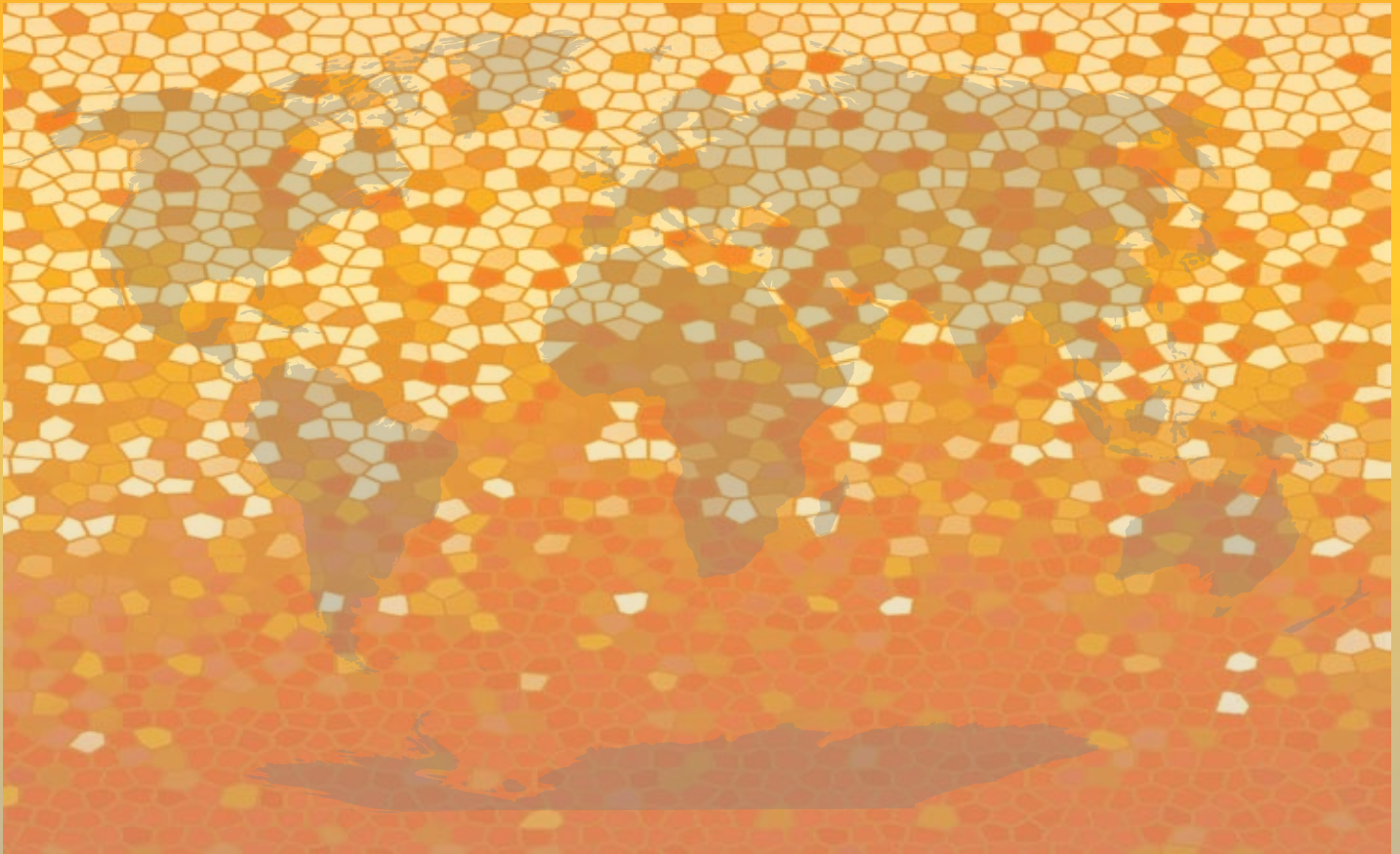


HANDBOOK ON POPULATION AND HOUSING CENSUS EDITING REVISION 1



United Nations

Department of Economic and Social Affairs
Statistics Division

Studies in Methods

Series F No. 82/Rev.1

Handbook on Population and Housing Census Editing

Revision 1



United Nations
New York, 2010

Department of Economic and Social Affairs

The Department of Economic and Social Affairs of the United Nations Secretariat is a vital interface between global policies in the economic, social and environmental spheres and national action. The Department works in three main interlinked areas: (i) it compiles, generates and analyses a wide range of economic, social and environmental data and information on which States Members of the United Nations draw to review common problems and to take stock of policy options; (ii) it facilitates the negotiations of Member States in many intergovernmental bodies on joint courses of action to address ongoing or emerging global challenges; and (iii) it advises interested Governments on the ways and means of translating policy frameworks developed in United Nations conferences and summits into programmes at the country level and, through technical assistance, helps build national capacities.

Note

The designations employed and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The term “country” as used in this publication also refers, as appropriate, to territories or areas.

The designations “developed regions” and “developing regions” are intended for statistical convenience and do not necessarily express a judgement about the stage reached by a particular country or area in the development process.

Symbols of United Nations documents are composed of capital letters combined with figures.

ST/ESA/STAT/SER.F/82/Rev.1

UNITED NATIONS PUBLICATION

Sales No. E.09.XVII.11

ISBN 978-92-1-161530-2

Copyright © United Nations, 2010
All rights reserved
Printed in United Nations, New York

Preface

The United Nations has, over the years, issued a series of handbooks and technical reports intended to assist countries in planning and carrying out improved and cost-effective population and housing censuses. These handbooks and reports have been reviewed from time to time and repeated to reflect new developments and emerging issues in census-taking. The present publication is part of a series of handbooks that have been developed to assist countries in preparing for the 2000 and future rounds of censuses. The other handbooks in the series include:

- (a) *Handbook on Census Management for Population and Housing Censuses*, Series F, No. 83 (United Nations Publication, Sales No. E.00.XVII.15);
- (b) *Handbook on Geographic Information Systems and Digital Mapping*, Series F, No. 79 (United Nations publication, Sales No. E.00.XVII.12);
- (c) *Guide for the Collection of Economic Characteristics in Population Censuses* (forthcoming).

The *Principles and Recommendations for Population and Housing Censuses, Revision 2* (United Nations, 2008) examines the merits of a quality control and improvement system at an early stage of the census, which is crucial to the success of overall census operations, and the importance of the editing plan, which should be developed as part of the overall census programme and integrated with other census plans and procedures. Users of the *Handbook on Population and Census Editing* will find it particularly helpful to refer to the *Principles and Recommendations*, which provides considerable background information for the editing procedures outlined in chapters III, IV and V.

The purpose of the present publication is to provide countries with a broad overview of census and survey data editing methodology and to provide information for concerned 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, promote communication among subject matter and data processing specialists, and document the activities undertaken in the current census or survey in order to avoid duplicating effort during the next census or survey.

The *Handbook* reviews the advantages and disadvantages of manual and computer-assisted editing. In large censuses, manual correction is rarely economically feasible. The conditions for such corrections are usually specified in specially-designed computer programs that perform automatic error scrutiny and imputation based on other information for that person or household or for other persons or households. The bulk of the *Handbook* deals with the automatic correction of data.

Computer edits play an important role in error detection and correction. At the computer editing stage, detailed consistency checks can be established in consultation with subject matter specialists. The errors detected can be corrected either by reference to original schedules or automatically. While automatic editing speeds up data processing, it demands careful control over the quality of incoming data.

The publication is divided into an introduction and five chapters. The introduction describes the census process and the various types of errors that occur in a census. Chapter I covers the basics of census editing. Chapters II to V present procedures and techniques for editing census data at various stages of processing. Technical consideration, particularly those pertinent to programming, are covered in the annexes.

Although this *Handbook* focuses on editing for population and housing censuses, many of the concepts and techniques also apply to survey operations.

The contribution of Michael J. Levin, Harvard Center for Population and Development Studies, to the drafting of the *Handbook*, is gratefully acknowledged. Appreciation is also extended to Michael Bankier, Wesley Benjamin, Marcel Bureau, Sean Crowe, Sylvain Delisle and Darryl Janes from Statistics Canada, who reviewed the draft and provided valuable comments with respect to finalizing the publication.

Contents

	<i>Page</i>
Preface.....	iii
Introduction.....	1
A. Purpose of the <i>Handbook</i>	1
B. The census process	1
C. Errors in the census process.....	2
1. Coverage errors.....	2
2. Content errors	3
(a) Errors in questionnaire design.....	3
(b) Enumerator errors	3
(c) Respondent errors.....	3
(d) Coding errors.....	4
(e) Data entry errors.....	4
(f) Errors in computer editing	4
(g) Errors in tabulation	4
D. Structure of the <i>Handbook</i>	5
I. Editing in censuses and surveys	7
A. Editing in historical review	7
B. The editing team	8
C. Editing practices: edited versus unedited data	9
D. The basics of editing	11
1. How over-editing is harmful.....	13
(a) Timeliness	13
(b) Finances	14
(c) Distortion of true values	14
(d) A false sense of security	14
2. Treatment of unknowns.....	14
3. Spurious changes	14
4. Determining tolerances.....	15
5. Learning from the editing process.....	15
6. Quality assurance	16
7. Costs of editing.....	16
8. Imputation.....	16
9. Archiving.....	17
II. Editing applications.....	19
A. Coding considerations	20
B. Manual versus automatic correction	24
C. Guidelines for correcting data.....	26
D. Validity and consistency checks	29
1. Top-down editing approach	29
2. Multiple-variable editing approach	30

	<i>Page</i>
E. Methods of correcting and imputing data.	33
1. Static imputation or “cold deck” technique.	33
2. Dynamic imputation or “hot deck” technique.	34
3. Dynamic imputation (hot deck) issues.	37
(a) Geographical considerations.	37
(b) Use of related items.	37
(c) How the order of the variables affects the matrices.	37
(d) Complexity of the imputation matrices.	38
(e) Imputation matrix development.	38
(f) Standardized imputation matrices.	39
(g) When dynamic imputation is not used.	41
(h) How big should the imputation matrices be?	43
(i) Problems that arise when the imputation matrix is too big.	43
(ii) Understanding what the imputation matrix is doing.	43
(iii) Problems that arise when the imputation matrix is too small.	43
(iv) Items that are difficult for imputation matrices.	44
4. Checking imputation matrices.	44
(a) Setting up the initial static matrix.	44
(b) Messages for errors.	44
(c) Custom-made error listings.	45
(d) How many times to run the edit?	46
5. Imputation flags.	47
F. Other editing systems.	49
III. Structure edits.	51
A. Geography edits.	52
1. Location of living quarters (locality) (P1H).	52
2. Urban and rural residence (P1L).	52
B. Coverage checks.	53
1. De facto and de jure enumeration.	53
2. Hierarchy of households and housing units.	53
3. Fragments of questionnaires.	54
C. Structure of housing records.	54
D. Correspondence between housing and population records.	54
1. Vacant and occupied housing.	54
(a) Choosing to leave a housing unit vacant.	55
(b) Revisiting the housing unit several times to complete questionnaires.	55
(c) Substituting another housing unit for missing persons.	55
2. Duplicate households and housing units.	55
3. Missing households and housing units.	55
4. Correspondence between the number of occupants and the sum of the occupants.	56
(a) When the number of occupants is greater than the sum of the occupants.	56
(b) Checking numbers of persons by sex.	56

	<i>Page</i>
(c) Sequence numbering	57
5. Correspondence between occupants and type of building/ household	57
E. Duplicate records	57
F. Special populations	58
1. Persons in collectives	58
(a) When collectives are a different record type	58
(b) When a variable distinguishes collectives from other records	58
(c) When the “type of collective” code is missing	58
(d) When the collective code is present, but all of the persons are related	59
(e) Distinguishing various types of collectives	59
2. Groups difficult to enumerate	59
(a) Seasonal migrants	59
(b) Homeless persons	59
(c) Nomads and persons living in areas to which access is difficult	59
(d) Civilian residents temporarily absent from the country ...	60
(e) Civilian foreigners who do not cross a frontier daily and are in the country temporarily, including undocumented persons, or transients on ships in harbour at the time of the census	60
(f) Refugees	60
(g) Military, naval and diplomatic personnel and their families located outside the country and foreign military, naval and diplomatic personnel and their families located in the country	60
(h) Civilian foreigners who cross a frontier daily to work in the country	61
(i) Civilian residents who cross a frontier daily to work in another country	61
(j) Merchant seamen and fishermen resident in the country but at sea at the time of the census (including those who have no place of residence other than their quarters aboard ship)	61
G. Determining head of household and spouse	61
1. Editing the head of household variable	61
(a) The order of the relationships	61
(b) When the head is not the first person	62
(i) Assigning a pointer for the head’s record	63
(ii) Making the first person the head	63
(iii) Reassigning relationship codes to make the first person the head	63
(c) More than one head	64
(d) No head	64
2. Editing the spouse	64
(a) When exactly one spouse is found in monogamous societies	64

	<i>Page</i>
(b) When more than one spouse is found in monogamous societies.	64
(c) Spouses in polygamous societies	65
(d) Other characteristics of heads and spouses	65
H. Age and birth date	65
1. When date of birth is present, but age is not	65
2. When the age and date of birth disagree	65
I. Counting invalid entries.	65
IV. Edits for population items	67
A. Demographic characteristics	68
1. Relationship	68
(a) Relationship edits	69
(b) When the head must appear first	69
(c) When the relationships are coded upside down	69
(d) When polygamous spouses are present.	69
(e) When multiple parents appear	70
(f) When censuses collect sex-specific relationships	70
(g) When relationship and marital status do not match.	70
2. Sex	70
(a) When the sex code is valid but the head and spouse are of the same sex	71
(b) When a male has fertility information or an adult female does not	71
(c) When the sex code is invalid and a spouse is present.	71
(d) When the sex code for spouse is invalid	71
(e) When the sex code is invalid and female information is present	71
(f) When the sex code is invalid and this person is spouse's husband	72
(g) When the sex code is invalid and there is insufficient information to determine sex	72
(h) Note on imputed sex ratios	72
3. Birth date and age	72
(a) Age and date of birth	72
(b) Relationship between date of birth and age	73
(c) When calculated age falls above the upper limit	73
(d) Age edit.	74
(e) Age edit when the head of household and spouse are present.	74
(f) Age edit for head when the head's spouse is absent, but child is present.	74
(g) Age edit for head when head's parent is present	75
(h) Age edit for head when head's grandchild is present	75
(i) Age edit for head when no other ages are available.	75
(j) Age edit for spouse when head's age is already determined.	75
(k) Age edit for other married couples in the household when the age of one of the persons is known	76
(l) Age edit for child when head's age is already determined ..	76

	<i>Page</i>
(m) Age edit for parent when head's age is already determined.....	76
(n) Age edit for grandchild when head's age is already determined.....	76
(o) Age edit for all other persons	77
4. Marital status	77
(a) Marital status edit.....	77
(b) Marital status assignment when dynamic imputation is not used.....	77
(c) Marital status assignment when dynamic imputation is used	78
(d) Spouse should be married	78
(e) Spouse of a married-couple pair.....	78
(f) If spouse, head should be married	78
(g) Head, no spouse, without children	78
(h) If all else fails, impute.....	78
(i) Relationship of age to marital status for young people.....	78
5. Age at first marriage.....	79
(a) Age at marriage for never-married persons should be blank.....	79
(b) Ever-married persons should have an entry.....	79
6. Fertility: children ever born and children surviving	79
(a) Fertility items collected.....	80
(b) General rules for the fertility edit.....	80
(c) Relationship between children born and children surviving.....	81
(d) Edit when only children ever born is reported	81
(e) Edit when children ever born and children surviving are reported	81
(f) Edit when children ever born, children surviving, and children who died are reported.....	83
(i) When all three items are reported.....	83
(ii) When two items are reported.....	84
(iii) When one item is reported	84
(iv) When none of the items are reported	84
(g) Edit when both children ever born, children living at home, children living away and children who died are reported	84
(i) When all four items are reported.....	84
(ii) When three of the four items are reported.....	85
(iii) When two of the four items are reported	85
(iv) When only one item is reported.....	86
(v) When none of the items are reported	87
(h) Special case of five or more items.....	87
(i) Importance of a single donor source for all fertility items..	87
(j) Relationship of own children to children in the house and children surviving	88
7. Fertility: date of birth of last child born alive and births in the 12 months before the census	88

	<i>Page</i>
8. Fertility: age at first birth	89
9. Mortality	90
(a) Age and sex of the deceased	90
(b) Cause of death	90
(c) Maternal mortality	90
(d) Infant mortality	91
10. Maternal or paternal orphanhood (P5G) and mother's line number	91
B. Migration characteristics	92
1. Place of birth	92
(a) Relationship of entries for country of birth and years lived in district	93
(b) Assigning "unknown" for invalid entries for birthplace ...	93
(c) Using static imputation for birthplace	93
(d) Using dynamic imputation for birthplace	93
(e) Assigning birthplace when a person's mother is present ...	93
(f) Assigning birthplace for child of head	94
(g) Assigning birthplace for child, but not of head	94
(h) Assigning birthplace for adult females with husband	94
(i) Assigning birthplace for adult females with no husband ...	94
(j) Assigning birthplace for males	94
2. Citizenship	95
(a) Citizenship edit	95
(b) Relationship of ethnicity/race to citizenship	95
(c) Relationship of naturalization to citizenship	95
(d) Relationship of duration of residence to citizenship	95
3. Duration of residence	96
(a) Edit for duration of residence	96
(b) De facto/de jure residence and duration	96
(c) Relation of age to duration of residence	96
(d) Relation of birthplace to duration	96
(e) For persons who have always lived here	96
(f) Person's duration from mother's duration	97
(g) Person's duration from child's duration	97
(h) Person's duration when no other information is available	97
4. Place of previous residence	98
(a) Previous residence edit	98
(b) Previous residence when boundaries have changed	98
(c) When person has not moved since birth	98
(d) Use of other persons in unit	98
(e) No appropriate other person for previous residence	98
5. Place of residence at a specified date in the past	99
6. Year of arrival	99
(a) Relation of age to year of arrival	99
(b) Relation of birthplace to year of arrival	100
(c) For persons who have always lived here	100
(d) Person's year of arrival from mother's year of arrival	100
(e) Child's year of arrival from head's year of arrival	100

	<i>Page</i>
(f) Person's year of arrival when no other information is available	101
7. Relationship of duration of residence to year of arrival	101
8. Usual residence	102
C. Social characteristics	102
1. Ability to read and write (literacy) (P6A)	102
2. School attendance (P6B)	103
(a) School attendance edit	103
(b) Full-time or part-time enrolment	103
(c) Consistency between school attendance and economic activity	104
(d) Assignment for invalid or inconsistent entries for school attendance	104
3. Educational attainment (highest grade or level completed)	104
(a) Edit for educational attainment	104
(b) Minimum age for educational attainment	104
(c) Relationship of age to educational attainment	104
4. Field of education and educational qualifications	105
5. Religion	105
(a) Religion edit	105
(b) No religion for head of household, but religion present for someone else in the unit	105
(c) No religion for head, or for anyone else in unit	105
(d) For person other than head, without religion	106
6. Language	106
(a) Language edit	106
(b) Language edits: head of household	106
(c) Language edits: persons other than head of household	106
(d) Language edits: use of ethnic origin or birthplace	106
(e) Language edit: mother tongue	106
(f) Language edits: ability to speak a designated language	107
7. Ethnicity and indigenous peoples	107
(a) Ethnicity edit	107
(b) Ethnicity edit: for head of household	107
(c) Ethnicity edit: persons other than head of household	108
(d) Ethnicity edit: use of language and birthplace	108
8. Disability	108
(a) Disability census questions	108
(b) Disability edit	109
(c) Multiple disabilities	109
(d) Cause of disability edit	109
D. Economic characteristics	109
1. Activity status	110
(a) Categories related to activity status	110
(i) Unemployed population	110
(ii) Seeking work	111
(iii) Not currently active	111
(iv) Why not looking for work	111
(b) Editing for economic activity status	112

	<i>Page</i>
(i) Employed persons	112
(ii) Economic activity of unemployed persons	112
(iii) Economic activity of students and retired persons	112
(iv) When economic activity is not valid and employed variables are reported	112
(v) When economic activity is not valid and the unemployed variables are reported	112
(vi) When economic activity is not valid and none of the economic variables are reported	113
2. Time worked	113
3. Occupation	113
4. Industry	114
5. Status in employment	114
6. Income	115
7. Institutional sector	115
8. Employment in the informal sector	116
9. Place of work	116
V. Housing edits	117
A. Core and additional topics	119
1. Living quarters—type of (core topic)	119
2. Location of living quarters (core topic)	120
3. Occupancy status (core topic)	120
4. Ownership—type of (core topic)	121
5. Rooms—number of (core topic)	121
6. Bedrooms—number of (additional topic)	122
7. Useful floor space (additional topic)	122
8. Water supply system (core topic)	122
9. Drinking water—main source of (core topic)	123
10. Toilet—type of (core topic)	124
11. Sewage disposal (core topic)	124
12. Bathing facilities (core topic)	124
13. Kitchen—availability of (core topic)	125
14. Fuel used for cooking (core topic)	126
15. Lighting and/or electricity—type of (core topic)	126
16. Solid waste disposal—main type of (core topic)	127
17. Heating—type and energy used for (additional topic)	127
18. Hot water—availability of (additional topic)	128
19. Piped gas—availability of (additional topic)	128
20. Use of housing unit (additional topic)	128
21. Occupancy by one or more households (core topic)	128
22. Number of occupants (core topic)	129
23. Building—type of (core topic)	129
24. Year or period of construction (additional topic)	129
25. Dwellings in the building—number of (additional topic)	130
26. Construction material of outer walls (core topic)	130
27. Construction material of floor, roof (additional topic)	130
28. Elevator—availability of (additional topic)	131
29. Farm building (additional topic)	131

	<i>Page</i>
30. State of repair (additional topic)	131
31. Characteristics of head or other reference member of household (core topic)	132
32. Tenure (core topic)	132
33. Rental and owner-occupant housing costs (additional topic)	132
34. Furnished or unfurnished (additional topic)	133
35. Information and communications technology devices—availability of (core topic)	133
36. Cars—number of (additional topic)	134
37. Durable household appliances—availability of (additional topic)	134
38. Outdoor space—availability of (additional topic)	134
B. Occupied and vacant housing units	135

Annexes

I. Derived variables	137
A. Derived variables for housing records	137
1. Household income	137
2. Family income	138
3. Family nucleus	138
4. Type of household	138
5. Household composition	139
6. Family composition	139
7. Household and family status	140
8. Impact of HIV/AIDS on household structure	141
9. Related persons	141
10. Workers in the family	141
11. Complete plumbing	141
12. Complete kitchen	142
13. Gross rent	142
14. Wealth index	142
B. Derived variables for population records	143
1. Economic activity status	143
2. Own children	143
3. Parents in the house	143
4. Current year in school	144
5. Months since last birth	144
II. Relationship of questionnaire format to keying	145
III. Scanning versus keying	149
A. Entering the data	149
1. Scanning	149
2. Heads-down keying	150
(a) Heads-down keying without skip patterns	150
(b) Heads-down keying with skip patterns	150
3. Interactive keying	151
B. Verification	152
1. Dependent verification	153
2. Independent verification	153

	<i>Page</i>
C. Editing considerations with scanned data.	153
D. Conclusions	154
IV. Sample flow charts.	155
V. Imputation methods	161
VI. Computer editing packages	165
Glossary.	169
References	175

Tables

1. Sample population by 15-year age group and sex, using unedited and edited data	9
2. Population and population change by 15-year age group with unknowns: 2000 and 2010	11
3. Population and population change by 15-year age group without unknown data: 2000 and 2010.	11

Figures

1. Examples of common codes for selected items	23
2. A typical hypothetical household including relationships, sex and fertility of the members	27
3. Example of household with head and spouse of the same sex	27
4. Example of household with ages of some household members	28
5. Example of household with potential inconsistencies in age reporting ...	28
6. Example of rules for a multiple-variable edit of selected population characteristics	31
7. Example with head and spouse of same sex in an unedited dataset and its resolution	31
8. Sample editing specifications to correct sex variable, in pseudocode	32
9. Example of multiple-variable edit analysis for very young widow with three children	32
10. Sample household as example of input for dynamic imputation	35
11. Initial static matrix for age based on sex and relationships.	36
12. Example of a dynamic imputation matrix after one change	36
13. Example of a dynamic imputation matrix after multiple changes	37
14. Example of head of household and head's father without assigned language.	39
15. Initial values for a dynamic imputation matrix for language	40
16. Example of members of a household without an assigned language.	40
17. Example of head of household and child with child's age missing.	42
18. Example of head of household and child with child's age and grade missing	42
19. Sample set of values for an imputation matrix and sample imputation code	44
20. Example of a summary report for number of imputations per error.	45
21. Sample report for errors in a questionnaire	45

	<i>Page</i>
22. Example of supplementary error listing by questionnaire including multiple variables	45
23. Sample population records with flags for imputed values	48
24. Example of a flag for a young female with fertility blanked and flag added	48
25. Example of household with head of household listed as first person.	62
26. Example of household with head of household listed as fifth person	63
27. Illustration of household with fertility information	82
28. Initial values for determining children surviving when age and value for children ever born are valid	83
29. Sample imputation matrices to be developed for pairs of known information	86
 <i>Boxes</i>	
1. What census editing should do.	8
2. Major guidelines for correcting data	26
3. Guidelines for structure edits	51
4. Editing and imputation for age.	73
 <i>Annex Figures</i>	
A.II.1. Sample questionnaire form with person pages	145
A.II.2. Example of flow within a questionnaire with person pages	145
A.II.3. Sample questionnaire, household page with all persons on same page.	146
A.II.4. Example of flow for a questionnaire with household pages, with multiple persons per page	146
A.II.5. Example of a household page with multiple persons, without keying problems	147
A.II.6. Example of a household page with multiple persons, with potential keying problems.	148
A.IV.1. Sample flow chart to determine head of household	156
A.IV.2. Sample flow chart to determine presence of spouse in household	157
A.IV.3. Sample flow chart to edit sex variable for head of household and spouse.	158

Abbreviations

CD	children who died
CEB	children ever born
CLA	children living away
CLH	children living at home
CS	children surviving
EA	enumeration area
GIS	Geographic Information Systems
ILO	International Labour Organization
ICIDH	International Classification for Impairments, Disabilities and Handicaps
NIM	Nearest-neighbour Imputation Methodology
OCR	optical character reading
OMR	optical mark reading
PES	post-enumeration survey
SAS	Statistical Analysis System
SNA	System of National Accounts
SPSS	Statistical Package for the Social Sciences
UNESCO	United Nations Educational, Scientific and Cultural Organization

Introduction

A. Purpose 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.

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.

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. The *Handbook on Population and Housing Census Editing* is designed to bridge this knowledge gap in census and survey data editing methodology and to provide information for concerned 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.

4. The *Handbook* is a reference for both subject matter² and data processing specialists or methodologists 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.

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.

5. The introduction describes the census process, the various types of errors that occur in a census and the fundamentals of census editing. Subsequent chapters present procedures and techniques for editing census data at various stages of processing. Although this handbook focuses on editing for population and housing censuses, many of the concepts and techniques also apply to survey operations.

B. The census process

6. A population and/or housing census is the total process of collecting, compiling, evaluating, analysing and releasing demographic and/or housing, economic and social data pertaining to all persons and their living quarters (United Nations,

2007). Traditionally, censuses are conducted at specified times in an entire country or a well-delimited part of it. Recently, some countries have started carrying out continuing surveys to cover the whole country, using a “long” form, to provide complete coverage over time. In either scenario, the census provides a snapshot of the population and housing at a given point in time.

