head of household | 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

359. 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.

360. 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.

361. 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.

362. 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. 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.

```
  if ((AGE in 15:16 and CEB <= 1) or
       (AGE in 17    and CEB <= 2) or
       (AGE in 18:19 and CEB <= 3) or
       (AGE in 20    and CEB <= 4) or
       (AGE in 21:22 and CEB <= 5) or
       (AGE in 23    and CEB <= 6) or
       (AGE in 24:25 and CEB <= 7) or
       (AGE in 26    and CEB <= 8) or
       (AGE in 27:28 and CEB <= 9) or
       (AGE in 29    and CEB <= 10) or
       (AGE in 30:31 and CEB <= 11) or
       (AGE in 32    and CEB <= 12) or
       (AGE in 33:34 and CEB <= 13) or
       (AGE in 35    and CEB <= 14) or
       (AGE in 36:37 and CEB <= 15) or
       (AGE in 38:98 and CEB <= 16)) THEN
     {Age and fertility check, do nothing}
   else
     errmsg ("*P30c-02* Female too young for number of children she's had, pn = [%2d],age = [%2d],
         Total children = [%2d]",PERSON_NUMBER,AGE,CEB) denom = denomPOP summary;
     {Need program here to impute the fertility variables – see below for specifications}
   endif;
endif;
```

363. 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) nor 4 children surviving (age difference of 11) would be acceptable.

**Figure 28. 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 |

364. 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.

365. *Edit when children ever born, children surviving, and children who died are reported.* "Children ever born" is the sum of "children surviving" and "children who died". Any inconsistency may be resolved as explained below.

366. (i) When all three items are reported. If all three pieces of information are present, the program needs to determine:

a) 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;
b) Whether each of the three items is consistent with the age of the female;
c) Whether the number of children ever born is consistent with number born in the last year (or the last birth), if collected.

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.

```
If CEB in 0:20 and CS = 0:20 and CD = 0:20 then
  If CEB = CS + CD then
    If [CEB and age agree]  then
      {see above for relationship between
        age and children ever born}
      If CEB >= BIRTHLASTYEAR then
        AFERT (AGE,1) = CEB;
        AFERT (AGE,2) = CS;
        AFERT (AGE,3) = CD;
        AFERT (AGE,4) = BIRTHLASTYEAR;
      Else
        [impute some or all items];
      Endif;
    Else
        [impute some or all items];
    Endif;
  Else
        [impute some or all items];
  Endif;
endif;
```

367. (ii) When two items are reported. 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.

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. See below for an example.

368. (iii) When one item is reported. 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. See below for an example.

369. (iv) When none of the items is reported. 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. See below for an example.

370. 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. 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.

```
If CEB in 0:20 and CS = 0:20 and CD = 0:20 then
  If CEB = CS + CD then
    If [CEB and age agree]  then
      {see above for relationship between
        age and children ever born}
      If CEB >= BIRTHLASTYEAR then
        AFERT (AGE,1) = CEB;
        AFERT (AGE,2) = CS;
        AFERT (AGE,3) = CD;
        AFERT (AGE,4) = BIRTHLASTYEAR;
      Else
        [impute some or all items];
      Endif;
    Else
        [impute some or all items];
    Endif;
  Else
        [impute some or all items];
  Endif;
Else
  If CEB in 0:20 then
    If CS in 0:20 then
      If CEB >= CS then
        Impute (CD,CEB-CS);
        [check for last births]
      Else
        [Impute the items]
      Endif;
    Else
      If CD in 0:20 then
        If CEB >= CD then
          Impute (CS,CEB-CD);
          [check for last births]
        Else
          [Impute the items]
        Endif;
      Else
          [Impute the items]
      Endif;
    Endif;
  Else
    If CS in 1:20 then
      [check for CD, then do as above]
    Endif;
  Endif;
Endif;
```

371. *Edit when children ever born, children living at home, children living away and children who died are reported.* The edit will differ somewhat when all four items are collected:

372. (i) When all four items are reported. If all four pieces of information are present, the program needs to determine:

a) 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;
b) whether each of the four items is consistent with the age of the female;
c) whether children ever born "is consistent with" children born in the last year (or last birth), if collected.

```
{ Children ever born not sum of the parts}

   if MCEB <> (MHH + MELSE + MDEAD) and MHH <> NOTAPPL and MELSE <> NOTAPPL and MDEAD <> NOTAPPL then
     errmsg ("*P30b-01* Total boys not boys pres + boys else + boys dead, PN = [%2d],Boys pres = [%2d],Boys else = [%2d],
            Boys dead = [%2d] Total = [%2d]",PERSON_NUMBER,MHH,MELSE,MDEAD,MCEB) denom = denomPop summary;
     impute (MCEB,(MHH+MELSE+MDEAD));
   endif;

   if FCEB <> (FHH + FELSE + FDEAD) and FHH <> NOTAPPL and FELSE <> NOTAPPL and FDEAD <> NOTAPPL then
     errmsg ("*P30b-02* Total girls not girls pres + girls else + girls dead, PN = [%2d],Girls pres = [%2d],Girls else = [%2d],
            Girls dead = [%2d] Total = [%2d]",PERSON_NUMBER,FHH,FELSE,FDEAD,FCEB) denom = denomPop summary;
     impute (FCEB,(FHH + FELSE + FDEAD));
   endif;
```

373. 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.

374. (ii) When three of the four items are reported. 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.

```
{Children present not known, but others are}

   if MHH = NOTAPPL and MCEB <> NOTAPPL and MDEAD <> NOTAPPL and MELSE <> NOTAPPL then
     errmsg ("*P30b-03* Boys and Girls present from other information") denom = DenomPOP summary;
     i = MCEB - MDEAD - MELSE;
     impute (MHH,i);
   endif;
```

375. (iii) When two of the four items are reported. 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.

376. 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.

377. 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.

378. 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).

379. 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.

380. 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 29.

**Figure 29. 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 |

381. 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 the best path to follow based on cultural circumstances.

382. (iv) When only one item is reported. When only one of the four items is 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.

383. (v) When none of the items is reported. 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.

384. 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.

385. 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.

386. *Special case of 5 or more items.* 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.

387. *Importance of a single donor source for all fertility items.* 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.

388. *Relationship of own children to children in the house and children surviving.* 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.

### *7. Fertility: date of birth of last child born alive and Births in the 12 months before the Census*

389. 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.

390. 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.

391. 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.

392. 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.

393. 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).

394. 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.

395. 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. Many countries choose to develop a recode for months since last birth. This variable can be divided up into those born in the 12 months before the census, 13 to 24 months before, and so forth.

396. 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.

397. 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 olds should be checked from mothers and the last births should be checked against the household listing.

398. 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 olds 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.

```
if AGE >= 12 then
  if CEB = 0 then
    if AGE_FIRST_BIRTH <> NOTAPPL then  {No births}
       errmsg("*P25-1* PN %d, age %d: Age at first birth) changed from %2d to not applicable",
         curocc(Person), AGE, AGE_FIRST_BIRTH) denom = AgeGE12 summary;
       impute (AGE_FIRST_BIRTH, notappl);  {Must be blank for no births}
     endif;
```

```
  else {some births}
    if CEB > 0  then {Age at first birth is greater than 11 but less than or equal to current age}
      if AGE_FIRST_BIRTH > 11 and
         AGE_FIRST_BIRTH <= AGE then
           AAGEFIRSTBIRTH (AGE) = AGE_FIRST_BIRTH;
      else {Age at first birth is out of range}
        if not AGE_FIRST_BIRTH in 12:95 then
          errmsg("*P25-2* PN %d, age %d: Age at first birth changed from %2d to 98 (Not reported)",
            curocc(Person), AGE, AGE_FIRST_BIRTH) denom = AgeGE12 summary;
          impute(AGE_FIRST_BIRTH, AAGEFIRSTBIRTH (AGE));
        else {Age at first birth is too low}
          if AGE_FIRST_BIRTH < 12 then
            errmsg("*P25-3* PN %d, age %d: Age at first birth) %2d imputed", curocc(Person),
             AGE, AGE_FIRST_BIRTH) denom = AgeGE12 summary;
            impute(AGE_FIRST_BIRTH, AAGEFIRSTBIRTH (AGE));
          else {Age at first marriage is too high}
            if AGE_FIRST_BIRTH > AGE then
              errmsg("*P25-4* PN %d, age %d: Age at first birth %2d imputed)", curocc(Person),
                AGE, AGE_FIRST_BIRTH) denom = AgeGE12 summary;
              impute(AGE_FIRST_BIRTH, AAGEFIRSTBIRTH (AGE));
            endif;
          endif;
        endif;
      endif;
    endif;
  endif;
endif;
```

*8. Fertility: age at first birth*

399. 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. 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.

400. 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 labour force status is not necessarily the same as status at her first birth) and level of educational attainment.

```
If SEX = MALE or Age < 15 or CEB = 0 then
  If AGE_AT_FIRST_BIRTH  <> NOTAPPL
    Impute (AGE_AT_FIRST_BIRTH,NOTAPPL);
  Endif;
else
  If AGE_AT_FIRST_BIRTH > AGE or
     AGE_AT_FIRST_BIRTH > AGE_AT_LAST_BIRTH  then
    Impute (AGE_AT_FIRST_BIRTH,AAGE_AT_FIRST_BIRTH (AGE));
  Else
    AAGE_AT_FIRST_BIRTH (AGE) = AGE_AT_FIRST_BIRTH;
  Endif;
Endif;
```

*9. Mortality*

401. 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.

402. 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 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.

*(a) Age and Sex of the Deceased*

403. The *Principles and Recommendations* suggest collecting name, age and sex, and day, month and year of death for persons who died in the year before the census. 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 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.

*(b) Cause of death*

404. 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 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.

```
If CAUSE_OF_DEATH = NOTAPPL then   {age and sex of deceased already edited}
   Impute (CAUSE_OF_DEATH,ACAUSE_OF_DEATH (DEATHAGEX,DEATHSEX));
Else
   ACAUSE_OF_DEATH (DEATHAGEX, DEATHSEX);
Endif;
```

*(c) Maternal mortality*

405. 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.

```
If DEATHSEX = 2 and DEATHAGE in 15:50 then
    If PREGNANCY_DEATH in 1:2 then
        APREG_DEATH (DEATHAGE) = PREGNANCY_DEATH;
    Else
        Impute (PREGNANCY_DEATH,APREG_DEATH (DEATHAGE));
    Endif;
Else
     If PREGNANCY_DEATH <> NOTAPPL then
        Impute (PREGNANCY_DEATH,NOTAPPL);
     Endif;
Endif;
```

*(d) Infant mortality*

406. Finally, the *Principles and Recommendations* 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” (rather than “deaths in the last months”) is used, then this item probably should not be used.

407. 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 having 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.

408. 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. (The pseudocode for this item appears in the section on fertility.)

*10. Maternal or paternal orphanhood (P5G) and mother’s line number*

409. 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. Because there is usually more than one surviving child who will respond on orphanhood status, it is necessary to devise questions to overcome duplications in respect of parents reported by siblings. For this purpose, two additional questions should be asked: (c) if the respondent is the oldest surviving child of his or her mother; and (d) if the respondent is the oldest surviving child of his or her father. Tabulations should be made in reference to the oldest surviving child only.

410. The edits for “mother living” and “mother’s line number” items are interrelated 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.

Malawi 2008

```
if AGE > 18 then
  {For people 18 years and over, this should be blank}
  if MOTHER <> NOTAPPL then
    errmsg ("*P14b-1* Mother in house not blank [%2d] for
      someone over 18 [%2d]", MOTHER, AGE)
      denom = orphcnt summary;
    impute (MOTHER, NOTAPPL);
  endif;
else
 {For people under 18 years, this should have a value}
  if MOTHER in 1:2 then
    AMOTHERINHOUSE (AGE5A,SEX) = MOTHER;
    if MOTHER = 1 then
      N97 = 0;
      if RELATIONSHIP = 3 then
        do varying N98 = 1 until N98 > totocc (POPULATION)
          if RELATIONSHIP (N98) in 1:2 and
            {child so looking for head or spouse}
            SEX (N98) = 2 then
            MPN = N98;
            break;
          endif;
        enddo;
      endif;
    endif;
  else
   {If mother not alive BUT in the house,
    make her not in the house}
    if MOTHERALIVE = 2 then
      errmsg ("*P14b-2* Mother in house not known [%2d] but
        mother dead [%2d]", MOTHER, MOTHERALIVE)
        denom = orphcnt summary;
      impute (MOTHER, 2);
    else
    {Mother is alive}
      N97 = 0;
      if RELATIONSHIP = 3 then
        do varying N98 = 1 until N98 > totocc (POPULATION)
          if RELATIONSHIP (N98) in 1:2 and
            {child so looking for head or spouse}
             SEX (N98) = 2 then
            N97 = 1;
          endif;
        enddo;
        if N97 = 1 then
          errmsg ("*P14b-3* Mother in house [%2d] from
            relationship [%2d]", MOTHER, RELATIONSHIP)
            denom = orphcnt summary;
          impute (MOTHER, 1);
          MPN = N98;
        else
          errmsg ("*P14b-4* Mother in house not known [%2d]
            and not female head/spouse, so imputed [%2d]",
            MOTHER, RELATIONSHIP) denom = orphcnt summary;
          impute (MOTHER, AMOTHERINHOUSE (AGE5A,SEX));
        endif;
      else
          errmsg ("*P14b-5* Mother in house not known [%2d]
            and not child, so imputed [%2d]", MOTHER,
            RELATIONSHIP) denom = orphcnt summary;
          impute (MOTHER, AMOTHERINHOUSE (AGE5A,SEX));
      endif;
    endif;
  endif;
endif;
```

411. 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. Note that, in the structure edits, if the head is not the first person, and then is moved to the first person, that mother's line number may need to be adjusted for one or more people.

```
SOUTHERN SUDAN
if Q13_FATHER in 1 then  {Father alive}
  if Q12_MOTHER in 1 then
    MOFA = 1;  {Father alive, mother alive}
  else
    MOFA = 2; {Father alive, mother dead}
  endif;
else   {Father dead}
  if Q12_MOTHER in 1 then
    MOFA = 3;  {Father dead, mother alive}
  else
    MOFA = 4; {Father dead, mother dead}
  endif;
endif;
```

## B. MIGRATION CHARACTERISTICS

412. The demographics of a country change over time as a result of natural increase (fertility and mortality) and net migration. Migration can be long-term migration (since birth) or short-term migration, measured by previous residence and duration or at a 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.

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

414. 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 is placed on international migration as well. 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. For international migration, country of birth and year of entry into the country are of concern.

### *1. Place of birth*

415. 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. 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.

416. *Relationship of entries for country of birth and years lived in district.* 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.

417. *Assigning "unknown" for invalid entries for birthplace*. 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.

```
if BIRTHPLACE in 1:999 then
```

```
else
  impute (BIRTHPLACE,999);
endif;
```

418. *Using static imputation for birthplace.* The entry for country of birth should be altered only if it is out of range. If the code for years in district is "always", the code for the country 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 district 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". 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.

```
If not BIRTHPLACE in [valid entries] then
  Impute (BIRTHPLACE, 99); {make unknown rather than imputing}
Endif;
```

419. *Using dynamic imputation for birthplace*. As before, the entry for country of birth should be altered only if it is out of range. If the entry for years in district 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.

```
If not BIRTHPLACE in [valid entries] then
  If DURATION = AGE   {or DURATION = ALWAYS}  then
    If BIRTHPLACE <> CURRENT_RESIDENCE then
      Impute (BIRTHPLACE,CURRENT_RESIDENCE);
    Endif;
  Else
    Count (BIRTHX where BIRTHPLACE = BIRTHPLACE (1));
    {See if everyone except this person has same birthplace}
    If BIRTHX = totocc (POP) – 1 then
      Impute (BIRTHPLACE, BIRTHPLACE (1));
    Else
      Impute (BIRTHPLACE, 99); {make either unknown or impute}
    Endif;
  Endif;
Else
  {if hot deck is used, update the hotdeck here}
Endif;
```

420. *Assigning birthplace when a person's mother is present.* 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 district 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 district, but this person's age is less than or equal to the number of years that the mother has lived in the district, 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 district, 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.

421. *Assigning birthplace for child of head.* 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 district 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 district 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 district 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 district, the program should assign the head of household's country of birth if it has a valid code for country of birth.

422. *Assigning birthplace for child, but not of head.* 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.

423. *Assigning birthplace for adult females with husband*. 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 district should be looked at. If the husband's "years in district" is coded "always", the woman's country of birth should be assigned "this country". If the husband's "years in district" is not "always", then the woman's country of birth should be imputed by age and sex.

424. *Assigning birthplace for adult females with no husband*. 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 district 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 district are less than or equal to the child's age, the program should assign her the child's country of birth.

425. *Assigning birthplace for males*. To obtain the birthplace for a male, the editing program can try to find his wife, or 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 district are less than or equal to hers, the wife's country of birth is assigned to the man's record. If the man's years in the district 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 district 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 district is greater than his child's age, the program should impute by age.

*2. Citizenship*

426. Information on citizenship should be collected so as 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. 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. 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 Standard Country or Area Codes for Statistical Use. 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.

427. *Citizenship edit*. 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.

428. *Relationship of ethnicity/race to citizenship*. 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.

429. *Relationship of naturalization to citizenship*. 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.

430. *Relationship of duration of residence to citizenship*. 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.