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

8. All censuses and surveys share certain major features that include (a) preparatory work; (b) enumeration or collection; (c) data processing, including data entry (keying or scanning), editing and tabulating, (d) databases construction and dissemination of results; (e) evaluation of the results; and (f) analysis of the results.

9. The preparatory work includes many elements such as determining the legal basis for the census; budgeting; developing the calendar; administrative organization; cartography; creating a listing of dwelling units; developing of the tabulation program; preparing the questionnaire; and developing plans and training staff for enumeration, pretests, data processing, and dissemination.

10. The enumeration process depends on the method of enumeration selected, the timing and length of the enumeration period, the level of supervision and whether and how a sample is used. After the data are collected, they must be coded, captured, edited and tabulated. Data processing produces both microdatabases and macrodatabases. National statistical/census offices use these databases for tabulations, time series analysis, graphing and mapping operations, and Geographic Information Systems (GIS) for thematic mapping and other dissemination techniques. The results are evaluated for both content and coverage using a variety of methods, including demographic analysis and post-enumeration surveys. Finally, results are analysed in a variety of ways, including descriptive summaries of results, policy-oriented analyses of census results and detailed analytical studies of one or more aspects of the demographic and social situation in the country.

C. Errors in the census process

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

1. Coverage errors

12. Coverage errors arise from omissions or duplications of persons or housing units in the census enumeration. The sources of coverage error include incomplete or inaccurate maps or lists of enumeration areas, failure by enumerators to canvass all the units in their assignment areas, duplicate counting, omission of persons who are not willing to be enumerated, erroneous treatment of certain categories of persons

such as visitors or non-resident aliens and loss or destruction of census records after enumeration. Coverage errors should be resolved, to the greatest extent possible, in the field. The office editing process eliminates actual duplicate records. However, care must be taken to determine whether these are duplicate persons or households. Twins, for example, may have identical information, except for sequence number. Hence, the editing rules applied during this process determine when to accept and when to reject seemingly duplicate information, and when to make changes through imputation.

13. Structure edits, described in chapter III, check households for the correct number of person records, correct sequencing, and the existence of duplicate persons.

2. Content errors

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

(a) Errors in questionnaire design

15. Poorly phrased questions or instructions are one source of content errors. The type of questionnaire, its format and the exact wording and arrangement of the questionnaire items merit the most careful consideration, since the faults of a poorly designed questionnaire cannot be overcome during or after enumeration. Pretesting should be used to minimize errors that may arise due to poorly designed questionnaires. If, for example, skip patterns are not clear or are not placed appropriately, the enumerator may erroneously skip sections of the questionnaire and fail to collect all the relevant information.

(b) Enumerator errors

16. Enumerators and respondents interact, unless the census is conducted using a self-administered questionnaire. The enumerator can err when asking the questions, either by abridging or changing the wording of the questions or by not fully explaining the meaning of the questions to the respondent. The enumerator may also add errors in recording the responses. The quality of enumerators and enumerator training are crucial factors in the quality of data collected. Enumerators must be properly trained in all aspects of census procedures. They should be made to understand why their role in the census process is important and how the enumeration fits in with the other stages of the census. Moreover, since enumerators come from many different backgrounds and have varying levels of education, training must be developed to make certain that enumerators know how to ask the questions to obtain an appropriate response.

(c) Respondent errors

17. Errors may be introduced into the data when respondents misunderstand certain items. Errors may also occur as a result of deliberate misreporting, or proxy responses (when someone other than the person to whom the information pertains provides the responses to the questionnaire). The quality of individual responses can be improved through publicity for the census as well as by training enumerators to

explain the purpose of the census and the reasons for the various questions. Some countries use self-administered questionnaires, so no enumerator-respondent interaction exists. For self-administered forms, errors occur when the respondents misunderstand the questions or instructions.

18. Respondent and enumerator errors are best addressed at the enumeration stage while the forms, the respondents and the enumerators are still available. Supervisors must be able to train enumerators. The supervisors must also be able to check data collected by enumerators regularly during enumeration to ensure that enumerators do not introduce systematic bias into the data. Supervisors should deal with enumerator and respondent errors in the field before the questionnaires are sent to the regional or central offices.

(d) Coding errors

19. Errors may arise in the course of coding since the coder may miscode information. Miskeyings may introduce errors into the data during data entry. In general, lack of supervision and verification at this stage delay the release of data, as error detection and correction become more difficult later. Manual edits often occur before or during the coding operation.

(e) Data entry errors

20. Range checks and certain basic consistency checks can be built into data entry software to prevent invalid entries. An intelligent data entry system ensures that the value for each field or data item is within the permissible range of values for that item. This system increases the chance that the data entry operator will key reasonable data and relieves some of the burden of data editing at later stages of the data preparation process. These checks may, however, slow down the speed of data entry. Therefore, the amount of consistency checking during data entry must be carefully weighed against the need to maintain a reasonable speed of data entry. A balance needs to be established beforehand, so that the data entry clerks do not spend too much time on these efforts. Verification of keying inevitably improves the quality of the data. Keyed forms may be verified by rekeying the same information, often on a sample basis. The relationship of questionnaire format to keying is discussed in annex II and scanning versus keying considerations are discussed in annex III.

(f) Errors in computer editing

21. One of the crucial steps in census data processing is editing. The editing process changes or corrects invalid and inconsistent data by imputing non-responses or inconsistent information with plausible data. Paradoxically, any of these editing operations can introduce new errors.

(g) Errors in tabulation

22. Errors can occur at the tabulation stage owing to data processing errors or the use of information that is “unknown” (not supplied). Errors at this stage are difficult to correct without introducing new errors. Lack of inter-tabulation checking and printing errors produce errors at the publication stage. Rather than trying to correct the tables themselves it is important to maintain the processing system so that additional editing is done when inconsistencies in the tables appear. If errors are carried through all stages of the process to publication, they will be apparent and the results will be of questionable value. If “corrections” are made at the tabulations stage, say a

few miscellaneous unknowns are found and placed in the “totals” but not in the distribution, the tables cannot be replicated by other analysts, and have less overall value. It is wiser to see census processing as a feedback system so that changes to the dataset are made during editing and not during the tabulation phase. Before the release of tables, it is essential to conduct a thorough check to ensure that all planned tabulations are prepared for all intended geographical units. While range checks and consistency checks introduced at the editing stage can reduce most of the errors, an aggregate check after tabulation—sometimes called a “macroedit”—is essential. Trained and experienced persons should go through the different tables to check whether the reported numbers in different cells are consistent with the known local situation. In a limited number of cases, a quick reference to census schedules may indicate coding errors. Calculation of selected ratios and growth rates and comparison with previous census figures or other figures published by sample surveys can also be useful. However, comparison with other survey-based or administrative data figures should be attempted only if the concepts used are comparable. If errors are found in the final tabulations, corrections should be made first on the microdataset, partly to permit other data processors in the national statistical/census office to make comparable tables. In addition, since national statistical/census offices sometimes release parts of the microdata files to researchers and other users in the public and private sectors, tables need to be replicable.

23. As indicated above, the census process involves a number of sequential, interrelated operations, and errors may occur in each operation. It is important to remember that computer edits are part of a feedback system, and that the computer edit not only feeds forward to tabulations, but also links backward to collection and field processing. The best way a national statistical/census office can prevent problems with the computer edit is to maximize the field edit. The national statistical/census office also needs to make sure that coding and data entry are accurate, and should have continuing feedback among all operations, including entry, editing and tabulations.

D. Structure of the *Handbook*

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

Chapter I

Editing in censuses and surveys

A. Editing in historical review

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

26. Census editing changed markedly with the introduction of computers. Computers detected many more inconsistencies than manual editing. Editing specifications became increasingly sophisticated and complicated. Automated imputation became possible, with concomitant rules for the process (Nordbotten, 1963; Naus, 1975). At the same time, the process allowed for more and more contact with respondents, or at least with the completed questionnaires of these respondents. Many editing teams began to feel that the more editing the better, and the more the sophisticated the edit, the more accurate the results. Programs produced thousands of error messages, requiring manual examination of the original forms or, for some surveys, re-interviews of the respondents.³

27. With computers it became increasingly easier to make changes in the dataset. Sometimes these changes corrected records or items. Many records passed through the computer multiple times, with errors and inconsistencies reviewed by different persons each time (Boucher, 1991; Granquist, 1997).

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

29. A major advance in census editing came in the 1980s when national statistical/census offices began to use personal computers to enter, edit and tabulate their data. Suddenly, data processors could perform edits online at the data entry stage or soon after. For surveys and small country censuses, staff could develop programs to catch errors during collection or while entering data directly into the machine. Computer edits allowed more, continuous contact with respondents to resolve problems encountered in the editing process (Pierzchala, 1995).

30. In the early years, the process of making increasingly sophisticated and thorough checks on census and survey data seemed to be very successful. Editing teams created ever more complicated editing specifications, and data processing specialists spent months developing flow charts or decision charts and program code. Analysts

³ Annex V contains a discussion of various kinds of imputation methods; annex VI discusses computer-editing packages.

seldom evaluated the packages. It seemed that editing could correct any problems arising from earlier phases of data collection, coding, and keying. Nevertheless, it also became apparent to many analysts that in many cases, all of this extra editing harmed the data, or at least delayed the results or caused bias in the results. Sometimes the program made so many passes through the data, correcting first one item, and then another item, that the results were far removed from the initial, unedited data.

31. For many censuses and large surveys, such extensive editing caused considerable delay in the census or survey process. Clerks spent much time searching for forms manually; data processing specialists continued to develop applications that look at very small numbers of cases. Granquist (1997) notes that many studies have shown that for much of this extra work, “the quality improvements are marginal, none or even negative; many types of serious systematic errors cannot be identified by editing”.

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

B. The editing team

33. 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.

Box 1

What census editing should do

Census editing should achieve the following:

- ☞ Give users high-quality census data;
- ☞ Identify the types and sources of error;
- ☞ Provide adjusted census results.

34. 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 turnaround 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.

35. 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.

36. 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.

37. 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.

C. Editing practices: edited versus unedited data

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

39. The national statistical/census 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 statistical/census office disseminates an unedited table, such as that on the left side of table 1, both the analysts and the policymakers 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, sex⁴ was not reported, and for 15, age was not reported. These omissions may have resulted from non-responses or from keying errors. Two cases reported neither sex nor age.

⁴ "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			

40. Most users would make their own decisions about what to do with the unknowns. A logical, possibly naive, approach would be to distribute the unknowns in the same proportion as the known values. If the national statistical/census 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.

41. 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.

42. Another alternative the national statistical/census 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.

43. 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.

44. Statistical offices should make every effort to maintain the original collected data. A complete set of the original keyed data should be archived, not only as part of the historical record, but also for reference if staff make decisions about reediting any part of the dataset from the beginning. However, original values for crucial items, like age, sex and fertility, should be kept somewhere on each record to allow demographers and others to analyse the results of the edits.

45. 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.

46. Here, the national statistical/census 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 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 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

47. Revised table 3 shows the unknowns distributed, either proportionally or through some method of imputation. Here, it is much easier to see both the numerical 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 ensure that the edits are consistent between the two censuses and/or surveys, as well as internally consistent. The row for “not reported” is dropped.

D. The basics of editing

48. Editing comprises 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 statistical/census offices do not edit the census or survey results. Editing reduces distorted estimates, facilitates processing, and increases user confidence. Further, according to Pullum, Harpham and Ozsever (1986): “The primary achievements of editing or cleaning are, first, to detect whether the various responses are consistent with one another and with the basic format of the survey instrument.”

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

49. The raw data files in a census contain errors of many kinds. Data processing categorizes the errors into two types: those that may block further processing and those that produce invalid or inconsistent results without interrupting the logical flow of subsequent processing operations. As noted in *Principles and Recommendations for Population and Housing Censuses, Revision 2* (United Nations, 2008, para. 1.311), all errors of the first kind must be corrected and as many as possible of the second kind. The basic purpose of census editing at the processing stage, therefore, is to identify as many errors as possible and make changes to the dataset so that data items are valid and consistent. Nevertheless, processing cannot correct all census errors, including questionnaire responses that are internally consistent but are in fact instances of misreporting on the part of respondents or misrecording on the part of enumerators.

50. Edits tend to fall into two categories: (1) fatal edits, which identify errors with certainty, and (2) query edits, which point to suspicious data items (Granquist and Kovar, 1997, p. 420). Fatal edits identify data items that are certainly in error, while query edits point to data that are likely to be invalid or inconsistent. Fatal errors, the errors detected by fatal edits, include invalid or missing entries as well as errors due to inconsistencies. By contrast, query edits identify data items that fall outside predominantly subjective edit bounds, items that are relatively high or low as compared with other data on the same questionnaire, and other suspicious entries. In order to maintain confidence in the census, particularly when the national statistical/census office decides to disseminate microdata, the editing process must detect and handle fatal edits. Query edits are more difficult to correct, have fewer benefits than the detection and resolution of fatal edits, and add more to the cost of the total process.

51. Since all items in a census are included specifically because planners and policymakers need them, relatively more of the query edits must be resolved during census editing and imputation than for surveys. Nonetheless, in determining the final edits for a census, subject matter personnel should investigate the edits developed 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. As Granquist and Kovar (1997, p. 422) note, data “on hit rates, that is the share of the number of flags that result in changes to the original data, are rarely reported in evaluations or studies of editing processes”.

52. Another set of techniques and terminology relates to microediting and macroediting. As noted, census and survey editing detects errors in and between data records. This *Handbook* describes microediting, which concerns the ways to ensure the validity and consistency of individual data records and relationships between records in a household. Another method, macroediting, checks aggregated data to make sure that they are also reasonable. In this method, tables are run on the edited data and checked against predicted frequencies and tolerances to identify various problems with the data; if “errors” are found, the macroedit can make a global change to the aggregated data, send a unit record back for reprocessing, or add new microedits to rectify the problem. For example, a country may have a very large percentage of persons without a reported age. After imputing for age to obtain a complete dataset, checks at the macro, or aggregate level could make sure that selective under-reporting by older persons does not skew imputed values. The editing team could choose to take measures to alleviate the risk of potential skewing, depending on the results of the analysis. Both macroediting and microediting require thorough testing before they are used.

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

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

54. Sometimes the source of error is outside the processing office. Banister (1980, p. 2) notes that if “we know that a high proportion of some subgroup did not answer a particular census question, it means that they did not understand the question, that they resisted answering it or that they were apathetic about cooperating with the census”. Hence, she argues that non-response rates for subgroups should be included on census storage media and in published tables. National statistical/census offices are now more likely to preserve these data on compact disk or other media for researchers.

55. More and more evidence exists that no amount of computer editing can take the place of high quality census data collection. National statistical/census offices know that at some point computer editing is not only limited, but becomes counter-productive: the edit adds more errors to the dataset 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.

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

57. 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 statistical/census offices subject matter specialists have in fact, become editing enthusiasts. Hence, offices now perform many consistency tests that were difficult in the past, particularly those involving inter-record checking and inter-household checks. Unfortunately, this feature of microcomputers has also led to many problems, and the greatest of these is over-editing.

1. How over-editing is harmful

58. Over-editing has a negative impact on the editing process in several ways, including timeliness, cost and the distortion of true values. It also gives a false sense of security regarding data quality. These concerns are reviewed below.

(a) Timeliness

59. The more editing a national statistical/census 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.

(b) Finances

60. Similarly, the costs of the census process increase as the time increases. Each national statistical/census 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.

(c) Distortion of true values

61. 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: miscommunication 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 statistical/census 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.

(d) A false sense of security

62. Over-editing gives national statistical/census 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 statistical/census 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 statistical/census 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 statistical/census offices cannot edit all pairs of variables, editing teams should check the most important combinations. When editing teams find inconsistencies, correction procedures should be available.

2. Treatment of unknowns

63. The editing team must decide early in census planning how to handle “not stated” or unknown cases. As noted earlier, columns or rows of unknowns in tables are neither informative, nor useful, so planners in most countries prefer 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. In addition, some informative invalid values cannot be treated as blanks but must be resolved. For example, “South America” as place of birth must be resolved to a country (such as Peru). The editing team needs to decide how to deal with the unknowns systematically.

3. Spurious changes

64. Although they could do so, national statistical/census 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, workforce participation and other areas may develop. Hence, this rule should be tested to see the results before being fully implemented.

4. Determining tolerances

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

66. To reduce missing responses to a minimum, the national statistical/census 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 statistical/census office may decide either not to use the item or to use it only with cautionary notes attached.

67. The question arises as to who should determine the tolerance level for an item. The editing team, including both subject matter and data processing specialists, may have to decide on tolerance levels. The subject matter personnel must use the items over time and therefore have a professional stake in making sure they obtain the highest quality data. The data processing specialists, however, may find that they cannot 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.

5. Learning from the editing process

68. 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 statistical/census offices detect errors, the more likely they will be to correct them.

6. Quality assurance

69. 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 (Granquist and Kovar 1997; Statistics Canada, 1998).

7. Costs of editing

70. 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. As Granquist and Kovar (1997, p. 418) note that, even “in the 1990s, editing is essentially as expensive as it was in the 1970s, although the process has been largely rationalized by continuous exploitation of technological developments”. For most countries, editing activities take a disproportionate amount of time and funding, so each country must determine the return on its investment. According to the same authors, the cost of editing household surveys was about 20 per cent of the total census budget worldwide in the early 1990s.

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

8. Imputation

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

- (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.

9. Archiving

73. Part of the quality assurance process of the census or survey is to document all processes and then to archive that documentation. National statistical/census 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 as in other parts of the country, and outside the country as well.

74. 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.

75. 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.

76. The two sets of errors should be provided at logical geographical levels, certainly for the major civil divisions; on the other hand, providing error listings at lower levels of geographical levels could assist in targeting problems in enumerator training, quality control, and other issues connected with the enumeration.

Chapter II

Editing applications

77. The present 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 providing an edited dataset.⁵ Examples have been selected to illustrate which kinds of problems unedited data may present for users and why edited data are more useful. The chapter considers issues of keying and coding as related to the preliminary editing process. The chapter also considers general issues in computer editing along with guidelines on topics such as checking for validity and consistency. The two generic types of computer editing, entailing static imputation (cold deck) and dynamic imputation (hot deck) techniques, are reviewed in detail.

78. Whether a census dataset 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, hence the dataset 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. In some cases, the scanned data also require a second automated coding operation so that items like birthplace, industry and occupation can be filled in.

79. In either case, the unedited data should appear in a form that allows the computer programs to develop the structure edits (as described in more detail in chapter III). The structure edit entails a check to ensure that all of the major civil divisions are presented in geographical or numerical order and that, within each major civil division, each minor civil division appears, and in geographical or numerical order. Then, within each minor civil division, each locality must appear, and within geographical or numerical order. This procedure continues down to the lowest geographical level. As described in the next chapter, appropriate procedures must be instituted to ensure that each housing unit appears once and only once in the dataset.

80. The structure edit must also ensure that all record types are present when appropriate, and that no record types are repeated when they should not be. Thus, for a population and housing census, either the population or the housing records must come first, and that convention must be respected throughout the whole dataset. In most cases, only housing records should be present, so that surplus records must be dealt with; programmers must also 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 absent for vacant units.

81. With regards to the structure's being it is important to note that, inevitably, the structure edit will be revisited during the content edit, and often beyond, as glitches appear during the various census processes. This is normal census procedure, and should be expected; consequently, time, personnel and equipment requirements should be built into the total system.

⁵ 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* was being written, new technologies were emerging, entailing the use both of the use of personal digital assistants (PDAs) and the Internet for data collection and interactive editing (see, for example, Ireback (2000) for Internet use in Sweden). Just as technologically developing countries had problems with scanning in the early 2000s, so are many countries currently finding that PDA usage still needs to be refined.

82. The content edit can then begin. Each population and housing item must be considered alone and, usually, also in combination to determine the considered validity of each item, and the best fit among the items. Chapters IV and V cover the various population and housing items in *United Nations Principles and Recommendations for Population and Housing Censuses, Revision 2*.

83. When the content editing has been completed, a completely edited dataset 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 end 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 revisited as well in order that any specific problems resulting from particular cross-tabulations may be resolved.

84. 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 for dealing 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.

85. 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 statistical/census offices should allocate sufficient resources to obtain the highest quality census data.

86. To implement the computer editing phase of the process the editing team prepares written editing instructions or specifications, decision tables, flow charts and pseudocode. Pseudocode is a set of written editing instructions or specifications as shown in figure 8.

87. Flow charts help the subject matter specialists to understand the various linkages among the variables and make it easy to write editing instructions. Sample flow charts are given in annex IV. 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.

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

A. Coding considerations

89. During much of the second half of the twentieth century, as noted earlier, countries keyed their data. Although most countries now scan their censuses, fre-

quently they do continue to key their surveys. Even when forms are scanned, certain variables still need to be translated from words into numbers. The process of generating machine-readable numbers and alphanumerics is called coding.

90. Some editing packages can easily accept and work with alphanumeric data, but most packages experience some problems categorizing, summing, and deriving percentages, medians, etc., when non-numerical data are included.

91. Codes that are composed completely of alphabetic characters or alphabetic characters combined with numbers (called alphanumerics) should be avoided whenever possible. When forms are scanned, alphanumerics do not pose a significant problem; on the other hand, for many computer packages, their use demands considerable manipulation or at least considerable attention. Many editing programmes can process alpha characters only if they are placed between quotation marks, or distinguished in some other manner.

92. When developing a coding scheme, census and survey staff must consider the returns on each investment of time, energy and funds. Coding considerations are reasonably insignificant for small countries or small surveys since the amount of processing required is much smaller than for a census. Also, data that are scanned are not affected as much by problems associated with additional columns of information.

93. On the other hand, 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 if there was a single column of information; that is, to say with codes 1 through 9, the scanner may pick up an alpha character, or a blank, or a stray mark converted to some readable character. However, these issues are readily handled in the edit, as described further on in this publication.

94. When there are two columns, however, say, codes 1 to 10, then there is a risk of introducing a whole new spectrum of errors. Instead of legal values 1 to 9, there would now be incoming values 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 begin to make strategic decisions about what to do with this value. Was it meant to be 3, and is the 1 therefore erroneous? Was it meant to be 10, and is the 3 therefore incorrect? In most cases, the subject specialists provide the edit specifications for the item, but the presence of these values automatically increases the time and complexity of the edit and could decrease the quality of the final dataset.

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

96. Therefore, 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 actually have more than nine people in a par-

ticular category; but, as always in census work, the statistical office will have to strike a relative balance between errors and good data.

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

- 1 Head of household (or householder)
- 2 Spouse
- 3 Child
- 4 Sibling
- 5 Parent
- 6 Grandchild
- 7 Other relative
- 8 Non-relative

For most countries, this set of standard codes covers the majority of relationships. Some countries add a “0” code for head of household and can then add a tenth category to the others.

98. These codes can be used to obtain household composition, as shown in annex I 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, the statistical office will require additional two-digit codes to carry out its mission.

99. When a country decides on multiple columns, it also needs to decide on how to use them. In the example given above, it is assumed 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 ascribe significance to the columns themselves. For example:

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

100. This scheme codes for 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. While these values could be useful in family reconstruction, the office staff and certain general users might find them cumbersome.

101. This type of coding should, however, be considered for certain social and economic variables. With regard to ethnicity, for example, the first digit would denote the major tribal or ethnic grouping, the minor tribal or ethnic grouping. When more

than 10 minor groupings exist, two digits in the first column would obviously need to be used.

102. Similarly, for items entailing three or four digits, like occupation or industry, the first digit would denote the major occupation or industry; the second digit, the minor occupation or industry; and the third digit, the specific occupation or industry. Since most international coding schemes already have the levels embedded in the codes, the statistical office does not have to do any additional work.

103. As national statistical/census 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, workplace), 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 statistical/census 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.

104. 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.

105. 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.

106. 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

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

must decide, on a case-by-case basis, when to use other items directly for assignment, and when to use other available variables.

107. Another problem occurs when definitions differ between censuses (or between a census and a survey) for variables such as work or ethnicity. The national statistical/census 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.

108. 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.

B. Manual versus automatic correction

109. 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.

110. In the early years of computer entry, no editing on entry was possible; that is to say, all corrections either had to be made manually, as part of the office operations of coding and checking, or were part of the computer operations following the keying of the data. Newer packages have built-in edit functions, which ensure that invalid entries are not be entered unless forced by the keyers, and that inconsistencies can be flagged, for correction by the keyers, manually, or by a computer programmer. As scanning became more prevalent, this evolution has been repeated: in the early years of scanning, no edit during entry was possible; but recently, validity edits and data conversions and recodes have been built into the scanning systems.

111. 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.

112. 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.

113. 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 statistical/census 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.

114. 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 statistical/census 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.

115. 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 workforce 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.

116. 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 statistical/census office must either assign “unknown” or use some set of values to assign the age item.

117. 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.

118. 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.

119. 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.

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

C. Guidelines for correcting data

121. 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.

122. 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.

123. 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.

124. 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 statistical/census office would end up with four errors:

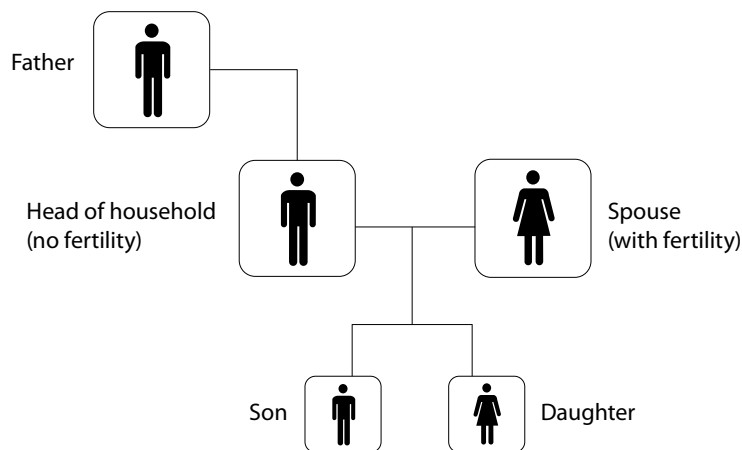
Box 2

Major guidelines for correcting data

- ☞ As editing procedures are established, it may be helpful to keep in mind the following suggestions for correcting data:
- ☞ Make the fewest required changes possible to the originally recorded data;
- ☞ Eliminate obvious inconsistencies among the entries;
- ☞ 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.

Figure 2

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



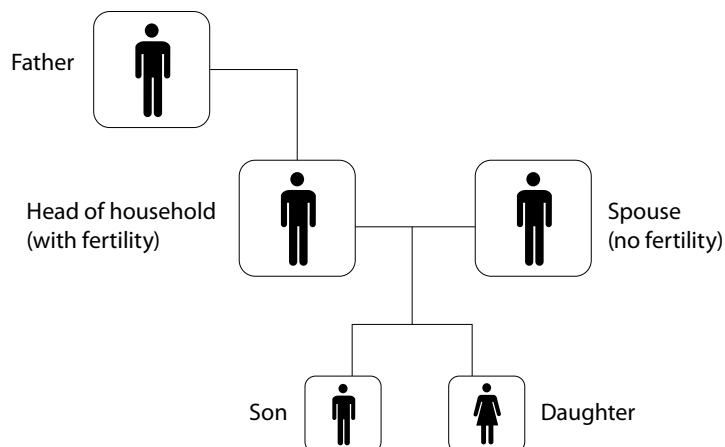
- (a) The head of household's sex would be wrong;
- (b) The spouse's sex would be wrong;
- (c) The head of household would lose her fertility information;
- (d) The male spouse would erroneously be assigned fertility.

This is clearly not good editing procedure.

125. 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.

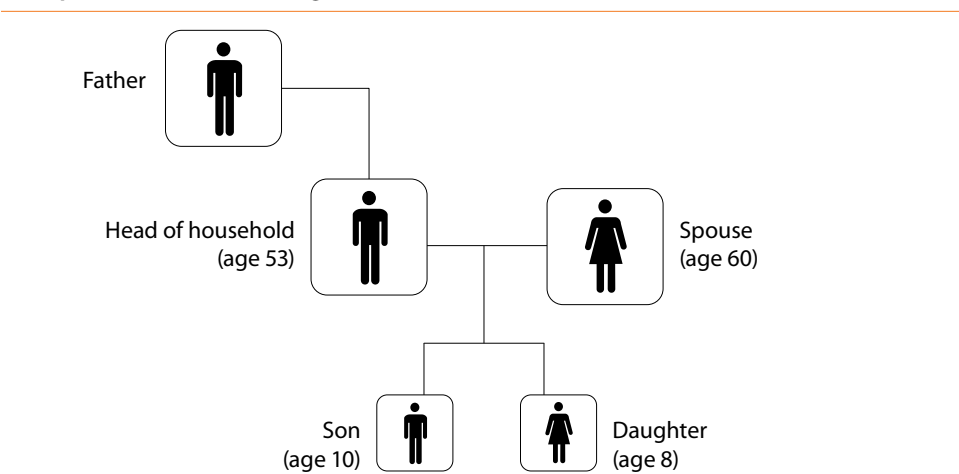
Figure 3

Example of household with head and spouse of the same sex



126. Another example, in figure 4, also illustrates the point. Most countries consider the age for childbearing to be between 15 and 49 years. 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

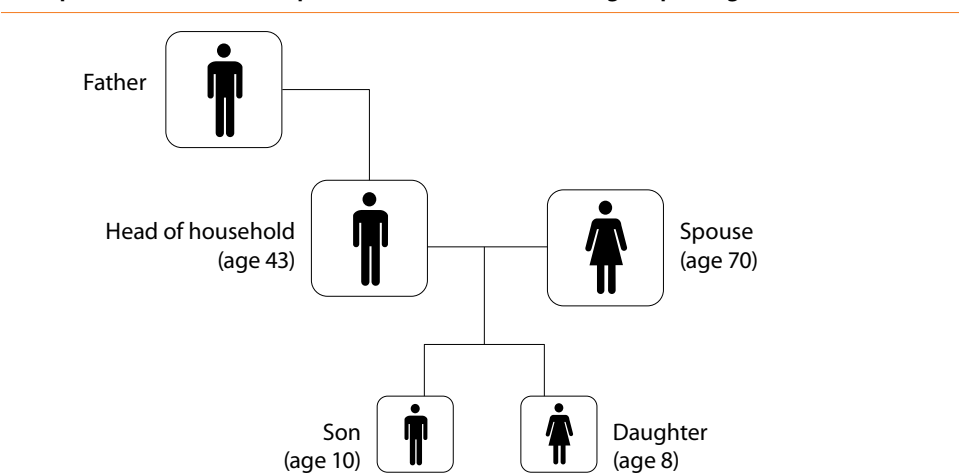
Figure 4
Example of household with ages of some household members



mother and a biological child should probably not exceed 50 though for adopted children, the age differences could be larger). 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 dataset if the age itself is misreported. Again, the editing team must decide the appropriateness of reported ages for particular variables.

127. 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

Figure 5
Example of household with potential inconsistencies in age reporting



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. In the first case, a problem still remains with respect to the second child's age, which also requires editing.

128. 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.

D. Validity and consistency checks

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

130. 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.

1. Top-down editing approach

131. 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. The edits usually start with sex and age because of their importance, particularly in dynamic imputation. While the top-down approach does not completely preserve the relationships among the data items, it does provide an adequate framework within which to complete the edit.

132. 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.

133. 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”.

134. 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 annex [I](#).

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

2. Multiple-variable editing approach

136. 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.

137. Another approach is multiple-variable editing, which is based on the Fellegi-Holt system. This approach requires more computing expertise and computer power but probably obtains results that are closer to “reality”. Different kinds of multiple-variable editing appear in annex V, entitled “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.

138. 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 than others for data presentation.

139. Figure 6 shows a simple case where, for some reason, both the head and the spouse have the same sex—as it turns out, both are male and one is a male with fertility. It is quite 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 of age or older					
2	Spouse should be 15 years of age 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 of opposite sexes	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” indicate that two or more items are inconsistent. For example, in item 5, since the head and the spouse are of the same sex, the edit fails for relationship and sex, and the 1 appear in these cells.

140. 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 dataset 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

141. 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" to be done 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.]
    End-if
  End-if
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" to be done 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.]
  End-if
End-if
End-if

```

142. Figure 9 below provides 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, what results is the following:

Figure 9
Example of multiple-variable edit analysis for very young widow with three children

No.	Rule	Relation	Sex	Age	Marital status	Fertility
1	Head of household should be 15 years of age or older	1		1		
2	Spouse should be 15 years of age 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 of opposite sexes					
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		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	1

143. 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. As she is 13 years old, the boxes for "relationship" and "age" are marked, since the inconsistency is between these two variables. She is not a spouse, hence neither rule 2 nor rule 3 is triggered. Rules 4 and 5 are not triggered for the same reason—they apply only to the spouse. However, since

under rule 6, a person less than 15 years old (in this case 13 years old) should be never married and our 13 year old widow is “widowed”, the rule is violated. Inasmuch as rule 7 applies to males, it is not triggered. Under rule 8, females less than 15 years of age should have no fertility; inasmuch as this person does have fertility, the rule is violated. Rules 9, 10 and 11 do not apply to this person.

144. 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.

E. Methods of correcting and imputing data

145. 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.

146. 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, as illustrated in the annexes.

1. Static imputation or “cold deck” technique

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

148. Static imputation is a stochastic method, as is dynamic imputation, but the values do not change over time. This approach is described in annex V.

149. 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

150. 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.

151. 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”.

152. The values stored in the hot deck represent information about the “nearest neighbours” with similar information. Note that the nearest neighbour is usually the nearest *previous* neighbour because, especially in the top-down approach described elsewhere, housing units and people in those units are considered only once, and the program then moves on. Therefore, 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 a high prevalence of HIV/AIDS or other phenomena that give rise to unusual statistical activity. Similarly, housing characteristics are more likely to be similar within a village or set of villages than between a village or set of villages and other parts of the country.

153. 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.

154. 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.

155. The data contained in figure 10 below cover a household comprising a set of 10 individuals. Blanks, as signified by “X” and “XX”, indicate missing data. Often, the numbers 9 and 99 are used to indicate missing information, in this case on sex (a 9 for a single digit) and on age (99 for two digits). Sometimes, however, the value 9 is

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	X	10
5	4	2	40
6	4	1	XX
7	4	2	13
8	5	X	XX
9	5	1	44
10	5	2	36

Note: X and XX = missing information.

required to stand for another, real value, for example, in a limited number of relationship codes; therefore, these values should be used very sparingly and in fact, if another value like “X”, “.” 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 brief example.

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

157. 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

158. 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.

159. 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.

160. 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

161. 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

162. 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

163. 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

164. 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.

165. 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).

166. 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.

167. 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.

168. 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

(a) Geographical considerations

169. 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; therefore, the editing team must look at variables on a case-by-case basis. Also, as will be discussed later, there are variables that are not applicable in certain areas—like central heating in very warm places—and the edit should take this into account.

(b) Use of related items

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

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

171. National statistical/census 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.

172. 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. A possible

solution would consist in using imputation flags to discourage using imputed data to match a donor to a failed unit

(d) Complexity of the imputation matrices

173. The national statistical/census 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.

174. Nonetheless, the procedure described above can create new problems. The national statistical/census 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.⁶ 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.

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

(e) Imputation matrix development

176. 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.

177. 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

⁶ The best editing practice does not entail the use of edited values in hot decks. Sometimes this practice is difficult to follow, because of the factor of timing for results or difficulty in the computer programming. In these cases, one of several variables would be imputed, and its value placed in the appropriate hot decks and then used to impute subsequent variables.

to reconcile them. When imputation matrices do not use standard conventions, staff must consider each one separately.

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

(f) Standardized imputation matrices

179. 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.

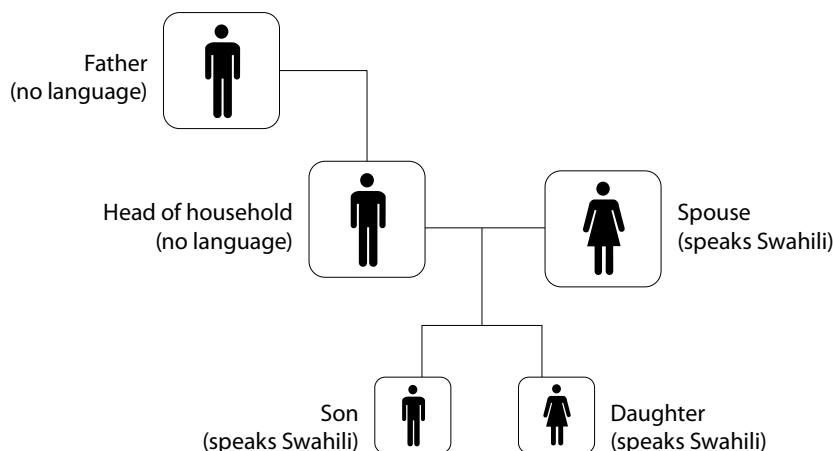
180. For example, the national statistical/census office may want to develop an imputation matrix to determine a code for language when none is given. The first place for the editing program to look will almost certainly be within the household for another person reported as speaking a given language. 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.

181. 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 14 is female.

182. 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

Figure 14

Example of head of household and head’s father without assigned language



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

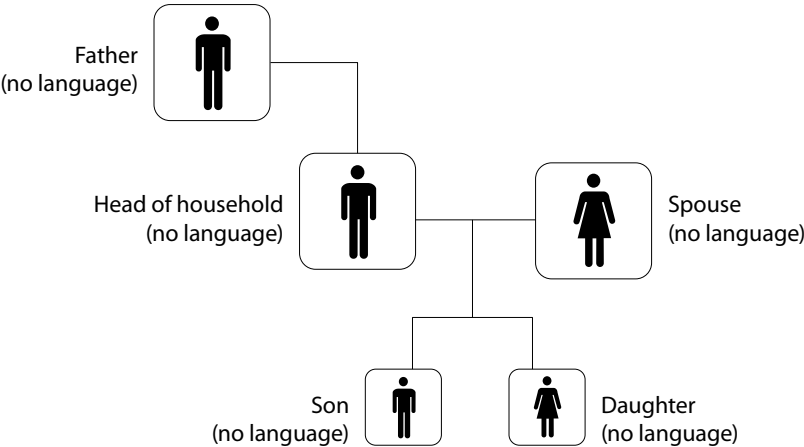
183. If it is decided to impute, the program assigns the head of household a language based on age group and sex. In this case, the entries in the imputation matrix will be for previous heads of household only, since all other persons in a given household receive the same language code as the head of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

184. At this point, if the household still has no one who reports speaking a defined language, the editing program uses the imputation matrix to assign a language to the head of household based on the head of household’s age and sex. The language assigned is the most recent one in the data file spoken by another head of household of the same age and sex. Since the imputation matrix is “updated” continuously as acceptable cases are encountered, the assigned language is likely to be a language spoken in the general community (see figure 16).

185. 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.

186. 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

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



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.

187. This explanation assumes a top-down, sequential approach. Editing teams using sophisticated approaches such as the Fellegi-Holt method and the Nearest-neighbour Imputation Methodology (NIM) (see annex V) apply all related edits concurrently. The present procedure also assumes the existence of an appropriate order for the edits.

188. 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.

189. 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.

190. 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.

191. 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.

192. 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.

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

(g) When dynamic imputation is not used

194. 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

195. 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).

196. 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.

197. 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.

198. 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.

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

199. If the editing team decides to impute all or most of its items, it should develop a strategy for building the edit in a logical way. For population items, the edit should begin by considering all items potentially having unknowns. Editing teams should use information from surveys and administrative records, earlier censuses, the pilot for the census under consideration, and other information available to help determine each item's inclusion in the first, and subsequent, imputation matrices. While

development of the details of imputation matrices is very country-specific, all national statistical/census 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.

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

(h) How big should the imputation matrices be?

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

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

202. One of the biggest problems that some national statistical/census 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 “labour-force participation”, “number of children ever born” must also be included for females. The addition of “number of children” ever born may provide a slightly better age estimate, but the increased complexity of the programming may not justify it. Editing teams have to decide how many imputation matrix dimensions will give the best results, in terms of both accuracy and efficiency. Imputation matrices that are too big (with too many cells) cannot be updated thoroughly, and cold deck values may inappropriately be used instead.

(ii) *Understanding what the imputation matrix is doing*

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

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

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

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

205. 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**

206. 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**

207. 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).

(b) **Messages for errors**

208. 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.

209. One source of information is the display of a message, as seen below in figure 19. This command generates specific messages and summary counts (the total

Figure 19

Sample set of values for an imputation matrix and sample imputation code

Initial values for the imputation matrix A01-AGE-FM-SEXRL (2,6)

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