[Lesotho 2006]

```
PROC CITIZENSHIP

if RELATIONSHIP in 1 then
  if CITIZENSHIP in 10,15,20,15,30,35,40,45,50,55,60,65,70,75,80,85 then
    ACITIZENSHIP (AGE10,SEXI) = CITIZENSHIP;
  else
    do varying N01 = 1 while N01 <= totocc (POP_EDT)
      if CITIZENSHIP (N01) in 10,15,20,15,30,35,40,45,50,55,60,65,70,75,80,85 then
        errmsg ("*B14-1* Head's citizenship from other in the house, PN = [%2d], Citz = [%d]", PERSNUM (N01),CITIZENSHIP (N01))
           denom = denomPOP summary;
        F1F2();
        write ("*B14-1* Head's citizenship from other in the house, PN = [%2d], Citz = [%d]", PERSNUM (N01),CITIZENSHIP (N01));
        impute (CITIZENSHIP,CITIZENSHIP (N01));
        exit;
      endif;
    enddo;
        errmsg ("*B14-2* Head's citizenship from other head, PN = [%2d], Citz = [%d]", PERSNUM,CITIZENSHIP) denom = denomPOP summary;
        F1F2();
        write ("*B14-2* Head's citizenship from other head, PN = [%2d], Citz = [%d]", PERSNUM,CITIZENSHIP);
        impute (CITIZENSHIP,ACITIZENSHIP (AGE10,SEXI));
  endif;
else
  if not CITIZENSHIP in 10,15,20,15,30,35,40,45,50,55,60,65,70,75,80,85 then
    if CITIZENSHIP (HEADPT) in 10,15,20,15,30,35,40,45,50,55,60,65,70,75,80,85 then
      errmsg ("*B14-3* Non-head without citizenship, so obtained from head, PN = [%2d], Citz = [%d]",PERSNUM,CITIZENSHIP)
         denom = denomPOP summary;
      F1F2();
      write ("*B14-3* Non-head without citizenship, so obtained from head, PN = [%2d], Citz = [%d]",PERSNUM,CITIZENSHIP);
      impute (CITIZENSHIP,CITIZENSHIP(HEADPT));
    else
      do varying N01 = 1 while N01 <= totocc (POP_EDT)
        if CITIZENSHIP (N01) in 10,15,20,15,30,35,40,45,50,55,60,65,70,75,80,85 then
          impute (CITIZENSHIP,CITIZENSHIP (N01));
          exit;
        endif;
      enddo;
      if not CITIZENSHIP in 10,15,20,35,40,45,50,55,60,65,70,75,80,85 then
        impute (CITIZENSHIP,10);
      endif;
    endif;
  endif;
endif;
```

*3. Duration of residence*

431. 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.

432. *Edit for duration of residence*. 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".

433. *De facto/de jure residence and duration*. 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 at their residence on census night. 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.

434. *Relation of age to duration of residence.* 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.

435. *Relation of birthplace to duration.* 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.

436. *For persons who have always lived here*. 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.

437. *Person's duration from mother's duration*. 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 the program can find the person's mother, this information can help assign 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.

438. *Person's duration from child's duration.* 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. 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.

439. *Person's duration when no other information available*. 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.

*4. Place of previous residence*

440. 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.

441. *Previous residence edit*. 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.

442. *Previous residence when boundaries have changed*. 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 continent of residence, the second digit the region within the continent and the third digit the country within the region.

443. *When person has not moved since birth*. 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.

444. *Use of other persons in unit*. 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.

445. *No appropriate other person for previous residence*. 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 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.

*5. Place of residence at a specified date in the past*

446. 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.

447. Some countries, however, may have to use a different time reference than either one year or five years preceding the census because both of 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.

448. "Place of residence at a specified date in the past" is similar to 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.

*6. Year of Arrival*

449. 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.

450. *Relation of age to year of arrival*. 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.

451. 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.

452. *Relation of birthplace to year of arrival*. 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.

453. *For persons who have always lived here*. If the response for the number of years since arrival for a person has 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.

454. *Person's year of arrival from mother's year of arrival*. 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 the program can find the mother's record, it can help assign 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.

455. *Child's year of arrival from head's year of arrival***.** 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.

456. *Person's year of arrival when no other information available*. 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.

*7. Relationship of Duration of Residence to Year of Arrival*

457. 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.

458. 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.

459. 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.

460. 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.

*7. Usual Residence*

461. When countries collect *de jure* census data, the enumeration is by "usual residence", compared to *de facto* collection, where the enumeration takes place using current residence. Hence, countries taking *de jure* censuses should not be asking a separate item on usual residence. 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 current residence, so missing information can be filled directly. But, when the data show movement, the situation becomes more complicated. Usually, countries assume that when this item is left blank, the usual residence and the current residence are the same, and the enumerator and/or respondent simply left the information out.

462. However, when the data show evidence, by duration of residence, or year of arrival, or some evidence of changing residence, then the statistical staff may want to try to develop methods of obtaining best guesses 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. 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.

### C. SOCIAL CHARACTERISTICS

463. Social characteristics vary from country to country, but are generally items that describe various aspects of socio-cultural conditions in the country. Educational items, including literacy, school attendance and educational attainment as well as field of education and educational qualifications, can be classified according to the categories of the 1997 revision of the International Standard Classification of Education (ISCED), developed by the United Nations Educational, Scientific and Cultural Organization (UNESCO).

*1. Ability to read and write (literacy)*

464. Data on literacy should be collected for all persons 10 years of age and over. In some of countries, however, certain persons between 10 and 14 years of age may be becoming literate through schooling so 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.

465. 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.

466. The edit for literacy first checks the highest grade completed; if highest grade has an entry of "literate" based on specifications, the code for "yes" should be assigned. 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."

```
Lesotho 2006
if LITERACY = 1 then    {Read and Write with ease}
  if EDUCATION in 7:27 then
    impute (LITERACY2,1);   {People who read and write with ease and have completed Standard 4}
  else
    if EDUCATION in 4:6 then
      impute (LITERACY2,2);  {People who read and write with ease and completed standards 4 to 6}
    else
      impute (LITERACY2,3);  {Not literate}
    endif;
  endif;
else
  impute (LITERACY2,4);  {not literate because cannot read and write}
endif;
```

*2. School attendance*

467. 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 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 programmes organized for adults in productive and service enterprises, community-based organizations and other non-educational institutions, the age range may be adjusted as appropriate.

468. *School attendance edit*. 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, 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. Entries for all other persons must be changed. 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 secondary or tertiary schooling 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.

469. *Full-time or part-time enrolment.* 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.

470. *Consistency between school attendance and economic activity*. Consistency edits with other major items, such as major economic activity, should be performed first. If attending school is one of the entries for major economic activity, and a person reported his or her major activity as going to school, the code for "yes" should be assigned to school

attendance and major economic activity should be “student”. That is, the responses should be consistent. In all other cases, any valid response should be accepted.

```
If ECONOMIC_ACTIVITY = STUDENT then
  If SCHOOL_ATTENDANCE <> YES then
    Impute (SCHOOL_ATTENDANCE,YES);
  Endif;
Else
  If SCHOOL_ATTENDANCE = YES then
    [Impute ECONOMIC_ACTIVITY = Student when country does not have students who also work]
    [Otherwise subject matter specialists must make a decision.]
  Endif;
Endif;
```

471. *Assignment for invalid or inconsistent entries for school attendance*. 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. 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).

*3. Educational attainment (highest grade or level completed)*

472. *Edit for educational attainment*. 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. See section of “Derived Variables” for suggestions for a recode for “current grade” based on school attendance and highest grade attended.

473. *Minimum age for educational attainment*. 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.

474. *Relationship of age to educational attainment.* 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 a large number of responses that are truly inconsistent. Therefore, an entry can be assigned when the original entry is out of range or inconsistent with age. 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. 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.

Lesotho 2006

```
if AGE < 2 then
  if SCHOOL <> NOTAPPL then
    errmsg ("*P13 -02* Age less than 2 but
      Schooling reported") denom = denomPOP summary;
    impute (SCHOOL,NOTAPPL);
  endif;
  if EDATTNMT <> NOTAPPL then
    errmsg ("*P13 -03* Age less than 2 but EDATTNMT
      reported") denom = denomPOP summary;
    impute (EDATTNMT,NOTAPPL);
  endif;
  exit;
endif;

if AGE >= 2 then
  SCHOOLAGE = AGE;
  if SCHOOLAGE > 60 then SCHOOLAGE = 60; endif;
  {Preliminary check for SCHOOL attendance}
    if SCHOOL in 1:3 then
      ASCHOOL (SCHOOLAGE) = SCHOOL;
    else
      {SCHOOL attendance is not 1 to 3}
      errmsg ("*P13 -04* Schooling not reported, so imputed")
        denom = denomPOP summary;
      if AGE in 2:4 then impute (SCHOOL,1); endif;
      if AGE in 5:20 then impute (SCHOOL,2); endif;
      if AGE in 21:59 then impute (SCHOOL,3);
         endif;
      if AGE >= 60 then impute (SCHOOL,1); endif;
  endif;
 {Old people 'attending' made 'left SCHOOL'}
 if SCHOOL = 2 then
   if ((EDATTNMT=0 and AGE>6) or
     {Pre-SCHOOL}
       (EDATTNMT = 1 and AGE > 13) or {Standard 1}
       (EDATTNMT = 2 and AGE > 14) or
       (EDATTNMT = 3 and AGE > 15) or
       (EDATTNMT = 4 and AGE > 16) or
       (EDATTNMT = 5 and AGE > 17) or
       (EDATTNMT = 6 and AGE > 18) or
       (EDATTNMT = 7 and AGE > 19) or {Standard 7}
       (EDATTNMT = 8 and AGE > 20) or
       (EDATTNMT = 9 and AGE > 21) or
       (EDATTNMT = 10 and AGE > 22) or
       (EDATTNMT = 11 and AGE > 20) or {Form 1}
       (EDATTNMT = 12 and AGE > 21) or
```

```
        (EDATTNMT = 13 and AGE > 22) or
        (EDATTNMT = 14 and AGE > 23) or
        (EDATTNMT = 15 and AGE > 24) or {Form 5}
        (EDATTNMT = 29 and AGE > 6)) then {None}
      errmsg ("*P13 -6a* Currently attending
        SCHOOL [%2d] but 60 or more [%2d]",
        SCHOOL,AGE) denom = denomPOP summary;
      impute (SCHOOL,3);
    endif;
  endif;

  if AGE >= 60 then
    if SCHOOL = 2 then
      errmsg ("*P13 -6b* Currently attending
        SCHOOL [%2d] but 60 or more [%2d]",
        SCHOOL,AGE) denom = denomPOP summary;
      impute (SCHOOL,3);
    endif;
  endif;

  {If they never attended SCHOOL,
   make them have no EDATTNMT}
  if SCHOOL = 1 then
    if EDATTNMT <> 29 then
      errmsg ("*P13 -07* person never attended SCHOOL, made no
EDATTNMT") denom = denomPOP summary;
      impute (EDATTNMT,29);
    endif;
    exit;
  endif;

 {The following for legal EDATTNMTal attainment}
 if EDATTNMT in 0,1:7,11:15,20:30 then
   {Currently attending}
   if SCHOOL = 2 then
     if ((EDATTNMT = 0 and AGE in 2:5) or  {Pre-SCHOOL}
        (EDATTNMT = 1 and AGE in 6:10) or {Standard 1}
        (EDATTNMT = 2 and AGE in 7:11) or
        (EDATTNMT = 3 and AGE in 8:12) or
        (EDATTNMT = 4 and AGE in 9:13) or
        (EDATTNMT = 5 and AGE in 10:14) or
        (EDATTNMT = 6 and AGE in 11:15) or
        (EDATTNMT = 7 and AGE in 12:16) or{Standard 7}
        (EDATTNMT = 8 and AGE in 13:17) or
        (EDATTNMT = 9 and AGE in 14:18) or
        (EDATTNMT = 10 and AGE in 15:16) or
        (EDATTNMT = 11 and AGE in 13:17) or {Form 1}
        (EDATTNMT = 12 and AGE in 14:18) or
        (EDATTNMT = 13 and AGE in 15:19) or
        (EDATTNMT = 14 and AGE in 16:20) or
        (EDATTNMT = 15 and AGE in 17:21) or {Form 5}
        (EDATTNMT in 20:25 and AGE in 18:59) or {Certificates}
        (EDATTNMT = 26 and AGE in 18:59) or    {Graduate}
        (EDATTNMT = 27 and AGE in 23:59) or    {Post-graduate}
        (EDATTNMT = 28 and AGE in 6:59) or    {Non formal}
        (EDATTNMT = 29 and AGE in 5:18) or    {None}
        (EDATTNMT = 30 and AGE in 5:50)) then  {Other}

        AEDATTNMT2 (AGE) = EDATTNMT;
      else
        errmsg ("*P13 -08* For currently enrolled, EDATTNMT
[%2d] and age [%2d] disagree",EDATTNMT,AGE) denom = denomPOP
summary;
        impute (EDATTNMT,AEDATTNMT2 (AGE));
      endif;

      {Not currently attending ro never attended}
    else
      if ((EDATTNMT = 0 and AGE >= 2) or
          (EDATTNMT = 1 and AGE >= 6) or  {Standard 1}
          (EDATTNMT = 2 and AGE >= 7) or
          (EDATTNMT = 3 and AGE >= 8) or
          (EDATTNMT = 4 and AGE >= 9) or
          (EDATTNMT = 5 and AGE >= 10) or
          (EDATTNMT = 6 and AGE >= 11) or
          (EDATTNMT = 7 and AGE >= 12) or
          (EDATTNMT = 11 and AGE >= 13) or   {Form 1}
          (EDATTNMT = 12 and AGE >= 14) or
          (EDATTNMT = 13 and AGE >= 15) or
          (EDATTNMT = 14 and AGE >= 16) or
          (EDATTNMT = 15 and AGE >= 17) or
          (EDATTNMT   in   20:25   and   AGE   >=   18)   or
{Certificate/diploma}
          (EDATTNMT = 26 and AGE >= 18) or      {Graduate}
          (EDATTNMT = 27 and AGE >= 23) or            {Post-
graduate}
          (EDATTNMT in 28:30 and AGE >= 2)) then
        if AGE in 1:98 then
         AEDATTNMT1 (AGE) = EDATTNMT;
        endif;
      else
        errmsg ("*P13 -09* For left SCHOOL, EDATTNMT [%2d] and
age [%2d] disagree",EDATTNMT,AGE) denom = denomPOP summary;
        if AGE in 1:98 then
          impute (EDATTNMT,AEDATTNMT1 (AGE));
        else
          impute (EDATTNMT,1);
        endif;
      endif;
    endif;
  else {EDATTNMTal attainment reported is illegal}
    errmsg   ("*P13   -10*   Schooling   reported   but   EDATTNMT
illegal") denom = denomPOP summary;
    if AGE in 1:98 then
      errmsg   ("*P13   -11*   Schooling   reported   but   EDATTNMT
illegal") denom = denomPOP summary;
      impute (EDATTNMT,AEDATTNMT (AGE));
      if not EDATTNMT in 0:30 then
        if AGE > 30 then
          impute (EDATTNMT,3);
        endif;
      endif;
    else
      impute (EDATTNMT,1);
    endif;
  endif;
endif;
```

### *4. Field of education and educational qualifications*

475. 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 labour market. It is equally important for planning and regulating the production capacities of different levels, types and branches of educational institutions and training programmes. 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.

```
{.*************************************************************************
. **************        Edit P14b - Field of Education          ***************
. *************************************************************************
.}
{14b      If high school graduate or above, What was ...'s major in academic college or vocational school?
   01 Basic prog            42 Life science          72 Health
   08 Literacy/numeracy     44 Physical science      76 Social service
   09 Personal dvlopmt      46 Math/stats            81 Personal service
   14 Education training    48 Computing             84 Transport service
   21 Arts                  52 Engineering           85 Environmental protect
   22 Humanities            54 Manufacturing         86 Security service
   31 Social/behavior sci   58 Architect/bldg        99 Unknown
   32 Journalism/Info       62 Agrc/fishing
   34 Business/administ     64 Veterinary}
```

```
if EDUC_ATTAINMENT < 20 then
  if FIELD_OF_EDUCATION <> NOTAPPL then
    errmsg ("*P14b-01* Field of education [%2d] for less than high school grad [%2d],
      PN = [%2d]",FIELD_OF_EDUCATION,EDUC_ATTAINMENT,PERSON_NUMBER) denom = denomPOP summary;
    impute (FIELD_OF_EDUCATION,NOTAPPL);
  endif;
else
  if FIELD_OF_EDUCATION in 1,8:9,14,21:22,31,32,34,42,44,46,48,52,54,58,62,64,72,76,81,84,85,86,99 then
    AFIELD_OF_EDUCATION (AGE10,SEX) = FIELD_OF_EDUCATION;
  else
    errmsg ("*P14b-02* Field of education [%2d] imputed, education = [%2d], PN = [%2d]",
      FIELD_OF_EDUCATION,EDUC_ATTAINMENT,PERSON_NUMBER) denom = denomPOP summary;
    impute (FIELD_OF_EDUCATION,AFIELD_OF_EDUCATION (AGE10,SEX));
  endif;
endif;
```

*5. Religion*

476. For census purposes, religion may be defined as either (a) religious or spiritual belief of preference, regardless of whether or not an organized group represents this belief, 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.

477. *Religion edit*. Religion is one of the variables fitting the standard edit examples introduced in chapter II.2. 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 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:

478. *No religion for head of household, but religion present for someone else in the unit.* The first step is to determine if anyone else in the housing unit has a valid religion, and assign the first valid religion.

479. *No religion for head, or for anyone else in unit.* 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 of household with similar characteristics including age and sex as well as language, birthplace and other variables as appropriate, considering the circumstances.

480. *For person other than head, without religion.* If this person is not the head of household and reports no religion, assign the head's religion.