```

.
.
imputation code
if AGE = 0:98
  let A01-AGE-FM-SEXRL (SEX,RELATIONSHIP) = AGE
else
  message 'Age is unknown, so imputed' AGE
  write 'Age is unknown, so imputed, Age = ' AGE
  impute AGE = A01-AGE-FM-SEXRL (SEX,RELATIONSHIP)
  message 'AGE is now known' AGE
end-if
.
.
.

```

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	2669
2	14-4	Girls present not stated	2678
33	14-5	Month last birth not stated	2723
7	15-6	No children ever born; age difference between mother and child OK	2892

Note: Here “14” simply refers to item 14 in a given series; errors are numbered sequentially.

210. 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, as 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, if desired.

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	

211. Of course, while it makes sense to list all individual errors on sample tests or small, selected datasets, in production runs the size of output would be extremely large and cumbersome (and result in meaninglessness after awhile). To obviate this, a trigger should be set to turn off all or some of the individual questionnaire problems for the complete census. The summary statistics would, of course, remain.

(c) Custom-made error listings

212. 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

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

214. 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:

- (a) The data processors introduce a single error into each household, in sequence, to correspond to the sequence of the editing specifications and the editing program;
- (b) The analyst then checks all of the paths early in the editing process;
- (c) Once the edit follows all paths properly, data processors run a sample of the whole dataset, looking for idiosyncrasies in the actual dataset and making modifications as necessary;
- (d) Finally, the data processors run the whole dataset.

215. 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 datasets 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 22.

216. 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.

217. 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.

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

(d) How many times to run the edit?

219. As noted, as soon as the questionnaire is set, development and testing of edit specifications and programs should begin. Individual items should be developed separately when a top-down approach is used, but even when several variables are to be edited at the same time, edits for individual items will need to be tested on small

parts of the whole dataset. The edit specifications should be developed by the subject matter specialists, and individual edit programs should then be implemented by the programmers. The total edit can then be built up, and run on larger and larger parts of the dataset, while being refined along the way.

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

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

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

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

5. Imputation flags

224. Imputation flags are one method used to retain information about unedited data. As mentioned previously, many editing teams are concerned about the loss of potential information when unedited responses are changed. In cases where a value is changed because of an inconsistency, the editing teams may wish to save the original value or values in order to carry out further demographic or error analysis after the census. Both subject matter specialists and programmers will want to analyse various aspects of the missing, invalid or inconsistent data. Members of the editing team need to make sure that the imputed and unimputed distributions are consistent, to see if any systematic error appears in the editing and imputation plan. For example, sometimes data processing specialists accidentally use only cold deck values because the program neglects to update the imputation matrix. If the country conducted a census pretest, the editing team may need to investigate the relationships between some of the variables after the pretest in order to finalize the questionnaire. In prior censuses, before micro-computers with large hard disks were common, many statistical offices did not have the space on their tapes or other storage media to maintain extra data; however, these days, for most countries, keeping information about unedited data is no longer a problem.

225. Some countries choose to maintain a simple, binary accounting variable as a flag for each item. This method is simple and takes up a single byte for each variable. For example, the United States Census Bureau places imputation flags for each variable at the end of each record, for both housing and population records. For each housing variable, for example, the variable for the flag was initially “0”, but was changed to “1” if

the original item is changed in any way. The program does not retain the original value, although offices sometimes compile these, either for each record or in the aggregate.

226. Other methods are available to save unedited responses. In the example in figure 23, the national statistical/census office has changed a spouse's age from 70 to 40 using an imputation matrix. The national statistical/census office can easily put the pre-imputation value, in this case 70, in the area reserved for imputation flags and reserve the variable used for published tabulations for the allocated value, in this case 40. In order to examine changes in the dataset, 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 dataset, 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 containing only the edited data is likely to be tabulated more rapidly.

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	

227. 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 dataset 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

228. 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.

229. 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 dataset will be approximately twice as large as the unedited data-

set. For many countries, this would be unacceptable for long-term storage. However, the original data and the edits could be stored for later reconstruction.

230. 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 dataset 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.

F. Other editing systems

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

232. 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.).

233. 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.

234. 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.

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

- (a) The imputed household should closely resemble the failed edit household;
- (b) The imputed data for a household should come from a single donor, if possible, rather than two or more donors. In addition, the imputed household should closely resemble that single donor;

- (c) Equally good imputation actions, based on the available donors, should have a similar chance of being selected to avoid falsely inflating the size of small but important groups in the population (Bankier, Houle, and Luc, n.d.).

236. Under the Nearest-neighbour Imputation Methodology, 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.).

237. The present chapter has discussed general editing and tabulation procedures. Chapter III covers structure edits, encompassing the first computer editing task and the most important, since it establishes that each housing unit does appear, and appears only once and in its proper place in the hierarchy of the country.

Chapter III

Structure edits

238. 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.

239. 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

Box 3

Guidelines for structure edits

Structure edits should manage the following tasks:

- ☞ 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;
- ☞ Make sure that every housing unit is included; and that all households in an EA are entered;
- ☞ Merge the households into their appropriate EAs, and merge the EAs into the appropriate higher level of geography;
- ☞ 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;
- ☞ Assign each individual record to its valid record type;
- ☞ Handle group quarters or collective housing records separately from housing units;
- ☞ 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;
- ☞ Eliminate duplicate records both within households (duplicate persons) and between households (duplicate households, or parts of households) to avoid over-coverage;
- ☞ Handle blank records within a record type;
- ☞ Handle missing housing units.

item validity and the relationship of items between and within records. Chapters IV and V deal with specific individual population and housing items.

A. Geography edits

1. Location of living quarters (locality) (P1H)

240. A locality, according to the *Principles and Recommendations for Population and Housing Censuses, Revision 2* (United Nations, 2008, para. 2.78) 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 paragraphs 2.78-2.88 of *Principles and Recommendations*. It is essential for those concerned with carrying out housing censuses to study this information, as the geographical concepts used to describe the location of living quarters when carrying out a housing census are extremely important, both for the execution of the census and for the subsequent tabulation of the census results (United Nations, 2008, para. 2.455).

241. 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 (P1L)

242. 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 and usually a higher standard of living than that are found in rural areas. In many industrialized countries, this distinction has become blurred, and the principal difference between urban and rural areas in terms of the circumstances of living tends to be a matter of the degree of concentration of population. Although the differences between urban and rural ways of life and standards of living remain significant in the developing countries, rapid urbanization in these countries has created a great need for information related to different sizes of urban areas (United Nations, 2008, para. 2.82).

243. Most countries determine which geographical areas are “urban” and which are “rural” before the census and make needed adjustments after census data are collected. If the country attributes codes for urban and rural residence (such as 1 for urban and 2 for rural), these codes can be entered during keying 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.

244. 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.

B. Coverage checks

1. De facto and de jure enumeration

245. *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 where he or she has been for some time or intends to stay for some time. Generally, most individuals enumerated have not moved for some time and thus defining their place of usual residence is clear-cut. For others, the application of the definition can yield 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 in accordance with which the usual residence would then be either: (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 or (b) the place at which the person has lived continuously for at least the last 12 months, not including temporary absences for holidays or work assignments, or intends to live for at least 12 months (United Nations, 2008, para. 1.461-463).

246. Given this definition of usual residence, national statistical/census 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. Tabulations on a *de facto* basis use only (1) and (2) if all three types are present; tabulations on a *de jure* basis use only (1) and (3) if all three types are present.

247. National statistical/census organizations implementing the utilization of these categories must be very careful in their use at every stage—during data collection, data processing, dissemination and analysis. In particular, tabulations of the dataset should be conducted giving consideration to the desired category of population. Users of these three categories must be familiar with the selected population, since analysing all of the dataset will result in the inclusion of 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 dataset for processing. For later tabulations, file documentation should state explicitly how to handle the various possibilities. Multiple files might be more appropriate.

248. 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.

2. Hierarchy of households and housing units

249. Chapter V examines the relationships between households, housing units and living quarters. Implementation of these concepts depends on the individual national statistical/census 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

250. 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.

C. Structure of housing records

251. One of the topics that may be included in the collection of information through national housing censuses or surveys is the number of dwellings in a building. In this case, the unit of enumeration is a building and information is collected on the number of conventional and basic dwellings in it (see United Nations, 2007, para. 2.524).

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

D. Correspondence between housing and population records

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

1. Vacant and occupied housing

254. 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 use this information in the determination; otherwise, an imputation matrix may be needed.

255. If no population records appear for what is supposed to be an occupied unit, then the editing team must decide whether to count it as a vacant unit or substitute persons from another unit. If the unit is vacant, imputation can easily change the variable for vacancy status. If the unit is occupied, however, then the editing team must decide whether and how to assign persons from another unit with the same number of persons,

with similar characteristics, if possible. Since it is impossible to know the characteristics of missing persons, this method should be used, if at all, only when the editing team decides it has no other alternative. Three possible alternatives are outlined below:

(a) Choosing to leave a housing unit vacant

256. 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 chapter V.

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

257. The national statistical/census 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 chapter IV is applied, or data from donor records are supplied for “missing” persons, as described above.

(c) Substituting another housing unit for missing persons

258. 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

259. 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 statistical/census 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.

260. 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

261. 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 means are available for adding missing households when it is clear that they are in fact missing and need to be supplied. One method is to simply duplicate the previous household. However, if the number of people in the household is known, as is often the case (even if their characteristics are not known), it is possible to work backwards 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 of assistance in obtaining a substitute house. It is not advisable 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

262. 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

263. 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.

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

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

(b) Checking numbers of persons by sex

266. 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.

(c) Sequence numbering

267. 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 statistical/census 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.

5. Correspondence between occupants and type of building/household

268. 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.

E. Duplicate records

269. Duplicate line numbers are not likely to appear in optically read or other scanned questionnaires. For forms that are to be keyed, the national statistical/census 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.

270. 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.

271. 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:

⁷ Traditionally, duplicate records were tracked down and corrected manually, but more and more, these are being at least partially automated. In a recent paper, Winkler (2006) begins to look at automating structure edits and content edits together.

- (a) If the line number for two different records is identical and the number of characteristics that differ is 2 or less, the edit will eliminate one of the records since it is a likely duplicate.
- (b) If 3 or more characteristics are different, the line number will be changed.⁷

F. Special populations

1. Persons in collectives

272. 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.

(a) When collectives are a different record type

273. When the national statistical/census office chooses to use 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.

(b) When a variable distinguishes collectives from other records

274. 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 by referring to the variable itself, which notes which records are persons in collectives. Variables unique to the collectives, such as type of collective, can still be edited and imputed separately. Variables that are excluded from the collective records easily can be checked to make sure they are actually blank by referring to the code for collectives. A more compact file results, since the additional records for persons in collectives are not needed but are simply included as population records with a different code for the variable for household/collectives. During editing and imputation, the program will have to check only population records, and not both population and collective records, for some items.

(c) When the “type of collective” code is missing

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

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

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

(e) Distinguishing various types of collectives

277. 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

(a) Seasonal migrants

278. 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.

279. 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 statistical/census office and the editing team must decide how to handle various types of situations. For example, some persons spend part of each year in another home, such as those who live in a colder part of a country in the warm parts of the year and in a warmer part of the country in the cold parts of the year. Another case is that of nomads who travel for part of the year but are sedentary for a part of the year—perhaps the part of the year when the country chooses to do its census.

(b) Homeless persons

280. 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.

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

281. Again, as for the homeless, a structure edit may be very difficult. Certain countries will collect some “housing” information, so that 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 it would be for persons living in standard housing units, and edited in the regular way, following the guidelines below.

(d) Civilian residents temporarily absent from the country

282. In de jure censuses, civilian residents temporarily absent from the country, but living in households that can report them, should be included in the standard population edits. For the de jure census, some indicator should identify persons who are temporarily absent to allow for the determination of both de jure and de facto populations. The housing edit will not differ because of the absentees. Obviously, however, these people will not be included in a de facto census; therefore, they will not appear in the population edits.

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

283. Since, for a de facto census, everyone resident in the country at the time of the census should be included, these persons should be included as well. Individuals should be included in their place of residence at the time of the census, and edited 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 carried out for these individuals either. If ships in harbours are considered housing units, then the housing characteristics should be described, and edited, using the information for other ships for the hot decks.

284. Foreign persons in the country only temporarily are presumably not included in de jure censuses. Undocumented persons would be included, particularly in those countries that do not distinguish between documented and undocumented persons separately in the census (although so distinguishing 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 usually still reside in the country. If a ship is usually harboured in the country, then presumably the persons on the ship would be included as usual residents and be edited as such.

(f) Refugees

285. 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.

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

286. 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. When information is limited, the use of hot decks will be difficult, and it is likely that errors will be introduced into the dataset; therefore, it is usually preferable 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; therefore, these housing units should be included in the regular edit, though there should be an indicator for the

special status of the housing unit. As housing units outside the country may not be enumerated in the standard way, will care be needed in assessing whether or not to include these units in the edits? (They could still be included in some of the tabulations.)

287. 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.

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

288. 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 did 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* and *de facto* censuses.

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

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

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

290. 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 residence, but it should be possible to use standard edits for population items when the country's regular questionnaire is used on the ships.

G. Determining head of household and spouse

1. Editing the head of household variable

291. 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 (United Nations, 2007, para. 2.114).

(a) The order of the relationships

292. 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 IV). 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.

(b) When the head is not the first person

293. 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.

294. 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.

295. 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.

296. Obviously as illustrated by these two households, the edit paths based on different designated heads of household would be different. Three different possibilities exist for determining the actual head of household for the rest of the edits and tabulations: (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.

Figure 25

Example of household with head of household listed as first person

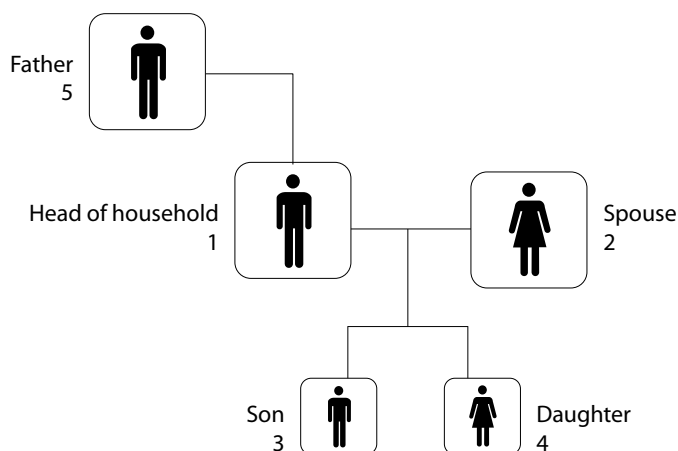
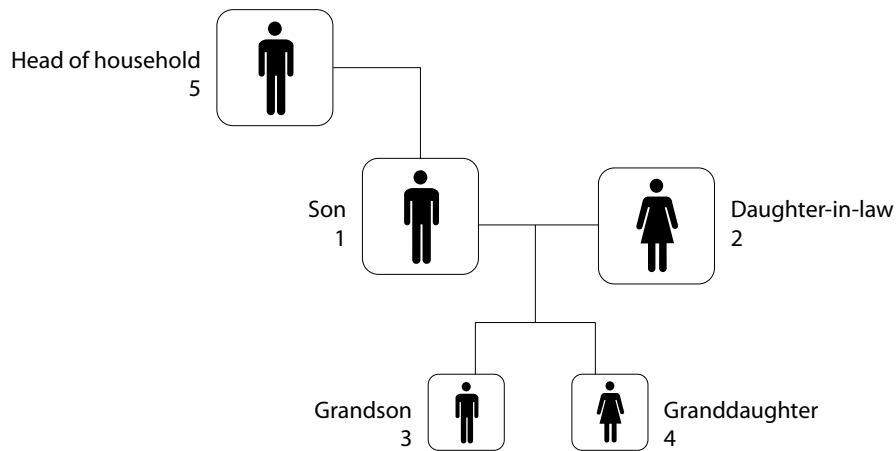


Figure 26
Example of household with head of household listed as fifth person



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

297. 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”.

(ii) *Making the first person the head*

298. 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 dataset. 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.

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

299. 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.

300. 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.

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

(c) More than one head

302. 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.

303. A special case exists in those countries permitting "co-heads" either because of socio-economic conditions (like the frequent departure of male heads for mining or other activities which entails 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 United Nations (2008, para. 2.117) includes a provision for joint heads. If a country chooses to include co-heads, these must be maintained in the edit; however, many of the suggested subsequent edits in this *Handbook* would also have to be modified: when the co-heads practise different religions, belong to different tribes or have different other demographic and social characteristics, a single person can no longer be used in the imputation procedures.

(d) No head

304. 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 in order to determine the most likely head. A flow chart for a sample edit for head of household appears in annex IV.

2. Editing the spouse

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

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

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

306. 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.

(c) Spouses in polygamous societies

307. 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.

(d) Other characteristics of heads and spouses

308. Good editing practice entails imputing 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 of 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 obtain “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, it is important to have the head’s information in place before editing the other people in the unit. A sample flow chart for an edit for spouse appears in annex IV.

H. Age and birth date

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

309. 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 statistical/census offices choose to obtain the age based on the year of the census and the year of birth only, giving a value with potential deviation. If year and month are used, the age will be more accurate, but using day, month and year will give the most accurate results.

2. When the age and date of birth disagree

310. 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.

I. Counting invalid entries

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

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

Chapter IV

Edits for population items

313. Chapter IV 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.

314. 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.

315. 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.

316. All population records should have serial numbers to assist in data processing. The structural edits described in chapter III check for correspondence between the sequence number and the order of serial numbers.

317. 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.

318. 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.

319. 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.

320. 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.

321. 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.

322. 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.

A. Demographic characteristics

323. 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.

324. The multiple-variable (Fellegi-Holt) approach to editing population and housing data was introduced in chapter II of this *Handbook*. Since the demographic variables are integral to all census planning, this approach should be used if time and expertise permit. The quality of the overall dataset is almost certain to benefit from a priority edit looking at age and sex and other selected variables to determine errors or inconsistencies. The items most in error are edited first, followed by those items less in error or inconsistent.

1. Relationship

325. 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.

(a) Relationship edits

326. 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 (see annex I for “family type”, and subfamily number and subfamily relationship recodes).

327. 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.

(b) When the head must appear first

328. 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.

(c) When the relationships are coded upside down

329. 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.

330. The procedure for inverting the relationships usually involves running a “lookup” file containing the original relationships and the inverted relationships, taking into account the sex of the respondents.

(d) When polygamous spouses are present

331. 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, the editing team should decide when polygamous relationships are permitted and when they are not. Sometimes households that seem to have polygamous relationships are actually mistakes.

332. 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.

(e) When multiple parents appear

333. 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.

(f) When censuses collect sex-specific relationships

334. 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. The edit is more important in some cases, such as that of husband and wife, than in others, like that of a young child. Note that it is not advisable to use sex-specific relationships, since redundancy does not clarify but only obscures the nature of the relationships, and necessitates additional editing.

(g) When relationship and marital status do not match

335. 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).

336. 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.

337. 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 “room-mate” or “non-relative”. 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.

2. Sex

338. 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. Annex IV contains an example of a flow chart for editing sex of head and spouse.

339. If a person is not the head of household or the spouse of the head, no other persons exist to refer to; therefore, other items within the person's record should be checked. If sufficient fertility items occur, the code for female should be assigned. However, if this person's sex is missing or invalid, for example, but a spouse exists, for whom sex is indicated, the edit can impute opposite sex to this person.

(a) When the sex code is valid but the head and spouse are of the same sex

340. 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.

(b) When a male has fertility information or an adult female does not

341. 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.

(c) When the sex code is invalid and a spouse is present

342. 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.

(d) When the sex code for spouse is invalid

343. 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.

(e) When the sex code is invalid and female information is present

344. Numerous clues in the questionnaire indicate whether a respondent is female. If the program has not yet determined the person's sex and any female indicators are present, then the record for this person should be assigned female sex. For example, if the person being edited includes sufficient fertility items, then sex can be assigned as female. The fertility items include children ever born, children living in this household, children living elsewhere, children dead and children born alive in the last 12 months. Another possibility is that this person could be the mother of someone else in the household, so that this person's line number equals the line number of the mother of another person in the household.

(f) When the sex code is invalid and this person is spouse's husband

345. If the person is the husband of someone else in the household, based on an item showing the husband's line number, the entry for sex should be assigned male.

(g) When the sex code is invalid and there is insufficient information to determine sex

346. 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 one, using the opposite sex for the second invalid entry and continuing in this fashion.

(h) Note on imputed sex ratios

347. 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.

3. Birth date and age

348. 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.

(a) Age and date of birth

349. The structure edit calculates age from date of birth. First, however, it is useful to review the difference between age and birth date. As stated in *Principles and Recommendations for Population and Housing Censuses, Rev. 2* (United Nations, 2008, para. 2.136), 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.

350. 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 can be substituted. The question on date of birth is appropriate when people know their birth date, which may be established in accordance with the solar calendar or a lunar calendar, or expressed in years numbered or identified in traditional folk culture by names within a regular cycle.

351. It is extremely important, however, that a clear understanding should exist between the enumerator and the respondent about which calendar system the date of birth is based on. If there is a possibility that some respondents will reply with refer-

ence to a calendar system different than that of other respondents, provision must be made in the questionnaire for noting the calendar system that was used. It is not advisable for the enumerator to attempt to convert the date from one system to another. The needed conversion can be best carried out as part of the computer editing work (United Nations, 2008, para. 2.137).

352. The direct question on age is likely to yield less accurate responses for a number of reasons. Even if all responses are based on the same method of reckoning age, the respondent may not understand whether the age wanted is that at the last birthday, the next birthday or the nearest birthday. In addition, other problems can occur: age may be rounded to the nearest number ending in zero or five; estimates may not be identified as such, and deliberate misstatements can be made with comparative ease (United Nations, 2008, para. 2.138).

Box 4

Editing and imputation for age

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.

353. Many national statistical/census offices collect either date of birth or age, but not both. As noted in the *Principles and Recommendations for Population and Housing Censuses, Rev. 2* (United Nations, 2007), 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.

(b) Relationship between date of birth and age

354. 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.

355. 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.

(c) When calculated age falls above the upper limit

356. 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.

(d) Age edit

357. 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.

358. The edit should begin with a check for validity. If the age is valid, specialists might want to check to see whether this person's age is consistent with his/her mother's age (if the person's mother is found in the household) and with the age of this person's children (if this person is a woman and has children in the household). If the ages are inconsistent, this person's age should be noted, and the age should be changed later.

(e) Age edit when the head of household and spouse are present

359. 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.

360. To ensure that this computed age is consistent with other characteristics, the imputation matrix should also include marital status, duration of residence and highest grade of school completed. Exclusion of those variables can result in a computed age that is less than the number of years the person has lived in the place, or less than the level of schooling implies. For example, the imputation matrix may give an age of 8, but the person may have recorded that they lived in the place for 10 years. Without the other variables, when the editing program carries out the years-in-place edit, another imputation matrix will change the years in residence from a correct value to an incorrect value.

(f) Age edit for head when the head's spouse is absent, but child is present

361. 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.

(g) Age edit for head when head's parent is present

362. When a person does not fall into one of the categories described above, the program can search for the person's parent in the household. If the person's parent is found, an age can be computed with an imputation matrix using the difference in age. The difference in age between child and parent generally varies much more than that between husband and wife. For this reason, the program applies this edit only after the husband/wife age difference technique fails. The computed age should take into account the educational characteristics, the highest grade of school completed and the years lived in the district, marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or reports economic activity of any kind.

(h) Age edit for head when head's grandchild is present

363. 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 of school completed, including the years lived in the district, marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or participates in economic activity of any kind.

(i) Age edit for head when no other ages are available

364. When a person does not fall into one of the categories described above, the program can search for another relative or a non-relative 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 non-relatives 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 participates in economic activity of any kind.

(j) Age edit for spouse when head's age is already determined

365. 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 participates in economic activity of any kind.

(k) Age edit for other married couples in the household when the age of one of the persons is known

366. 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.

367. To ensure that this computed age is consistent with other characteristics, the imputation matrix should also include marital status, duration of residence and highest level of educational attainment. Exclusion of those variables could result in a computed age that is less than the number of years the person has lived in the place, or less than the level of schooling implies.

(l) Age edit for child when head's age is already determined

368. 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 completed, years lived in the district, and the marital status, fertility and economic activity. The program should presume that a person has at least the minimum acceptable age if he/she has ever married, has children or participates in economic activity of any kind.

(m) Age edit for parent when head's age is already determined

369. 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.

(n) Age edit for grandchild when head's age is already determined

370. 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.

(o) Age edit for all other persons

371. The editing team should determine appropriate imputation matrices for other related and non-related 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.

4. Marital status

372. In *Principles and Recommendations for Population and Housing Censuses, Rev. 2* (para. 2.144-2.151), marital status is defined as the personal status of each individual in relation to the marriage laws or customs of the country. The categories of marital status to be identified include, but are not limited to, the following: (a) single, (never married); (b) married; (c) 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.

(a) Marital status edit

373. 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 statistical/census 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.

(b) Marital status assignment when dynamic imputation is not used

374. Although marital status should be tabulated only for persons aged X years and older, where X is the earliest age at first marriage, editing teams must determine whether and how much to edit. If the country uses only "not stated" or "unknown" for invalid or inconsistent responses, then when invalid or inconsistent entries are found, the code for "not stated" should replace the inappropriate response. If, for persons

under age X, the response “never married” is missing, it should be imputed; since statistical offices release samples of data to the public, it is important that items like marital status always have entries.

(c) Marital status assignment when dynamic imputation is used

375. If dynamic imputation is used, the edit for marital status should (a) impute a value when an entry is out of range and (b) check for consistency between reported marital status and relationship and age.

(d) Spouse should be married

376. All persons coded “spouse” in the relationship category should be coded as married.

(e) Spouse of a married-couple pair

377. 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.