Malawi 2008

```
if RELIGION in 1:4 then
    if RELATIONSHIP = 1 then
      ARELIGION (AGE5A,SEX) = RELIGION;
      HEADRELIG = RELIGION;
    endif;
  else
    if RELATIONSHIP = 1 then
      if totocc (POPULATION) = 1 then
        errmsg("*P09-0* Religion of head [%2d] imputed for
          single person house", RELIGION)
          denom = POPCNT summary;
        impute (RELIGION, ARELIGION (AGE5A,SEX));
      else
      N98 = 0;
      do varying N01 = 1 until N01 > totocc (POPULATION)
        if RELIGION (N01) in 1:4 then
          N98 = N01;
          break;
        endif;
      enddo;
      if N98 = 0 then
        errmsg("*P09-1* Religion of head [%2d] imputed",
        RELIGION) denom = POPCNT summary;
        impute (RELIGION, ARELIGION (AGE5A,SEX));
        HEADRELIG = RELIGION;
      else
        errmsg ("*P09-2* Religion of head from other person
          [%2d] in house", RELIGION) denom = POPCNT summary;
        impute (RELIGION, RELIGION (N98));
        HEADRELIG = RELIGION;
      endif;
    endif;
  else
    if HEADRELIG in 1:4 then
      errmsg ("*P09-3* Religion of other member [%2d] from
        head [%2d]", RELIGION,HEADRELIG)
        denom = POPCNT summary;
      impute (RELIGION, HEADRELIG);
    else
      errmsg ("*P09-4* Religion of other member [%2d]
        imputed for visiting head [%2d]", RELIGION,
        HEADRELIG) denom = POPCNT summary;
      impute (RELIGION, ARELIGION (AGE5A,SEX));
    endif;
  endif;
endif;
```

*6. Language*

481. Three types of language data can be collected in censuses, namely:

- Mother tongue, defined as the language usually spoken in the individual's home in his or her early

childhood;

- 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.

482. *Language edit*. Of the three 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.

483. *Language edits: head of household*. 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 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 not imputing or impute a language from the most recent head of household with similar characteristics, including age and sex as well as other language variables, birthplace and other variables as appropriate under these circumstances.

484. *Language edits: persons other than head of household*. If the person is not the head of household and the language is invalid, then assign the head of household's language.

485. *Language edits: use of ethnic origin or birthplace*. 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.

486. *Language edit: Mother tongue*. If the mother tongue is unknown, but 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 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.

```
if RELATIONSHIP in 1 then
  if MOTHER_TONGUE in 1:4 then
    AMOTHER_TONGUE (AGE10,SEX) = MOTHER_TONGUE;
  else
    do varying N01 = 1 while N01 <= totocc (POPULATION_EDT)
      if MOTHER_TONGUE (N01) in 1:4 then
        errmsg ("*P17a-01* Head's MOTHER_TONGUE from other in
           the house, PN = [%2d], Citz = [%d]", PN (N01),
           MOTHER_TONGUE (N01)) denom = denomPOP summary;
        impute (MOTHER_TONGUE,MOTHER_TONGUE (N01));
        exit;
      endif;
    enddo;
        errmsg ("*P17a-02* Head's MOTHER_TONGUE from other
           head, PN = [%2d], Citz = [%d]", PN,
           MOTHER_TONGUE) denom = denomPOP summary;
        impute (MOTHER_TONGUE,AMOTHER_TONGUE (AGE10,SEX));
  endif;
else
```

```
  if not MOTHER_TONGUE in 1:4 then
    if MOTHER_TONGUE (HEADPT) in 1:4 then
      errmsg ("*P17a-03* Non-head without MOTHER_TONGUE, so
         obtained from head, PN = [%2d], Citz = [%d]",
         PN,MOTHER_TONGUE) denom = denomPOP summary;
      impute (MOTHER_TONGUE,MOTHER_TONGUE(HEADPT));
    else
      do varying N01 = 1 while N01 <= totocc (POPULATION_EDT)
        if MOTHER_TONGUE (N01) in 1:4 then
          impute (MOTHER_TONGUE,MOTHER_TONGUE (N01));
          exit;
        endif;
      enddo;
      if not MOTHER_TONGUE in 1:4 then
        impute (MOTHER_TONGUE,1);
      endif;
    endif;
  endif;
endif;
```

487. *Language edits: Ability to speak a designated language*. The ability to speak a designated language is a third variable fitting the examples presented in chapter II. Again, the head 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.

```
if SPEAKING_ENGLISH in 1:2 then
  AENGLISH (AGEX,SEX) = SPEAKING_ENGLISH;
else
  if RELATIONSHIP = HEAD then
```

```
    count (ENGSPEAKERS where SPEAKING_ENGLISH = YES);
    if ENGSPEAKERS > 1 then
      impute (ENGLISH_SPEAKING,YES);
    else
      impute (ENGLISH_SPEAKING,AENGLISH (AGEX,SEX));
    endif;
  else
    impute (ENGLISH_SPEAKING,ENGLISH_SPEAKING (HEADPTR));
   endif;
endif;
```

*7. Ethnicity and Indigenous peoples*

488. 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, colour, 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.

489. The *Principles and Recommendations* 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 so that “indigenous” is identified uniquely. This identification will allow appropriate edits and tabulations to 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.

490. *Ethnicity edit.* 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.

491. *Ethnicity edit: for head of household*. Ethnic origin also fits the example introduced in chapter II. Editing teams should follow consider the scheme already described for the other social items. The head of household should be edited first. If the person with an invalid or unknown ethnic origin is the head of household, 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 with similar characteristics (age and sex as well as language, birthplace and other variables that may be appropriate, considering the circumstances).

492. *Ethnicity edit: persons other than head of household.* If the person is not the head of household and ethnic origin is invalid, then assign the head of household’s ethnic origin.

```
if ETHNICITY in 1:9 then
    if RELATIONSHIP = 1 then
      AETHNICITY (AGE5A,SEX) = ETHNICITY;
      HEADETHNIC = ETHNICITY;
    endif;
  else
    if RELATIONSHIP = 1 then
      if totocc (POPULATION) = 1 then
        impute (ETHNICITY, AETHNICITY (AGE5A,SEX));
      else
        N98 = 0;
        do varying N01 = 1 until N01 > totocc (POPULATION)
          if ETHNICITY (N01) in 1:9 then
            N98 = N01;
            break;
          endif;
        enddo;
        if N98 = 0 then
          impute (ETHNICITY, AETHNICITY (AGE5A,SEX));
          HEADETHNIC = ETHNICITY;
        else
          impute (ETHNICITY, ETHNICITY (N98));
          HEADETHNIC = ETHNICITY;
        endif;
      endif;
    else
      if HEADETHNIC in 1:9 then
        impute (ETHNICITY, HEADETHNIC);
      else
        impute (ETHNICITY, AETHNICITY (AGE5A,SEX));
      endif;
    endif;
endif;
```

493. *Ethnicity edit: use of language and birthplace.* 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. For example, if ethnic origin is unknown, but 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 particular circumstances, including geography (such as urban or rural residence), age or other items.

```
{Example for Filipino ethnicity}
If not ETHNICITY in 1:9 then
  If LANGUAGE = FILIPINO or BIRTHPLACE = PHILIPPINES then
    Impute (ETHNICITY,FILIPINO);
  Endif;
Endif;
```

*8. Disability*

494. Disability status characterizes the population into those with and without a disability. A person with a disability should be defined as a person who is at greater risk than the general population in experiencing restrictions in performing specific tasks or participating in role activities. The UN recommends inclusion of four domains in assessing disability status: (1) walking, (2) seeing, (3) hearing, and (4) cognition.

495. The question used to identify persons with disabilities should list broad categories of disabilities so that each person can check the presence or absence of each type of disability. Use of the following list of disabilities based on the International Classification for Impairments, Disabilities and Handicaps (ICIDH) 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.

496. A census format offers only limited space and time for questions on any one 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. Based on the World Programme of Action concerning Disabled Persons, three major classes of purposes for measuring disability need to be considered in questionnaire formation: (1) to provide services, including specific programs, (2) to monitor the level of functioning in the population, and (3) to assess equalization of opportunities.

497. *Disability census questions*. The UN recommends asking about each domain separately. The language should be clear, unambiguous and simple, without negative terms, and should be addressed of each household member separately.

498. *Disability edit*. When persons do 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 they were unwilling 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. Alternatively, the specialists may decide that only those responses specifying that a disability is present should be accepted, and that any invalid response should be "no disability". In the latter case, dynamic imputation would not be used.

```
if not WALKING in 1:2 then impute (WALKING,2); endif;
if not SEEING in 1:2 then impute (SEEING,2); endif;
if not HEARING in 1:2 then impute (HEARNG,2); endif;
if not COGNITION in 1:2 then impute (COGNITION,2); endif;
```

or

```
if WALKING in 1:2 then
    AWALKING (AGEX,SEX) = WALKING;
Else
    Impute (WALKING,AWALKING (AGEX,SEX);
Endif;
```

499. *Multiple disabilities*. 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. See the section on derived variables below for when multiple disabilities must be derived on the basis of the collected individual disabilities.

500. *Cause of disability edit.* 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.

```
If DISABILITY = YES then
    If DISABILITY_CAUSE in 1:9 then
       ADISABILIY_CAUSE (AGEX,SEX) = DISABILITY_CAUSE;
    Else
       Impute (DISABILITY_CAUSE,(ADISABILITY_CAUSE (AGEX,SEX));
    Endif;
Endif;
```

### II.4.4. ECONOMIC CHARACTERISTICS

501. 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.

502. Tabulations 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 endeavour 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. This calls for close attention when measuring the economically active population. A maximum age limit for measurement of the economically active population should normally not be used, as a considerable number of elderly persons beyond retirement age may be engaged in economic activities, either regularly or occasionally.

503. Each country must determine a minimum age for participation in economic activity. Countries interested in collecting data on child labour may need to choose a low minimum age, but must remember that some noise will occur when children who are not in the labour force are erroneously enumerated as being in the labour 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 all tabulations, any responses that may have been entered for children under age X should be eliminated.

*1. Activity status*

504. Economic activity status is made up of several economic variables, some of which are described below. These variables are satisfactory for data collection, but may need to be re-categorized for data processing and analysis. "Current activity status" is the relationship of a person to economic activity, based on a brief reference period such as one week or one day. 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.

505. According to the United Nations the **employed** comprise all persons above a specified age who, during a short reference period of either one week or one day, were in one of the following categories:

(a) *Paid employment*. Paid employment is of two types:

(1) At work: persons who during the reference period performed some work for wage or salary, in cash or in kind;

(2) 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*. Self employment is also of two types:

(1) At work: persons who during the reference period performed some work for profit or family gain, in cash or in kind;

(2) 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.

506. The population that is "not currently active" comprises all persons not classified either as employed or as unemployed. It is recommended that the "not usually active" population should be classified into the following four groups:

(1) *Students*: persons of either sex, not classified as "usually economically active", who attended any regular educational institution, public or private, for systematic instruction at any level of education during the reference period;

(2) *Homemakers*: persons of either sex, not classified as "usually economically active", who were engaged in household duties in their own home, for example, housewives and other relatives responsible for the care of the home and children (domestic employees, working for pay, however, are classified as "economically active");

(3) *Pension or capital income recipients*: persons of either sex, not classified as "usually economically active", who receive income from property or investments, interests, rents, royalties or pensions from former activities, and who cannot be classified as students or homemakers;

(4) *Others*: persons of either sex, not classified as "usually economically active", who are receiving public aid or private support, and all other persons not falling into any of the above categories.

507. *Categories related to activity status.* The following categories are related to activity status:

508. (i) Unemployed population. The **unemployed** population comprises, according to the United Nations, all persons above a specified age who, during the reference period, met the following conditions:

~~(1)~~ Without work: they were not in paid employment or self-employment;

(2) Currently available for work: they were available for paid employment or self-employment during the reference period;

(3) 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 licences. It is useful to distinguish first-time job-seekers from other job-seekers in the classification of the unemployed.

509. In general, to be classified as unemployed, a person must satisfy all three of the above criteria. However, in situations where the conventional means of seeking work are of limited relevance, where the labour market is largely unorganized or of limited scope, where labour absorption is, at the time, inadequate, or where the labour force is largely self-employed, the standard definition of unemployment may be applied by relaxing the criterion "seeking work". Such a relaxation is aimed primarily at those developing countries where the criterion does not capture the extent of unemployment in its totality. With this relaxation of the criterion of "seeking work", which permits in extreme cases the criterion's complete suppression, the two basic criteria that remain applicable are "without work" and "currently available for work".

510. 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.

511. (ii) Looking for work. 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.

512. (iii) Not currently active. The population that is "not currently active" or persons that are "not in the labour force", comprise all persons who were neither "employed" nor "unemployed" during the short reference period used to measure current activity. They may, according to their reasons for not being "currently active", be classified in any of the following groups:

(1) Attending an educational institution;
(2) Performing household duties;
(3) Living on a pension or capital income;
(4) Not worthy for other reasons, including disability or impairment.

The edits for "not currently active" have been incorporated into the above edits for economic activity.

213. (iv) Why not looking for work. 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.

514. *Editing for economic activity status*. Economic activity 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 labour force.

515. For this variable, the first four possibilities are for persons who are economically active, and the second four categories are for persons who are not economically active. Persons who are "at work" (categories 1 and 3) are employed, those who are "not at work" (categories 2 and 4) may be unemployed or not in the labour force, depending on the responses to the unemployment items ("on layoff"; "looking for work"; "year last worked").

516. (i) Employed persons. If one of the categories for economically active persons is selected (categories 1 to 4), the variables for time worked, occupation, industry, economic activity status, 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.

517. (ii) Economic activity of unemployed persons. If category one of the categories for persons who are not economically active (5 to 9) is selected, the variables for "on layoff", "looking for work" and "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 BLANK.

518. (iii) Economic activity of students and retired persons. 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.

519. (iv) When economic activity is not valid and employed variables are reported. If the entry for economic activity is not valid, and if some of the variables for time worked, occupation, industry and workplace are reported, the respondent's economic activity should be coded with a value from 1 to 4. An imputation matrix will probably be needed to select the appropriate response.

520. (v) When economic activity is not valid and the unemployed variables are reported. If any of the variables for "on layoff", "looking for work" and "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.

521. (vi) When economic activity is not valid and none of the economic variables are reported. If no response appears for any of the economic activity items, the subject-matter specialists will probably want to use imputation matrices to determine the most appropriate response and then impute the other economic items.

Sierra Leone 2004

```
{Economic activity}
if AGE < 10 then
  if $ <> NOTAPPL then
    errmsg ("*P23-1* Too young for economic activity") denom = popdenom summary;
    impute ($,NOTAPPL);
  endif;
else
  if ECONOMIC_ACTIVITY in 1:9 then
    if ECONOMIC_ACTIVITY in 1:3 then
      if not OCCUPATION in 0:9 then
        errmsg ("*23-2* Occupation imputed because working") denom = popdenom summary;
        impute (OCCUPATION,AOCCUPATION (AGE10,SEXY));
      endif;
      if not INDUSTRY in 1:21 then
        errmsg ("*P23-3* Industry imputed because working") denom = popdenom summary;
        impute (INDUSTRY,AINDUSTRY (AGE10,SEXY));
      endif;
    else
      if OCCUPATION in 0:9 or INUDSTRY in 1:21 then
        errmsg ("*P23-4* Occupation or industry present so working imputed") denom = popdenom summary;
        impute (ECONOMIC_ACTIVITY,1);
        if not OCCUPATION in 0:9 then
          errmsg ("*P23-5* Occupation imputed because working") denom = popdenom summary;
          impute (OCCUPATION,AOCCUPATION (AGE10,SEXY));
        endif;
        if not INDUSTRY in 1:21 then
          errmsg ("*P23-6* Industry imputed because working") denom = popdenom summary;
          impute (INDUSTRY,AINDUSTRY (AGE10,SEXY));
        endif;
      else
        if OCCUPATION <> NOTAPPL then
          errmsg ("*P23-7* Occupation not applicable because not working") denom = popdenom summary;
          impute (OCCUPATION,NOTAPPL);
        endif;
        if INDUSTRY <> NOTAPPL then
          errmsg ("*P23-8* Industry not applicable not working") denom = popdenom summary;
          impute (INDUSTRY,NOTAPPL);
        endif;
      endif;
    endif;
  else
    if OCCUPATION in 0:9 or P25 in 1:21 then
        errmsg ("*P23-9* Occupation or industry present so working imputed") denom = popdenom summary;
        impute (ECONOMIC_ACTIVITY,1);
        if not OCCUPATION in 0:9 then
          errmsg ("*P23-10* Occupation imputed because working") denom = popdenom summary;
          impute (OCCUPATION,AOCCUPATION (AGE10,SEXY));
        endif;
        if not INDUSTRY in 1:21 then
          errmsg ("*P23-11* Industry imputed because working") denom = popdenom summary;
          impute (INDUSTRY,AINDUSTRY (AGE10,SEXY));
        endif;
    else
        if OCCUPATION <> NOTAPPL then
          errmsg ("*P23-12* Occupation not applicable because not working") denom = popdenom summary;
          impute (OCCUPATION,NOTAPPL);
        endif;
        if INDUSTRY <> NOTAPPL then
          errmsg ("*P23-13* Industry not applicable not working") denom = popdenom summary;
          impute (INDUSTRY,NOTAPPL);
        endif;
    endif;
  endif;
endif;
```

*2. Time worked.*

522. Time worked is the total time actually spent producing goods and services, within regular working hours and as overtime, during the reference period adopted for economic activity in the census. It is recommended that if the reference period is short, for example, the week preceding the census, time worked should be measured in hours. In this case, time worked may be measured by requesting separate information for each day of the week. If the reference period is long, for example, the 12 months preceding the census, time worked should be measured in units of weeks, or in days where feasible, or in terms of larger time intervals. 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. This item should be edited only for persons whose response for economic activity was "employed, at work" or "self-employed, at work". 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. If dynamic imputation is used, the minimal variables for the imputation matrix include age groups and sex, but other variables such as educational attainment, occupation or industry major categories can also be used.

```
 if HOURS_WORK in 1:99 then
   AHOURS_WORK (AGE10,SEX) = HOURS_WORK;
 else
   errmsg ("*P19 -05* Work last week [%2d] but hours worked [%2d] invalid",WORK_WEEK,HOURS_WORK) denom = denomPOP summary;
   FPOP();
   write  ("*P19 -05* Work last week [%2d] but hours worked [%2d] invalid, PN = [%2d]",WORK_WEEK,HOURS_WORK,PERSON_NUMBER);
   impute (HOURS_WORK, AHOURS_WORK (AGE10,SEX));
 endif;
```

*3. Occupation.*

523. 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. This item should be edited only for persons whose economic activity is "employed, at work" or "self-employed, at work". If dynamic imputation is not used, blank, zero or invalid responses should be changed to "not reported". Codes for industry tend to be developed so that different digits represent major and minor occupation codes. Write-ins, which are almost unavoidable for occupation, will add to the coding burden. 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.