(f) If spouse, head should be married

378. 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”.

(g) Head, no spouse, without children

379. 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.

(h) If all else fails, impute

380. 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.

(i) Relationship of age to marital status for young people

381. 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 per-

son 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.

5. Age at first marriage

382. According to *Principles and Recommendations for Population and Housing Censuses, Rev. 2* (para. 2.192), “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.

383. The date of first marriage should be entered for all ever-married persons. 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.

(a) Age at marriage for never-married persons should be blank

384. 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.

(b) Ever-married persons should have an entry

385. For the year of first marriage, countries not using dynamic imputation can assign “not stated” or “unknown”. Countries using dynamic imputation can use other variables, such as age of spouse or age differences between spouses, number of children and children born in the last year, to determine an appropriate year of first marriage.

6. Fertility: children ever born and children surviving

386. “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.

387. The universe for which data should be collected for each of the topics included in this section consists of women 15 (or some other minimum acceptable age) years of age and over, regardless of marital status or of particular subcategories such as ever-married women. In countries that do not collect or tabulate data for women 50 years of age and over, efforts should be concentrated on collecting data from women between 15 and 50 years of age only; in the investigation of recent fertility it may be appropriate in some countries to reduce the lower age-limit by several years (United Nations, 2007 para. 2.170).

(a) Fertility items collected

388. In *Principles and Recommendations for Population and Housing Censuses*, the United Nations (2007) 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.

389. Censuses and surveys collect information on fertility from all females, using a country-defined minimum age and sometimes a maximum age as well.

(b) General rules for the fertility edit

390. Females younger than the designated earliest age for fertility, and all males, should be checked, and any present fertility information should be blanked out.

391. 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 section A.8 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.

(c) Relationship between children born and children surviving

392. 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.

393. 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).

(d) Edit when only children ever born is reported

394. If the country does not use dynamic imputation, an invalid or missing value for “children ever born” should be assigned as “unknown”. In countries using dynamic imputation, the specialists must decide whether they want to use dynamic imputation for all items. If the specialists use this method, children ever born can be obtained based on single year of age of the female and at least one other characteristic. It is also possible to use a single dimensional array for single year of age of mother only. The other characteristics might be items such as educational attainment or religion, since it is known that in many countries differential fertility exists for various levels of educational attainment or different religious affiliations.

(e) Edit when children ever born and children surviving are reported

395. 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 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.

396. 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

397. 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 miskeyed 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.

398. 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.

399. 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.

400. 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.

401. 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.

402. 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

Figure 28

Initial values for determining children surviving when age and value for 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

the imputed values for children surviving. Sometimes the responses will be appropriate, but sometimes they will not. If the program encounters a 19-year-old female with 5 children ever born, the value of 5 children surviving should probably pass the age difference criteria (an age difference of 15, based on children surviving and reported age). However, for a 15-year-old, neither the 5 children ever born (age difference of 10) or 4 children surviving (age difference of 11) would be acceptable.

403. 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.

(f) Edit when children ever born, children surviving, and children who died are reported

404. “Children ever born” is the sum of “children surviving” and “children who died”. Any inconsistency may be resolved as explained below.

(i) When all three items are reported

405. 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.

406. If all of these are consistent, the edit is finished. However, if any are inconsistent, the edit must resolve them. The three items may not be internally consistent: for example, a female may have 5 children ever born, but only two children surviving and two deceased. The editing team should decide which variable takes precedence over the others. In many cases, the female is likely to remember all of the children she has ever borne, although she may forget the exact number who died. Then, the editing team may choose to accept the number of children ever born and those surviving, and subtract to obtain a new, consistent value for deceased children.

(ii) When two items are reported

407. 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:

- (a) If CEB and CS are known, $CD = CEB - CS$.
- (b) If CS and CD are known, $CEB = CS + CD$.
- (c) 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.

(iii) When one item is reported

408. 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 dynamic imputation, editing teams need to determine a method of getting at least one more item and the third item should then be obtained through subtraction or addition. A two-dimensional matrix can be used to get the second fertility value, based on the first item and single year of age for the females. If children ever born is known, for example, children surviving can be obtained from the imputation matrix, as described above, and then dead children should be obtained by subtraction. Similarly, if children surviving is known, children ever born is obtained from the imputation matrix of single year of age of female and children surviving, and number of dead children is obtained by subtraction.

(iv) When none of the items are reported

409. 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.

410. 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.

411. Once the first item is determined, to obtain the second fertility item it is possible to follow the steps outlined above for editing when only one item is reported. Then, the third item can be obtained from the first two items. The three items should be compatible because the imputation matrices should be updated only when all items are compatible. The fertility obtained should also be compatible with other females in the geographical area since information from those females is used to update the imputation matrix.

*(g) Edit when both children ever born, children living at home, children living away and children who died are reported**(i) When all four items are reported*

412. 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.

413. 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 described above, all four items may not be internally consistent. Again, the editing team should decide which variable takes precedence over the others. In many cases, the female respondent is likely to remember all of the children she has ever borne, although she may forget some of those who moved away or the exact number who died. Then, the editing team may choose to accept the number of children ever born and those surviving (the sum of the children living away and the children living at home), and subtract to obtain new, consistent values for other variables. The editing team may need to develop algorithms for various combinations of events.

(ii) When three of the four items are reported

414. The children ever born (CEB) is the sum of the children living at home (CLH), the children living away (CLA) and the children who died (CD). If any three of the four pieces of information are available, the computer program can determine the fourth variable:

If CEB, CLH and CLA are known, $CD = CEB - CLH - CLA$.

If CLH, CLA and CD are known, $CEB = CLH + CLA + CD$.

If CEB, CLH and CD are known, $CLA = CEB - CLH - CD$.

If CEB, CLA and CD are known, $CLH = CEB - CLA - CD$.

(iii) When two of the four items are reported

415. 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.

416. 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.

417. 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 chil-

dren 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: $CLA = CEB - CLH - CD$.

418. 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).

419. 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.

420. 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.

421. In each case, two of the four items are available. The third item is obtained by dynamic imputation, and the fourth item by subtraction or addition. Editing teams must decide which is the best path to follow based on cultural circumstances.

(iv) When only one item is reported

422. 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

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

information is missing. If dynamic imputation is used, the first imputation matrix would, as noted above, use an item such as single year of age of female and the one known item to create a two-dimensional matrix for imputation of any one of the other items. Once two items are determined, the other two remain unknown, by definition. Hence, continuing to use dynamic imputation for the third item should not create an incompatibility with the other items since they are unknown. The scheme discussed above for two known items and two unknown items, is used to obtain a third item. Then, the fourth item is obtained by subtraction. All four items should be compatible.

(v) When none of the items are reported

423. 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.

424. 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.

425. Once the first item is determined, the approach used above when only one item is known can be used to obtain the second fertility item. Then, the third item can be obtained from the first two items, and the fourth item can be obtained by subtraction. The four items should be compatible because the imputation matrices should only be updated when all items are compatible. The fertility obtained should also be compatible with other females in the geographical area since information from these females is used to update the imputation matrix.

(h) Special case of five or more items

426. 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. Moreover, as noted, it is a good idea to have a single array line for each age of woman, with complete fertility information entering in when all items are valid, internally consistent and consistent with her age; then, in cases where fertility information is inconsistent (including with age), imputation can be applied using the whole, appropriate line.

(i) Importance of a single donor source for all fertility items

427. Therefore, it is very important, if at all possible, 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 indicated in paragraph 426 above, it is better not to impute item by item, but when several items are amiss, to use another woman’s total information.

(j) Relationship of own children to children in the house and children surviving

428. 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. As very few countries use this method to assist in the edit, it remains experimental; however, the results look promising.

7. Fertility: date of birth of last child born alive and births in the 12 months before the census

429. Information on last births assists in providing estimates of current fertility just prior to a census or survey. One approach is to obtain the date of birth (day, month and year) of the last child born alive and the child's sex and to determine whether the child is alive or not. A second approach is to collect information on births in the 12 months before the census; it is easier for enumerators and respondents to make use of this second approach because only a single "yes" or "no", rather than an exact date, is needed.

430. During processing, an estimate of the number of children born alive in the 12 months immediately preceding the census date can be derived from information on date of birth of last child born alive (and then kept as a recode). For estimating current age-specific fertility rates and other fertility measures, the data provided by this approach are more accurate than information on the number of births to a woman during the 12 months immediately preceding the census (United Nations, 2008, paras. 2.188-2.191).

431. It should be noted that information on the date of birth of the last child born alive does not produce data on the total number of children born alive during the 12-month period. Even if there are no errors in reporting the data on the last live-born child, this item only ascertains the number of women who had at least one live-born child during the 12-month period, not the number of births, since a small proportion of women will have had more than one child in a year (United Nations, 2007, para. 2.189).

432. The information needs to be collected only for women between 15 and 50 years of age who have reported having at least one live birth during their lifetime. In addition, the information should be collected for all the marital-status categories of women for whom data on children ever born by sex are collected. If the data on children ever born are collected for a sample of women, information on current fertility should be collected for the same sample (United Nations, 2007, para. 2.190).

433. 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).

434. 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 demographers, want actual age of mothers at birth of their children as a recode for fertility analysis, then at the very least, the month of last birth should probably be imputed if it is not present. The recode can then be obtained.

435. Similarly, some demographers want to analyse 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 be used in the determination of the recode for months since last birth. (See annex I for a method of obtaining a recode for months since last birth.)

436. 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.

437. 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 entails checking within the household for a child or children zero years old, and using the relationships of the mother and that child or those children (or mother’s person number for the child, if collected) to determine whether the child is reported as a last birth for the mother. The checks should go both ways: the zero-year-olds should be checked against mothers and the last births should be checked against the household listing.

438. Those countries also collecting deaths in the year before the census or survey may decide to include a check of deaths of 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 be feasible if the mother is not in the house because of death or movement, or if, for some reason, the child is not reported, some proportion of infant deaths could still be checked in this manner.

8. Fertility: age at first birth

439. The age of the mother at the time of the birth of her first live-born child is used for the indirect estimation of fertility based on first births and to provide information on the onset of childbearing. If the topic is included in the census, information should be obtained for each woman who has had at least one child born alive (United Nations, 2008, para. 2.193).

440. 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.

441. 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.

9. Mortality

442. Information on deaths in the past 12 months is used to estimate the level and pattern of mortality by sex and age in countries that lack satisfactory continuous death statistics from civil registration. In order for estimates derived from this item to be reliable, it is important that deaths in the past 12 months by sex and age be reported as completely and as accurately as possible. The fact that mortality questions have been included extensively in the census questionnaire in the past decades has resulted in an improvement in the use of indirect estimation procedures for estimates of adult mortality (United Nations, 2007, para. 2.194).

443. 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 (United Nations, 2007, para. 2.195).

(a) Age and sex of the deceased

444. The *Principles and Recommendations for Population and Housing Censuses, Rev. 2* (United Nations, 2007) suggests 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 subnational geographical levels, can be of use in developing the most appropriate edit.

(b) Cause of death

445. 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 the 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 entail to assumption that, if the information is not collected or is invalid, the value will normally be “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.

(c) Maternal mortality

446. In the current census round, more and more countries are also asking if the deceased person was female, and 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 chooses imputation, the hot deck would be applied to females only, obviously, and only for likely ages of pregnancy—probably 12-54 years—and probably by single year, rather than by 5-year age groups.

(d) Infant mortality

447. Finally, the *Principles and Recommendations for Population and Housing Censuses, Rev. 2* (para. 2.191) suggests collecting information on deaths among children born “in the past 12 months”. Normally, this question would be asked only in conjunction with the item on births in the 12 months before the census. If the other current fertility item, date of last birth, is used, then this item probably should not be used.

448. 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, an edit check of last births who died in the year before the census against deaths of zero-year-olds in that year is an appropriate one and could provide useful information on infant mortality.

449. 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 to conduct. With several females in the unit, care must be taken to ensure that the right children are connected to the appropriate women.

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

450. 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 (United Nations, 2007, para. 2.199).

451. The edits for “mother living” and “mother’s line number” items are inter-related 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.

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

B. Migration characteristics

453. 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 previous, specified point in time. Since these items are often inter-related, 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.

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

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

456. 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.

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

1. Place of birth

458. The place of birth is, in the first instance, the country in which the person was born. It should be noted that the country of birth is not necessarily related to citizenship, which is a separate topic (see United Nations, 2007, para. 2.92 to 2.102 and the section on country of birth in this *Handbook*). For persons born in the country where the census is taken (natives), the concept of place of birth also includes the specified type of geographical unit of the country in which the mother of the individual resided at the time of the person's birth. In some countries, however, the place of birth of natives is defined as the geographical unit in which the birth actually took place. Each country should explain which definition it has used in the census (United Nations, 2007, para. 2.57).

(a) Relationship of entries for country of birth and years lived in district

459. 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.

(b) Assigning “unknown” for invalid entries for birthplace

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

(c) Using static imputation for birthplace

461. 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”.

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

(d) Using dynamic imputation for birthplace

463. 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.

(e) Assigning birthplace when a person’s mother is present

464. 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.

(f) Assigning birthplace for child of head

465. 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.

(g) Assigning birthplace for child, but not of head

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

(h) Assigning birthplace for adult females with husband

467. 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.

(i) Assigning birthplace for adult females with no husband

468. 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.

(j) Assigning birthplace for males

469. 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

470. 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.

471. 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 (United Nations, 1999). 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 subregional classifications identified in the above-mentioned publication is recommended (United Nations, 2008, para. 2.97).

(a) Citizenship edit

472. 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.

(b) Relationship of ethnicity/race to citizenship

473. Some countries also collect "ethnicity" or "race" which may give additional information to be used in determining citizenship, particularly when the collected response is invalid. For many countries, first and second generation migrants should have almost complete consistency between their ethnic origin and their citizenship. For countries with a long history of international immigration, this characteristic may be less valuable, but still might be considered with other variables.

(c) Relationship of naturalization to citizenship

474. 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 fulfilment 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 fulfil the duration of residence requirements for naturalization would be assigned as "foreign", using the cold deck method of imputation.

(d) Relationship of duration of residence to citizenship

475. The item "duration of residence" may not appear on the questionnaire or may be ambiguous in determining citizenship, or the editing team may choose not to use it. Then, if the value for citizenship is invalid or inconsistent with birthplace, "unknown" should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics

(and one should probably be birthplace) to obtain “known” information from similar persons in the geographical area.

3. Duration of residence

476. The duration of residence is the interval of time up to the date of the census, expressed in complete years, during which each person has lived in (a) the locality that is his or her usual residence at the time of the census, and (b) the major or smaller civil division in which that locality is situated (United Nations, 2008, para 2.64).

(a) Edit for duration of residence

477. 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”.

(b) De facto/de jure residence and duration

478. 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.

(c) Relation of age to duration of residence

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

(d) Relation of birthplace to duration

480. 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.

(e) For persons who have always lived here

481. 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.

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

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

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

483. 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.

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

484. 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

485. The place of previous residence is the major or smaller civil division, or the foreign country, in which the individual resided immediately prior to migrating to his or her present civil division of usual residence (United Nations, 2007, para. 2.67).

(a) Previous residence edit

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

(b) Previous residence when boundaries have changed

487. 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.

(c) When person has not moved since birth

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

(d) Use of other persons in unit

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

(e) No appropriate other person for previous residence

490. 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

491. 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.

492. 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 (United Nations, 2008, para. 2.69).

493. “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

494. *Principles and Recommendations for Population and Housing Censuses, Revision 2*, divides migration variables into internal migration and international migration. 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 for along with its complement, that is to say, place of residence before arrival in this country (United Nations, 2008, para. 2.103).

(a) Relation of age to year of arrival

495. 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 number of years in locality or civil division is greater than age, and country of birth is this country. If number of years since arrival in the locality or civil division is greater but country of birth is not this country, one method for dealing with this case 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.

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

(b) Relation of birthplace to year of arrival

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

(c) For persons who have always lived here

498. If the response for the number of years since arrival that 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 of arrival in locality or civil division. Since the length of time that a person lived in the locality or civil division cannot be greater than the person’s age, for this situation, age will be assigned as comprising the years in locality or civil division.

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

499. 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 she is found, her record can provide information helpful in assigning missing values. If the person’s mother has always lived in the locality or civil division, and her country of birth is “this country” (as it should be), the program will assign “always” to this person’s years in locality or civil division category. If the mother’s country of birth is not “this country”, even though the entry for her years in the locality or civil division is “always”, this indicates the presence of an error connected 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 the number of years since the mother’s arrival in locality or civil division, the program will assign “mother’s arrival year in locality or civil division” to this person’s year of arrival in locality or civil division.

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

500. If the person in question is a child (son or daughter), the editing program should check the head of household’s record for information that might aid in assign-

ing 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 still 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 that child. If the child’s age is less than that determined by 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.

(f) Person’s year of arrival when no other information is available

501. 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

502. 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 will therefore include both items.

503. 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 to say, the variables for age, duration of residence, and year of arrival must be considered together to ensure that sum of the duration of residence *and* time since arrival is not greater than age. Hence, programmers will need to consider all three variables at the same time when this occurs.

504. When dynamic imputation is used, the statistical staff may need to use a hot deck that includes multidimensional 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 five-year group, for example, may cause a conflict during imputation.

505. Also, great care is needed when grouped data for duration of residence or year of entry or both are collected in the course of 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 in this situation.

8. Usual residence

506. In general, “usual residence” is defined for census purposes as the place at which the person lives at the time of the census, and where he or she has stayed for some time or intends to stay for some time (United Nations, 2008, paras. 1.461-1.463). It is recommended that countries apply a threshold of 12 months when considering place of usual residence, according to one of the following two definitions:

- (a) The place at which the person has lived continuously for most of the last 12 months (that is to say, 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.

507. Countries conducting *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; thus, missing information can be filled in directly.

508. However, when the data show movement, the situation becomes more complicated. Usually, when this item is left blank, countries assume that the usual residence and the current residence are the same and that the enumerator and/or respondent simply left the information out.

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

510. If the enumerator is instructed to leave the entry blank when the usual residence is the same as the place of enumeration, the code for the place of enumeration should be placed, during edit, in the item for usual residence. 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

511. Social characteristics vary from country to country, but are generally items that describe various aspects of sociocultural 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) (United Nations, 2008, paras. 2.202-2.230).

1. Ability to read and write (literacy) (P6A)

512. Data on literacy should be collected for all persons 10 years of age and over. In a number of countries, however, certain persons between 10 and 14 years of age may be about to become literate through schooling and the literacy rate for this age group may be misleading. Therefore, in an international comparison of literacy, data

on literacy should be tabulated for all persons 15 years of age and over. Where countries collect data on younger persons, tabulations for literacy should at least distinguish between persons under 15 years of age and those 15 years of age or over (United Nations, 2008, para. 2.202).

513. 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. However, the item should be filled during the edit to assist researchers and others using the public-use data.

514. 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”.

2. School attendance (P6B)

515. 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 (United Nations, 2008, para. 2.209).

(a) School attendance edit

516. 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.

(b) Full-time or part-time enrolment

517. Some countries may want to obtain information on part-time or full-time attendance in school. If this item is included, it may need to be part of the school attendance edit, or it may be a separate edit.

(c) Consistency between school attendance and economic activity

518. 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.

(d) Assignment for invalid or inconsistent entries for school attendance

519. 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.

520. 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)

(a) Edit for educational attainment

521. 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 (United Nations, 2008, para. 2.215). Annex I should be consulted for suggestions regarding a recode for “current grade” based on school attendance and highest grade attended.

(b) Minimum age for educational attainment

522. 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.

(c) Relationship of age to educational attainment

523. 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, for cases where the original entry is out of range or inconsistent with age, an entry can be assigned. For countries not using dynamic imputation, “not stated” can be entered. For those using dynamic imputation, an entry can be obtained based on age (including single year of age for persons of school age), sex and school attendance. 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.

4. Field of education and educational qualifications

524. 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 (United Nations, 2008, para. 2.223).

525. 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.

5. Religion

526. For census purposes, religion may be defined as either (a) religious or spiritual belief of preference, regardless of whether or not this belief is represented by an organized group, or (b) affiliation with an organized group having specific religious or spiritual tenets. Each country that investigates religion in its census should use the definition most appropriate to its needs and should set forth, in the census publication, the definition that has been used (United Nations, 2008, para. 2.152).

(a) Religion edit

527. Religion is one of the variables fitting the standard edit examples introduced in chapter II. For the items on religion, unlike others, a “non-response” is significant and needs accounting for: 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:

(b) No religion for head of household, but religion present for someone else in the unit

528. The first step is to determine if anyone else in the housing unit has a valid religion, and assign the first valid religion.

(c) No religion for head, or for anyone else in unit

529. 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.

(d) For person other than head, without religion

530. If this person is not the head of household and reports no religion, the editing team may decide to assign the head's religion.

6. Language

531. Three types of language data can be collected in censuses (United Nations, 2008, para. 2.156), 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.

(a) Language edit

532. 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.

(b) Language edits: head of household

533. 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 dynamic imputation is not used 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.

(c) Language edits: persons other than head of household

534. If the person is not the head of household and the language is invalid, then assign the head of household's language.

(d) Language edits: use of ethnic origin or birthplace

535. 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.

(e) Language edit: mother tongue

536. 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.

(f) Language edits: ability to speak a designated language

537. 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.

7. Ethnicity and indigenous peoples

538. The decision, in a country census, to collect and disseminate information on ethnic or national groups of a population is dependent upon a number of considerations and national circumstances, including, for example, the national needs for such data, and the suitability and sensitivity of asking ethnicity questions in that census. Identification of the ethnocultural characteristics of a country's population has increasing importance in the context of migration, integration and policies affecting minority groups (United Nations, 2008, para. 2.160).

539. *Principles and Recommendations for Population and Housing Censuses, Rev. 2* (para. 2.163) suggests taking particular care in identifying indigenous peoples, usually as a subset of the ethnicity item. Care must be taken in developing the code lists to ensure that indigeneity is identified uniquely so that appropriate edits and tabulations can be developed to assist in planning and policy formation for indigenous peoples. For example, separate codes may be needed for the same group when they are nomadic compared with when they have settled into residential areas. Special edits can be developed, partially through "lookup" files for particular groups of indigenous peoples, to ensure that they are properly and fully identified for subsequent tables. Special imputation procedures can be developed for these groups, or additional categories within existing hot decks can be used.

(a) Ethnicity edit

540. 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.

(b) Ethnicity edit: for head of household

541. Ethnic origin also fits the example introduced in chapter II. Editing teams should follow the scheme already described for the other social variables. 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).

(c) Ethnicity edit: persons other than head of household

542. If the person is not the head of household and ethnic origin is invalid, then assign the head of household’s ethnic origin.

(d) Ethnicity edit: use of language and birthplace

543. 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.

544. 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.

8. Disability

545. Disability status separates the population into those with and those without a disability. A person with a disability should be defined as a person who is at greater risk than the general population for experiencing restrictions in respect of performing specific tasks or of participating in role-related activities. The United Nations recommends inclusion of four domains in assessing disability status: (1) walking, (2) seeing, (3) hearing and (4) cognition (United Nations, 2008, paras. 2.351-2.352 and 2.367-2.371).

546. The question used to identify persons with disabilities should list broad categories of disabilities so that each person can check the presence or the absence of each type of disability. Disability may be monitored through use of the following components based on the International Classification of Impairments, Disabilities and Handicaps (ICIDH): (1) functioning and disability, including body functions and body structures (impairments) and activities (limitations) and participation (restrictions) and (2) contextual factors, including environmental factors and personal factors (United Nations, 2008, para. 2.354).

(a) Disability census questions

547. It is recommended that special attention be paid in designing census questions to measuring disability. The wording and the construct of questions greatly affect their precision in identifying people with disabilities. Each domain should be associated with a separate question.⁸ The language used should be simple, clear and unambiguous. Negative terms should always be avoided. The disability questions should be addressed to each single household member and general questions on the presence

⁸ When domains are combined, through asking, for example, a question about seeing or hearing, respondents are frequently confused, assuming that they need to have difficulty in both domains in order to answer yes. Furthermore, having numbers for those with specific limitations is useful both for internal planning and for cross-national comparisons.

of persons with disabilities in the household should be avoided. If necessary, a proxy respondent can be found to report for the family member who is incapacitated. What is most important is that each family member be accounted for individually rather than surveyed through a blanket question. Scaled response categories can also improve the reporting of disability (United Nations, 2008, para. 2.373).

(b) Disability edit

548. When someone does not respond to the disability-related questions asked, it is difficult to determine whether the item is left blank because of no disability or because of an unwillingness on the part of the respondent to answer, for whatever reason. A country's editing team must decide whether they want to edit the item in the usual way, by assigning unknowns when dynamic imputation is not used, or by using the responses of other individuals when dynamic imputation is used. 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.