Southern Sudan 2008

```
if Age < 10 then {already made blank}
else
  if Activity in 1:3 then
    if not Occupation in 1:3,11:14,21:26,31:35,41:44,51:54,61:63,71:75,81:83,91:96 then
        if Activity in 1:3 then {in work or looking for work, so set Occupation to Not Reported. }
           errmsg("*P21-1* PN %d, age %d: Q21 (Occupation) changed from %2d to 98 (Not reported)", curocc(Person),
             Age, Occupation) denom = AgeGE10 summary;
           impute(Occupation, AOCCUPATION (AGE5,SEX));
        elseif Occupation <> notappl then {Set Occupation to Not Applicable }
           errmsg("*P21-2* PN %d, age %d: Occ (Occupation) changed from %2d to not applicable", curocc(Person),
              Age, Occupation) denom = AgeGE10 summary;
           impute(Occupation, notappl);
        endif;
    else { valid occupation. Is Activity in work/looking for work? If not, then set Occupation to not applicable }
        if not Activity in 1:3 then
           errmsg("*P21-3* PN %d, age %d: Occupation changed from %2d to not applicable", curocc(Person), Age,
             Occupation) denom = AgeGE10 summary;
           impute(Occupation, notappl);
       else
         AOCCUPATION (AGE5,Q03_SEX) = Occupation;
        endif;
    endif;
  else
    if occupation <> notappl then
       errmsg("*P21-4* PN %d, age %d: Occupation %1d not blank for worker %1d", curocc(Person), Age,
```

```
          occupation,Activity) denom = AgeGE10 summary;
        impute(Q21_occupation, notappl);
    endif;
  endif;
endif;
```

524. Sometimes enumerated results of the census have inconsistencies arising from difficulty in interpreting certain variables, like occupational classification. Many times the inconsistencies can be corrected during the edit. In the example below, education took precedence over occupation, so when a particular occupation was inconsistent with educational level, an occupation in the same series but consistent with the educational attainment was assigned:

Southern Sudan 2008

```
  {Engineers}
  if Occupation  in 21 then
    if Attainment  >= 9 then  {Secondary 4}
    {21 - Science and engineering professionals = minimum S3 or greater}
    elseif Attainment  in 5:8 then   {P8}
      {31 - Science and engineering associate professionals = P8 or greater}
      errmsg ("*P21-101* Engineers [%2d] changed by education [%2d], PN = %2d",Occupation,
        Attainment, PERSON_NUMBER) denom = PERSON_COUNT summary;
      impute (Q21_Occupation,31);
    elseif Q18_Attainment  in 1:4 then
      {72 - Metal, machinery and related trades workers = Any education}
       errmsg ("*P21-102* Engineers [%2d] changed by education [%2d]",Occupation ,Attainment )
          denom = PERSON_COUNT summary;
       impute (Q21_Occupation,72);
    else
      {93 - Labourers in mining, construction, manufacturing and transport = Any and no education}
       errmsg ("*P21-103* Engineers [%2d] changed by education [%2d]",Occupation ,Attainment )
          denom = PERSON_COUNT summary;
      impute (Occupation,93);
    endif;
  endif;
```

*4. Industry.*

525. According to the United Nations “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). This item should be edited only for persons whose economic activity was “employed, at work” or “self-employed, at work”. If dynamic imputation is not used, blank, zero or invalid responses should be changed to “not reported”. Codes for industry tend to be developed so that different digits represent major and minor industry codes. Write-ins, which are almost unavoidable for this item, will add to the coding burden. 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. The edit for industry will look similar to the one for occupation above.

*5. Status in employment.*

526. Status in employment refers to the status of an economically active 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 economically active person is classified by status in employment based on the same job(s) as used for classifying the person by occupation, industry and sector. The economically active population should be classified by status in employment, 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.

527. 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. This item should be edited only for persons whose economic activity is “employed, at work” or “self-employed, at work”. 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.

Southern Sudan 2008

```
{ ************************************************************************************************
                              Employment Status
  ************************************************************************************************}
{ [Persons 10 years and over]
  Q23. for those who worked or have worked before, what was (name's)
       employment status?
       1 Paid employee
       2 Employer
       3 Own account worker
       4 Unpaid family worker
       5 Unpaid working for others
  *********************************************}

if AGE < 10 then {if age < 10 then Q19 - Q23 must be blank }
else
  if Activity in 1:3{,5} then
    if not Status in 1:5 then {Employment status is blank or invalid }
        if Activity in 1:3 then {in work or looking for work, so set Industry to Not Reported. }
           errmsg("*P23-1* PN %d, age %d: Status changed from %1d to 8 (Not reported)", curocc(Person),
             Age, Status) denom = AgeGE10 summary;
           impute(Status, ASTATUS (AGE5,Sex));
        elseif Status <> notappl then {Set Occupation to Not Applicable }
            errmsg("*P23-2* PN %d, age %d: Status changed from %1d to not applicable", curocc(Person),
               Age, Status) denom = AgeGE10 summary;
            impute(Status, notappl);
        endif;
    else {valid employment status. Is work or looking for work? If not, then set Industry to not applicable }
        if not Activity in 1:3 then
           errmsg("*P23-3* PN %d, age %d: Status changed from %1d to not applicable", curocc(Person),
             Age, Q23_Status) denom = AgeGE10 summary;
           impute(Q23_Status, notappl);
        else
            ASTATUS (AGE5,Q03_SEX) = Q23_Status;
        endif;
    endif;
  else
    if Status <> notappl then
      errmsg("*P23-4* PN %d, age %d: Emp status %1d not blank for worker %1d", curocc(Person),
         Age, status, Activity) denom = AgeGE10 summary;
      impute(Status, notappl);
    endif;
  endif;
endif;
```

*6. Income*

528. The census topics relating to economic characteristics of the population presented in *Principles and Recommendations for Population and Housing Censuses* 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 (SNA). The economically active population comprises all persons of either sex who provide or are available to provide the supply of labour for the production of economic goods and services, as defined by the SNA, during a specified time-reference period.

529. 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.

530. *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. For individual income, the program might look something like this:

```
if AGE < 15 then
  if TOTAL_INCOME_LAST_YEAR <> NOTAPPL then
    errmsg ("*P29 -01* Income last year [%2d] for young person [%2d], PN = [%2d]",TOTAL_INCOME_LAST_YEAR,AGE,PERSON_NUMBER)
       denom = denomPOP summary;
    impute (TOTAL_INCOME_LAST_YEAR,NOTAPPL);
  endif;
else
  if TOTAL_INCOME_LAST_YEAR in 0:999 then
    ATOTAL_INCOME_LAST_YEAR (AGE10,SEX) = TOTAL_INCOME_LAST_YEAR;
  else
    errmsg ("*P29 -02* Income last year imputed [%2d] age [%2d], PN = [%2d]",TOTAL_INCOME_LAST_YEAR,AGE,PERSON_NUMBER)
       denom = denomPOP summary;
    impute (TOTAL_INCOME_LAST_YEAR,ATOTAL_INCOME_LAST_YEAR (AGE10,SEX));
  endif;
endif;
```

531. Household income is as 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, nevertheless, 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 the Recodes section.

```
HHINCOME = sum (TOTAL_INCOME_LAST_YEAR);
```

*7. Institutional sector*

532. The Institutional sector of employment relates to the legal organization and principal functions, behaviour and objectives of the enterprise with which a job is associated. 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. 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.

*8. Employment in the Informal Sector*

533. Many workers participate in the informal sector at the same time, or in lieu of, participation in the formal sector. Some countries use the items on activity status and other economic items to identify the informal sector. Other countries ask specific questions about participation in the informal sector. The edit for informal sector should be straight-forward. If participation in the informal sector is independent of participation in the formal sector, "unknown" can be assigned for blank or invalid entries, or a hot deck based on age and sex can be used. If participation in the information sector is not independent of participation in the formal sector, than an additional variable in the hot deck matrix could be included to indicate whether the person was also in the formal sector as well.

```
If AGE < 10 then
 If INFORMAL_SECTOR <> NOTAPPL then impute (INFORMAL_SECTOR,NOTAPPL); endif;
else
   if INFORMAL_SECTOR in 1:2 then   {whether working in informal sector known}
       AINFORMAL_SECTOR (AGEX,SEX) = INFORMAL_SECTOR;
   Else
       Impute (INFORMAL_SECTOR,(AINFORMAL_SECTOR (AGEX,SEX));
   Endif;
Endif;
```

Or, if a person in paid employment can not also be in the informal sector:

```
if AGE < 10 then
  if INFORMAL_SECTOR <> NOTAPPL then impute (INFORMAL_SECTOR,NOTAPPL); endif;
else
   if INFORMAL_SECTOR in 1:2 then   {whether working in informal sector known}
      if PAID_WORK = 1 then
         impute (INFORMAL_SECTOR,2);
      else
         AINFORMAL_SECTOR (AGEX,SEX) = INFORMAL_SECTOR;
```

```
      endif;
   else
      if PAID_WORK = 1 then
         impute (INFORMAL_SECTOR,2);
      else
         Impute (INFORMAL_SECTOR,(AINFORMAL_SECTOR (AGEX,SEX));
      endif;
   endif;
endif;
```

*9. Place of work*

534. "Place of work" is the location in which a currently employed person performs his or her job, and where a usually employed person performs the primary job used to determine his/her other economic characteristics such as occupation, industry and status in employment. While the information on place of work can be used to develop area profiles in terms of the employed labour force (as opposed to demographic profiles by place of residence), the primary objective is to link place-of-work information to place of residence.

535. 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 analyse only the "known" cases.

```
    if PLACE_OF_WORK in 1:999 then
       APLACE_OF_WORK (AGE10,SEX) = PLACE_OF_WORK;
     else
       errmsg ("*P19 -06* Work last week [%2d] but place of work [%2d] invalid", WORK_WEEK, PLACE_OF_WORK)
         denom = denomPOP summary;
       impute (PLACE_OF_WORK, APLACE_OF_WORK (AGE10,SEX));
     endif;
```

536. 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.

537. 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.

538. This section looked at the population variables recommended in the *Principles and Recommendations.* No country should be using all of these variables, and the selected variables and their spatial relationships with the other variables should be thoroughly tested in hot house and pre-census survey situations for reliable and complete responses. Unlike the housing variables, the population items are usually cross-tabulated in many different combinations, so thorough testing is needed.

## II.6. HOUSING EDITS

539. The specifications for housing edits take into account 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.

540. As with population items, for missing 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. 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. The editing team may want to specify the circumstances in which an entry should be supplied for a blank from a previous housing unit with other similar characteristics.

541. 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 III). 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.

542. 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 actually implement selective editing.

543. 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 the "shotgun" approach, and to remove cases that may not actually be extraneous. For example, although wealthy individuals in an area may purchase gas generators to use when electricity is not otherwise available, the editing team may decide not to include these cases in the data set.

544. 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.

545. 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.

546. 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. 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.

547. 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.

548. 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.

549. 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. See the section on Tenure below for sample pseudo-code.

550. 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.

551. 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.[7]

552. *Finding at least one variable to use in the Hot Decks.*

```
{.
. ****************************************************************************
. **************                                                ****************
. **************        Lighting                                ****************
. **************                                                ****************
. ****************************************************************************
.}
if LONGFORM = 1 then
  totpop = TOTOCC (POP) + 1;
  if totpop > 9 then totpop = 9 endif;

  If !($ in 1:9,NOTAPPL) THEN
    errmsg("H11-1 Lighting invalid %d",S5Q11),denom=denomHouse summary;
    impute ($,ALIGHTING (totpop));
  else
    ALIGHTING (totpop) = $;
  endif;
endif {LongQuest}
```

Ethiopia 2007

```
PROC BATHING_FACILITIES
{.
. ****************************************************************************
. **************        Bathing facilities                      ****************
. ****************************************************************************
.}
if LONGFORM = 1 then
  If !($ in 1:6) then
     errmsg("H7-1 Bathing facilities invalid %d",$),denom=denomHouse summary;
     F1F2();
     write ("H7-1 Bathing facilities invalid %d",$);
     impute ($,ABATHING (Lighting));
  else
     ABATHING (Lighting) = $;
  endif;
endif {LongQuest}
```

553. *Living quarters: type of living quarters (Core topic).* The classification outlined below describes a system of three-digit codes designed by the United Nations 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:

[7] Housing, more than population, can use "standard" edits with little likely harm to the data set since most of the variables are independent of each other. And, as noted, tenure is often a good variable to use in the hotdeck. The form would be something like:

```
If VARIABLE in 1:9 then
   AVARIABLE (TENURE) = VARIABLE;
Else
   Impute (VARIABLE,AVARIABLE (TENURE));
Endif;
```

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 Improvised 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

554. 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.

555. *Living quarters: location of living quarters (Core topic).* Location of living quarters is a geographical variable and is presented with the structure edits in Chapter II.3.

556. *Living quarters: occupancy status (Core topic)* The decision to record living quarters whose occupants are temporarily absent or temporarily present as "occupied" or "vacant" will need to be considered in relation to whether a *de jure* or *de facto* population census is being carried out. In either case, it would seem useful to distinguish as far as possible living quarters 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. 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

557. If the housing unit is occupied, the number of occupants and the count of population records must not be zero. 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". 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.

558. *Living quarters: type of ownership (Core topic)* This topic refers to the type of ownership of the living quarters themselves and not of that of the land on which the living quarters stand. Type of ownership should not be confused with tenure. Information should be obtained to show whether the living quarters are owned by the public sector (central Government, local Government, public corporations) or whether the living quarters 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. 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.
The classification of living quarters 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

Lesotho 2006

```
PROC TENURE

{.
. ****************************************************************************
. **************                                                ****************
. **************        Edit H9 - Tenure                        ****************
. **************                                                ****************
. ****************************************************************************
.}
  N01 = TOTOCC (INDATA_EDT) + 1; {JH - what is TYPE supposed to be here?}
  if N01 > 10 then
     N01 = 10;
  endif;
  if TENURE in 1:7 then
    ATENURE (N01) = TENURE;
  else
    {F8F9();}
    errmsg ("*H09-1* Tenure from number of persons, atenure = %02d tenure = %01d", ATENURE (N01), TENURE) denom = denomHOUSE summary;
    FHOUSE();
    write ("*H09-1* Tenure from number of persons, atenure = %02d tenure = %01d", ATENURE (N01), TENURE);
    impute (TENURE, ATENURE (N01));
  endif;



PROC WALLS
{. ****************************************************************************
. **************        Edit H55 - Walls                        ****************
. ****************************************************************************
.}
  if WALLS in 1:7 then
    AWALLS (TENURE) = WALLS;
  else
    errmsg ("*H55-1* Walls from Tenure, walls = %02d, tenure = %01d", WALLS, TENURE)  denom = denomHOUSE summary;
    FHOUSE();
    write ("*H55-1* Walls from Tenure, walls = %02d, tenure = %01d", WALLS, TENURE);
    impute (WALLS, AWALLS (TENURE));
  endif;
```

559. *Living quarters: number of rooms (Core topic)* 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. 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.

560. *Living quarters: number of bedrooms (Additional Topic).* 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. Sometimes enumerators report a value for the number of bedrooms that is greater than the value for the number of rooms. If this occurs and if the country uses "not stated" only for invalid or inconsistent responses, "not stated" should 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. 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. (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.)

```
If BEDROOMS in 0:9 then
  If BEDROOMS <= ROOMS then
    ABEDROOMS (TENURE,ROOMS) = BEDROOMS;
  Else
    Impute (BEDROOMS,ABEDROOMS (TENURE,ROOMS));
  Endif;
Else
   Impute (BEDROOMS,ABEDROOMS (TENURE,ROOMS));
Endif;
```

561. *Living quarters: useful floor space (Additional Topic)* 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. 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. Suggested edit follows standard housing edit.

562. *Living quarters: water supply system (Core topic).* According to the United Nations, 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 water is provided to the living quarters 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, 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. No piped water available (including piped water from a source beyond a distance of 200 metres from the living quarters)

563. 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. 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. 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. Suggested edit follows standard housing edit.

564. *Drinking water – main source (Core topic)* 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. 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. Suggested edit follows standard housing edit.

565. *Living quarters: Toilet facilities and Sewerage Disposal (Core topics)* 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. Nonetheless, these items should be edited together, and use the same dynamic imputation matrices, if possible. 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 recommendations for classification of housing unit by toilet facilities include the following:

1 With toilet within housing unit
1.1 Flush/pour flush104 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 latrine105
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, and so forth

566. The type of toilet facilities and sewerage 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. Suggested edit follows standard housing edit.

567. *Living quarters: Bathing facilities (Core topic)* According to the United Nations, information should be obtained on whether or not a fixed bath or shower is installed within the premises of each set of living quarters. 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.

568. 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 as a rule and then water supply, construction material of walls, tenure or vacancy status, to obtain "known" information from similar housing units in the geographical area. Suggested edit follows standard housing edit.

569. *Living quarters: Availability of Kitchen (Core topic)* According to the *Principles and Recommendations* 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

570. 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. Suggested edit follows standard housing edit.

571. *Living quarters: cooking fuel (Core topic)* In the context of the need to monitor closely the use of natural resources, a number of 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. Response for type of cooking fuel should be edited with those for cooking facilities. 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. Suggested edit follows standard housing edit.

572. *Living quarters: Lighting and/or electricity – type of (Core topic).* Information should be collected on the type of lighting in the living quarters, 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. 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. Suggested edit follows standard housing edit.

573. *Living quarters: Solid waste disposal – main type of (Core topic)* According to *Principles and Recommendations*, 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

574. 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. Suggested edit follows standard housing edit.

575. *Living quarters: type of heating and energy used for heating (Additional Topic)* This topic refers to the type of heating of living quarters and the energy used for that purpose. The units of enumeration are all living quarters. This topic is irrelevant for a number of countries where, owing to their geographical position and climate, there is no need to provide heating in living quarters. 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 living quarters or serving a set of living quarters, or it may not be central, with the heating provided separately within the living quarters 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. 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 take into account 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 to obtain "known" information from similar housing units in the geographical area. These two characteristics might include type of housing unit, construction material of walls, tenure, and vacancy status. Suggested edit follows standard housing edit.

576. *Living quarters: availability of hot water (Additional Topic)* 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. 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. Suggested edit follows standard housing edit.

577. *Living quarters: piped gas (Additional Topic)* This topic refers to the availability of piped gas in the living quarters. 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. 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. Suggested edit follows standard housing edit.

578. *Living quarters: use of housing unit (Additional Topic)* "Use of a housing unit" indicates whether a housing unit is being used wholly for habitation or residential purposes or not. The housing unit can be used for habitation as well as for commercial, manufacturing or other purposes. "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. Suggested edit follows standard housing edit.

579. *Living quarters: occupancy by one or more households (Core topic)* 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. It is important to note that this edit must come after the structure edit determines the household head. Suggested edit follows standard housing edit.

580. *Living quarters: number of occupants (Core topic)* 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. "Number of occupants" is related to the number of population records and the two should be identical. 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. Suggested edit follows standard housing edit.

581. *Building: building description (Core topic)* The following classification by type is recommended by the United Nations for buildings in which some space is used for residential purposes. 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, in order to obtain "known" information from similar housing units in the geographical area.

1 Buildings containing a single housing unit
1.1 Detached
1.2 Attached
2 Buildings containing more than one housing unit
2.1 Up to 2 floors
2.2 From 3 to 4 floors
2.3 From 5 to 10 floors
2.3 Eleven floors or more
3 Buildings for persons living in institutions
4 All others

Suggested edit follows standard housing edit.

582. *Building: year or period of construction (Additional Topic)* The year of 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 immediately preceding intercensal period 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. 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. Suggested edit follows standard housing edit.

583. *Building: number of dwellings (Additional Topic)* Editing for the number of housing units in a building is explained in Chapter III as part of the structure edits.

584. *Building: construction material of outer walls (Core topic).* This topic refers to the construction material of the external (outer) walls of the building 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. 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. See also combined edit for walls, roof, and floor below.

Lesotho 2006

```
PROC WALLS
{. ***************************************************************************
. *************        Edit H55 - Walls                        ****************
. ***************************************************************************
.}
   if WALLS in 1:7 then
      AWALLS (TENURE) = WALLS;
   else
    errmsg ("*H55-1* Walls from Tenure, walls = %02d, tenure = %01d", WALLS, TENURE)  denom = denomHOUSE summary;
    impute (WALLS, AWALLS (TENURE));
  endif;

{Changing walls for Rontabole 10/2/09}
  if TYPE_OF_HOUSE = 1 then  {Rontabole}
    if WALLS = 5 then        {Walls are corrugated iron}
      errmsg ("*H55-2* Walls for Rontabole made stone instead of corrugated iron")  denom = denomHOUSE summary;
      impute (WALLS, 6);
    endif;
  endif;
```

585. *Building: construction material of roof (Additional Topic)* 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. 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 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. See combined edit for walls, roof, and floor below.

586. *Building: construction material of floor (Additional topic).* 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. The following edit is for walls, roof, and floor combined.

```
{Check for various illegal combinations to prepare for the edit.
   One example below}
If WALLS = 5 and ROOF = 1 then  {Thatch walls and concrete roof}
   WALLS = 9;
```

```
Endif;

If WALLS in 1:5 then
  If ROOF in 1:5 then
    If FLOOR in 1:3 then
      AWALLSFROMROOFFLOOR (ROOF,FLOOR) = WALLS;
      AFLOORFROMWALLSROOF (WALLS,ROOF) = FLOOR;
      AROOFFROMWALLSFLOOR (WALLS,FLOOR) = ROOF;
      AWALLSFROMROOFTENURE (ROOF,TENURE) = WALLS;
      AROOFFROMWALLSTENURE (WALLS,TENURE) = ROOF;
      AWALLSFROMFLOORTENURE (FLOOR,TENURE) = WALLS;
      AWALLSFROMTENURE (TENURE) = WALLS;
    Else
      Impute (FLOOR,AFLOORFROMWALLSROOF (WALLS,ROOF));
    Endif;
  Else
    If FLOOR in 1:3 then
      Impute (ROOF,AROOFFROMWALLSFLOOR (WALLS,FLOOR));
    Else
      Impute (ROOF,AROOFFROMWALLSTENURE (WALLS,TENURE));
      Impute (FLOOR,AFLOORFROMWALLSROOF (WALLS,ROOF));
    Endif;
  Endif;
Else
  If ROOF in 1:5 then
    If FLOOR in 1:3 then
      Impute (WALLS,AWALLSFROMROOFFLOOR (ROOF,FLOOR));
    Else
      Impute (WALLS,AWALLSFROMROOFTENURE (ROOF,TENURE));
      Impute (FLOOR,AFLOORFROMWALLSROOF (WALLS,ROOF));
    Endif;
  Else
    If FLOOR in 1:3 then
      Impute(WALLS,AWALLSFROMFLOORTENURE (FLOOR,TENURE));
      Impute (ROOF,AROOFFROMWALLSFLOOR (WALLS,FLOOR));
    Else
      Impute (WALLS,AWALLSFROMTENURE (TENURE));
      Impute (ROOF,AROOFFROMWALLSTENURE (WALLS,TENURE));
      Impute (FLOOR,AFLOORFROMWALLSROOF (WALLS,ROOF));
    Endif;
  Endif;
Endif;
```

587. *Building: elevator (Additional topic).* 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. If the building has only one storey 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 stories or the fact that an elevator is present. If the elevator takes precedence, the number of stories must be changed, either by making the value "unknown" or by using dynamic imputation to obtain another value. If the number of stories takes precedence, and the building has only one storey, the response on "presence of an elevator" must be changed to "no". When an elevator is present, if it requires electricity, a check should be made to be certain that electricity exists in the building. 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 and construction material of outer walls, to obtain "known" information from similar housing units in the geographical area. Suggested edit follows standard housing edit.

588. *Building: Farm (Additional topic)* 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. 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. Suggested edit follows standard housing edit.

589. *Building: state of repair (Additional topic)* This topic indicates whether the building is in need of repair and identifies the kind of repair needed. The unit of enumeration is a building. The classification of buildings according to the state of repair may include "repair not needed", "in need of minor repair", "in need of moderate repair" or "in need of serious repair" and "irreparable". 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. 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. 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, construction of outer walls and type of housing unit, to obtain "known" information from similar housing units in the geographical area. Suggested edit follows standard housing edit.

590. *Occupants; characteristics of head of household (Core topic)* The characteristics of the head of household are usually obtained from the population records to assist in developing cross-tabular information for planning and analysis. These items, including ethnic origin, religion or income, assist in determining differential social status or need. Since these characteristics will already have been edited for the population items, no further editing should be needed here.

```
if $ = 1 then
    AGEHEAD  = P03_AGE;
    SEXHEAD  = P04_SEX;
    MARITALHEAD  = P08_MARITAL_ST;
    DISABILITYHEAD  = P21_ANY_DISABILITY;
    EDUCHEAD  = P29_LEVEL_EDUC;
    EMPSTATHEAD  = DER01_VESO;
 endif;
```

591. *Occupants: tenure (Core topic)* According to the United Nations, 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 Member of household owns housing unit (code 1)
2 Member of household rents all or a part of housing unit
2.1 Member of household rents all or a part of housing unit as a main tenant (code 2)
2.2 Member of household rents a part of housing unit as a subtenant (code 3)
3 Occupied free of rent (code 4)
4 Other arrangement (code 5)

Units occupied free of cash rent, with or without the permission of the owner, especially where this practice is prevalent, should be considered separately. Tenure may relate to type of ownership (H12), 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.

```
{If the distribution for tenure is somewhat symmetrical, it is a good variable to use for hot decks}
POPINHH = totocc (POP);
If POPINHH = 0 then POPINHH = 1; endif; {If vacant units are included}
If POPINHH > 10 then POPINHH = 10; endif; {If more than 10 people in the unit}
If TENURE in 1:5 then
  ATENURE (POPINHH) = TENURE;
Else
  Impute (TENURE,ATENURE (POPINHH));
Endif;
```

592. *Occupants: rental and owner-occupied housing costs (Additional Topic)* 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. 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 to obtain "known" information from similar housing units in the geographical area. Suggested edit follows standard housing edit.

593. *Occupants: furnished or unfurnished (Additional Topic)* 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. Suggested edit follows standard housing edit.

594. *Information and Communication technology (ICT) devices – availability of (Core topic)* 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. The recommended classification is:

1. Household having radio
2. Household having television set
3. Household having fixed-line telephone
4. Household having mobile cellular telephone(s)
5. Household having personal computer(s)
6. Household accessing the Internet from home
7. Household accessing the Internet from elsewhere other than home

8. Household without access to the internet

```
If not RADIO in 1:2 then impute (RADIO,2); endif;  {Assumes if radio not reported, household has no radio.}
If not TV    in 1:2 then impute (TV,2); endif;  {Assumes if TV not reported, household has no TV.}
If not FIXED_PHONE in 1:2 then impute (FIXED_PHONE,2); endif;
If not MOBILE_PHONE in 1:2 then impute (MOBILE_PHONE,2); endif;
If not COMPUTER in 1:2 then impute (COMPUTER,2); endif;  {Assumes if computer not reported, household has no
computer.}
If not INTERNET in 1:3 then impute (INTERNET,3); endif; {If neither home nor outside, then not present}
```

595. 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. 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. 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. Suggested edit follows standard housing edit.

596. *Occupants: number of cars (Additional Topic)* "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. 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 particular 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. Suggested edit follows standard housing edit.

597. *Occupants: durable appliances (Additional Topic)* Information is collected on the availability of such durable appliances as washing machines, dishwashing machines, refrigerators, deep freezes, television sets, personal computers, depending on national circumstances. 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. 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). Suggested edit follows standard housing edit.

598. *Occupants: outdoor space available for household use (Additional Topic)* 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 may refer to the outdoor space available as part of a housing unit (for example, the backyard in the case of a detached house), the outdoor space available adjacent to a building (for example, backyards and playgrounds placed next to an apartment building), the outdoor space available as part of common recreational areas within a 10-minute walk from the housing unit (for example, parks, sports centres and similar sites), or if outdoor space is not available within a 10-minute walk. 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. Suggested edit follows standard housing edit.

### C. OCCUPIED AND VACANT HOUSING UNITS

599. The edits described above 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 particular attention to the imputation matrix variables since these are most likely to differ.

600. This section looked at the housing variables recommended in the *Principles and Recommendations.* No country should be using all of these variables, and the selected variables and their spatial relationships with the other variables should be thoroughly tested in hot house and pre-census "pilot" survey situations 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.

## II.7 DERIVED VARIABLES

601. 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. For example, the item on economic activity status is already a combination of several collected variables on the census. Rather than having to develop a program to recode the information each time the national census/statistical office wants a special tabulation, data processing specialists can write a program to make the recode once, store the recoded information on the person's record, and then use it for further tabulations. National census/statistical offices need to decide how often the recodes will be used in tabulations and how relevant a particular recode will be when they determine whether or not to produce and store the information. It is important to remember that the recodes also take up room on the person records. The larger the population size, the more space will be used.

602. 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 each individual's income and placing the sum on the housing record for later use.

603. 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 economically active–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.

604. *Multiple occurrences – when recoding is not appropriate.* One of the most common is multiple entries for the same item, which is sometimes covered by the term "occurrences". For example, on a housing record, a country may collect age and sex of deaths in the household in the year before the census. If a country allows for up to 6 deaths, for example, the item may be repeated 6 times (with DeathSex1, DeathSex2, DeathSex3, etc.), or may occur as a series of deaths, with Deathsex (1), Deathsex (2), Deathsex (3), for example. In this type of display, it would be rare to combine the variables as recodes.

### A. DERIVED VARIABLES FOR HOUSING RECORDS

605. *Household income.* 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. The editing team must take into account 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 particular person's income.

```
HHINCOME = sum (TOTAL_INCOME_LAST_YEAR);
```
[8]

606. *Family income.* 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 definition will depend on the particular country's situation. 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. The editing team must take into account the situation where one or more persons in the family has negative income because of a business failure or other reasons. In such a case, the total family income will be decreased, rather than increased, by this particular person's income.

```
FAMINCOME = sum (TOTAL_INCOME_LAST_YEAR where RELATIONSHIP in 1:9);
```

607. *Family type.* "Family type" is useful 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. As stated in *Principles and Recommendations for Population and Housing Censuses, Revision 1*, definitions of family vary from country to country. One definition is that a family consists of a head of household and one or more other persons living in the same household, related to the head of household by birth, marriage or adoption. All persons in a household related to the head of household are members of his or her family. However, not all households contain families since a household might comprise a group of unrelated persons or one person living alone. Subject-matter specialists might classify families by type as either "married-couple families" or "other families" according to the sex of the head of household and the presence of relatives. The United States of America, for example, has used the following codes based on data on family type, which were derived from answers to questions on sex and relationship:

(a) Codes 1 and 2: **Married-couple family.** A family in which the head of household and his or her spouse were enumerated as members of the same household; code 1 is used when the head of household is male, code 2 when the head of household is female.
(b) Code 3: **Other family: male head of household, no wife present.** A family with a male head of household and no spouse of head of household present.
(c) Code 4: **Other family: female head of household, no husband present.** A family with a female head of household and no spouse of head of household present.
(d) Codes 5 and 6: **Non-family household.** A household which is not a family, so no spouse or other relative of head of household is present; code 5 is used when the head of household is male, code 6 when the head of household is female.
(e) Codes 7 and 8: **Single person household.** A single person living alone is considered a household, but not a family, since no other relatives are present. Code 7 is used when the head of household is male, code 8 when the head of household is female.

Ethiopia 2007

[8] As in the other sections, we present pseudo-code for the derived variables. These can be cut and pasted into CSPro programs, but adjustments must be made for the specific variables in the dictionary. The error messages and write statements are usually dropped here.

```
PROC HOUSEHOLD_TYPE

   RELATED = 0;
   UNRELATED = 0;
   do varying i = 1 until i > TOTOCC (POP)
     if S3Q4 (i) in 3:8 then
       RELATED = RELATED + 1;
     endif;
     if S3Q4 (i) = 9 then
       UNRELATED = UNRELATED + 1;
     endif;
   enddo;


if RELATED > 0 then
  if NumSpouses >= 1 then
    if S3Q5 (1) = 1 then
      errmsg ("R1-1 MC,male head"), summary;
      impute (HOUSEHOLD_TYPE,1);
    else
      errmsg ("R1-2 MC,female head"), summary;
      impute (HOUSEHOLD_TYPE,2);
    endif;
  else
    if S3Q5 (1) = 1 then
      errmsg ("R1-3 No spouse,male head"), summary;
      impute (HOUSEHOLD_TYPE,3);
    else
      errmsg ("R1-4 No spouse,female head"), summary;
      impute (HOUSEHOLD_TYPE,4);
    endif;
  endif;
else
  if not TOTOCC (POP) = 1 then
    if S3Q5 (1) = 1 then
      errmsg ("R1-5 Non-family,male head"), summary;
      impute (HOUSEHOLD_TYPE,5);
    else
      errmsg ("R1-6 Non-family,female head"), summary;
      impute (HOUSEHOLD_TYPE,6);
    endif;
  else
    if S3Q5 (1) = 1 then
      errmsg ("R1-7 Male alone"), summary;
      impute (HOUSEHOLD_TYPE,7);
    else
      errmsg ("R1-8 Female alone"), summary;
      impute (HOUSEHOLD_TYPE,8);
    endif;
  endif;
endif;
```

608. A simpler method of identifying a portion of the categories above is to obtain a derived variable called “head of household, married with spouse present”. The marital status of the head of household can be recoded according to whether the head of household’s spouse is present 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 present will be identified during tabulation.

609. *Family nucleus.* For household composition, the *Principles and Recommendations* 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 co-heads 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 present, with at least one unmarried child determined as above)
4. A mother with one or more unmarried children (female household, no husband present, with at least one unmarried child determined as above)

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.

```
COUNT (UNMARRIED_CHILDREN where (RELATIONSHIP = 3 and MARITAL_STATUS = NEVER_MARRIED));
If SPOUSE_PTR > 0 then
  If UNMARRIED_CHILDREN = 0 then
    FAMILY_NUCLEUS = 1;  {Married couple, no unmarried children}
  Else
    FAMILY_NUCLEUS = 2;  {Married couple with unmarried children}
  Endif;
Else
  If UNMARRIED_CHILDREN > 0 then
    If SEX (HEADPTR) = 1 then
      FAMILY_NUCLEUS = 3;  {Father with unmarried children}
    Else
      FAMILY_NUCLEUS = 4;  {Mother with unmarried children}
    Endif;
  Else
    FAMILY_NUCLEUS = 5;  {All other cases}
  Endif;
Endif;
```

610. *Type of household.* The *Principles and Recommendations* include 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. A first recode could identify type of household. The following items describe the types of households. The definitions are in parentheses, but 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 family – 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 (other types of households)

```
COUNT (NUCLEAR where RELATIONSHIP in 1:3);
COUNT (EXTENDED where RELATIONSHIP > 4);
If totocc (POP) = 1 then
  TYPE_OF_FAMILY = 1;  {Single person household}
Else
```

```
  If NUCLEAR > 1 then
    If EXTENDED = 0 then
      TYPE_OF_FAMILY = 2; {Nuclear family}
    Else
      TYPE_OF_FAMILY = 3; {Extended family}
    Endif;
  Else
    TYPE_OF_FAMILY = 4;  {Other types of family}
  Endif;
Endif;
```

611. *Household composition.* The Principles and Recommendations suggests a recode for household composition. The recode has 4 types of households:

1. *Single person households* are households, but not families, so should be included as a separate category in the household composition recode.
2. *Nuclear family households*. Nuclear family households can be divided into (and received individual codes for: (1) married-couple family with children, (2) married-couple family without children, (3) partners in consensual union with children, (4) partners in consensual union without children, (5) fathers with children, and (6) mothers with children. 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.
3. *Extended family households.* Extended families 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. 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.
4. *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 plus other persons, some of whom are related to the nucleus and some of whom are not, (42) a single family nucleus plus 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. Once again, a series of searches and summaries will permit the appropriate designation for each type of household.

```
PROC NUCLEAR
{Determines the type of family:
    1 = Nuclear family - head, spouse and children ONLY
    2 = Extended family - head and other relatives
    3 = other type of family
    4 = single person household}

if totocc (INDATA_EDT) = 1 then
  $ = 4;  {Single person household}
else
  N02 = 0;{set for nuclear family -- head, spouse, kids only}
  N03 = 0;  {set for extended family -- head &other relatives}
  do varying N01 = 1 until N01 > totocc (INDATA_EDT)
    if RELAT (N01) in 1:3 then
      N02 = N02 + 1;  {Add for nuclear family}
      N03 = N03 + 1;  {Add for extended family}
    endif;
    if RELAT (N01) in 4:7 then
      N03 = N03 + 1;  {Add for extended family only}
    endif;
  enddo;

  if N02 = totocc (INDATA_EDT) then
    $ = 1;   {Nuclear family}
  else
    if N03 = totocc (INDATA_EDT) then
      $ = 2;  {extended family}
    else
      $ = 3;  {other type of household}
    endif;
  endif;
endif;
```

612. *Family composition.* 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*.)