(c) Multiple disabilities

549. 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.

(d) Cause of disability edit

550. 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.

D. Economic characteristics

551. 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.

552. 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 (United Nations, 2008, para. 2.241).

553. 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

554. 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.

555. "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 (United Nations, 2008, para. 2.248).

556. According to the United Nations (2008, para. 2.253) the employed comprise all persons above a specified age who, during a short reference period of either one week or one day (a) performed some work for pay, profit or family gain, in cash or in kind; or (b) were temporarily absent from a job in which they had already worked and to which they maintained a formal attachment or from a self-employment activity such as a business enterprise, a farm or a service undertaking.

557. The population *not currently active* or, equivalently, persons not in the *labour force*, comprises all persons who were neither *employed* nor *unemployed* during the short reference period used to measure current activity, including persons below the minimum age specified for measurement of the economically active population (United Nations, 2008, paras. 2.278-2.279).

(a) Categories related to activity status

(i) Unemployed population

558. The unemployed population comprises, according to the United Nations (1998, para. 2.271), all persons above a specified age who, during the reference period, met the following conditions:

- (a) Without work: they were not in paid employment or self-employment;
- (b) Currently available for work: they were available for paid employment or self-employment during the reference period;
- (c) Seeking work: they took specific steps in a specified recent period to seek paid employment or self-employment. The specific steps may have included regis-

tration 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.

559. 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" (United Nations, 2008, para. 2.272).

560. The edits for unemployment—"on layoff", "seeking 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.

(ii) *Seeking work*

561. The edit for "seeking work" should be carried out 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.

(iii) *Not currently active*

562. 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 (United Nations, 2008, para. 2.278). They may, according to their reasons for not being "currently active", be classified in any of the following groups:

- (a) Attending an educational institution;
- (b) Performing household duties;
- (c) Retiring on pension or capital income;
- (d) Other reasons.

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

(iv) *Why not looking for work*

563. This item should be edited only for persons who were recorded as "not looking for work"; all others should have a blank entry. Alternatively, if a valid entry

appears in occupation, industry and status in employment, the code for “with a job but not at work” should be entered. This code designates economically active persons who were employed but were not at work during the reference period. In all other cases, if dynamic imputation is not used, “unknown” can be assigned. For countries using dynamic imputation, an entry can be allocated using age, sex and major activity.

(b) Editing for economic activity status

564. Economic activity generally encompasses the following categories:

Economically active

- (1) Employed
- (2) Unemployed

Not economically active

- (1) Student
- (2) Homemaker
- (3) Pension or capital income recipient
- (4) Other

(i) Employed persons

565. If category 1 (“Employed”) is selected, the variables for time worked, occupation, industry, economic activity status, and workplace 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 category 1 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.

(ii) Economic activity of unemployed persons

566. If category 2 (“Unemployed”) is selected, the variables for “on layoff”, “seeking 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 2 through 6 are selected, the variables for time worked, occupation, industry, economic activity status, and workplace should be blank. If they are filled, they should be BLANK.

(iii) Economic activity of students and retired persons

567. If category 3, 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 5, pensioner, is selected, the subject matter personnel need to decide whether persons must be of a certain age to be retired.

(iv) When economic activity is not valid and employed variables are reported

568. 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 1, employed. An imputation matrix will probably be needed to select the appropriate response.

(v) When economic activity is not valid and the unemployed variables are reported

569. 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 2 to 6. If the person is attending school, that value should probably be 3. If the person is elderly, the value should probably be 5. Otherwise, the subject matter specialists may decide to use an imputation matrix to allocate an appropriate response.

(vi) *When economic activity is not valid and none of the economic variables are reported*

570. 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.

2. Time worked

571. 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. (United Nations, 2008, para. 2.323).

572. 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.

573. If dynamic imputation is used, the minimal variables for the imputation matrix includes age groups and sex, but other variables such as educational attainment, occupation or industry major categories can also be used.

3. Occupation

574. Occupation refers to the type of work performed in a job by the person employed (or the type of work done previously, if the person is unemployed), irrespective of the industry or the status in employment in which the person should be classified (United Nations, 2008, para. 2.301).

575. This item should be edited only for persons whose economic activity is “employed”. If dynamic imputation is not used, blank, zero or invalid responses should be changed to “not reported”.

576. 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.

577. 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.

4. Industry

578. According to *Principles and Recommendations for Population and Housing Censuses, Revision 2* (para. 2.306) “industry” refers to the activity of the establishment in which an employed person worked during the time-reference period established for data on economic characteristics (or last worked, if unemployed). For guidance on the selection of the job/activity to be classified, see paragraph 2.307 thereof.

579. This item should be edited only for persons whose economic activity was “employed”. If dynamic imputation is not used, blank, zero or invalid responses should be changed to “not reported”.

580. 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.

581. If dynamic imputation is used, minimal variables for the imputation matrix include age groups and sex, but other variables such as educational attainment or industry major categories can also be used.

5. Status in employment

582. 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 (United Nations, 2008, para. 2.310).

583. The economically active population should be classified by status in employment as follow (United Nations, 2008, para. 2.311):

- (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’ cooperatives;
- (f) Persons not classifiable by status.

584. 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.

585. This item should be edited only for persons whose economic activity is “employed”. If dynamic imputation is not used, blank, zero or invalid responses can be changed to “not reported”. If dynamic imputation is used, minimal variables for the

imputation matrix include age groups and sex, but other variables such as educational attainment or industry major categories can also be used.

6. Income

586. 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) (United Nations, 2008, para. 2.331). 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 (United Nations, 2008, para. 2.237).

587. 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.

588. *Principles and Recommendations for Population and Housing Censuses, Rev. 2* identify 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.

589. 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 annex I.

7. Institutional sector

590. 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 (United Nations, 2008, para. 2.335).

591. 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.

592. 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

593. Where informal sector activities play an important role in employment creation and income-generation, 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 (United Nations, 2008, para. 2.337).

594. The edit for the informal sector should be straightforward. 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 informal sector is not independent of participation in the formal sector, then an additional variable in the hot deck matrix could be included to indicate whether the person was in the formal sector as well (United Nations, 2008, para. 2.343).

9. Place of work

595. “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 main 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 (United Nations, 2008, para. 2.346).

596. 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.

597. 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.

598. 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.

599. The present chapter has 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 hothouse and pre-census survey situations for reliable and complete responses. Since the population items, unlike the housing variables, are usually cross-tabulated in many different combinations, thorough testing is needed.

Chapter V

Housing edits

600. 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.

601. 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.

602. 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 chapter [III](#), which 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.

603. 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.

604. Nonetheless, for aesthetic more than for technical reasons, and particularly for housing items, as editing has become more sophisticated and detailed, greater emphasis is now placed on ensuring that 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 geographical areas

so as to ensure that no anomalies slip through into the final dataset. It is probably best to adopt the “shotgun” approach and to remove cases that may actually be extraneous. For example, although there may be cases where wealthy individuals in an area purchase gas generators for use when electricity is not otherwise available, the editing team can decide not to include those cases in the dataset.

605. 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.

606. 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.

607. 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.

608. 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.

609. 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.

610. 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.

611. 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.

A. Core and additional topics

612. 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.

1. Living quarters—type of (core topic)

613. The classification outlined below describes a system of three-digit codes designed by the United Nations (2008, paras. 2.412-2.454) to group, in broad classes, housing units and collective living quarters with similar structural characteristics. The distribution of occupants (population) among the various groups supplies valuable information about the housing accommodations available at the time of the census. The classification also affords a useful basis of stratification for sample surveys. The living quarters may be divided into the following categories:

- 1 Housing units
 - 1.1 Conventional dwellings
 - 1.1.1 Has all basic facilities
 - 1.1.2 Does not have all basic facilities
 - 1.2 Other housing units
 - 1.2.1 Semi-permanent housing units
 - 1.2.2 Mobile housing units
 - 1.2.3 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 (for example, 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

614. Editing teams should develop edits that ensure that all collective living quarters and housing units have internally consistent information. If the value for type of living quarters is unknown or invalid, editing teams might want to develop an edit that looks at other variables to assign type of living quarters. Otherwise, if the value is invalid, “unknown” should be assigned when dynamic imputation is not used. National statistical/census offices choosing dynamic imputation for invalid values should use at least two characteristics, such as type of building, tenure, number of rooms, floor space or vacancy status, to obtain “known” information from similar housing units in the geographical area.

2. Location of living quarters (core topic)

615. Location of living quarters is a geographical variable and is presented with the structure edits in chapter III.

3. Occupancy status (core topic)

616. The decision to record housing units 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 housing units used as a primary residence from those that are used as a secondary residence. This is particularly important if the secondary residence has markedly different characteristics from the primary residence, as is the case, for example, when persons in agricultural households move during certain seasons of the year from their permanent living quarters in a village to rudimentary structures located on agricultural holdings (United Nations, 2008, para. 2.466). The recommended classification of occupancy status for conventional 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

617. 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”.

618. If the value is invalid, the value for number of occupants is zero and no population records are present, “unknown vacant” should be assigned when dynamic imputation is not used. If the value is invalid, but the number of occupants is not zero or population records are present, “occupied” should be assigned. Countries choosing dynamic imputation for invalid values (to impute type of vacancy) should use at least two characteristics to obtain “known” information from similar housing units in the geographical area, or, alternatively, “unknown vacant” can be assigned.

4. Ownership—type of (core topic)

619. The present topic refers to the type of ownership of the housing unit itself and not of that of the land on which it stands (United Nations, 2008, para. 2.467). Type of ownership should not be confused with tenure. Information should be obtained to show whether the housing units are owned by the public sector (central Government, local government, public corporations) or are privately owned (by households, private corporations, cooperatives, housing associations or other). The question is sometimes expanded to show whether the housing units are fully paid for, are being purchased in instalments or are mortgaged. The classification of housing units by type of ownership is as follows:

- 1 Owner-occupied
- 2 Non-owner-occupied
 - 2.1 Publicly owned
 - 2.2 Privately owned
 - 2.3 Communally owned
 - 2.4 Cooperatively owned
 - 2.5 Other

620. If ownership is related to tenure, this should be taken into account in developing the edit; if it is not related, then the type of ownership is probably independent of other housing variables. If the value for “type of ownership” is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics which might include construction material of walls, tenure, type of housing unit and number of rooms, in order to obtain “unknown” information from similar housing units in the geographical area.

5. Rooms—number of (core topic)

621. A room is defined as a space in a housing unit or other living quarters enclosed by walls reaching from the floor to the ceiling or roof covering, or to a height of at least two metres, of an area large enough to hold a bed for an adult, that is, at least four square metres. The total number of types of rooms therefore includes bedrooms, dining rooms, living rooms, studies, habitable attics, servants’ rooms, kitchens, rooms used for professional or business purposes and other separate spaces used or intended for dwelling purposes, so long as they meet the criteria concerning walls and floor space. Passageways, verandas, lobbies, bathrooms and toilet rooms should not be counted as rooms, even if they meet the criteria. Separate information may be collected for national purposes on spaces of less than four square metres that conform in other respects to the definition of “room” if it is considered that their number warrants such a procedure (United Nations, 2008, para. 2.472).

622. Since the number of rooms may be independent of the other housing variables, if the value is invalid, “unknown” should be assigned when dynamic imputation

is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, construction material of walls, tenure and vacancy status, to obtain “known” information from similar housing units in the geographical area.

6. Bedrooms—number of (additional topic)

623. In addition to enumerating the number of rooms, a number of national censuses collect information on the number of bedrooms in a housing unit, which is the unit of enumeration for this topic. A bedroom is defined as a room equipped with a bed and used for night rest (United Nations, 2008, para. 2.475).

624. Sometimes enumerators report a value for the number of bedrooms that is greater than the value for the number of rooms.⁹ 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.

625. Otherwise, if the value for bedrooms is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics (with one of them being number of rooms) to obtain “known” information from similar housing units in the geographical area.

7. Useful floor space (additional topic)

626. 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 (United Nations, 2008, para. 2.476).

627. Floor space may relate to number of rooms and/or number of bedrooms, so country editing teams may want to take these into account when developing the edits. Other useful items for dynamic imputation include number of occupants and occupants per room. For the most part, floor space is independent of other housing edits. A unit of measurement, such as square metres, may need to be specified. If the value is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including type of housing unit, construction material of walls, tenure and vacancy, to obtain “known” information from similar housing units in the geographical area.

8. Water supply system (core topic)¹⁰

628. According to the United Nations (2008, para. 2.479), 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

⁹ 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.

¹⁰ For the following variables, the unit of enumeration is actually the housing unit’s water supply system, toilet and sewerage facilities, bathing facilities, cooking facilities, lighting and solid waste disposal.

water is provided to the housing unit 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 an individual 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 an individual source
 - 2.2.1 For exclusive use
 - 2.2.2 Shared
- 3 Other

629. 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.

630. 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.

631. If the value for water system is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics. These might include as a rule, type of housing unit, and then toilet and sewerage facilities, and bathing facilities, to obtain “known” information from similar housing units in the geographical area.

9. Drinking water—main source of (core topic)

632. Drinking water should be edited with water system. Many of the criteria described above also apply here. Since bottled and other non-traditional sources of drinking water will normally be included on the questionnaire, they must also be included in the edit (United Nations, 2008, para. 2.483).

633. If the value for drinking water is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics. These might include, as a rule, type of housing unit, and then water system, toilet and sewerage facilities, and bathing facilities, to obtain “known” information from similar housing units in the geographical area.

10. Toilet—type of (core topic) and

11. Sewage disposal (core topic)

634. Toilet facilities and sewerage should be edited together with the other plumbing variables to obtain the most consistent results. Although *Principles and Recommendations for Population and Housing Censuses, Rev. 2* combined the two variables, they have been separated for 2010. Nonetheless, these items should be edited together, and, if possible, the same dynamic imputation matrices should be used.

635. 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 (2008, para. 2.487) recommendations for classification of housing unit by toilet facilities include the following:

- 1 With toilet within housing unit
 - 1.1 Flush/pour flush toilet
 - 1.2 Other
- 2 With toilet outside housing unit
 - 2.1 For exclusive use
 - 2.1.1 Flush/pour flush toilet
 - 2.1.2 Ventilated improved pit latrine
 - 2.1.3 Pit latrine without ventilation with covering
 - 2.1.4 Holes or dug pits with temporary coverings or without shelter
 - 2.1.5 Other
 - 2.2 Shared
 - 2.2.1 Flush/pour flush toilet
 - 2.2.2 Ventilated improved pit latrine
 - 2.2.3 Pit latrine without ventilation with covering
 - 2.2.4 Holes or dug pits with temporary coverings or without shelter
 - 2.2.5 Other
- 3 No toilet available
 - 3.1 Service or bucket facility (excreta manually removed)
 - 3.2 Use of natural environment, for example, bush, river, stream and so forth

636. 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.

12. Bathing facilities (core topic)

637. According to the United Nations (2008, para. 2.490), information should be obtained on whether or not a fixed bath or shower is installed within the premises

of each set of housing units. The unit of enumeration for this topic is also a housing unit. Additional information may be collected to show if the facilities are for the exclusive use of the occupants of the living quarters and if there is a supply of hot water for bathing purposes or cold water only. However, in some areas of the world the distinction proposed above may not be the most appropriate for national needs. Instead, it may be important, for example, to distinguish in terms of availability among a separate room for bathing in the living quarters, a separate room for bathing in the building, an open cubicle for bathing in the building and a public bathhouse. The recommended classification of housing units by availability and type of bathing facilities is as follows:

- 1 With fixed bath or shower within housing unit
- 2 Without fixed bath or shower within housing unit
 - 2.1 Fixed bath or shower available outside housing unit
 - 2.1.1 For exclusive use
 - 2.1.2 Shared
 - 2.2 No fixed bath or shower available

638. Type of bathing facilities should be part of a joint edit with other water-related items. Values such as “private”, “shared”, or “exclusive use” can be used to determine whether values are consistent, and, if they are not, to establish the edit paths to follow to fix the problem. When one or more other water-related variables is present, an estimate for unknown or inconsistent information may be developed without resorting to use of “unknown” or dynamic imputation. However, when all else fails, if the value is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics. These include, as a rule, type of housing unit and then water supply, construction material of walls, tenure or vacancy status, to obtain “known” information from similar housing units in the geographical area.

13. Kitchen—availability of (core topic)

639. According to *Principles and Recommendations for Population and Housing Censuses, Revision 2* (United Nations, 2008, para. 2.494), 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

640. 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.

14. Fuel used for cooking (core topic)

641. 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 (United Nations, 2008, para. 2.496).

642. 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.

15. Lighting and/or electricity—type of (core topic)

643. Information should be collected on the type of lighting in the housing unit, 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 enquiry (United Nations, 2008, para. 2.497).

644. 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.

16. Solid waste disposal—main type of (core topic)

645. According to *Principles and Recommendations for Population and Housing Censuses, Rev. 2* (para. 2.500), 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 of solid waste in a river/sea/creek/pond
- 9 Occupants compost solid waste
- 10 Other arrangement

646. 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.

17. Heating—type and energy used for (additional topic)

647. This topic refers to the type of heating of housing units and the energy used for that purpose. The units of enumeration are all housing units. This topic is irrelevant for 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 (United Nations, 2008, para. 2.501).

648. 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, such as type of housing unit, construction material of walls, tenure and vacancy status, to obtain “known” information from similar housing units in the geographical area.

18. Hot water—availability of (additional topic)

649. This topic concerns the availability of hot water in housing units. Hot water denotes water heated to a certain temperature and conducted through pipes and tap to occupants. The information collected may indicate whether hot water is available within the living quarters or outside the living quarters, for exclusive or shared use, or not at all (United Nations, 2008, para. 2.502).

650. The availability of hot water may be related to the means for heating the water, although the use of solar energy for heating water may not be related to other housing items. The editing teams must decide on the appropriate edits, depending on other housing items and geographical location. In the end, if the value is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as those for piped water, to obtain “known” information from similar housing units in the geographical area.

19. Piped gas—availability of (additional topic)

651. This topic refers to the availability of piped gas in the housing unit. 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 there is either a lack of sources of natural gas or no developed pipeline system (United Nations, 2008, para. 2.503).

652. Piped gas is not related to other housing items except for type of lighting and cooking fuel. Editing teams must determine the appropriate editing path as well as how to check for consistency. If the value remains invalid or inconsistent, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as energy used for heating, type of building, type of housing unit, construction material of walls, tenure and vacancy status, to obtain “known” information from similar housing units in the geographical area.

20. Use of housing unit (additional topic)

653. “Use of a housing unit” refers to whether a housing unit is being used wholly for habitation (residential) purposes or not. The housing unit can be used for habitation and for commercial, manufacturing or other purposes (United Nations, 2008, para. 2.504).

654. “Use of housing unit” is independent of the other housing items. If the value is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, construction material of walls, tenure and ownership, to obtain “known” information from similar housing units in the geographical area.

21. Occupancy by one or more households (core topic)

655. 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 determining the household head.

22. Number of occupants (core topic)

656. Each person usually resident in a housing unit or set of collective living quarters should be counted as an occupant. Therefore, the units of enumeration for this topic are living quarters. However, since housing censuses are usually carried out simultaneously with population censuses, the applicability of this definition depends upon whether the information collected and recorded for each person in the population census indicates where he or she was on the day of the census or whether it refers to the usual residence. For persons occupying mobile units, such as boats, caravans and trailers, care should be exercised to distinguish those who use them as living quarters from persons who use these units as a means of transportation (United Nations, 2008, para. 2.510).

657. “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.

23. Building—type of (core topic)

658. The following classification by type is recommended by the United Nations (2008, para. 2.514) for buildings in which some space is used for residential purposes.

- 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

659. 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.

24. Year or period of construction (additional topic)

660. The year or period of construction refers to the age of the building in which the sets of living quarters are located. It is recommended that the exact year of construction be sought for buildings constructed during the intercensal period immediately preceding if it does not exceed 10 years. Where the intercensal period exceeds 10 years or where no previous census has been carried out, the exact year of construction should be sought for buildings constructed during the preceding 10 years. For buildings constructed before that time, the information should be collected in terms of periods that will provide a useful means of assessing the age of the housing stock. Difficulty may be experienced in collecting data on this topic because in some cases the occupants may not know the date of construction (United Nations, 2007, para. 2.519).

661. Some countries, even those using dynamic imputation, accept an “unknown” response for the item on year or period of construction. When this occurs, the country may choose not to use dynamic imputation for this item, even if it uses imputation matrices for other variables. Countries choosing dynamic imputation for invalid values should use at least two characteristics, including type of building, construction material of outer walls and/or type of housing units in the building, to obtain “known” information from similar housing units in the geographical area.

25. Dwellings in the building—number of (additional topic)

662. Editing for the number of housing units in a building is explained in chapter III as part of the structure edits.

26. Construction material of outer walls (core topic)

663. 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 (United Nations, 2008, para. 2.525).

664. 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.

27. Construction material of floor, roof (additional topic)

665. In some cases the materials used for the construction of roofs and floors may be of special interest and can be used to further assess the quality of dwellings in the building. This topic refers to the material used for roof and/or floor (although, depending on the specific needs of a country, it may refer to other parts of the building as well, such as the frame or the foundation). The unit of enumeration is a building. Only the predominant material is enumerated and, in the case of a roof, it may be tile, concrete or metal sheeting, palm, straw, bamboo or similar plant material; or mud, plastic sheeting or some other material (United Nations, 2008, para. 2.528).

666. 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.

667. 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.

28. Elevator—availability of (additional topic)

668. This topic refers to the availability of an elevator (an enclosed platform raised and lowered to transport people and freight) in a multi-storey building. The information is collected on the availability of an elevator for most of the time: in other words, one that is operational for most of the time, subject to regular maintenance (United Nations, 2008, para. 2.529).

669. 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 storeys or the fact that an elevator is present. If the elevator takes precedence, the number of storeys must be changed, either by making the value “unknown” or by using dynamic imputation to obtain another value. If the number of storeys takes precedence, and the building has only one storey, the response on “presence of an elevator” must be changed to “no”.

670. When an elevator is present, if it requires electricity, a check should be made to be certain that electricity exists in the building.

671. 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.

29. Farm building (additional topic)

672. Some countries have found it necessary for their national censuses to specify if an enumerated building is a farm building or not. A farm building is one that is part of an agricultural holding and is used for agricultural and/or housing purposes (United Nations, 2008, para. 2.531).

673. 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.

30. State of repair (additional topic)

674. 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 repair; they have so many serious structural defects that it is deemed more appropriate to tear the buildings down than to undertake repairs. This term is most often used for buildings with only the frame left standing, without complete external walls and/or a roof (United Nations, 2008, para. 2.532).

675. 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.

31. Characteristics of head or other reference member of household (core topic)

676. 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 sex, age, 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 (United Nations, 2008, para. 2.533).

32. Tenure (core topic)

677. According to the United Nations (2007, para. 2.536), 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
- 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
 - 2.2 Member of household rents a part of housing unit as a subtenant
- 3 Occupied free of rent
- 4 Other arrangement

678. Tenure may relate to type of ownership, so the editing team may need to consider the relationship between the two items when developing the edits. Otherwise, if the value for tenure is invalid, “unknown” should be assigned when dynamic imputation is not used. Countries choosing dynamic imputation for invalid values should use at least two characteristics, such as type of housing unit, rent and vacancy status, to obtain “known” information from similar housing units in the geographical area.

33. Rental and owner-occupant housing costs (additional topic)

679. 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 (United Nations, 2008, para. 2.540).

34. Furnished or unfurnished (additional topic)

680. 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 utilized to resolve invalids or inconsistencies (United Nations, 2008, para. 2.542).

35. Information and communications technology devices—availability of (core topic)

681. The importance of the availability of information and communications 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 its use for planning purposes by government and the private sector so as 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 than home
- 8 Household without access to the Internet

682. Information and communications technology devices are new items. Items requiring electricity should be available only where electricity is available. As solar power, wind power and other “renewables” become used more frequently, 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 conduct of 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 (United Nations, 2008, para. 2.543).

683. These topics refer to the availability of the item within the housing unit. For example, a telephone denotes a telephone line rather than a physical telephone, since more than one telephone can be connected to a single telephone line (United Nations, 2008, para. 2.547-548). 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.

36. Cars—number of (additional topic)

684. “Number of cars” refers to the number of cars and vans normally available for use by the members of the household. 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, and those provided by an employer if available for use by the household, but excludes vans used solely for carrying goods (United Nations, 2008, para. 2.551).

685. 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.

37. Durable household appliances—availability of (additional topic)

686. Information is collected on the availability of such durable appliances as laundry washing machines, dishwashing machines, refrigerators, deep freezers and so forth, depending on national circumstances (United Nations, 2008, para. 2.552).

687. 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).

38. Outdoor space—availability of (additional topic)

688. 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 (United Nations, 2007, para. 2.553).

689. 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.

B. Occupied and vacant housing units

690. 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 statistical/census 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.

691. The present chapter has looked at the housing variables recommended in *Principles and Recommendations for Population and Housing Census, Revision 2*. No country should be using all of these variables, and the variables selected and their spatial relationships with the other variables should be thoroughly tested in hothouse and pre-census survey situations for the purpose of ensuring complete and reliable responses. Housing variables are important on their own for use in making up part of a wealth index for assessing well-being in all or parts of a country.

Annex I

Derived variables

1. In order to get the best use out of their census or survey data, countries often need variables that are combinations and variations of other variables. For example, the item on economic activity status (see chapter IV, sect. D.1) 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 statistical/census 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 statistical/census 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.

2. 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.

3. 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.

A. Derived variables for housing records

1. Household income

4. 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.

5. 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.

2. Family income

6. 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.

7. 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.

3. Family nucleus

8. For household composition, *Principles and Recommendations for Population and Housing Censuses, Revision 2* developed a code for family nucleus, defined as one of the following, with recode suggestions in parentheses:

- 1 Married couple (or couple living in consensual union) without children (householder and spouse or co-heads or a couple living in consensual union)
- 2 Married couple (or couple living in consensual union) with one or more unmarried children (as above, but, through a search of 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 householder, no husband present, with at least one unmarried child determined as above)

4. Type of household

9. The *Principles and Recommendations for Population and Housing Censuses, Revision 2* includes general features of various types of households to assist in developing a recode for household composition. Countries may choose to carry out a single recode, or a series of recodes, depending on potential use of the data.

10. A first recode could identify type of household, as represented by the following items, including definitions. The suggested recodes follow in the next section.

- 1 One-person household
- 2 Nuclear household: a single family nucleus, i.e., 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)

5. Household composition

11. *Single-person households* are households, rather than families, and should therefore be included as a separate category in the household composition recode.