613. *HIV/AIDS Household Structure*. The impact of the HIV/AIDS epidemic affects so many countries, a recode can assist 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 can 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., can assist government planners in fully describing the impact of the HIV/AIDS situation.

```
PROC HIVAIDS
if RELATIONSHIP = 1 then    {do this only for the head of household}
  if AGE > 17 then
    SPOUSE_PTR = 0;                         {Determine if there's a spouse}
    do varying N01 = 1
    until N01 > TOTOCC (POP)
       if RELATIONSHIP (N01) = 2 then
          SPOUSE_PTR = 1;
       endif;
    enddo;
       KIDS = 0; GRCH = 0; NIECES = 0; OREL = 0;
       do varying N01 = 1 until
         N01 > TOTOCC (POP)
         if RELATIONSHIP (N01) in 3 then KIDS = KIDS + 1; endif; {count children}
         if RELATIONSHIP (N01) = 6  then NIECES = NIECES + 1; endif;   {count of nieces and nephews}
         if RELATIONSHIP (N01) = 7  then GRCH = GRCH + 1; endif;   {count of grandchildren}
         if RELATIONSHIP (N01) in 4:5,8 then OREL = OREL + 1; endif; {parent or sibling or other}
       enddo;

    if SPOUSE_PTR <> 0 then                  {This part is for married couple families}
         if GRCH = 0 and NIECES = 0 and OREL = 0 then   {nuclear family}
            errmsg ("11 Nuclear family: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 11; endif;
         if GRCH > 0 or NIECES > 0 or OREL > 0 then   {nuclear family and other relatives}
            errmsg ("12 Nuclear family and others: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 12; endif;
         if KIDS = 0 and GRCH > 0 and NIECES = 0 and OREL = 0 then {missing generation}
            errmsg ("13 Missing generation: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 13; endif;
         if KIDS = 0 and GRCH > 0 and NIECES = 0 and OREL > 0 then {missing generation but other relatives}
            errmsg ("14 Missing gen but ORELs: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 14; endif;
         if KIDS = 0 and GRCH = 0 and NIECES > 0 and OREL = 0 then {nieces and nephews}
            errmsg ("15 Nieces/nephews only: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 15; endif;
         if KIDS = 0 and GRCH = 0 and NIECES > 0 and OREL > 0 then {nieces and nephews and other relatives}
            errmsg ("16 Nice/nephew & others: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 16; endif;
         if KIDS = 0 and GRCH > 0 and NIECES > 0 and OREL = 0 then {nieces and nephews and grandchildren}
            errmsg ("17 Niece/nephew, grch: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 17; endif;
         if KIDS = 0 and GRCH > 0 and NIECES > 0 and OREL > 0 then {nieces, nephews, grandchildren and other relatives}
            errmsg ("18 Niece/nephew, GRCH,ORELS: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 18; endif;
         if KIDS = 0 and GRCH = 0 and NIECES = 0 and OREL > 0 then {other realtives only}
            errmsg ("19 Other relatives only: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 19; endif;
    else                                     {This part is for single head of household}
      if S3Q25 = 5 then             {Single head of household who is a widow or widower}
         if TOTOCC (POP) = 1 then
            errmsg ("20 Widow only: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 20; endif;
         if GRCH = 0 and NIECES = 0 and OREL = 0 then   {nuclear family}
            errmsg ("21 Widow & kids: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 21; endif;
         if GRCH > 0 or NIECES > 0 or OREL > 0 then   {nuclear family and other relatives}
            errmsg ("22 Widow, kids & others: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 22; endif;
         if KIDS = 0 and GRCH > 0 and NIECES = 0 and OREL = 0 then {missing generation}
            errmsg ("23 Widow, missing gen: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 23; endif;
         if KIDS = 0 and GRCH > 0 and NIECES = 0 and OREL > 0 then {missing generation but other relatives}
            errmsg ("24 Widow, ORELS: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 24; endif;
         if KIDS = 0 and GRCH = 0 and NIECES > 0 and OREL = 0 then {nieces and nephews}
            errmsg ("25 Widow, Niece/nephew: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 25; endif;
         if KIDS = 0 and GRCH = 0 and NIECES > 0 and OREL > 0 then {nieces and nephews and other relatives}
            errmsg ("26 Widow,N/N,ORELS: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 26; endif;
         if KIDS = 0 and GRCH > 0 and NIECES > 0 and OREL = 0 then {nieces and nephews and grandchildren}
            errmsg ("27 Widow, N/N, GrCh: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 27; endif;
         if KIDS = 0 and GRCH > 0 and NIECES > 0 and OREL > 0 then {nieces, nephews, grandchildren and other relatives}
            errmsg ("28 Widow, N/N, GrCh, ORELs: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 28; endif;
         if KIDS = 0 and GRCH = 0 and NIECES = 0 and OREL > 0 then {other realtives only}
            errmsg ("29 Widow, ORELS only: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
            HIVAIDS = 29; endif;
      else    {These are households headed by single parent who is not a widow}
         if TOTOCC (POP) = 1 then HIVAIDS = 30; endif; {Single person household}
         if GRCH = 0 and NIECES = 0 and OREL = 0 then   {nuclear family}
            errmsg ("31 Single, kids: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
```

```
          HIVAIDS = 31; endif;
       if GRCH > 0 or NIECES > 0 or OREL > 0 then   {nuclear family and other relatives}
          errmsg ("32 Single, kids, ORELS: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
          HIVAIDS = 32; endif;
       if KIDS = 0 and GRCH > 0 and NIECES = 0 and OREL = 0 then {missing generation}
          errmsg ("33 Single: Missing Gen: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
          HIVAIDS = 33; endif;
       if KIDS = 0 and GRCH > 0 and NIECES = 0 and OREL > 0 then {missing generation but other relatives}
          errmsg ("34 Single, missing gen, ORELS: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
          HIVAIDS = 34; endif;
       if KIDS = 0 and GRCH = 0 and NIECES > 0 and OREL = 0 then {nieces and nephews}
          errmsg ("35 Single, N/N: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
          HIVAIDS = 35; endif;
       if KIDS = 0 and GRCH = 0 and NIECES > 0 and OREL > 0 then {nieces and nephews and other relatives}
          errmsg ("36 Single, N/N, ORELS:  kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
          HIVAIDS = 36; endif;
       if KIDS = 0 and GRCH > 0 and NIECES > 0 and OREL = 0 then {nieces and nephews and grandchildren}
          errmsg ("37 Single, N/N, GRch: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
          HIVAIDS = 37; endif;
       if KIDS = 0 and GRCH > 0 and NIECES > 0 and OREL > 0 then {nieces, nephews, grandchildren and other relatives}
          errmsg ("38 Single: N/N,Grch,ORELS: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
          HIVAIDS = 38; endif;
       if KIDS = 0 and GRCH = 0 and NIECES = 0 and OREL > 0 then {other realtives only}
          errmsg ("39 Single: ORELS: kids = %d grch = %d nieces = %d others = %d", KIDS, GRCH, NIECES, OREL),  summary;
          HIVAIDS = 39; endif;
    endif;
  endif;
else                                    {The following for heads less than 18 years old}
      N02 = 0;
      do varying N01 = 1 while N01 <= totocc (POP)
        if AGE (N01) > 17 then N02 = N02 + 1; endif;
      enddo;
      if N02 = 0 then            { Everyone in house is less than 18 head less than 18 yrs }
         if ECONOMIC_ACTIVITY = 1 then HIVAIDS = 41;   {in the labor force}
           errmsg ("41 head < 17, in LF: age = %d others 17+ = %d in work force = %d ", AGE, N02, ECONOMIC_ACTIVITY),  summary;
         else HIVAIDS = 42;          {not in the labor force}
          errmsg ("42 head < 17, not in LF:  age = %d others 17+ = %d in work force = %d ", AGE, N02, ECONOMIC_ACTIVITY),  summary;
         endif;
     else                        { head less than 18 years old and some people 18 years and ove}
         if ECONOMIC_ACTIVITY = 1 then HIVAIDS = 43;    {in the labor force}
           errmsg ("43 head<17,some 18+,in LF: age = %d others 17+ = %d in work force = %d ", AGE, N02, ECONOMIC_ACTIVITY),  summary;
         else HIVAIDS = 44; {not in the labor force}
           errmsg ("44 head<17,some 18+,not LF:age = %d others 17+ = %d in work force = %d ", AGE, N02, ECONOMIC_ACTIVITY),  summary;
         endif;
     endif;
  endif;
endif;
```

614. *Related persons.* Related persons are those persons who are related to the head of household in some way. The derived variable for related persons is the sum of all persons related to the 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." 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 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. "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 head of household, except the spouse of the head of household. Related children may or may not include foster children since they are not related to the head of household, but this decision would depend on the country's situation.

```
RELATED_PERSONS = SUM (RELATIONSHIP where RELATIONSHIP in 1:9);
RELATED_0_TO_5 = SUM (RELATIONSHIP where RELATIONSHIP in 1:9 and AGE in 0:5);
RELATED_6_TO_17 = SUM (RELATIONSHIP where RELATIONSHIP in 1:9 and AGE in 6:17);
RELATED_0_TO_17 = SUM (RELATIONSHIP where RELATIONSHIP in 1:9 and AGE in 0:17);
```

615. *Workers in family.* Sometimes countries want to compare household variables by number of workers, such as income distributions by household size and workers per dependant. The country might 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.

```
WORKERS_IN_FAMILY = SUM (WORK_LAST_WEEK = 1);
```

616. *Complete plumbing.* 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. 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."

```
COMPLETE_PLUMBING = 2;  {Assume "no" until checked}
if PIPED_WATER = PIPED and TOILET = FLUSH and SHOWER = YES then
   COMPLETE_PLUMBING = 1;
endif;
```

617. *Complete kitchen* 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 located in the same building as the living quarters being enumerated. They need not be in the same room. 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 is present, but the equipment is located in a different building; some, but not all of the facilities are present; or none of the three specified kitchen facilities is present in the same building as the living quarters being enumerated.

```
COMPLETE_KITCHEN = 2;  {Assume "no" until checked}
if STOVE = YES and REFRIGERATOR = YES and SINK = YES then
   COMPLETE_KITCHEN = 1;
endif;
```

618. *Gross rent.* 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. The derived variable for gross rent is obtained by summing the amount of rent paid and the amount paid for utilities, if these are collected separately.

```
GROSS_RENT = CASH_RENT + UTILITIES + FUELS;
```

619. *Wealth index.* 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 'wealthier' the household. 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 (actually three sets of binary variables). The various items might be weighted when summing. 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. (CSPro is not set up for the factor analysis and weighting needed to develop the wealth index; the index is usually constructed in STATA or another statistical package, with quintiles being developed from the results and recorded for use in the tables. See the Africa Census Tabulation Handbook for examples.)

620. *Households with at least one orphan, by type of orphan.* Sometimes countries, particularly those wanting to assess the extent of the effects of HIV/AIDS on the population and housing, will want to determine the number of households with orphans. The program below illustrates one method of determining the number of orphans in the household, and by type of orphan.

Lesotho 2006

```
PROC WITH_AN_ORPHAN
N01 = 0;  {both parents dead}
N02 = 0;  {Father only dead}
N03 = 0;  {Mother only dead}
N04 = 0;  {Father or mother or both}
do varying i = 1 while i <= totocc (INDATA_EDT)
 if AGE (i) in 0:17 then
  if FATHER_ALIVE (i) = 2 then   {Father is dead}
    if MOTHER_ALIVE (i) = 2 then {Mother is dead}
       N01 = N01 + 1;
       N02 = N02 + 1;
       N03 = N03 + 1;
       N04 = N04 + 1;
    else
```

```
      N02 = N02 + 1;
      N04 = N04 + 1;
    endif;
  else  {Father alive}
    if MOTHER_ALIVE (i) = 2 then {Mother is dead}
      N04 = N04 + 1;
      N03 = N03 + 1;
    endif;
  endif;
 endif;
enddo;



{Any orphan}
if N04 > 0 then
  WITH_AN_ORPHAN = 1;
  NUMBER_OF_ORPHANS = N04; {either or both parents dead}
 else
  WITH_AN_ORPHAN = 0;
  NUMBER_OF_ORPHANS = 0; {either or both parents dead}
endif;

{Double orphans}
if N01 > 0 then
  DOUBLE_ORPHANS = 1;
else
  DOUBLE_ORPHANS = 0;
endif;

{Father dead}
if N02 > 0 then
  WITH_ORPHAN_FA_DEAD = 1;
else
  WITH_ORPHAN_FA_DEAD = 0;
endif;

{Mother dead}
if N03 > 0 then
  WITH_ORPHAN_MO_DEAD = 1;
else
  WITH_ORPHAN_MO_DEAD = 0;
endif;
```

621. *Characteristics of household head or householder.* Sometimes countries need to cross-tabulate housing characteristics by characteristics of the head. These crosses might by age and sex or ethnicity or educational attainment. So, the characteristics of the head would be put on each housing record:

```
HEADS_SEX = SEX (HEADPTR);
HEADS_AGE = AGE (HEADPTR);
HEADS_ETHNICITY = ETHNICITY (HEADPTR);
HEADS_EDUCATION = EDUCATIONAL_ATTAINMENT (HEADPTR);
```

### B. DERIVED VARIABLES FOR POPULATION RECORDS

*1. Economic Activity Status or Economic Status Recode (ESR)*

622. A derived variable for economic status can be very useful for the tabulations, but it requires information from several variables. Since the various classifications of economic activity are used in many of the related tables, the editing team should consider inserting a derived variable into the data records rather than having the data processors reclassify economic 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 particular requirements of the edit or tabulation. Specialists in economic characteristics should prepare the specifications for the derived variable. 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:

**Economically active**
- *Employed*
  1. At work
  2. With job, not at work
  3. In Armed Forces
- *Unemployed*
  4. Looking for work
  5. Discouraged worker

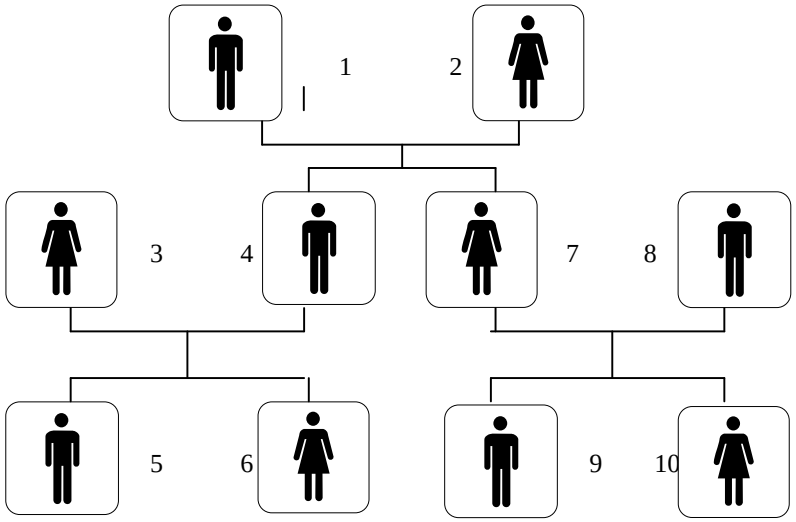
**Not economically active**
6. Homemaker
7. Student
8. Unable to work
9. Other

```
If ECONACTV in 1 then
  If ECONACTV = 1 then
    If OCCUPATION = ARMED_FORCES then
      ESR = 3; {Armed forces}
    else
      ESR = 1; {Working}
    Endif;
  Else
    If TEMPORARILY_NOT_WORKING = 1 then
      ESR = 2; {Temporarily not working}
    Else
       If LOOKING = 1 then
         ESR = 4; {Unemployed and looking for work}
       Else
         ESR = 5; {Unemployed and not looking for work}
       Endif;
    Endif;
  Endif;
Else  {Economically inactive}
  Recode REASON_NOT_WORKING => ESR;
    1 => 6;  {Homemaker}
    2 => 7;  {Student}
    3 => 8;  {Unable to work}
    4 => 9;  {Other}
```

```
    => 9;  {Other}                                  Endif;
Endrecode;
```

623. *Subfamily number and relative in subfamily (family nuclei)* All countries have extended as well as nuclear families. Consider the following extended family, with the triangles representing males and the octagons representing females. The household has a head of household 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 10.

**Figure A.I.1. Illustration of an extended family**




624. In most censuses, if the editing team wants to study the structure of a family such as the one illustrated in figure A.I.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 head of household. Recodes for subfamily and subfamily member will permit a more detailed analysis of the family structure.

625. One definition of a subfamily would be "a married couple (husband and wife enumerated as members of the same household) with or without never-married children under 18 years old." The editing team might want to add to this "one parent with one or more never-married children under 18 years old, living in a household and related to, but not including, either the head of household or the head of household's spouse." The number of subfamilies is not included in the count of families, since subfamily members are counted as part of the head of household's family.

626. The derived variables for subfamilies, 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 subfamilies based on relationships within the household. As each subfamily is determined—a non-head of household/spouse pair (with or without children), or a non-head of household/parent and child—numbers are assigned in order to each subfamily. Code numbers then can be assigned to the various relationships: family "head of household" is code 1, "spouse" of family "head of household" is code 2, and child of family "head of household" is code 3. In order for a subfamily to exist, at least one pair of subfamily relatives must exist: either a "head of household" and "spouse" (codes 1 and 2), or a "head of household" and "child" (codes 1 and 3). When a family head of household, spouse and child are all living together, codes 1, 2 and 3 will all be present for the subfamily. Subfamilies are classified by type: married-couple subfamilies, with or without own children; mother-child subfamilies; and father-child subfamilies. Lone parents include people maintaining either one-parent families or one-parent subfamilies. Married couples include husbands and wives in both married-couple families and married-couple subfamilies.

627. In developing the derived variables, the relationship to head of household 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 subfamilies. 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 subfamily. 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 head of household" with "spouse-in-law" and with "niece/nephew".

628. The example given in figure A.I.2. shows a household with two subfamilies: subfamily 1 consists of person 3 (head of household of subfamily 1), person 4 (spouse), and persons 5 and 6 (children in subfamily 1). Subfamily 2 consists of person 7 (head of household of subfamily 2), person 8 (spouse), and persons 9 and 10 (children in subfamily 2).

**Figure A.I.2. Sample household with two subfamilies**

| *Person number* | *Relationship* | *Sex* | *Subfamily* | |
|---|---|---|---|---|
| | | | *Number* | *Relation* |
| 1 | Head of household | M | | |
| 2 | Spouse | F | | |
| 3 | Son | F | 1 | 1 |
| 4 | Daughter-in-law | M | 1 | 2 |
| 5 | Grandchild | M | 1 | 3 |
| 6 | Grandchild | F | 1 | 3 |
| 7 | Daughter | F | 2 | 1 |
| 8 | Son-in-law | M | 2 | 2 |
| 9 | Grandchild | M | 2 | 3 |
| 10 | Grandchild | F | 2 | 3 |

```
{.************************************************************************
. **************                                                ***************
. **************    Subfamily number and relationship           ***************
. **************                                                ***************
. ************************************************************************
.}

do varying N01 = 1 while N01 <= totocc (POPULATION_EDT)
  subfamily (N01) = 0;
  subrelation (N01) = 0;
enddo;

i = 0;

if TOTOCC (POPULATION_EDT) > 1 then
do varying N01 = 1 while N01 <= totocc (POPULATION_EDT)
  if RELATIONSHIP (N01) in 3,4 then    {A child or adopted
child has been found}
    {Looking for spouse for subfamily}
    do varying N02 = 1 while N02 <= totocc (POPULATION_EDT)
      if RELATIONSHIP (N02) = 8 then   {A son/daughter-in-law
is found}
        if SUBFAMILY (N01) = 0 and
           SEX (N01) <> SEX (N02) then
           i = i + 1;
          impute (SUBFAMILY(N01),i);
          impute (SUBFAMILY(N02),i);
          impute (SUBRELATION (N01),1);
          impute (SUBRELATION (N02),2);
          j = N02;
        endif;
      endif;
    enddo;
    {Looking for children in subfamily}
    do varying N02 = 1 while N02 <= totocc (POPULATION_EDT)
      if RELATIONSHIP (N02) = 8 then   {A son/daughter-in-law
is found}
        if SUBFAMILY (N02) = 0     {child is not already in a
subfamily}
          then
          impute (SUBFAMILY(N02),i);
          impute (SUBRELATION (N02),3);
        endif;
      endif;
    enddo;
  endif;
enddo;

{Following looks for subfamilies of two parents/parents-in-law
of head}

do varying N01 = 1 while N01 <= totocc (POPULATION_EDT)
  {looking for parents}
  if RELATIONSHIP (N01) = 6 then
    do varying N02 = 1 while N02 <= totocc (POPULATION_EDT)
      if RELATIONSHIP (N02) = 6 and              {This is a
parent, and apparent}
        N02 <> N01  and                  {This person is not
already idetified as a parent}
        SEX (N02) <> SEX (N01)   then    {The two people are
opposite sex to each other}
         i = i + 1;
         impute (SUBFAMILY(N01),i);
         impute (SUBFAMILY(N02),i);
         impute (SUBRELATION (N01),1);
         impute (SUBRELATION (N02),2);
      endif;
    enddo;
  endif;
enddo;
endif;
```

629. *Household and family status* 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. The *Principles and Recommendations* include 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 or male spouse)
1.2 Spouse (female head 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). The principles and recommendations divide this into two more groups – (1) living with relatives and (2) living with non-relatives

630. 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.

631. 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.

```
If SUBFAMILY = 0 then
  If totocc (POP) = 1 then HHSTATUS = 21; exit; endif;
  KIDCOUNT = count (POP where RELATIONSHIP = CHILD);
  ORELCOUNT = count (POP where RELATIONSHIP = OTHER_RELATIVE);
  If SPOUSEPTR = 0 and KIDCOUNT = 0 then HHSTATUS = 22: exit; endif;
  If SEX = MALE then
    If RELATIONSHIP = HEAD or RELATIONSHIP = SPOUSE then {Male head or male spouse}[9]
      If SPOUSEPTR = 0 then   {No spouse present}
        If KIDCOUNT = 0 then  {No children present}
          HHSTATUS = 11;
        Else   {Children present}
          HHSTATUS = 15;  {Lone father – no wife, but children}
        Endif;
      else
        HHSTATUS = 11;  {Male head or spouse}
      Endif;
    Elseif CONSENSUALPRT > 0 then
          HHSTATUS = 13;  {Partner}
    Elseif RELATIONSHIP = CHILD then
      If SPOUSEPTR > 0 then
        HHSTATUS = 16;    {Child, both parents present}
      Else
        If HEADPTR = MALE then
          HHSTATUS = 18; {Child, no mother present}
        Else
          HHSTATUS = 17; {Child, no father present}
      Endif;
    Else
      HHSTATUS = 19;    {Not a head, spouse or child}
    Endif;
  Else   {Person is female}
    If RELATIONSHIP = HEAD or RELATIONSHIP = SPOUSE then {Female head or female spouse}
      If SPOUSEPTR = 0 then   {No spouse present}
        If KIDCOUNT = 0 then  {No children present}
          HHSTATUS = 12;  {Female head or spouse}
        Else   {Children present}
          HHSTATUS = 14;  {Lone mother – no husband, but children}
        Endif;
      else
        HHSTATUS = 12;  {Female head or spouse}
      Endif;
```

[9] Note that these categories come directly from the *UN Principles and Recommendations.* The UN may reconsider the possibility of sexless heads for 2020.

```
    Elseif CONSENSUALPRT > 0 then
          HHSTATUS = 13;  {Partner}
    Elseif RELATIONSHIP = CHILD then
      If SPOUSEPTR > 0 then
        HHSTATUS = 16;    {Child, both parents present}
      Else
        If HEADPTR = MALE then
          HHSTATUS = 18; {Child, no mother present}
        Else
          HHSTATUS = 17; {Child, no father present}
      Endif;
    Else
      HHSTATUS = 19;    {Not a head, spouse or child}
    Endif;
  Else
    [Do for each subfamily as above]
  Endif;
```

632. The Principles and Recommendations also include criteria for family status. These include:

(1) whether the person is male or female in a householder-spouse pair and whether or not 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).

The determinations shown above for household status can be used for family status.

633. *Own children*. Sometimes countries want to produce information about "own children", who are the biological children of the head of household 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 head of household. Tables could show "own children" further classified as living with two parents or with one parent only. In this scheme, "own children" of the head of household living with two parents are by definition found only in married-couple families. In a subfamily, "own child" might refer to a never-married child less than 18 years old who is a son, daughter or stepchild; or an adopted child of (1) a mother in a mother-child subfamily, (2) a father in a father-child subfamily, or (3) either spouse in a married-couple subfamily. The derived variable for "own children" might be the sum of the number of own children of a particular person. That person is usually a female, but the decision must follow 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 especially to determine the characteristics of females in the labour force with own children.

```
If RELATIONSHIP = HEAD or RELATIONSHIP = SPOUSE then
   OWNCHILDREN = count (POP where RELATIONSHIP = CHILD and MARITAL_STATUS = NEVER_MARRIED and AGE < 18);
   OWN0to5 = count (POP where RELATIONSHIP = CHILD and AGE in 0:5);
   OWN6to17 = count (POP where RELATIONSHIP = CHILD and MARITAL_STATUS = NEVER_MARRIED and AGE in 6:17);
Endif;
```

634. *Parents in the house.* These data look at the characteristics of children in single-parent families compared to housing units in which both parents reside. The edit obtains this derived variable by determining how many parents of a particular 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 subfamilies to determine how many parents are living in the housing unit.

```
 {This edit determines whether one or both parents are in the
 house
    1  Both parents present
    2  Mother only present
    3  Father only present
    4  Neither parent present }
   j = 0;
 {  For head or sibling}
   if RELATIONSHIP in 1,5 then
     do varying i = 1 while i <= totocc (POPULATION)
       if RELATIONSHIP (i) = 6 then
         j = j + 1; k = SEX (i);
       endif;
     enddo;
   endif;
   { For child or stepchild}
   if RELATIONSHIP in 3,4 then
     do varying i = 1 while i <= totocc (POPULATION)
       if RELATIONSHIP (i) in 1,2 then
         j = j + 1; k = SEX (i);
       endif;
     enddo;
 endif;

 {  For grandchild}
 if RELATIONSHIP in 7 then
   do varying i = 1 while i <= totocc (POPULATION)
     if RELATIONSHIP (i) in 3,8 then
       j = j + 1; k = SEX (i);
     endif;
   enddo;
 endif;

 if j = 2 then
   $ = 1;    {both parents present}
 elseif j = 1 and k = 2 then
   $ = 2;    {mother only present}
 elseif j = 1 and k = 1 then
   $ = 3;    {father only present}
 else
   $ = 4;    {netiher parent present}
 endif;
```

635. *Employed Parents in the house.* These data look at the characteristics of children in single-parent families compared to housing units in which both parents reside and work. The edit obtains this derived variable by determining how many parents of a particular person are in the house and working, using the relationship codes and the employment status recode or other "work" variable. The program looks at the relationship code for each child and uses that information in combination with the information on subfamilies to determine how many parents are living and working in the housing unit.

```
{  This edit records whether one or both parents were working
during the last, including doing subsistence
   1  Both parents working
   2  Mother only working
   3  Father only working
   4  Neither parent working }
  j = 0;

{  For head or sibling}
  if RELATIONSHIP in 1,5 then
    do varying i = 1 while i <= totocc (POPULATION)
      if RELATIONSHIP (i) = 6 and
        WORK_WEEK (i) in 1:3 then
        j = j + 1;
        k = SEX (i);
      endif;
    enddo;
  endif;

{ as above for parents in house }

  if j = 2 then
    $ = 1;    {both parents working}
  elseif j = 1 and k = 2 then
    $ = 2;    {mother only working}
  elseif j = 1 and k = 1 then
    $ = 3;    {father only working}
  else
    $ = 4;    {netiher parent working}
  endif;
```

636. *Current year in school.* Some countries ask two questions about education: (1) if the person currently attends school, and (2) the highest level of educational attainment. In these countries, editing teams often find a mismatch between the two items when a person is actually attending school at the time of enumeration. Sometimes this 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 particular 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. 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." 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."

```
If SCHOOL = IN_SCHOOL then
  If EDUCATMT in 0 then CURRENT_GRADE = 1;
  If EDUCATMT in 1:11 then CURRENT_GRADE = EDUCATMT + 1; {Primary and secondary school}
  If EDUCATMT in 12   then CURRENT_GRADE = 21;   {First year academic or occupational college}
  If EDUCATMT in 21:24 then CURRENT_GRADE = EDUCATMT + 1; {University degree}
Else
  CURRENT_GRADE = NOTAPPL;  {not currently in school}
Endif;
```

637. *Months since last birth.* If the question of date of last birth – day, month and year or month and year of last birth – is collected, 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.

South Africa 2007

```
PROC MONTHS_LAST_BIRTH

if P45LASTYR = 2007 then
  if P45LASTMO >= 3 then
   $ = -1;
  endif;
  if P45LASTMO = 2 then
    if P45LASTDAY > 7 then
      $ = -1;
    else
      $ = 0;
    endif;
  endif;
  if P45LASTMO = 1 then
    if P45LASTDAY > 7 then
      $ = 0;
    else
      $ = 1;
    endif;
  endif;
endif;

  MONTHX = 13 - P45LASTMO;
  N01 = (2006 - P45LASTYR) * 12;
  $ = N01 + MONTHX;
  if P45LASTDAY in 1:7 then
    $ = $ + 1;
  endif;
  if $ > 99 then
    $ = 99;
  endif;
```

638. *Years since First Marriage.* Similarly, if the question of date of first marriage is collected, a recode can be created to get indirect estimates of year by year length of time since that marriage. The recode takes the year of enumeration and

then subtracts the year of first marriage. Of course, the value does not take into account marriages that dissolve, nor, since month of first marriage is not collected, do we get an exact number of years.

Southern Sudan 2008

```
PROC YEARSSINCE1STMAR
$ = 0;

{This variable holds the number of years since the first marriage}

if Q25_AGE_FIRST_MAR in 10:90 then
  $ = Q04_AGE - Q25_AGE_FIRST_MAR;
endif;
```

639. *Multiple disabilities reporting.* Traditionally, items on censuses and surveys expected a single response, as for sex, age, relationship to head, and marital status. However, recently, items with multiple entries have appeared, and must be handled during edit, and in the recodes. Multiple entries appear in several forms. The *Principles and Recommendations* now include several variables that require more consideration, and individual countries also now include items allowing more than one response. Items on disability now occur on many censuses and surveys. And item might appear as follows:

Disability. Does this person have a disability that limits his or her daily life? [Mark all that apply.]

1. Sight
2. Hearing
3. Speaking
4. Use of arm(s)
5. Use of leg(s)
6. Mental retardation

These items are presented for illustration only. As long as respondents select only one entry, a single digit is needed for entry (whether scanned or keyed); if more than 9 possibilities are included, of course, then two digits are needed.

640. However, when respondents can pick more than one entry, say, problems with both sight and hearing, then the data processors face problems. The first problem is how to capture the original information. The usual process, and the recommended process, is to capture the original information completely. For this, a single digit is reserved for all possibilities, so the information for sight will be captured or keyed as "Sight", one digit, "1" for yes, "2" for no; the next entry would "hearing", and so forth. While this procedure can take up quite a lot of space on the records, space is no longer of much consequence with larger and larger hard drives and faster and faster processing. Tallying on single items is simple. But, if multiple disability tallies are needed, either multiple tallying is required, or upfront recoding is needed to prepare for the tallying. Recodes usually occur at the end of the population or housing records; multiple disabilities within the household would appear on the housing record, multiple disabilities for one person would occur on that person's record. One recode would be whether the person actually had more than one disability: 0 for none, 1 for one disability, 2 for 2 disabilities, etc.

641. Then, the subject matter specialists and programmers must tackle the problem of whether to code all possible multiple entries, some multiple entries, or none of them. A country's use of the data will dictate the categories. For disabilities, for example, some have more specific requirements than others. For planning for access to buildings, for example, the type and degree of disability of the respondents will determine some of the planning for current changes in access and future building. So, a combination of individual entries might become a specific recode. Another method would be to make a recode for each pair of disability possibilities, with only the "top two" disabilities actually included in the recode. Each country must make this decision, based on its needs for planning and policy formation.

642. Other items would follow a similar pattern. If a country's census or survey collects information on various social grants, for example, early childhood programs, elderly programs, free or subsidized school lunches, health benefits and subsidies, etc., then combinations of these may be needed as recodes to get a full picture, both at the person or household level. For surveys, other items may require multiple entries. The recoding would depend on the needs of the specific survey results and the country short-term and long-term requirements. It is important to note that the programming required to develop recodes is relatively straight forward, but is not usually taught in introductory programming courses. Countries doing this type of recoding for the first time might want to consider working closely with personnel from other country offices who have already done this type of recoding, or with experts in this field.

643. *Characteristics of household head or householder.* As above for the housing characteristics, sometimes countries need to cross-tabulate characteristics of other household members by the characteristics of the head, but they do not want to bother finding the head – or if using CSPro for the crosstab, doing an extra operation to make the cross. These crosses

might by age and sex or ethnicity or educational attainment. So, the characteristics of the head would be put on each person's record:

```
HEADS_SEX = SEX (HEADPTR);
HEADS_AGE = AGE (HEADPTR);
HEADS_ETHNICITY = ETHNICITY (HEADPTR);
HEADS_EDUCATION = EDUCATIONAL_ATTAINMENT (HEADPTR);
```

## CONCLUSIONS

644. This volume of the Africa Census Processing Handbooks has covered census and survey computer editing. Computer Processing in the early years of census was done on Mainframe computers. At that time, programmers and subject-matter specialists had an almost adversarial relationship in many cases, with the computer becoming something of a "Black Box": the programmers knew how the program worked but often the subject matter specialists did not. In fact, many times the subject matter specialists simply gave the questionnaire to the programmers and said "do this" and waited for the results.

645. As this handbook has shown, much has changed in recent years. The hardware changed. Now, with microcomputers and laptops, size of hard drives is larger and easier to access. The software changed too. As we have seen in this handbook, pseudo-code can be written that both subject matter specialists and programmers can read, and use. Hence, the subject matter specialists can make sure that they get in the edit what they think they should get. And, both subject specialists and programmers can assess the quantity of programming needed, as well as the quality.

646. This handbook uses examples from previous African censuses and prototype materials to illustrate the various edits. It is meant to be a "cookbook", allowing easy understanding of editing concepts and easy use of the elements and programs. It is important to remember that not all programs have the error listings and summaries, but the examples do serve to show the various ideas involved in current editing philosophy.