12. *Nuclear-family households*. Nuclear-family households can be divided into (and may receive individual codes for): (1) married-couple families with children, (2) married-couple families 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 established, followed by searches of the household for a spouse and children. The type of nuclear household could be a two-digit code with the value 2 used for the first of the two digits (code 1 being reserved for single-person households); hence, code 21 would represent a married couple family with children.

13. *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 belong to 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 carried out.

14. *Composite households*. All other households would be composite households. Using the same scheme as before would yield 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, none of whom belong to a family nucleus, plus other unrelated persons; and (47) non-related persons. Once again, a series of searches and summaries will result in the appropriate designation for each type of household.

6. Family composition

15. Families are a subset of households; therefore, the recode for family composition will include those categories appropriate for families described above. Inasmuch as a one-person household does not constitute a family, it will not be included in the recode for family composition. Similarly, composite households are households but not families, hence they will also 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 for including nuclear-family households with extended families for all families).

7. Household and family status

16. Household and family status indicates how a person relates to other household or family members. The approach to establishing household and family status differs from the traditional approach of classifying household members solely according to their relationship to the head or reference person.

17. *Principles and Recommendations for Population and Housing Censuses, Revision 2* includes the following suggested coding scheme for household status. The first set of codes is for persons in households with at least one family nucleus (that is to say, households that are also families). Suggested determinations of the recode include:

- 1.1 Husband (male head or male spouse)
- 1.2 Wife (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's 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). *Principles and Recommendations for Population and Housing Censuses, Revision 2* divides this item into two groups : (1) living with relatives and (2) living with non-relatives

18. The second set for the recode is for persons in households without any family nucleus: persons living alone, and persons living with other relatives and/or non-relatives not including spouse or child of householder. The categories are as follows:

- 2.1 Persons living alone (single-person household)
- 2.2 Persons living with others (person living in a housing unit without the spouse or a child of householder). This category is further divided into persons living (1) with siblings, (2) with other, non-sibling relatives or (3) with non-relatives.

19. A single variable should be developed from these categories, since they are mutually exclusive. The variable would be a two-digit one. Some statistical agencies may want the first digit to be independent of the second digit; in this case, the first digit will indicate whether the household is a family nucleus or not, and the second will identify the kind of household status held by a person.

20. *Principles and Recommendations for Population and Housing Censuses, Revision 2* also includes categories for classification of a person by family status. These include (1) male or female in a householder-spouse pair with or without children, (2) lone parent, by sex, (3) child of householder, child of a married couple or child of a single parent, by sex of parent, and (4) non-member of the family nucleus (unrelated or related, and if related, how). The determinations shown above for household status can be used for family status.

8. Impact of HIV/AIDS on household structure

21. Given the impact of the HIV/AIDS epidemic on the household structure of so many countries, a recode can assist in describing different kinds of housing units. For example, a recode describing missing-generation households (grandparents and grandchildren only), households with heads under age 18, widow-headed households and so forth can be used to assess the social and economic impact of the epidemic, albeit very indirectly. Children in and out of the workforce, structure of the workforce within these households, etc., can assist government planners in fully describing the impact of HIV/AIDS.

9. Related persons

22. 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.”

23. 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.

24. “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.

10. Workers in the family

25. 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.

11. Complete plumbing

26. 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.

27. 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”.

12. Complete kitchen

28. 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.

29. 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.

13. Gross rent

30. 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.

31. 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.

14. Wealth index

32. The wealth index is a measure of well-being in a country or in parts of a country. In most cases, the index is constructed using the household assets. Often, factor analysis is utilized to obtain the best set of items and their variants. Usually, the items are assigned binary values—1 for “present” and 0 for “absent”—and the values are then summed. The higher the value, the greater the wealth. For example, having a television set would be coded as 1 for presence, 0 for absence. On the other hand, a toilet might be coded 1 for “outhouse”, 2 for “gravity flush”, or 3 for “flush” (involving three sets of binary variables). The various items might be weighted in the summing.

33. Quintiles can then be created by taking each fifth part of the distribution of the values of the wealth index. The lowest fifth would comprise the poorest households, and the highest fifth would comprise the wealthiest households.

B. Derived variables for population records

1. Economic activity status

34. A derived variable for economic status can be very useful for the tabulations, but it requires information from several variables. In following the categories of *Principles and Recommendations for Population and Housing Censuses Revision 2*, reconfiguration of several variables is necessary. The derived variable might encompass two categories, broken down into a total of six subcategories:

- 1 Economically active
 - 1.1 Employed
 - 1.2 Unemployed
- 2 Not economically active
 - 2.1 Student
 - 2.2 Homemaker
 - 2.3 Pension or capital income recipient
 - 2.4 Other

35. 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.

2. Own children

36. Sometimes countries want to produce information about “own children”, who are the biological children of the head of household and/or spouse. Tables could show “own children” further classified as living with two parents or with one parent only.

37. The derived variable for “own children” might be the sum of the number of own children of a particular person, usually a female, following the definitions selected by the editing teams. Sometimes users need more detailed information on own children by age. For the United States, for example, derived variables are developed for number of own children less than 6 years old and for those 6 to 17 years old. These values are placed on the records of all females. The information is used especially to determine the characteristics of females in the labour force with own children.

3. Parents in the house

38. 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.

4. Current year in school

39. Some countries ask two questions about education:

- (a) if the person currently attends school;
- (b) the highest level of educational attainment.

40. 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.

41. 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".

42. 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".

5. Months since last birth

43. If information on the question of date of last birth is collected, a recode can be created to obtain indirect estimates of year-by-year age-specific and total fertility. The recode takes the date of enumeration, usually the month and year, and converts this to all months, and the date of last birth and converts this 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.

Annex II

Relationship of questionnaire format to keying

1. The two most common questionnaire formats for population items in a census or survey are person pages and household pages.
2. Person pages contain one page or two facing pages of population information, with separate pages for each person. This method is useful because all of the information for one person appears on one page, making it easy to collect. Also, this format makes it easy to check for internal consistency during enumeration. Person pages may be combined in a bound booklet for ease of handling in the field as in figure A.II.1.
3. Coding and keying for items on person pages is basically a mechanical operation, in which the coder/data entry operator is not expected to evaluate the validity of the information supplied but rather assign its appropriate code or keystroke. Figure A.II.2. illustrates the flow of information for a given person recorded on a single page. It is easier to enter data on a single page for that person than to key by turning pages. Validity checks are implemented later during the computer edits.

Figure A.II.1
Sample questionnaire form with person pages

Person page for person X		Person page for person X+1	
Item 1	Item 10	Item 1	Item 10
Item 2	Item 11	Item 2	Item 11
.	.	.	.
.	.	.	.
.	.	.	.

Figure A.II.2
Example of flow within a questionnaire with person pages

Person page	
Item 1	Item 11
Item 2	Item 12
Item 3	Item 13
etc.	etc.
.	
.	
.	

4. Household pages have all information for a household on one page, if possible, or on a series of pages with all household members listed on each page. Listing the household members in this way is useful because the questionnaire items do not have to be printed for each individual, thus saving space. In addition, the enumerator can compare entries between household members as the data are collected (see figure A.II.3).

Figure A.II.3
Sample questionnaire, household page with all persons on same page

Household page					
Person	Item 1	Item 2	Item 3	Item 4	etc.
1					
2					
3					
4					
5					
.					
.					
.					

5. A third method is to have separate forms for each person, with the enumerator then assembling a loose booklet during or after enumeration. This method is efficient since the enumerator collects only the exact number of forms (pages) necessary for the household. The disadvantage is that the forms may separate during transfer or other handling, creating many potential editing and coverage problems if the census office is unable to reassemble them for the correct household.

6. The physical size of questionnaire pages is also a consideration, not only for enumeration, but also for keying. During coding and keying, the document must lie flat on the surface of the worktable, and coders or data entry operators must be able to locate and handle items on the form easily.

7. When all information is on a single page, staff can easily key the household pages as well, and it will obviously be faster since the data entry operator does not have to turn pages. Figure A.II.4 illustrates the flow of information on a household page.

Figure A.II.4
Example of flow for a questionnaire with household pages, with multiple persons per page

Household page					
Persons	Item 1	Item 2	Item 3	etc.	
1					
2					
3					
4					
5					
.					
.					
.					

8. Problems can occur in keying population or housing information that extends over more than one page. To solve the problems, the national statistical office is likely to take either of the two approaches outlined below.

9. Data may be entered one person at a time. The data entry operator may key the line of information for a person on the first page of the series of pages, and then turn to the second and subsequent pages. At the end of the first person's pages, the data entry operator then turns back to the first of the household pages for that household, and keys the second person, third person and so forth. This type of keying works as long as the data entry operator can remain on the proper line throughout the keying. Although computer editing programs can be created to disentangle information when person items are erroneously keyed on another person's line, the program itself is very difficult to prepare.

10. Data may be entered one page at a time. The data entry operator may key a whole page of information before moving on to the next page. Here, the data entry operator keys all information on the first page regardless of the number of persons. Then, the data entry operator turns the page and keys the next part of the information for all persons. Skip patterns may or may not be included here, depending on the type of keying (with or without computer editing). In any case, during the computer edit, the records from the various sets of keyed data will have to be assembled, and any miskeys of person numbers will have to be dealt with then.

11. In the following example (figure A.II.5), the household's demographic information poses no unusual keying problems since the census obtained a response for all items for all persons.

Figure A.II.5

Example of a household page with multiple persons, without keying problems

Household page				
Persons	Relation	Sex	Age	etc.
1	Head of household	M	40	
2	Spouse	F	35	
3	Child	F	18	
4	Child	M	12	
5	Sibling	M	35	
6	Sibling of spouse	F	30	
7	Sibling child	M	5	
8	Sibling child	F	3	
etc.				

12. However, a second page for the same household (figure A.II.6) could present some keying problem. For example, if the country chooses to collect language use only for persons 5 years of age and over, that information for the eighth person, a 3-year-old, will be blank. The data entry operator should leave the cell blank for this child, or else the computer edit will have to attempt to correct it later.

13. Similarly, other items should be blank, such as persons under the minimum age for labour-force participation, females under the minimum age for fertility and fertility for all males. In figure A.II.6, the data entry operator might incorrectly key person 6's information on children ever born (in this case, 4) in person 5's slot by mistake. The computer edit would then delete the male's fertility items and impute fertility for the female, but it might not impute the correct value.

Figure A.II.6
Example of a household page with multiple persons, with potential keying problems

Household page 2				
Persons	Language	Labour force	Children ever born	etc.
1	Language 1	Yes		
2	Language 1	No	3	
3	Language 1	No	0	
4	Language 1			
5	Language 1	Yes		
6	Language 1	No	4	
7	Language 1			
8				
etc.				

14. Many times a country must use the household form because of cost or space constraints. However, when the population is small, or the country can afford the additional expense, the form with person pages is likely to contain fewer matching errors through miskeying than occur with the household forms.

Annex III

Scanning versus keying

1. Many countries are using scanning equipment, either optical mark reading (OMR) or optical character recognition (OCR). Each of these has advantages over keying when the operation is smooth and efficient and when the costs are not great. However, many countries, even some that have committed to scanning, may not be able to afford the initial start-up costs or the continuing maintenance costs during and after enumeration. On the positive side, many countries use the scanners obtained for the census for a number of purposes, including other surveys and such administrative records as entry and exit forms. Countries may also resort to outsourcing for scanning operations, or rent the scanners for the time that they will be needed for the census.

2. One of the advantages of keying is that the skills learned during keying transfer to other activities in the national statistical/census offices and other government agencies. After the census develops expert data entry operators, these data entry operators then key data for various follow-up surveys. These surveys could include post-enumeration surveys (PES) and other surveys such as fertility or household income and expenditure surveys. Staff can also key administrative records, such as vital records and those for trade, immigration and emigration, and customs.

A. Entering the data

1. Scanning

3. Countries using optical or other scanning devices to capture their data will not normally correct their data as they are captured, although changes may depend on the skip patterns built into the system. Countries who choose to key their data, however, have several choices, depending on how quickly they need the data keyed and how much manual checking is needed. Each of the options depends on the requirements of the editing teams, the skills of the data entry operators and the sophistication of the editing program.

4. The quantity and type of data entry equipment required will depend on the method of data capture selected, the time available for this phase of the census, the size of the country, the degree of decentralization of the data capture operations and other factors. For keyboard data entry, the average input rates usually vary between 5,000 and 10,000 keystrokes per hour. Some operators stay well below that range, while others surpass it significantly. Among the factors that affect operator speed are (a) the supporting software and program; (b) the complexity of the operators' tasks; (c) the ergonomic characteristics, reliability and speed of the equipment; (d) the question whether work is always available; (e) the training and aptitude of the recruited staff; and (f) the motivation of the workers (United Nations, 2007, para. 1.193) .

2. Heads-down keying

5. Heads-down keying takes two forms. The first is keying all data items as they are encountered with no skip patterns. In this case, keying proceeds more quickly since data entry operators do not have to stop when invalid or inconsistent information is encountered. It may also be more accurate since keyers perform the task more mechanically. The second form of heads-down keying entails stopping to check the questionnaires for invalid or inconsistent results, so the process will go more slowly and will require much more expertise on the part of the keying staff. The high price in terms of speed must be seriously considered. Paradoxically, accuracy may also be improved with this method if data entry operators find that the data were recorded correctly but were miscoded. Miskeying itself may sometimes be immediately challenged because the editing package provides for automatic checking.

(a) Heads-down keying without skip patterns

6. When all entries are keyed, or skipped manually, a particular rhythm can be maintained, and certain skip patterns will not obviate valid but temporarily inconsistent information. For example, if a person is recorded as male, most editing teams will require that the whole section on fertility be skipped. In this case, a data entry operator will key through (use the space bar or arrow to move through a male's or young female's record) because all fields will be blank. However, this takes time, and the spacing may not be completely accurate. For example, the data entry operator might go too far or not far enough, and other items might be miskeyed because they are improperly aligned. If all fields are keyed in this way, then this information can be keyed when no skip patterns are present. For example, when the data entry operator encounters an adult female with fertility (a female for whom such items as children ever born, children surviving or children born in the last year have been collected and coded), all items are keyed. If the fertility information is keyed, the computer editing program can determine which item or set of items is valid and which must be changed. When the edit determines that the person is an adult female, but the fertility information is blank, then dynamic imputation or other appropriate means has to be used to obtain fertility information for the tabulations. If the actual information has been lost because of the skip patterns, the editing team must decide whether the loss is worth the increased efficiency and speed. If skip patterns are present, the data entry operators can still move backwards through the screens to the appropriate position for corrections. Although the data entry operators will waste some time spacing through items they do not key, with this form of data entry, inconsistencies between sex, age and fertility can be attacked during the edit rather than at the time of keying.

(b) Heads-down keying with skip patterns

7. A second method of heads-down keying involves keying with skip patterns in place. Again, if the editing team requires skip patterns, usually to represent the way the enumerators collected the data, keying is easier and faster if the skip patterns are easy to follow and if the data entry operators learn the keying patterns quickly. If the skip patterns are very complicated, data entry operators may become confused and persistently key in the wrong places. The most efficient keying with skip patterns occurs when limited patterns that cover large parts of the record being keyed are used.

8. The editing team will need to determine the appropriate skip patterns for their country's census or survey. For example, it makes sense, to skip all of the employment items for children, that is, persons below the country's defined age for potential

employment. Often, these are half of the population items, so it is efficient to skip them for children, except for special situations such as for children whose age is borderline, or when the country may be interested in child labour.

9. The editing team decides on an item-by-item basis which items will be included for which age groups. Staff can group the items to manage the skip patterns easily.

10. It is not always easy to have clear-cut decisions about skip patterns. For example, consider the following sequence:

- 1 What is this person's citizenship?
 - Born in this country (skip to item 3)
 - Naturalized
 - Not a citizen
- 2 What is this person's year of entry?
- 3 NEXT ITEM

11. A skip pattern could be created to skip from 1 to 3, that is, skip the item on year of entry, for persons born in the country. However, sometimes data entry operators violate the skip pattern, either because the enumerator or coder makes a mistake, or because of miskeying. The many factors involved include the skill level of the data entry operators, the cultural circumstances, the layout of the questionnaire and the layout of the screens. The editing team often works together to determine whether a skip pattern in a case such as this case is reasonable.

3. Interactive keying

12. Interactive keying may be used during census input but is more appropriate for surveys, particularly for small surveys where allocated items could affect the results of the survey. Interactive keying may involve manual or automatic corrections, depending upon the information available to make changes or corrections.

13. Consider the case of a small survey. For small surveys, every response is important. If a country takes a 1 per cent sample survey, for example, each response represents 100 persons, housing units, or agricultural holdings. A few invalid or inconsistent cases could have a considerable impact on the results of the survey. In these cases, the demographers and other social scientists usually want to have considerable control over the processing.

14. Control may be established in several ways. The demographers and other specialists may key the data themselves, checking for extraneous, invalid or inconsistent responses as they go along, using the information as recorded on the data collection forms. They can often resolve conflicts, miscodes, or other inconsistencies immediately, while looking directly at the collected information. Sometimes they may opt to send incomplete or invalid questionnaires back to the field. This type of interactive keying gives the best results since the demographer also serves as data entry operator, but it is by far the most expensive, and not many countries can afford this method.

15. The editing teams may develop very detailed edit rules to determine what data entry operators must do when particular cases occur during keying. For each unresolved invalid code, they can decide what the data entry operator will key. The editing team may resolve cases not covered by the detailed rules and may modify the rules (although at the risk of having inconsistencies between the first part and later parts of the keying).

16. Skip patterns which play an important role in heads-down keying, are important here, too. As with heads-down keying, data entry operators must be aware of and learn any skip patterns in use. As mentioned above, skip patterns can increase the speed of keying, but usually with some loss of quality. For interactive keying, a common rule of thumb is that the fewer the skips the better the quality.

17. After developing keying instructions, national statistical/census offices must have actual data entry operators test the keying instructions before deciding on the actual operation, whether or not heads-down keying is used. As the keying instructions are tested, bugs can be worked out of the system, and optimum keying can be obtained.

B. Verification

18. The national statistical/census office must also decide what level of verification is appropriate. For keyed data, many experts advise 100 per cent verification. In this case, all items are rekeyed (or keyed over the existing information) to make certain that the data collected are the data that enter the machine for computer processing. Often, however, total verification is not practical, because the country does not have the time, financial or human resources to rekey all of the data. The percentage sample verified should be larger for beginning keyers, but smaller for more experienced keyers. Also, if the tested error rate for the keying is very low, with the data entry operators making very few errors, then complete verification is probably not necessary.

19. In any verification operation, it is first necessary to determine what information is needed. Does the country wish to track individual operators or a team of keyers? Does it wish to determine whether skills are being acquired, or maintained? The units of control could also be important, including daily, weekly, or monthly reporting, and so forth, to determine the flow of work and the skills gained.

20. Finally, it is very important that verification be independent, and that a different set of keyers perform the verification from the data entry, or, at least, different members of the same team. Use of different sets of keyers allows for independence in the operations and therefore better results.

21. For scanned data, some countries also perform verification to ensure that the scanning was comprehensive and complete. Because scanning technology is still new, even when the systems are thoroughly tested with pilot or pretest data, issues of changes in paper quality, actual printing of forms in various places, storage, etc., can cause problems which would need to be addressed through verification.

22. If errors are systematic and can be removed through the edit program, keyers and verifiers should not be making judgements about correction. However, the keyers and verifiers are responsible for finding the errors. These errors could result from inadequate testing of the scanning equipment, causing systematic errors for certain items or combination items, confusion in reading certain digits (for example, interchanging 2s and 3s, or 8s and 9s), misreading of continuation check-off boxes, and so forth.

23. Misreading of continuation check-offs has been a continuing problem in recent years and can only sometimes be addressed in the edit. If the forms are not contiguous, other procedures will be needed, most likely during the structure edits, to resolve issues. As noted earlier, a completely sound, structured file needs to have been created before content ending begins.

1. Dependent verification

24. Techniques for verification are either dependent or independent. In dependent verification, data entry operators key over data previously keyed by other staff. When the key strokes differ, the software package informs the data entry operator, and, depending on the program, the data entry operator either overrides the previous data, or a note is made of the discrepancy. Since the data are keyed from the original questionnaires, usually the data entry operator himself or herself can make an informed decision about whether the original keying was in error.

2. Independent verification

25. In independent verification, data entry operators rekey the data from scratch; they create a completely independent file of the keyed data, using the original questionnaires. The two resulting files, the original keyed dataset and the verification dataset, are then compared, using a computer program, to test for discrepancies. Presumably, some manual operation is used to rectify invalid and inconsistent key strokes.

C. Editing considerations with scanned data

26. More and more countries now scan their data. In the early 2000s, many of these countries were surprised to find that scanning introduced types of errors that were different than those arising from keying. Part of the problem with editing scanned data involves lack of quality control during the scanning process. Because the technology was so new in the early part of the 2000s, many statistical offices did not have the background or facility needed to develop appropriate quality control for all items. Many of the countries that did develop appropriate quality control procedures ended up not developing them for all items; therefore, some of the items at the end of the question—particularly the fertility items—turned out to be invalid or inconsistent.

27. Of course, many of the inconsistencies found in keyed data also occur in scanned data. However, it is useful to discuss some of the special problems arising from the use of scanned data. Because scannable questionnaires require markers to assist the machine in reading them, items are often displayed in ways that may cause problems for enumerators and respondents during data collection. The problems connected with these items must be addressed systematically. When the items are closely related to other items, like religion to ethnicity, the regular edits described in the text can be used.

28. However, care must be taken when items needed for planning and policy may produce problems. Usually, the item for **sex** does not produce problems because it involves only two possibilities. However, as noted above, while the keyer is usually restricted to keying only a 1 or a 2 (or a code for “unknown”), any value can appear in the columns for sex—through the appearance of other digits, alpha-characters or other characters. Therefore, some type of edit needs to be added to what was once done for keyed data to account for these miscellaneous values.

29. Relationship codes provide a good illustration of the problem. If there is a single digit for relationship codes, as shown in the text, normally no problems will be experienced. However, if two digits are used, a problem will sometimes occur in scanning when the first digit is either coded incorrectly or picked up incorrectly by

the scanner. Normally, if codes 1 through 12 are used, the keyer will be restricted to keying only those codes, and the entry package will detect the entry of an illegal code. With scanning, almost anything will be accepted (although scanning packages can be programmed to detect as well.) In such cases, erroneous codes must be changed during edit; otherwise, they will cause all sorts of problems during the tabulation stage.

30. Age sometimes is an issue, particularly when three columns are used (to allow for people's age to be greater than 100); therefore, a digit-by-digit analysis may be needed—one that entails looking separately at the digits in the ones, the tens and the hundreds—to achieve a proper edit. Once it is established that where age has been properly captured, the regular edit can be used.

31. However, in cases where both age and date of birth are present, misleading information can cause problems when one item takes precedence over the others. Usually, subject matter specialists prefer the use of date of birth with the census or survey reference date to produce (by subtraction) an exact age for comparison with the reported age. When one or several digits are missing, care must be taken to ensure that all remaining digits are used properly to obtain the best estimate of the computed age for comparison. When the scan does not pick up a single digit, for example, the edit can take this into account so as to provide a best guess of what would have been there. This type of problem does not usually occur with keying.

32. The items with the largest problems in the early 2000s resulting from scanning involved fertility—both the numbers of children born and surviving as well as children born in the last year or over time. For most countries, the main problem has been lack of quality control during scanning, resulting in the presence of strange items in the data capture. Where the value for the item in a given country was 17, 18 or 19 dead female children, for example, the data, when unedited, were useless for planning purposes.

33. Mortality information also can present problems in the scanned data. For keyed data, if there is a series of items for deaths in the year before the census for example, (sex and age of the deceased, whether the person died a natural death and whether it was a maternal death), the keying proceeds even over erasures and strikeouts. However, with scanned data, erasures would normally not be read, and the scanner will leave items blank and then continue with the capture. The edit program must move the information into appropriate spaces for tabulation and subsequent analysis. It should be noted that newer scanning operations can perform these moves during and just after capture.

D. Conclusions

34. Since each country's problems unfortunately depend on the particular programming and functioning of the individual scanners, it becomes difficult to devise general guidelines. However, in all cases tackled thus far, scanning problems have been systematic; that is to say, when staffs determined the algorithm required to alleviate the problems, fully edited datasets could be produced.

Annex IV

Sample flow charts

1. One of the tasks of the editing team is to develop a relational structure for the variables used in the editing process. Flow charts facilitate the identification of various linkages among variables, and help in the development of clear and concise editing specifications. These specifications for relational linkages help both subject matter and data processing specialists to visualize the editing process and facilitate communication between the two groups.

2. Three sample flow charts are presented on the following pages:

- (a) Flow chart to determine the head of household;
- (b) Flow chart to determine a spouse in the households;
- (c) Flow chart to edit sex variable for head of household and spouse.

These sample flow charts are provided for illustrative purposes only and should be treated accordingly. The editing team may modify further the sample flow charts as necessary based on the situation in the country.

3. Editing flow charts should be developed for each variable in a census. The editing team should work together on the development of the flow charts, and the data processing specialists should use them with the editing specifications to develop computer programs to edit the census data. The flow charts and editing specifications should be properly documented for use in future census and survey data processing.

Figure A.IV.1
Sample flow chart to determine head of household

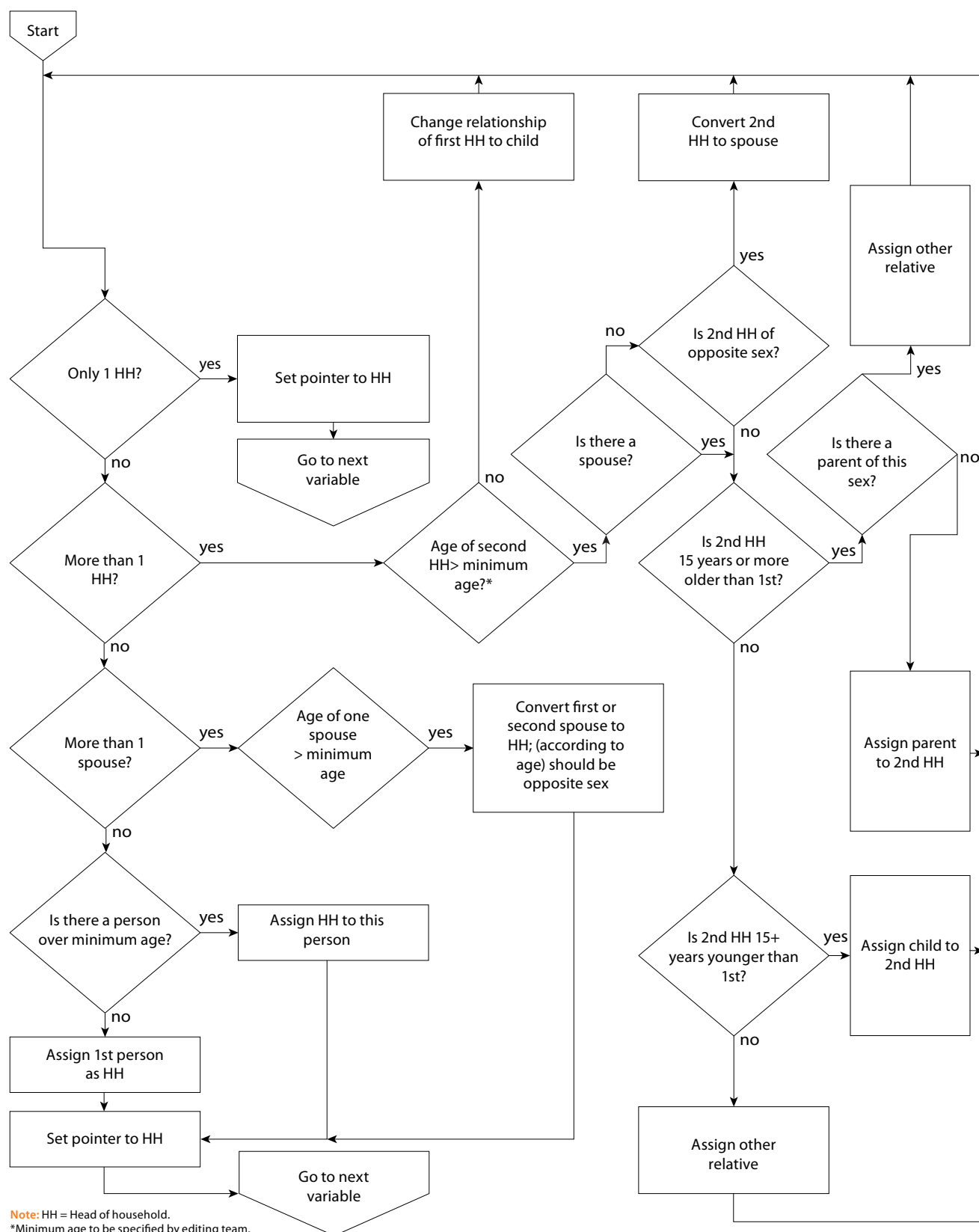
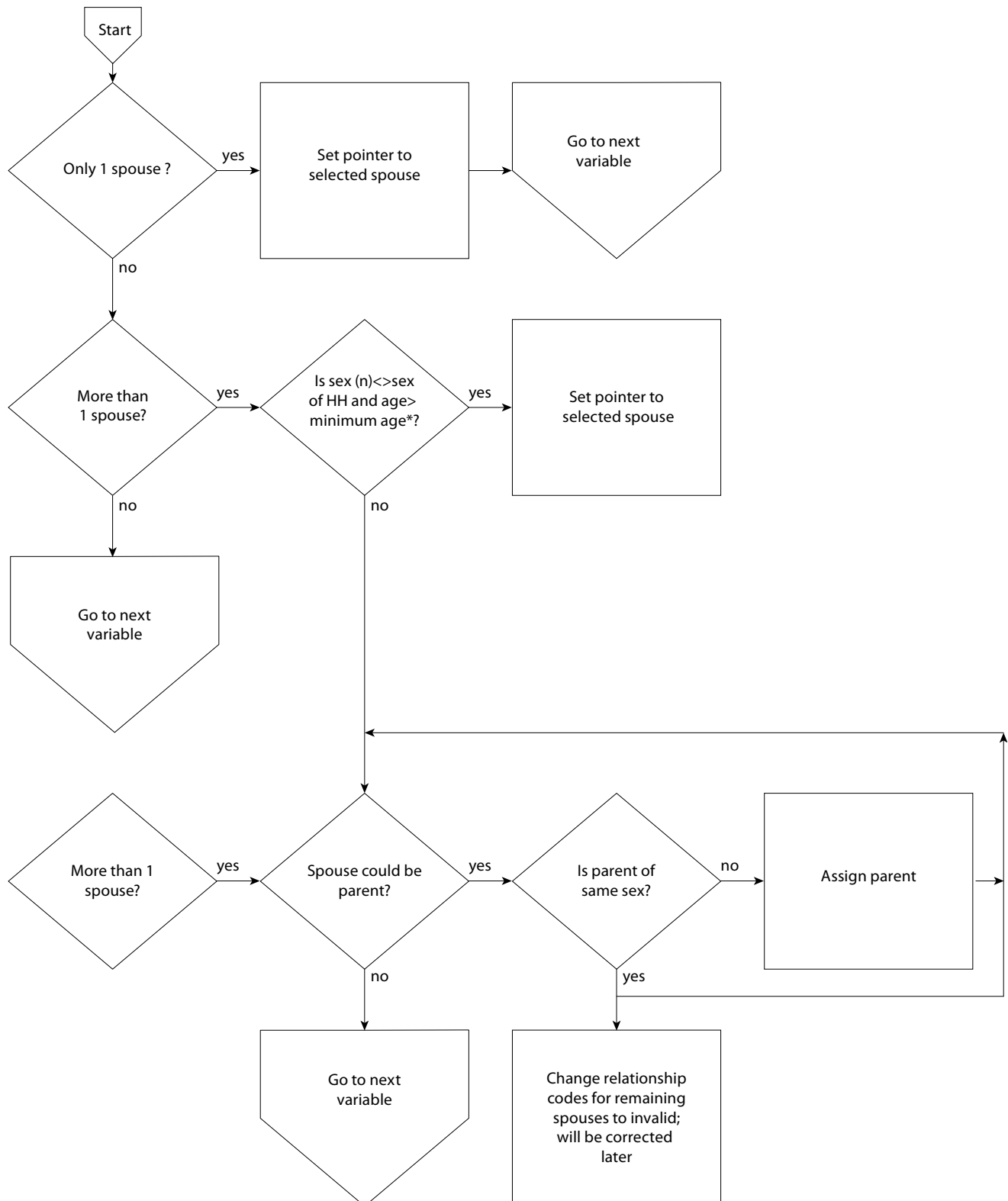


Figure A.IV.2
Sample flow chart to determine presence of spouse in household



Note: HH= Head of household.

* Minimum age to be specified by the editing team.

Figure A.IV.3

Sample flow chart to edit sex variable for head of household and spouse

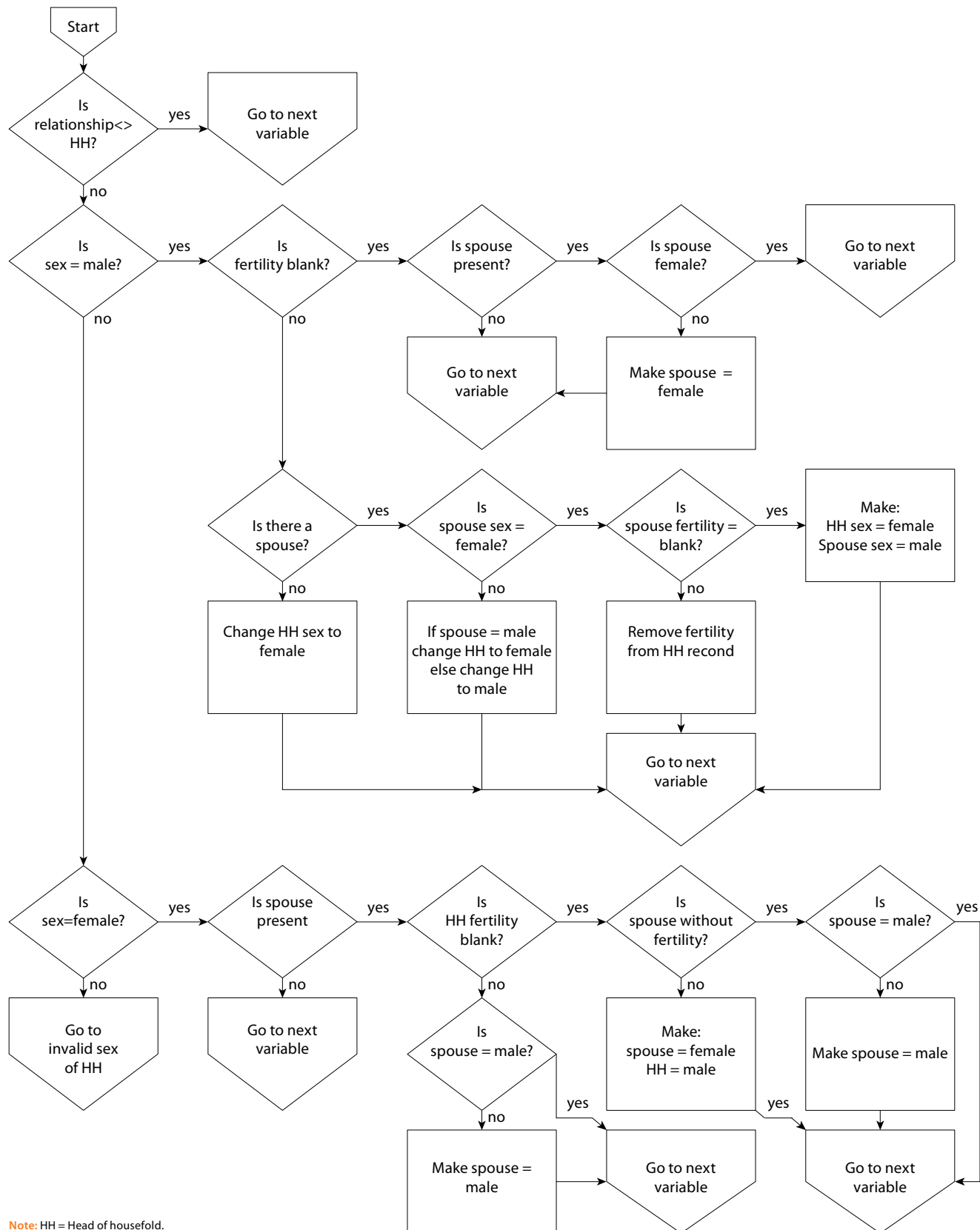
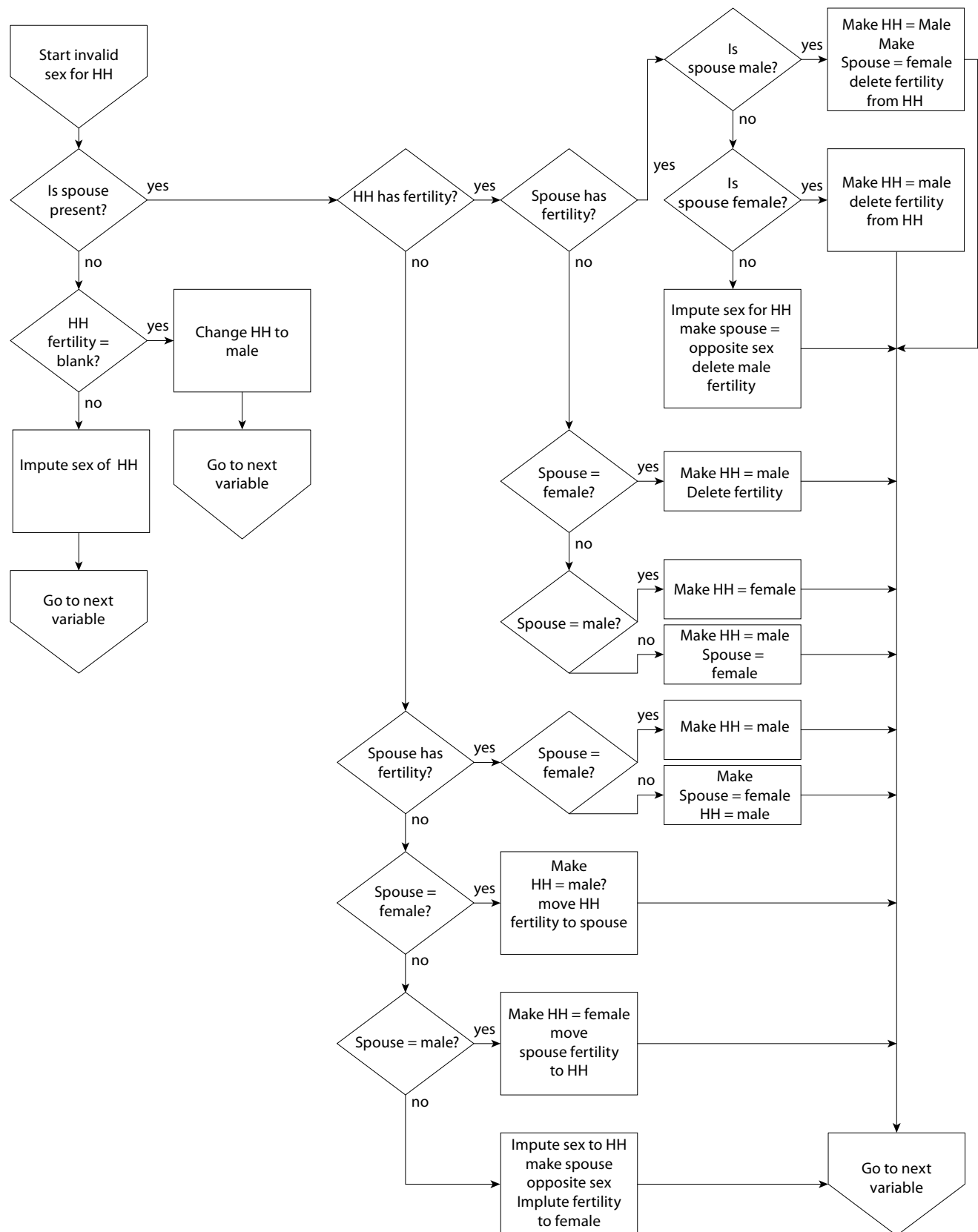


Figure A.IV.3 (continued)

Sample flow chart to edit sex variable for head of household and spouse



Annex V

Imputation methods

1. A number of imputation methods have been developed. Most methods described below are reviewed in papers by Kalton and Kasprzyk (1982, 1986); Sande (1982); and Särndal, Swensson and Wretman (1992).

2. Imputation methods can be classified as either stochastic or deterministic, depending upon the degree of randomness in the imputed data.

3. **Deterministic imputation methods** include deductive imputation; model-based imputation methods such as mean imputation and regression imputation; and (if appropriate) nearest-neighbour imputation.

4. **Deductive imputation** is a method that allows a missing or inconsistent value to be deduced with certainty. Often this will be based upon the pattern of responses given to other items on the questionnaire.

5. More commonly, the imputation technique must substitute a value that is not certain to be the true value. Some common imputation procedures are outlined in the following paragraphs.

6. With the exception of single donor dynamic imputation algorithms, the methods described below involve the imputation of one item at a time. So, within each imputation class, the items on the record are considered one after the other in a sequential fashion. Commonly, this is done by considering only those edits pertaining explicitly to the item in question or to a small set of closely related variables. Because there may be explicit or implicit edits that link the item(s) in question to other items to be considered later in the process, this procedure may cause an imputed value, while passing the edits currently being considered, to bring about failures on other edits to be considered later in the process. Only when a complete set of edits, including all implied edits, is considered can it be assured that imputed values will pass all edits. An implied edit is one that can be derived by logically combining two or more of the explicit edits.

7. In the following descriptions, “passed edit records” refers to those which have passed all edits pertaining to the item(s) in question. “Failed edit records” refers to those that have failed at least one edit pertaining to the item(s) in question.

8. **Overall mean imputation** assigns the item mean for passed edit records to the missing or inconsistent item for all failed edit records. This method may produce reasonable point estimates but is less appealing if variance estimates are to be computed using a standard variance estimator. Variance estimates can be severely underestimated unless the imputation rate is very low or a variance estimator modified to account for imputation is used.

9. **Class mean imputation** uses imputation classes defined to create groups of records having a degree of similarity. Within each class the item mean for passed edit records is imputed for the missing or inconsistent item for all failed edit records. This is much like overall mean imputation, but the impact upon the distribution and problems with variance estimation are likely to be less severe.

10. **Regression imputation** or, more generally, **model-based imputation** uses data from passed edit records to regress the variable for which imputation is required on a set of predictor variables. The predictors in the regression can be items from the questionnaire or auxiliary variables. The regression equation is then used to impute the values for the missing or inconsistent item values. This is a special case of model-based imputation. This method is frequently used for continuous variables in business survey applications where previous occasion data can often predict well current occasion values satisfactorily.

11. **Nearest-neighbour imputation** or **distance function matching** assigns an item value for a failed edit record from a “nearest” passed edit record where “nearest” is defined using a distance function in terms of other known variables. This method can be applied within imputation classes. It is usually considered appropriate for continuous variables but can also be applied with non-numeric variables.

12. **Stochastic imputation methods** include regression, or any other deterministic method, with random residuals added and hot deck or cold deck methods.