647. The first volume in this series is the Africa Census Data Capture Handbook, and showed how to capture census or survey data for computer editing. This volume covers the actual editing. And, the third volume, the Africa Census Tabulation Handbook, covers tabulations and dissemination. These volumes should assist countries as they work on their own census processing in the 2010 Round Censuses.

# APPENDIX

## SAMPLE DISABILITY EDIT FOR MULTIPLE DISABILITIES

```
PROC Q14_DISABILITY
{ ************************************************************************************************
  ************************************************************************************************
                               Disability
  ************************************************************************************************
  ************************************************************************************************}
errmsg (" ******************************** "), summary;
errmsg (" *******  Disability   ************ "), summary;
errmsg (" ******************************** "), summary;
{ ********************************************************************
  Q14. Does (Name) have any difficult  in moving, seeingm hearing,
       speaking, or learning?
       (Mark all that apply)
       1 Limited use of legs
       2 Loss of leg(s)
       3 Limited use of arms
       4 Loss of arm(s)
       5 Difficulty in hearing
       6 Deaf
       7 Difficulty in seeing
       8 Blind
       9 Difficulty in speaking
       10 Mute
       11 Mental disability
       12 No disability
       13 Don't know
  ********************************************************************}

{---------------------------------------(4)-------------------------------------------}
{4.       Variable Q 14 : Disability                                                  }
{    A. If   any one of codes 1-11 = 1                                               }
{        Do nothing                                                                  }
{    B. If    codes 1-11 (all of them)= blank/invalid and No Disability = 1          }
{        make codes 1-11 (all of them)= 2                                            }
{    C. If    codes 1-11 (all of them), No Disability
= blank/invalid and Don't know = 1}
{        make codes 1-11 (all of them)= 3                                            }
{    D. If    codes 1-11 (all of them), No Disability and Don't know = blank/invalid }
{        make codes 1-11 (all of them)= 8 Not Reported                               }
{        make No Disability = 8 Not Reported                                         }
{        make Don't know    = 8 Not Reported                                         }
{    E. If    codes 1-11 (at list one of them)=1 and  No Disability= blank/invalid   }
{        make No Disability = 1 Not Reported                                         }
{    F. If    codes 1-11 (all of them)= blank/invalid and No Disability not = 2      }
{        make No Disability = 2                                                      }
{    G. If    No Disability =1 or 2 or 8                                             }
{        make Don't know = 8                                                         }
{-------------------------------------------------------------------------------------}
{Note: need more clarification}
{DB - To be honest, it does not make much sense to me either as it stands }


  if Q14A in notappl and Q14B in notappl and Q14C in notappl and Q14D in notappl and Q14E in notappl and Q14F in notappl and
     Q14G in notappl and Q14H in notappl and Q14J in notappl and Q14K in notappl and Q14L in notappl then
      {Case 1. where none of the disability items are marked, but 'no disability' is}
       if Q14M = 1 then
          if Q14A <> 2 then Q14A = 2; endif;
          [Same for the others]
          if Q14N <> 2 then Q14N = 2; endif;  {This one because nothing is known, so we choose 'no disability' because it is marked}
          WriteCurrentCase(1);
          errmsg ("*P14-1A* Case 1: everything blank, but no disability filled") denom = PERSON_COUNT summary;
          write ("*P14-1A* Case 1: everything blank, but no disability filled, pn = %2d",PERSON_NUMBER);
       else
         {Case 1A. where none of the disability items are marked, but 'no disability' is NOT}
          if Q14A <> 2 then Q14A = 2; endif;
          [Same for the others]
          if Q14N <> 1 then Q14N = 1; endif;  {This one because nothing is known, so we choose 'Don't know'}
          WriteCurrentCase(1);
          errmsg ("*P14-1B* Case 1: everything blank, but no disability not filled") denom = PERSON_COUNT summary;
          write ("*P14-1B* Case 1: everything blank, but no disability not filled, pn = %2d",PERSON_NUMBER);
      endif;

  {Case 2: at least one disability}
  elseif Q14A in 1 or Q14B in 1 or Q14C in 1 or Q14D in 1 or Q14E in 1 or Q14F in 1 or
         Q14G in 1 or Q14H in 1 or Q14J in 1 or Q14K in 1 or Q14L in 1 then
       if Q14A in 1:2 and Q14B in 1:2 and Q14C in 1:2 and Q14D in 1:2 and Q14E in 1:2 and Q14F in 1:2 and
          Q14G in 1:2 and Q14H in 1:2 and Q14J in 1:2 and Q14K in 1:2 and Q14L in 1:2 and Q14M = 2 and Q14N = 2 then
         if Q14A = 1 and Q14B = 1 then Q14A = 2; endif; {Pain in legs and loss of legs}
         if Q14C = 1 and Q14D = 1 then Q14C = 2; endif; {Pain in arms and loss of arms}
         if Q14E = 1 and Q14F = 1 then Q14E = 2; endif; {Hearing}
         if Q14G = 1 and Q14H = 1 then Q14G = 2; endif; {Sight}
         if Q14J = 1 and Q14K = 1 then Q14J = 2; endif; {Speaking}
       else
{    A. If   any one of codes 1-11 = 1                                               }
{        Do nothing                                                                  }
     {Edit 4A. do nothing, but we will make the others 'no' just in case}
        if Q14A <> 1 then Q14A = 2; endif;
          [Same for the others]
        if Q14A = 1 and Q14B = 1 then Q14A = 2; endif; {Pain in legs and loss of legs}
```

```
        if Q14C = 1 and Q14D = 1 then Q14C = 2; endif; {Pain in arms and loss of arms}
        if Q14E = 1 and Q14F = 1 then Q14E = 2; endif; {Hearing}
        if Q14G = 1 and Q14H = 1 then Q14G = 2; endif; {Sight}
        if Q14J = 1 and Q14K = 1 then Q14J = 2; endif; {Speaking}
          WriteCurrentCase(1);
          errmsg ("*P14-2* Case 2: at least one disability") denom = PERSON_COUNT summary;
          write ("*P14-2* Case 2: at least one disability, %2d",PERSON_NUMBER);
        endif;
{    B. If    codes 1-11 (all of them)= blank/invalid and No Disability = 1            }
{        make codes 1-11 (all of them)= 2

  {Case 3: no disability only is filled}                                              }
  elseif (Q14A <> 1 and Q14B <> 1 and Q14C <> 1 And Q14D <> 1 and Q14E <> 1 and Q14F <> 1 or
        Q14G <> 1 and Q14H <> 1 and Q14J <> 1 and Q14K <> 1 and Q14L <> 1) and   {no type of disability reported}
        Q14M = 1   {No disability}
        then
         if Q14A in 2 and Q14B in 2 and Q14C in 2 and Q14D in 2 and Q14E in 2 and Q14F in 2 and
           Q14G in 2 and Q14H in 2 and Q14J in 2 and Q14K in 2 and Q14L in 2 and Q14M = 1 and Q14N = 2 then
         else
           if Q14A <> 2 then Q14A = 2; endif;
          [Same for the others]
          errmsg ("*P14-3* Case 3: everything not a particulat disability") denom = PERSON_COUNT summary;
         endif;

       {Case 4: Some but not all disabilities marked}
  elseif Q14A in 2 or Q14B in 2 or Q14C in 2 or Q14D in 2 or Q14E in 2 or Q14F in 2 or
         Q14G in 2 or Q14H in 2 or Q14J in 2 or Q14K in 2 or Q14L in 2 then
         if (Q14A in 2 and Q14B in 2 and Q14C in 2 and Q14D in 2 and Q14E in 2 and Q14F in 2 and
             Q14G in 2 and Q14H in 2 and Q14J in 2 and Q14K in 2 and Q14L in 2 and Q14M = 1 and Q14N = 2) or {Fact of No disability
checked}
            (Q14A in 2 and Q14B in 2 and Q14C in 2 and Q14D in 2 and Q14E in 2 and Q14F in 2 and
             Q14G in 2 and Q14H in 2 and Q14J in 2 and Q14K in 2 and Q14L in 2 and Q14M = 2 and Q14N = 1) {DOn't know ONLY checked}
              then
         else
           if Q14A <> 2 then Q14A = 2; endif;
          [Same for the others]
           errmsg ("*P14-4* Case 4: some are not disability, but not all") denom = PERSON_COUNT summary;
         endif;
{    C. If    codes 1-11 (all of them), No Disability = blank/invalid and Don't know = 1}
{        make codes 1-11 (all of them)= 3                                             }

{    D. If    codes 1-11 (all of them), No Disability and Don't know = blank/invalid    }
{        make codes 1-11 (all of them)= 8 Not Reported                                 }
{        make No Disability = 8 Not Reported                                          }
{        make Don't know    = 8 Not Reported                                          }
{    E. If    codes 1-11 (at list one of them)=1 and  No Disability= blank/invalid      }
{        make No Disability = 1 Not Reported                                          }
{    F. If    codes 1-11 (all of them)= blank/invalid and No Disability not = 2         }
{        make No Disability = 2                                                       }
{    G. If    No Disability =1 or 2 or 8                                              }
{        make Don't know = 8                                                          }
   else
     if Q14A = 1 and Q14B = 1 then Q14A = 2; endif; {Pain in legs and loss of legs}
     if Q14C = 1 and Q14D = 1 then Q14C = 2; endif; {Pain in arms and loss of arms}
     if Q14E = 1 and Q14F = 1 then Q14E = 2; endif; {Hearing}
     if Q14G = 1 and Q14H = 1 then Q14G = 2; endif; {Sight}
     if Q14J = 1 and Q14K = 1 then Q14J = 2; endif; {Speaking}
          WriteCurrentCase(1);
          errmsg ("*P14-5* Case 5: disabilities ok, but duplicates") denom = PERSON_COUNT summary;
          write ("*P14-5* Case 5: disabilities ok, but duplicates, pn = %2d",PERSON_NUMBER);

  endif;


  N01 = 0;
  if Q14A = 1 then N01 = N01 + 1; endif;
          [Same for the others]

  if N01 = 1 then
    if Q14A = 1 then DISABILITY2 =  1; endif;
    if Q14B = 1 then DISABILITY2 =  2; endif;
    if Q14C = 1 then DISABILITY2 =  3; endif;
    if Q14D = 1 then DISABILITY2 =  4; endif;
    if Q14E = 1 then DISABILITY2 =  5; endif;
    if Q14F = 1 then DISABILITY2 =  6; endif;
    if Q14G = 1 then DISABILITY2 =  7; endif;
    if Q14H = 1 then DISABILITY2 =  8; endif;
    if Q14J = 1 then DISABILITY2 =  9; endif;
    if Q14K = 1 then DISABILITY2 = 10; endif;
    if Q14L = 1 then DISABILITY2 = 11; endif;
  else
    if N01 = 0 then  DISABILITY2 = 12;  {No disability}
    else             DISABILITY2 = 14;  {Multiple disabilities}
    endif;
  endif;

if N01 = 2 then
  if (Q14A = 1 or Q14B = 1) then {legs}
    if (Q14C = 1 or Q14D = 1) then  {arms}    DISABILITY2 = 21; endif;
    if (Q14E = 1 or Q14F = 1) then  {hearing} DISABILITY2 = 22; endif;
    if (Q14G = 1 or Q14H = 1) then {seeing}   DISABILITY2 = 23; endif;
    if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 24; endif;
    if Q14L = 1 then {mental disability}      DISABILITY2 = 25; endif;
  endif;

  if (Q14C = 1 or Q14D = 1) then  {arms}
    if (Q14E = 1 or Q14F = 1) then  {hearing} DISABILITY2 = 26; endif;
    if (Q14G = 1 or Q14H = 1) then {seeing}   DISABILITY2 = 27; endif;
```

```
    if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 28; endif;
    if Q14L = 1 then {mental disability}      DISABILITY2 = 29; endif;
  endif;

  if (Q14E = 1 or Q14F = 1) then  {hearing}
    if (Q14G = 1 or Q14H = 1) then {seeing}   DISABILITY2 = 30; endif;
    if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 31; endif;
    if Q14L = 1 then      {mental disability} DISABILITY2 = 32; endif;
  endif;

  if (Q14G = 1 or Q14H = 1) then {seeing}
    if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 33; endif;
    if Q14L = 1 then      {mental disability} DISABILITY2 = 34; endif;
  endif;

  if (Q14J = 1 or Q14K = 1) then {speaking}
    if Q14L = 1 then    {mental disability}   DISABILITY2 = 35; endif;
  endif;
endif;

if N01 = 3 then
  if (Q14A = 1 or Q14B = 1) then {legs}
    if (Q14C = 1 or Q14D = 1) then  {arms}
      if (Q14E = 1 or Q14F = 1) then  {hearing} DISABILITY2 = 41; endif;
      if (Q14G = 1 or Q14H = 1) then {seeing}   DISABILITY2 = 42; endif;
      if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 43; endif;
      if Q14L = 1 then      {mental disability} DISABILITY2 = 44; endif;
    endif;
    if (Q14E = 1 or Q14F = 1) then  {hearing}
      if (Q14G = 1 or Q14H = 1) then {seeing}   DISABILITY2 = 45; endif;
      if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 46; endif;
      if Q14L = 1 then     {mental disability}  DISABILITY2 = 47; endif;
    endif;
    if (Q14G = 1 or Q14H = 1) then {seeing}
      if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 48; endif;
      if Q14L = 1 then      {mental disability} DISABILITY2 = 49; endif;
    endif;
    if (Q14J = 1 or Q14K = 1) then {speaking}
      if Q14L = 1 then     {mental disability}  DISABILITY2 = 50; endif;
    endif;

    if (Q14C = 1 or Q14D = 1) then  {arms}
      if (Q14E = 1 or Q14F = 1) then  {hearing}
        if (Q14G = 1 or Q14H = 1) then {seeing}   DISABILITY2 = 51; endif;
        if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 52; endif;
        if Q14L = 1 then      {mental disability} DISABILITY2 = 53; endif;
      endif;
      if (Q14G = 1 or Q14H = 1) then {seeing}
        if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 54; endif;
        if Q14L = 1 then      {mental disability} DISABILITY2 = 55; endif;
      endif;
      if (Q14J = 1 or Q14K = 1) then {speaking}
        if Q14L = 1 then    {mental disability}   DISABILITY2 = 56; endif;
      endif;
    endif;
  endif;

  if (Q14E = 1 or Q14F = 1) then  {hearing}
    if (Q14G = 1 or Q14H = 1) then {seeing}
      if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 57; endif;
      if Q14L = 1 then      {mental disability} DISABILITY2 = 58; endif;
    endif;
    if (Q14J = 1 or Q14K = 1) then {speaking}
      if Q14L = 1 then      {mental disability} DISABILITY2 = 59; endif;
    endif;
  endif;

  if (Q14G = 1 or Q14H = 1) then {seeing}
    if (Q14J = 1 or Q14K = 1) then {speaking}
      if Q14L = 1 then    {mental disability}   DISABILITY2 = 60; endif;
    endif;
  endif;
endif;

if N01 = 4 then
  if (Q14A = 1 or Q14B = 1) then {legs}
    if (Q14C = 1 or Q14D = 1) then  {arms}
      if (Q14E = 1 or Q14F = 1) then  {hearing}
        if (Q14G = 1 or Q14H = 1) then {seeing}   DISABILITY2 = 61; endif;
        if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 62; endif;
        if Q14L = 1 then      {mental disability} DISABILITY2 = 63; endif;
      endif;
      if (Q14G = 1 or Q14H = 1) then {seeing}
        if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 64; endif;
        if Q14L = 1 then      {mental disability} DISABILITY2 = 65; endif;
      endif;
      if (Q14J = 1 or Q14K = 1) then {speaking}
        if Q14L = 1 then     {mental disability}  DISABILITY2 = 66; endif;
      endif;
    endif;

    if (Q14E = 1 or Q14F = 1) then  {hearing}
      if (Q14G = 1 or Q14H = 1) then {seeing}
        if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 67; endif;
        if Q14L = 1 then     {mental disability}  DISABILITY2 = 68; endif;
      endif;
      if (Q14J = 1 or Q14K = 1) then {speaking}
        if Q14L = 1 then     {mental disability}  DISABILITY2 = 69; endif;
```

```
      endif;
    endif;
  endif;

  if (Q14C = 1 or Q14D = 1) then {arms}
    if (Q14E = 1 or Q14F = 1) then  {hearing}
      if (Q14G = 1 or Q14H = 1) then {seeing}
        if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 70; endif;
        if Q14L = 1 then     {mental disability}  DISABILITY2 = 71; endif;
      endif;
      if (Q14J = 1 or Q14K = 1) then {speaking}
        if Q14L = 1 then     {mental disability}  DISABILITY2 = 72; endif;
      endif;
    endif;
  endif;

  if (Q14E = 1 or Q14F = 1) then  {hearing}
    if (Q14G = 1 or Q14H = 1) then {seeing}
      if (Q14J = 1 or Q14K = 1) then {speaking}
        if Q14L = 1 then    {mental disability}   DISABILITY2 = 73; endif;
      endif;
    endif;
  endif;
endif;

if N01 = 5 then
  if (Q14A = 1 or Q14B = 1) then {legs}
    if (Q14C = 1 or Q14D = 1) then  {arms}
      if (Q14E = 1 or Q14F = 1) then  {hearing}
        if (Q14G = 1 or Q14H = 1) then {seeing}
          if (Q14J = 1 or Q14K = 1) then {speaking} DISABILITY2 = 81; endif;
          if Q14L = 1 then      {mental disability} DISABILITY2 = 82; endif;
        endif;
      endif;
    endif
  endif;

    if (Q14C = 1 or Q14D = 1) then  {arms}
      if (Q14E = 1 or Q14F = 1) then  {hearing}
        if (Q14G = 1 or Q14H = 1) then {seeing}
          if (Q14J = 1 or Q14K = 1) then {speaking}
            if Q14L = 1 then   {mental disability}  DISABILITY2 = 83; endif;
          endif;
        endif;
      endif;
    endif;
endif;

if N01 = 6 then DISABILITY2 = 91; endif;
```