13. For each deterministic method there is a stochastic counterpart. This can be achieved by adding a random residual from an appropriate distribution to the imputed value from the deterministic imputation. This procedure will help to better preserve the frequency structure of the data file. Kalton and Kasprzyk (1986) review some approaches to this technique.

14. **Hot deck and cold deck imputation** attempt to create a more realistic variability in the imputed values than deterministic methods can. Hot deck imputation procedures replace missing or inconsistent values with values selected (at random) from passed edit records in the current survey or census. Cold deck imputation procedures impute based on other sources, often historical data such as earlier occasions of the same survey or census. There are a number of different forms of hot deck and cold deck imputation.

15. **Random overall imputation** is the simplest form of hot deck imputation. For each failed edit record, one passed edit record is selected at random from the set of all passed edit records and its reported value for the item in question is imputed for the failed edit record.

16. **Random imputation within classes** again uses imputation classes to constrain the random selection of the donor record to a set considered to have some similarity to the record requiring imputation.

17. **Sequential hot deck imputation** also uses imputation classes and has the advantage that a single pass through the data file is sufficient to complete the imputation process. The procedure starts with a cold deck value for each imputation class and the records in the data file are considered in turn. When a passed edit record is detected, its value for the item in question replaces the stored value for the imputation class. When a failed edit record is detected its missing or inconsistent value is replaced by the stored value. The number of imputation classes cannot be excessively large as it must be ensured that donors are available in every imputation class. If the order of records in the data file is random, this method will be nearly equivalent to random imputation within classes. A disadvantage of this procedure is that it often leads to multiple uses of donors and so can adversely affect the item’s distribution and variance estimates.

18. **Hierarchical hot deck imputation** is an enhancement of sequential hot deck imputation in which a large number of imputation classes are used. When a donor cannot be found in the initial imputation class, classes are collapsed in a hierarchical fashion until a donor is found.

19. The objective of **single donor hot deck imputation** algorithms is to impute data for a failed edit record from a single donor. Hence they allow for the joint imputation of all item values on a record identified as problematic by the edits. Often in practice, the objective is to use a single donor for each section of closely related variables in the record. This approach provides the significant advantage of better maintaining not only marginal distributions, as the above hot deck imputation methods, but also joint frequency distributions. Another advantage of single donor hot deck imputation methods is that they reduce the problem of imputing values that will fail edits considered in subsequent sections of variables. In the context of single donor hot deck imputation methods, a passed edit record is one that has passed all edits applying to the section. A failed edit record is one that has failed at least one of those edits.

20. The **Fellegi-Holt edit and imputation method** (Fellegi and Holt, 1976) considers all edits concurrently. A key feature of the Fellegi-Holt edit and imputation method is that the imputation rules are derived from the corresponding edits without explicit specification. For each failed edit record, it first proceeds through a step of error localization in which it determines the minimal set of variables to impute as well as the acceptable range(s) of values to impute and then performs the imputation. In most implementations, a single donor is selected from among passed edit records by matching on the basis of other variables involved in the edits but not requiring imputation. The method searches for a single exact match and can be extended to take account of other variables not explicitly involved in the edits. Occasionally no suitable donor can be found and a default imputation method must be employed.

21. The **Nearest-neighbour Imputation Methodology (NIM)** (Bankier and others, 1996; Bankier, Lachance and Poirier, 1999) is similar to the Fellegi-Holt method in that it considers all edits concurrently, does not explicitly specify imputation actions and imputes from a single donor. For each failed edit record it identifies minimum-change imputation actions conditional on the potential donors available. This guarantees that a donor will be available. Unlike Fellegi-Holt, NIM first searches for donors and then determines minimum-change imputation actions. NIM searches for donors by matching, using all variables (including those potentially to be imputed) involved in the edits, and can be satisfied by near matches for numeric variables plus matches for most, but not necessarily all, other variables. Imputation actions based on each potential donor are determined and those that are minimum-change imputation actions are identified. The method also considers near minimum-change imputation actions; these can sometimes yield more plausible imputed records. Finally, one of the minimum-change and near minimum-change imputation actions is selected at random and the imputation is performed.

22. Although both Fellegi-Holt and NIM are computationally demanding, efficient algorithms are available so that their implementation and application are feasible with modern computers. This is particularly true for NIM, which can readily handle somewhat larger editing and imputation problems than can the Fellegi-Holt method.

23. All of the above imputation methods produce a single imputed value for each missing or inconsistent value. All will distort to some extent the usual distribution of values for the item in question and can lead to inappropriate variance estimates when standard variance estimators are used. The extent of distortion varies considerably depending on the amount of imputation and the method used.

24. **Multiple imputation** is a method, proposed by Rubin (1987), that addresses this problem by imputing several times (m) for each value requiring imputation. Then, from the completed dataset m estimates can be produced for the item. From these, a single combined estimate is produced along with a pooled variance estimate that will

express the uncertainty about which value to impute. A disadvantage of the multiple imputation method is that it requires more work for data processing and computation of estimates.

25. In most imputation systems a mix of imputation methods is used; typically, deductive imputation is used where possible and is followed by one or more other procedures. Most national statistical offices use some form of dynamic imputation method for census editing and imputation. Sequential hot deck imputation and the Fellegi-Holt method are currently the most commonly used. Of the national statistical offices presently using the Fellegi-Holt method, one is changing to NIM and a number of others are considering it. However, given the expected primary readership, this *Handbook* focuses on a form of sequential hot deck imputation.

Annex VI

Computer editing packages

1. With the availability of relatively inexpensive microcomputers, countries should be able to edit census and survey data thoroughly as well as in a timely manner.^a Until recently, each country had to write its own editing program, an undertaking requiring expensive debugging and processing time. With the advent of standard computer editing packages, however, the fulfilment of a country's editing needs may now be greatly facilitated, and with less data processing expertise.

2. One advantage of using a computer editing package is that when properly used, data will be consistent and clean so that tabulations can be produced in a more timely manner. Many computer packages, such as SAS and SPSS, or other high-level languages can be used to write editing programs. Or, a country can choose to use one of the computer packages written expressly for editing census and survey data. For most countries, general editing is faster with a package than with custom-written programs since the package will not require the same level of data processing knowledge as custom-written programs.

3. A good computer editing package will provide for communication between the subject matter specialists and programmers. A good editing package should allow the placement of narrative or pseudocode next to programming code, unless the programming code itself is apparent or transparent to the subject matter specialists. Demographers and other specialists should be able to go through the programme line by line and understand exactly what the program is doing.

4. Any existing computer editing software package that a country might consider using must be able to perform and produce reports for the various checks, tests and imputations required for editing census data. These requirements apply even when data processing specialists produce custom editing programs. The package must meet the following requirements:

Have the capacity to key and/or verify entry data. This package should provide for the addition of skip patterns. For example, the editing team may decide that fertility information must be skipped for males;

Perform structural edits, which will make it possible to determine whether the types of records that should be present are in fact present, including, for example, a housing record for each serial number;

- (a) Generate records if they are missing and/or add weights to existing records;
- (b) Determine that each variable has a valid value;
- (c) Store all or part of already edited records;
- (d) Test the consistency between two or more characteristics in the same record and between records. A subset of this is to test the consistency within households, checking responses with those of previous household mem-

^a It is important to note that this *Handbook* focuses on census editing. Because surveys have fewer respondents, and, usually more questions, more detailed edits are often prepared. Some of the packages discussed in the present annex are designed for surveys in more populated countries. However, the packages would also be appropriate for smaller countries that are conducting censuses. When files become larger, they utilize some statistical methods, such as regression and multivariate analyses, less easily. On the other hand, with an exhaustive file, the impact of a non-response is less than for a survey where each sampled individual "counts" for many. Hence, each country office must test various packages to determine the one that best suits its needs.

bers. Impute values by the hot deck technique, if the country chooses to use dynamic imputation;

- (e) Use several values within a record or from multiple records to construct a derived variable and insert the derived variable in the appropriate record;
- (f) Identify and eliminate duplicate records;
- (i) Produce a diary of errors and changes, by small geographical area.

5. Usually packages or programs edit one record at a time, but current packages also permit inter-record checking, particularly within housing units.

6. As noted in the text, until the introduction of the Fellegi-Holt method (1976) and its follow-ups, almost all editing used a top-down approach; that is to say, items were edited in order—usually, but not always, in the same order in which they had been collected. For example, since the first population item to be edited is usually “relationship”, “sex” would be edited on the basis of that item, then “age” on the basis of both sex and relationship, and so forth.

7. In the last few decades, several minimum change imputation systems based on Fellegi-Holt have been developed. Among them were CANEDIT and the General Edit and Imputation System (GEIS) at Statistics Canada, followed by the CANadian Census Edit and Imputation System (CANCEIS) (Bankier, 2005; Chen, 2007).^b The United States of America developed DISCRETE (Chen and others, 2000; Winkler, 1997a; Winkler, 1997b; Winkler and Chen, 2002) and Structured Programs for Economic Editing and Referrals (SPEER) (Greenberg and Surdi, 1984; Winkler and Draper, 1997) at the U.S. Census Bureau. Kovar and Winkler (2000) compared the Canadian and United States systems in some detail. Other Fellegi-Holt-derived editing packages include CHERRYPI, a Fellegi-Holt-based edit package developed by Statistics Netherlands (De Waal and Van de Pol, 1997).

8. 1996 Canada’s Census (and others) used a different approach, called the Nearest-neighbour Imputation Methodology (NIM). The 1996 version imputed responses for age, sex, marital status and relationship for all persons in a house simultaneously (Bankier, 1999). The method was improved and expanded for Canada’s 2001 and subsequent statistical activities (Bankier, Lachance and Poirier, 2000; Banquier, 2001).

9. The Nearest-neighbour approach searches for nearest-neighbour donors first and then determines the minimum change imputations based on these donors. While the Fellegi-Holt method involves imputing the fewest variables and preserving the integrity of the subpopulations, the NIM, which reverses the order of the operation—starting with the search for donors and then changing the minimum number of variables—provides computational advantage. However, NIM can carry out imputation using only donors, while Fellegi-Holt can be used with other methodologies (like top down). Statistics Canada incorporated NIM into its Canadian Census Edit and Imputation System for both the 2001 and 2006 censuses.

10. A series of meetings after 2000 brought many European statisticians together to consider various aspects of census and survey editing and imputation. Discussions surrounded what was called “The development and evaluation of new methods for editing and imputation” (EUREDIT) (Chambers, 2000; Poirier, 2000). Statistics Canada has also developed methods of measuring and reducing variability in imputation. Meant for surveys, their SIMPVAR system is designed to process the four main imputations methods (ratio, mean, hot deck and Nearest-neighbour) (Rancourt and others, 1997). (Istituto Nazionale di Statistica) (ISTAT) has described the editing and imputation system for the 2001 Census of Italy, using DIESIS (Data Imputation and Editing System—Italian Software) and other methods (Di Zio, 2002; Bianci and others, 2005).

^b In a review of CANCEIS, Banquier, Lachance and Poirier (2000, p. 10) summarize as follows: CANCEIS, with its highly efficient editing and imputation algorithms, shows great promise for solving very general imputation problems involving a large number of edit rules and a large number of qualitative and quantitative variables when minimum change donor imputation is appropriate. The Fellegi-Holt minimum change edit and imputation algorithm, however, should still be the method of choice for smaller imputation problems if there may not be sufficient donors available or if it is more appropriate to use another method to perform imputation.

11. Other methods for imputing unknowns besides that of using actual cases exist. Average measures are sometimes used. Some countries use regression models (Russian Federation, State Committee on Statistics, 2000). Regression was also used for imputing age on the 2000 United States short form (Williams, 1998).

12. Because of advances in interactive keying, some systems integrate coding, keying and editing into a single system, particularly for surveys. Among these is Brazil's CRIPTAX system (Hanono and Barbosa, n.d.) which entails a method of editing on entry. Other systems, including the Census and Survey Processing System (CSPro), have aspects of interactive editing. As noted elsewhere, country statistical offices must make decisions about returns on investments for machines, software, personnel, time and so forth.

13. As for the top-down approach, the United States Census Bureau developed the Integrated Microcomputer Processing System (IMPS) for the Censuses from 1980 through 2000. This package is in DOS, and includes data entry, editing and tabulation as well as other functions, and many countries continue to use it. During the late 1990s and the 2000s, the Census Bureau developed the Windows version, called CSPro (mentioned above), which performs many of the same tasks and is Windows compatible.^c CSPro works well for all surveys and censuses of small countries; the processing is so slow, however, that editing takes longer than in IMPS. Nonetheless, countries starting their edits from scratch should use the Windows version. Applications in both IMPS and CSPro can all be developed for the Fellegi-Holt approach; Nearest-neighbour would take much more work.

14. CSPro allows the user to create, modify, and run data entry, batch editing, and tabulation applications from a single integrated development environment. It processes data on a case basis (one or more questionnaires), where a case can consist of one or many data records. The data are stored in American Standard Code for Information Interchange (ASCII) text files described by data dictionaries. CSPro contains a powerful common procedure language to implement data entry control and edit rules.

15. Specifically, the batch editing function of CSPro identifies and reports structure, value and consistency errors in questionnaire data. The package can change (impute) data values based on simple or complex methods. It can produce summary or detailed reports of errors and corrections, access multiple lookup files, and read from, and write to, secondary files.

16. CSPro also provides tools for viewing data and other text files, viewing tables and thematic maps created by CSPro, converting IMPS and ISSA data dictionaries to and from CSPro, and converting Environmental Systems Research Institute (ESRI) shapefiles (maps) to CSPro map files. CSPro was developed jointly by the U.S. Census Bureau, Macro International, and Serpro, with major funding from the United States Agency for International Development. CSPro is in the public domain, is available at no cost and may be freely distributed, and is available for download at www.census.gov/ipc/www/cspro.

^c CSPro is a software package for entry, editing, tabulation and dissemination of census and survey data. CSPro combines the features of the Integrated Microcomputer Processing System (IMPS) and the Integrated System for Survey Analysis (ISSA) in a Windows environment.

Glossary

Array—A set of numeric values. Sometimes called a matrix, an array can be used to store numeric data of a repetitive nature.

Audit trail—A method of keeping track of changes to values in a field and the reason and source of each change. Audit trails are generally begun after the initial interview is completed.

Automated correction—The correction of data errors by computer without human intervention. One aspect of automated data editing (Pierzchala, 1995).

Between-record edit—Edits carried out on fields involving more than one record in the survey. Statistical edits are an example of between-record edits because distributions are generated on sets of fields over all the records in the survey (Pierzchala, 1995).

Class mean imputation—Uses imputation classes defined to create groups of records having a degree of similarity.

Clean record—Record which has no missing values and which passes all edits (Pierzchala, 1995).

Code list—List of all allowed (admissible) values of a data item.

Cold deck—The initial static matrix; a correction base for which the elements are given before correction starts and do not change during correction. For example, the correction base could be a prior year's data. A modified cold deck may adjust cold deck values according to (possibly aggregate) current information.

Complete set of edits—The union of explicit edits and implied edits. Necessary for the generation of feasible regions for imputation (if one wants imputations to satisfy edits) (Pierzchala, 1995).

Consistency edit—Checks for determinant relationships, such as parts adding to a total or “harvested acres” always less than “planted acres” (Pierzchala, 1995).

Data capture—The process by which collected data are put into a machine-readable form. Elementary edit checks are often performed in submodules of the software that captures the data.

Deductive imputation—A method in which, a missing or inconsistent value can be deduced with certainty, often based on the pattern of responses given to other items on the questionnaire.

Deterministic edit—An edit that, if violated, points to an error in the data with a probability of one. Example: age = 5 and status = mother. Contrast with stochastic edit (Pierzchala, 1995).

Deterministic imputation—This situation arises when only one value of a field will cause the record to satisfy all of the edits. Occurs in some situations (such as the parts of a total not adding to the total). The first solution to be checked for in the automated editing and imputation of survey data (Pierzchala, 1995).

Distance function—For numeric data, a function defined on the matching variables of both the recipient (candidate) and donor records and used to quantify the concept of similarity. Used to find matching records in hot deck imputation (Pierzchala, 1995).

Distance function matching—Assigns an item value for a failed edit record from a “nearest” passed edit record, where “nearest” is defined using a distance function in terms of other known variables.

Donor imputation—A method that pairs each record requiring imputation, the recipient or candidate record, with one record from a defined donor population as, for example, in hot deck imputation (Pierzchala, 1995).

Edit (definition 1)—Logical constraints on the values that each variable can assume (Pierzchala, 1995).

Edit (definition 2)—Rules that detect prohibited response combinations. (Pierzchala, 1995).

Edit trail—See “audit trail”.

Explicit edits—Those edits explicitly written by a subject matter specialist (contrast explicit edits with implied edits) (Pierzchala, 1995).

Failed edit records—In editing and imputation, records that have failed at least one edit pertaining to the item(s) in question.

Fellegi-Holt method for automatic correction—Automatic correction method in which the least possible number of data items are changed, and the Fellegi-Holt model is used to determine acceptable sets of values or ranges for the items that are imputed. Sequential or simultaneous imputation via cold deck or hot deck method may be applied.

Fellegi-Holt system—Refers to assumptions and editing and imputation goals put forth by Fellegi and Holt in their 1976 *Journal of the American Statistical Association* paper. A key feature of the Fellegi-Holt model is that it shows that implied edits are needed to assure that a set of values in data fields that are not imputed always lead to final (imputed) records that satisfy all [cases].

Flag—A flag is a variable used to note useful information about another variable or variables. For example, if an item is changed, from invalid to valid, the flag can be used to note either the original information or to note that the item’s value has changed.

Flow chart—Diagrammatic description of all of the functions that must be accomplished.

Hand edit—An edit performed by people before data are entered into the computer (see “manual edit”) (Pierzchala, 1995).

Heads down data entry—A style of data entry in which the data entry machine does not detect errors in the data as they are entered, allowing the operator to enter the data rapidly.

Heads-up data entry—A style of data entry in which the data entry machine detects errors in the data as they are entered, allowing the operator to correct the errors immediately (Pierzchala, 1995). See “interactive keying”.

Hot deck imputation—A method of imputation in which donor records are taken from the current deck of sample data (cold deck, in contrast, refers to the method of imputation in which the donor record comes from past survey data) (Pierzchala, 1995).

Implied edit—An unstated edit derived logically from explicit edits that were written by a subject matter specialist (Pierzchala, 1995).

Imputation—The assignment of a value to a field, either for non-response or for replacement of a recorded value determined to be inconsistent with a set of edits (Pierzchala, 1995).

Interactive keying—A style of data entry in which the data entry machine detects errors in the data as they are entered, allowing the operator to correct the errors immediately. See “heads up data entry”.

Internal consistency—This term pertains to relations among the variables for a given sample unit and is the reason for the edits in most survey procedures (Ford, 1983; Pierzchala, 1995).

Macro-edit—Detection of individual errors by (1) checks on aggregated data, or (2) checks applied to the whole body of records. The checks are based on the estimates (Granquist, 1987; Pierzchala, 1995).

Manual edit—An edit performed by people before data are entered into the computer (see “hand edit”) (Pierzchala, 1995).

Matching—In the hot deck imputation procedure, the act of matching a donor record to a recipient (candidate) record (Pierzchala, 1995).

Matching variables—Those variables used to find a match between a recipient (candidate) record and a donor record (Pierzchala, 1995).

Microedit—Traditional edit performed on record-level data. The logical antonym to macroedit (Pierzchala, 1995).

Micro-macro edit—An editing procedure whereby detailed micro-edits are replaced with a combination microedit and macro/statistical edit. The microedits in the combined procedure are less detailed than in the first. The idea is to “develop survey edits based upon an ‘impact on the estimates’ philosophy rather than a ‘catch all data inconsistencies’ philosophy” (Granquist, various dates; Pierzchala, 1995).

Minimal set—The smallest set of fields requiring imputation that will guarantee that all edits are passed (Pierzchala, 1995).

Model-based imputation—Uses data from passed edit records to regress the variable for which imputation is required on a set of predictor variables.

Multiple imputation—Imputes several times for each value requiring imputation and then produces an estimate for the item.

Multivariate edit—A type of statistical edit in which multivariate distributions are used to evaluate the data and find outliers (Pierzchala, 1995).

Nearest-neighbour imputation—Assigns an item value for a failed edit record from a “nearest” passed edit record where “nearest” is defined using a distance function in terms of other known variables.

Nearest-neighbour Imputation Methodology—Similar to the Fellegi-Holt method in that it considers all edits concurrently, does not explicitly specify imputation actions and imputes from a single donor. NIM looks at each failed edit record to identify minimum-change imputation actions conditional on the potential donors available.

Non-response—An incomplete questionnaire or a missing questionnaire (Pierzchala, 1995).

Outliers—Values of items that lie outside of some bound, according to some determination of the bound (Pierzchala, 1995).

Overall mean imputation—Assigns the item mean for passed edit records to the missing or inconsistent item for all failed edit records.

Passed edit records—During edit and imputation, records that have passed all edits pertaining to the items in question.

Pointer—A pointer is a variable used to mark an item or other variable for later reference. For example, pointers are used to note the line numbers of the “head” and the “spouse” to use later to make sure that the spouses are of the opposite sex and that both are married.

Production run—The act of processing large quantities of data after initial “bugs” have been removed from the edit or tabulation program.

Pseudocode—Written editing instructions or specifications.

Quality errors—Errors that may distort the quality of the data: for example, systematic errors that lead to bias (Granquist, 1984; Pierzchala, 1995).

Quantitative edits—Edits applied to fields measured on a continuous scale (Pierzchala, 1995).

Random imputation within classes—Uses imputation classes to constrain the random selection of the donor record to a set considered to have some similarity to the record requiring imputation.

Random overall imputation—For each failed edit record, one passed edit record is selected at random from the set of all passed edit records, and its reported value for the item in question is imputed for the failed edit record.

Record—A magnetically stored, computer-readable representation of survey data. Usually there is one record for each questionnaire, although it is possible to split data from one questionnaire into more than one record, such as population and housing (Pierzchala, 1995).

Regression imputation—Uses data from passed edit records to regress the variable for which imputation is required on a set of predictor variables.

Searching—In the hot deck imputation procedure, the act of searching for a donor record (Pierzchala, 1995).

Sequential hot deck imputation—Imputation that occurs when a series of variables are edited in sequence, with only edited values used as subsequent hot deck variables.

Similarity—In numeric data, a concept of closeness of two records based on prescribed matching variables. A distance function is used to quantify this concept according to some criteria (Pierzchala, 1995).

Single donor hot deck imputation—Imputes data for a failed record from a single donor, allowing for joint imputation of all item values on a record identified as problematic by the edits.

Statistical edit—A set of checks based on a statistical analysis of respondent data: for example, the ratio of two fields lies between limits determined by a statistical analysis of that ratio for presumed valid reporters (Greenberg and Surdi, 1984; Pierzchala, 1995).

Statistical imputation—An example of statistical imputation would be the use of a regression model in which the dependent variable is to be imputed, and the co-

efficients of the independent variables are derived from presumed valid responses (Pierzchala, 1995).

Statistical matching (in hot deck)—The act of matching a donor record with a recipient (candidate) record according to certain statistical criteria in order to transfer data from the donor to the recipient (Pierzchala, 1995).

Stochastic edit—An edit that, if violated, points to an error in the data with a probability of less than one (Pierzchala, 1995).

Structural edit—Checks based on a logical relationship between two or more edited fields. For example, a total must equal the sum of its parts: or, because of a skip pattern inherent in a questionnaire, two variables lying on disjoint paths cannot both be non-zero. A structural edit ensures that the structure of the questionnaire is maintained in the data record (Pierzchala, 1995).

Structural imputation—Structural imputation is used when a structural relationship holds between several variables. For example, a total must equal the sum of its parts: therefore, for a mother, children ever born must equal children alive plus children who have died (Pierzchala, 1995).

Validation edit—Edits checks that are made between fields in a particular record. This includes the checking of every field of every record to ascertain whether it contains a valid entry and the checking of entries in a certain predetermined combination of fields to ascertain whether the entries are consistent with each other (Pierzchala, 1995).

Weights—In the Fellegi-Holt school of edit and imputation, weights are assigned to fields based on reliability. The higher the weight the more likely a field will be imputed for (all other things being equal). Weights can also be assigned to edits (Pierzchala, 1995).

Within-record edit—Another name for a validation edit (Pierzchala, 1995).

References

- Banister, J. (1980). Use and abuse of census editing and imputation. *Asian and Pacific Census Forum*, vol. 6, No. 3, pp.1-20.
- Bankier, M. (1999). "Experience with the New Imputation Methodology used in the 1996 Canadian Census with Extensions for Future Censuses", Proceedings of the Workshop on Data Editing, UNECE, Italy (Rome).
- _____. (2005). Edit and imputation for the 2006 Canadian Census. Paper presented at the Work Session on Statistical Data Editing, Ottawa, 16-18 May, 2005.
- _____, A.-M. Houle and M. Luc (n. d.). Canadian census demographic variables imputation. Manuscript.
- Bankier, M., M. Lachance and P. Poirier (1999). A generic implementation of the new imputation methodology. In *Proceedings of the Section on Survey Research Methods*. Alexandria, Virginia: American Statistical Association forthcoming.
- _____. (2000). 2001 Canadian Census minimum change donor imputation methodology. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Cardiff, United Kingdom, 18-20 October 2000.
- Bankier, M., P. Poirier and M. Lachance (2001). Efficient methodology within the Canadian Census Edit and Imputation System (CANCEIS). Proceedings of the Annual Meeting of the American Statistical Association, 5-9 August 2001.
- Bankier, M., and others (1996). Imputing numeric and qualitative census variables simultaneously. In *Proceedings of the Section on Survey Research Methods*. Alexandria, Virginia: American Statistical Association, pp. 287-292.
- Bianchi, G., and others (2005). New procedures for editing and imputation of demographic variables. Paper presented at the Work Session on Statistical Data Editing, Ottawa, 16-18 May 2005.
- Boucher, L. (1991). Micro-editing for the annual salary of manufacturers: what is the value added? In *Proceedings of the Annual Research Conference*. Washington, D.C.: United States Bureau of the Census, pp. 765-781.
- Chambers, Ray (2000). Evaluation criteria for editing and imputation in EUREDIT. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Cardiff, United Kingdom, 18-20 October 2000.
- Chen, Bor-Chung (2007). CANCEIS experiments of edit and imputation with 2006 census test data. Statistical Research Division Study Series, No. 2007-1. Washington, D.C.: U.S. Bureau of the Census.
- _____, and others. (2000). Using the DISCRETE Edit System for ACS Surveys. Bureau of the Statistical Research Division Statistical Research Report Series, No. RR2000/03. Washington, D.C.: U.S. Bureau of the Census.

- De Waal, Tom, and Frank van de Pol (1997). A recipe for applying CHERRYPI in the edit process. Paper presented at the Work Session on Statistical Data Editing, Prague, 14-17 October 1997.
- Di Zio, M. (2002). Evaluating editing and imputation processes: the Italian experience. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Helsinki, 27-29 May 2002.
- Fellegi, I. P. and D. Holt (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, vol. 71, No. 353 (March), pp. 17-35.
- Ford, Barry L. (1983). An overview of hot deck procedures. In *Incomplete Data in Sample Surveys*, vol. 2, *Theory and Bibliographies*. William G. Madow, Ingram Olkin and Donald B. Rubin, eds.
- Granquist, L. (1984). Data editing and its impact on the further processing of statistical data. Paper presented at the Workshop on Statistical Computing, Budapest, 12-17 November 1984.
- _____ (1987). The short-term developing program for computer-supported editing at Statistics Sweden. Report presented at the Data Editing Joint Group Meeting, Madrid, 22-24 April 1987. Stockholm: Statistics Sweden.
- _____ (1997). The new view on editing. *International Statistical Review*, vol. 65, No. 3, New York: Academic Press, pp. 381-387.
- _____, and J. G. Kovar (1997). Editing of survey data: how much is enough? In *Survey Measurement and Process Quality*, Lyberg and others, eds. New York: Wiley and Sons, pp. 415-435.
- Greenberg, Brian, and Rita Surdi (1984). A flexible and interactive edit and imputation system for ratio edits. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*. Alexandria, Virginia: American Statistical Association, pp. 421-426.
- Hanono, Reina Marta, and Dulce Maria Rocha Barbosa (n.d.). Generalized environment for application development for capturing, editing and coding statistical survey's data. Rio de Janeiro: Brazilian Institute of Geography and Statistics (IBGE).
- Ireback, H. (2000). The impact of new information technologies on data collection at Statistics Sweden. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Cardiff, United Kingdom, 18-20 October 2000.
- Kalton, G., and D. Kasprzyk (1982). Imputing for missing survey responses. In *Proceedings of the Section on Survey Research Method*. American Statistical Association, pp. 23-31.
- _____ (1986). The treatment of missing survey data. *Survey Methodology*, vol. 12, pp. 1-16.
- Kovar, J., and W. Winkler (2000). Comparison of GEIS and SPEER for editing economic data. Bureau of the Census Statistical Research Division Statistical Research Report Series, No. RR 2000/04. Washington, D.C.: U.S. Bureau of the Census.
- Naus, J. I. (1975). *Data Quality Control and Editing*. New York: Marcel Dekker.

- Nordbotten, S. (1963). Automatic editing of individual statistical observations. Conference of European Statisticians, Statistical Standards and Studies, No. 2. New York: United Nations.
- Pierzchala, M. (1995). Editing systems and software. In *Business Survey Methods*. B. G. Cox, and others, eds. New York: John Wiley and Sons, pp. 425-411.
- Poirier, C. (2000). A prototype knowledge base on data editing and imputation. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Cardiff, United Kingdom, 18-20 October 2000.
- Pullum, T. W., T. Harpham and N. Ozsever (1986). The machine editing of large-sample surveys: the experience of the World Fertility Survey. *International Statistical Review*, vol. 54, 311-326.
- Rancourt, E., and others (1997). Estimation of variance in presence of imputation. Proceedings of Symposium 1997: New Directions in Surveys and Censuses. Ottawa: Statistics Canada, pp. 273-279.
- Rubin, D. B. (1987). *Multiple Imputation for Non-response in Surveys*. New York: Wiley.
- Russian Federation, State Committee on Statistics (2000). Data imputation based on regression models with variations of entropy. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Cardiff, United Kingdom, 18-20 October 2000.
- Sande, I. G. (1982). Imputation in surveys: coping with reality. *American Statistician*, vol. 36, pp. 145-152.
- Särndal, C. E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Statistics Canada (1998). *Statistics Canada Quality Guidelines*, 3rd edition. Ottawa: Statistics Canada.
- United Nations (1992a). *Handbook of Population and Housing Censuses, Part I: Planning, Organization and Administration of Population and Housing Censuses*. Studies in Methods, Series F, No. 54. Sales No. E.92.XVII.8.
- _____ (1992b). *Handbook of Population and Housing Censuses, Part II: Demographic and Social Characteristics*. Studies in Methods, Series F, No. 54. Sales No. E.91.XVII.9.
- _____ (1999). *Standard Country or Area Codes for Statistical Use*. Statistical Papers, Series M, No. 49/Rev.4. Sales No. M.98.XVII.9.
- _____ (2008). *Principles and Recommendations for Population and Housing Censuses, Revision 2*. Statistical Papers, Series M, No. 67/Rev.2. Sales No. E.07.XVII.8.
- United Nations Statistical Commission and Economic Commission for Europe (1994). *Statistical Data Editing*, vol. 1: *Methods and Techniques*. Conference of European Statisticians, Statistical Standards and Studies Series, No. 44. Sales No. 94.II.E.36.
- _____ (1997). *Statistical Data Editing*, vol. 2: *Methods and Techniques*. Conference of European Statisticians, Statistical Standards and Studies Series, No. 48. Sales No. 96.II.E.30.

- Williams, Todd R. (1998) Imputing person age for the 2000 census short form: a model-based approach. Washington, D.C. : Bureau of the Census Statistical Research Division, Statistical Research Report Series, No. RR98/07.
- Winkler, W. E. (1997a). Edit/imputation System for the U.S. Decennial Census. Paper presented at the Work Session on Statistical Data Editing, Prague, 14-17 October 1997.
- _____ (1997b). Set-covering and editing discrete data. Technical report. Washington, D.C.: U.S. Bureau of the Census.
- _____ (2006). Data quality: automated edit/imputation and record linkage. U.S. Census Bureau Statistical Research Division Research Report Series, No. 2006-7. Washington, D.C.: U.S. Bureau of the Census.
- _____, and B. C. Chen (2002). Extending the Fellegi-Holt Model of Statistical Data Editing. U.S. Bureau of the Census Statistical Research Division, Statistical Research Report Series, No. 2002-02.
- Winkler, W. E., and L. R. Draper (1997). The SPEER edit system. In United Nations Statistical Commission and Economic Commission for Europe, *Statistical Data Editing*, vol. 2, Method and Techniques. Conference of European Statisticians, Statistical Standards and Studies, No. 48. Sales No. E.96.II.E.30, pp. 56-62.